Class Notes in Operations Research, Statistics and Probability (Second Edition)

Edited by Roger Goodwin
Contact Information:
Roger L. Goodwin
3251 Old Lee Highway
Room 301
Fairfax, VA 22030
Phone: 703-877-8000, ext 120
Fax: 703-877-8044
Email: roger_goodwin@nass.usda.gov

Old Dominion Press
1989 Leetown Road
Summit Point, WV 25446
Email: rog982001@yahoo.com

USA

Editor's Note: In graduate school, it became too cumbersome for me to look-up equations, theorems, proofs, and problem solutions from previous courses. I had three boxes full of notes and was going on my fourth. Due to the need to reference my notes periodically, the notes became more unorganized over time. That's when I decided to typeset them. I have been doing this for over a decade. Later in life, some colleagues asked if I could make these notes available to others (they were talking about themselves). I did. These notes can be downloaded for *free* from the web site http://www.repec.org/ and can be found in the Library of Congress. I thank Winston for helping typeset the notes.

Contents

Chapter 7

SAS Programming

Dr. Morgan, Old Dominion University
Statistics 505, Fall 1996
Text used: Quick Start to Data Analysis with SAS, DiIorio, Frank C., and Kenneth A. Harding.

7.1 SAS Data Sets

When creating a SAS data set, you are specifying what the raw data means. SAS names are:

1. 1–8 characters long.
2. Can include numbers and letters.
3. Can include underscores.

Data sets end with "RUN;" PROC steps perform analysis on a data set. Every SAS job has 2 steps:

1. Data Set.
2. PROC steps.

Every line in SAS ends with a semicolon. "LABEL" means to associate the variable name with human readable text. SAS is not case sensitive. The "SET CASE MIXED" MVS command allows you to edit in lower case on the mainframe. Every data set has a name. Dots in raw data mean

information is missing. The "INFILE" statement tells SAS where to find the file. The "INPUT" statement lists the variable names in the "INFILE" statement file name. If no columns are specified, a blank is assumed to separate the data.

```
* ASSIGNMENT 1;
EDIT THE FILE YOU CREATED IN THE FIRST CLASS TO INCLUDE THE
CHANGES BELOW. RUN THIS JOB, THEN BRING THE OUTPUT FOR BOTH
VERSIONS TO THE NEXT CLASS. NOTE: BE CAREFUL TO DISTINGUISH
BETWEEN "ONES" AND "ELLS".;

* SIMPLE PROGRAM TO DEMONSTRATE SAS'S LOOK AND FEEL;

OPTIONS NODATE LINESIZE=72;

DATA STATES;
INFILE 'STATES1 RAW A';
INPUT NAME $ 4-19 HIGHTEMP 25-27 LOWTEMP 29-31
       POP80 POPSQMI FARMPCT PCTURB
       @83 PCAPINC 5.;
URB_POP = (PCTURB * POP80) / 100;
TEMPRNGE = HIGHTEMP - LOWTEMP;
IF  ((FARMPCT>50)&(PCAPINC>7000)) THEN RICHFARM = 'YES';
ELSE RICHFARM = 'NO';

FORMAT POP80 URB_POP COMMA11. FARMPCT PCTURB 6.2 PCAPINC DOLLAR8.;
LABEL NAME = 'State name"
      HIGHTEMP = 'Highest recorded temperature'
      LOWTEMP = 'Lowest recorded temperature'
      POP80 = 'Population, 1980'
      POPSQMI = 'Population density, 1980'
      FARMPCT = '% land devoted to agriculture'
      PCTURB = '% urban pop (areas over 2,500 pop)'
      PCAPINC = '1980 per captita income'
      URB_POP = 'Urban population, 1980'
      TEMPRNGE = 'Temperature range'
      ;
RUN;

PROC PRINT DATA=STATEST N;
VAR POP80 POPSQMI PCTURB URB_POP HIGHTEMP LOWTEMP TEMPRNGE RICHFARM;
ID NAME;
```

```
TITLE 'State-Level Data for Demo Program';
RUN;

PROC MEANS MAXDEC=2 N NMISS MIN MAX MEAN;
RUN;
```

Check the SAS "listing" file for error messages. To print 2 copies off the mainframe type: "lpr < filename >(form xrev copies 2)." To copy one file to another type: "copy < filename1 > < filename2 > ." To delete a file type "erase."

A SAS data step makes raw data into a meaningful format. SAS PROCS operate on the data sets, never on raw data. Before writing a SAS data step, consider:

1. What are you going to call the SAS data set?
2. Where is the raw data located?
3. How is the data laid out?

Every SAS data step begins with the DATA statement. DATA < filename > ; Next, if the data is stored in an external file, you need to tell SAS where that is. INFILE 'location'; Two options that can be used are;

- FIRSTOBS = n; * which record to start with;
- OBS = n; * which record to end with;

Example: INFILE 'HW1.SAS' FIRSTOBS=3 OBS=10;

Finally, tell SAS how the data is arranged with the INPUT statement.

1. Name the variables.
2. Tell SAS where the variables can be found.

Different forms of INPUT:

7.1.1 Reading Raw Data Assignment

```
* Demonstates reading raw data from a CMF file, and the input
  of raw data instream. The three approaches shown below create
  identical datasets and thus identical output(i.e. the same
  listing file). Create a new file containing this code, then
  run the job. Print the SAS file and the LISTING file;
```

```
* Approach 1;

DATA PRES1;
  INFILE 'MRPRES RAW A';
  INPUT NAME $ 1-20 PARTY $ 21-29 BORN 31-34 BORNST $ 37-38
        INAUG 43-46 AGEINAUG 51-52 AGEDEATH 56-57;
  IF (INAUG<1950) THEN DELETE;
RUN;

PROC PRINT DATA=PRES1;
TITLE 'Presidents elected after 1950';


* Approach 2;

FILENAME USLEADRS 'MRPRES RAW A';
DATA PRES2;
  INFILE USLEADRS;
  INPUT NAME $ 1-20 PARTY $ 21-29 BORN 31-34 BORNST $ 37-38
        INAUG 43-46 AGEINAUG 51-52 AGEDEATH 56-57;
  IF (INAUG<1950) THEN DELETE;
RUN;

PROC PRINT DATA=PRES2;
TITLE 'Presidents elected after 1950';

* Approach 3;

DATA PRES3;
  INPUT NAME $ 1-20 PARTY $ 21-29 BORN 31-34 BORNST $ 37-38
        INAUG 43-46 AGEINAUG 51-52 AGEDEATH 56-57;
  CARDS;
Eisenhower              rep     1890 tx 1953    62      78
Kennedy                 dem     1917 ma 1961    43      46
Johnson, L.             dem     1908 tx 1963    55      64
Nixon                   rep     1913 ca 1969    56      .
Ford                    rep     1913 ne 1974    61      .
Carter                  dem     1924 ga 1977    52      .
Reagan                  rep     1911 il 1981    69      .
Bush                    rep     1924 ma 1989    64      .
Clinton                 dem     .    ar 1993    .       .
;
```

```
RUN;

PROC PRINT DATA=PRES3;
TITLE 'Presidents elected after 1950';
```

7.1.2 Printing Example

```
OPTIONS NODATE LINESIZE=72;

DATA COASTAL;
INFILE 'COASTAL RAW A';
INPUT COAST $ 1-2 OCEAN $ 3-4 STATE $ 7-8 GENCST 9-14 TIDALCST 15-21;
RUN;

PROC PRINT DATA=COASTAL;
TITLE 'Take all defaults';
RUN;

PROC PRINT DATA=COASTAL SPLIT=' ' N;
VAR OCEAN STATE GENCST TIDALCST;
LABEL TIDALCST = 'Detailed outline';
SUM _NUMERIC_;
BY COAST;
FORMAT GENCST TIDALCST COMMA7.;
TITLE 'Use VAR, FORMAT, BY, And SUM statements';
TITLE2 'Use TIDALCST label for column header, print # of obs';
RUN;
```

7.1.3 Inputting Raw Data

The different ways to input data are as follow:

1. List — the variables seperated by a space. $ indicates a character variable. It can be upto 200 characters long. The $ follows the character variable name.

2. Column Input — identify exactly which columns to look-in for the variable's data.

3. Format — Example: @52 2. – the dot tells SAS this is a format. The @52 says start at column 52 to read the data. The 2. means to read two columns for numeric data. 4.0 would mean to read no decimal

places. 4.2 means to read two decimal places. Ex: 1956 is read as 19.56 with the 4.2 format.

The chart on page 24 in the text book summarizes the input formats. A "comma" lets SAS know that there are commas in the raw data and should be read as numeric.

7.1.4 INPUT Formats Assignment

This assignment is not that different from the first two, except that now you will write the code yourself. The problem is to take the raw data file NATLPARK RAW, read it into a SAS job which creates a SAS dataset, and then print it out. This will require one fairly simple DATA step and one PROC. Most of this material is covered in Chapter 3 of the text book and in the example from the first class and the first homework. For a brief description of the raw data, see the bottom of this assignment.

So that you will learn these techniques, in the INPUT statement read at least one variable with column input, read at least one variable with formatted input, and use the @col specification at least once.

Also, in the DATA step, assign a descriptive label to each variable. the LABEL statement is covered on page 43 of the text book. Make sure that the labels appear when you print the dataset(use the L option in PROC PRINT).

Variables		Raw Data				
Name	Label	Type	Cols	Format	Min	Max
PARK	Park Name	char	1-20	$20.		
ST	Principal State	char	22-23	$2.		
COAST	East/West	char	26-26	$1.		
YRESTAB	Year Established	num	27-30	4.	1872	1986
ACRES	Acres in Park	num	31-39	9.	5839	8331604

```
DATA NATIONAL;
INFILE 'NATLPARK RAW A';
INPUT PARK $ 1-20 ST $ 22-23 COAST $ 26-26 @27 YRESTAB 4. @31 ACRES 9.;
LABEL  PARK = 'PARK NAME'
       ST = 'PRINCIPAL STATE'
 COAST = 'EAST/WEST OF MISSISSIPPI RIVER'
 YRESTAB = 'YEAR ESTABLISHED'
 ACRES = 'ACRES IN PARK';
```

```
PROC PRINT DATA=NATIONAL L;
VAR PARK ST COAST YRESTAB ACRES;
TITLE 'SIZE AND HISTORY OF US NATIONAL PARKS';
RUN;
```

7.1.5 Permanent Data Sets

A SAS dataset has 2 major components:

1. The data — rectangular array where columns are the variables and rows are the observations.
2. Descriptor information — contains the number of observations, the size of the observations, date last modified, formats, labels, variable names, etc.

Temporary SAS datasets disappear when the job finishes running. Permanent SAS datasets are stored on disk and so are available for future SAS jobs without the need for a data step. Every SAS dataset has a library name in addition to the dataset name we have been using. Library names "WORK.*" are temporary. Change this with "DATA EX1HW1.STATES" which is LIBNAME.DATASETNAME. On CM, this file shows up as "STATES EX1HW1."

7.1.6 Options when Reading Datasets

Suppose you want to modify a SAS dataset that already exists. SAS dataset options can be used whenever the name of a dataset is invoked. They appear in parentheses directly following the dataset name.

- DROP — specifies variables to be dropped from the dataset(DROP = <var1> <var2> ...).
- KEEP — specifies variables to be kept in the dataset(KEEP = <var1> <var2> ...).
- RENAME — change the variable name in the dataset(RENAME = (OLDNAME = NEWNAME)).
- LABEL — gives a label to a dataset(LABEL = 'US PRESIDENTS').
- OBS and FIRSTOBS — (FIRSTOBS = n) specifies the first observation to be read. (OBS = n) specifies the last observation to read.
- WHERE — specifies a condition for an observation to be read into a dataset(WHERE = (conditions)).

7.1.7 Creating SAS Variables Assignment

```
OPTIONS LS = 132;
DATA VALID;

INFILE 'VALID DAT A';
INPUT ID 1-6 SATFATDR 8-15 SATFATFF 17-24 TOTFATDR 26-33
      TOTFATFF 35-42 ALCONDR 44-51 ALCONFFQ 53-60
      TOTCALDR 62-70 TOTCALFF 72-80;

DIFDRFFQ = SATFATDR - SATFATFF;
DIFTOT = TOTFATDR - TOTFATFF;
DIFFALCON = ALCONDR - ALCONFFQ;
DIFCAL = TOTCALDR - TOTCALFF;

* SUBJECT CONFORMS;

IF ((SATFATDR>=SATFATFF)AND(ALCONDR>=ALCONFFQ)AND(TOTFATDR>=TOTFATFF)
    AND(TOTCALDR>=TOTCALFF)) OR
   ((SATFATDR<=SATFATFF)AND(ALCONDR<=ALCONFFQ)AND(TOTFATDR<=TOTFATFF)
    AND(TOTCALDR<=TOTCALFF))
    THEN CONFORM = "CONFORM   ";

* WHICH SUBJECT CONFORMS?;

IF ((SATFATDR>=SATFATFF)AND(ALCONDR>=ALCONFFQ)AND(TOTFATDR>=TOTFATFF)
    AND(TOTCALDR>=TOTCALFF))
    THEN IDNT = "DR   ";

IF ((SATFATDR<=SATFATFF)AND(ALCONDR<=ALCONFFQ)AND(TOTFATDR<=TOTFATFF)
    AND(TOTCALDR<=TOTCALFF))
    THEN IDNT = "FFQ  ";

...etc...
```

7.2 PROC CONTENTS Assignment

PROC CONTENTS is used to display the descriptor information associated with every SAS dataset.

```
OPTIONS LS = 132;
DATA HW5.RATIOS;
```

```
SET HW5.VALID;

RATDRFFQ = SATFATFF/SATFATDR;
RATTOT = TOTFATFF/TOTFATDR;
RATALCON = ALCONFFQ/ALCONDR;
RATCAL = TOTCALFF/TOTCALDR;

LABEL RATDRFFQ = 'RATIO OF DR:FFQ FOR FAT'
      RATTOT = 'RATIO OF TOTAL FAT'
RATALCON = 'RATIO OF DR:FFQ FOR ALCOHOL'
RATCAL = 'RATIO OF TOTAL ALCOHOL';

PROC CONTENTS DATA = HW5.RATIOS;
RUN;
```

7.3 PROC MEANS

Here is a SAS example using PROC MEANS:

```
OPTIONS LS = 72;

DATA STATES;
INFILE 'STATES1 RAW A';
INPUT NAME $ 4-19 HIGHTEMP 25-27 LOWTEMP 29-31
      POP80 POPSQMI FARMPCT PCTURB @83 PCAPINC 5.;

URB_POP = (PCTURB*POP80)/100;
TEMPRNGE = HIGHTEMP - LOWTEMP;
IF ((FARMPCT>50)&(PCAPINC>7000)) THEN RICHFARM = 'YES';
ELSE RICHFARM = 'NO';

FORMAT POP80 URB_POP COMMALL. FARMPCT PCTURB 6.2 PCAPINC DOLLAR8.;
LABEL NAME = 'STATE NAME'
      HIGHTEMP = 'HIGEST RECORDED TEMPERATURE'
LOWTEMP = 'LOWEST RECORDED TEMPERATURE'
POP80 = 'POPULATION, 1980'
POPSQMI = 'POPULATION DENSITY, 1980'
FARMPCT = '% LAND DEVOTED TO AGRICULTURE'
PCTURB = '% URBAN POP(AREAS OVER 2,500 POP)'
PCAPINC = '1980 PER CAPITA INCOME'
URB_POP = 'URBAN POPULATION, 1980'
TEMPRNGE = 'TEMPERATURE RANGE';
```

```
RUN;

PROC MEANS DATA=STATES MAXDEC=1;
VAR HIGHTEMP LOWTEMP TEMPRNGE POP80;
RUN;

PROC SORT DATA=STATES;
BY RICHFARM;
RUN;

PROC MEANS DATA=STATES MAXDEC=1 N MEAN RANGE STD;
CLASS RICHFARM;
VAR HIGHTEMP LOWTEMP TEMPRNGE POP80;
RUN;
```

7.3.1 PROC MEANS Assignment

We will use the permanent dataset with 19 variables that you created in the previous assignment to try out some of the PROCs we have learned.

1. It has been discovered that observations #34 and #140 were erroneously recorded. So you need to delete them from your dataset. DO this first, creating a new permanent dataset with only 171 observations, but with all 19 variables. This will be the dataset that we work with from now on. You should be able to do this with a SET statement. Do NOT start from the beginning with the raw data!

2. Print the data so that it is grouped into the three categories CONFORM, SPLIT, and NONCONF that are given by the values of the first character variable that you created in assignment 4. Use subject's id number as an ID variable. Have the PRINT also show labels, and have it tell you how many measurements are in each of the three groups. Also, show the overall sum for each of four difference variables you created in assignment 4. What do these sums suggest?

   The difference between alcohol consumption DR and alcohol consumption FFQ is neglible. The difference between total fat seems to favor the DR variable some what. The differences of total calories seems to favor the DR variable significantly. The DR-FFQ difference sum is positive indicating favoring DR.

3. Now use the MEANS procedure to get the same information obtained with PRINT above, and obtain means as well as sums. However, not

get the values only for the four difference variables. Also, run a t-test for significance of the four difference variables, including p-values for the t-tests(the t-tests are for all of the data, not the grouped data). All of this can be done with two seperate PROC MEANS. What do you conclude from the t-tests?

The DR-FFQ differences, differences of total fat and differences of total calories are highly significant. The alcohol differences is negligible.

```
OPTIONS LS = 130;
DATA HW5.MISSING;
SET HW5.RATIOS;

IF (_N_ = 34)OR(_N_ = 140) THEN DELETE;

PROC SORT;
BY CONFORM;

PROC PRINT DATA = HW5.MISSING L N;
VAR SATFATDR SATFATFF TOTFATDR TOTFATFF ALCONDR ALCONFFQ
    TOTCALDR TOTCALFF DIFDRFFQ DIFTOT DIFALCON DIFCAL IDNT;
ID ID;
BY CONFORM;
SUM DIFDRFFQ DIFTOT DIFALCON DIFCAL;
TITLE 'FOOD FREQUENCY QUESTIONAIRE';
RUN;

PROC MEANS DATA=HW5.MISSING N MEAN SUM;
BY CONFORM;
ID ID;
VAR DIFDRFFQ DIFTOT DIFALCON DIFCAL;
TITLE 'FOOD FREQUENCY GROUPED BY CONFORMITY';
RUN;

PROC MEANS DATA=HW5.MISSING T PRT MEAN SUM;
VAR DIFDRFFQ DIFTOT DIFALCON DIFCAL;
TITLE 'FOOD FREQUENCY STUDENT T TESTS OF DIFFERENCES';
RUN;
```

7.4 PROC UNIVARIATE

The options for PROC UNIVARIATE are:

- DATA = <dataset> .
- FREQ — produces a frequency table for the data.
- NORMAL — test for normality of the data.
- PLOT — stem/leaf plot, boxplot, and Q-Q plot.
- NOPRINT — makes the procedure run, but does not print anything.

Here are 3 examples using PROC UNIVARIATE:

```
* A MANUFACTURER OF MICROWAVE OVENS TESTS PERIODICALLY
  FOR THE AMOUNT OF RADIATION THEY OMIT. A SAMPLE OF 42
  OVENS IS OBTAINED, AND THE EMISSIONS(WITH DOOR CLOSED)
  MEASURED FOR EACH. CAN IT BE CONCLUDED THAT THE AVERAGE
  EMMISSION FALLS BELOW THE STANDARD OF .15 UNITS?;

OPTIONS NODATE LINESIZE=72 PAGESIZE=40;

DATA MICRO;
INFILE 'MWOVEN DAT A';
INPUT EMIT;
EXCESS=EMIT-.15;
LABEL EMIT = 'RADIATION EMITTED WITH DOOR CLOSED'
      EXCESS = 'UNITS OF RADIATION OVER .15';

PROC UNIVARIATE DATA=MICRO FREQ PLOT NORMAL;
VAR EXCESS;
TITLE 'LEFTOVERS AGAIN?';
RUN;

*THE DATA USED HERE IS FROM A STUDY OF ANESTHETICS PERFORMED
 ON DOGS. EACH DOG WAS GIVEN THE DRUG PENTOBARBITOL. TWO
 MEASUREMENTS WERE THEN TAKEN, ONE WITH HALOTHANE
 ADMINISTERED, AND ONE WITHOUT. THE MEASURES ARE
 MILLISECONDS BETWEEN HEARTBEATS. DOES USE OF HALOTHANE
 MAKE FOR A MORE EFFECTIVE ANESTHETIC?;

OPTIONS NODATE LINESIZE=72 PAGESIZE=40;
```

```
DATA DOGS;
INFILE 'DOG DAT A';
INPUT X1 X2;
DIFF = X1-X2;
LABEL X1 = 'RESPONSE WITHOUT HALOTHANE'
      X2 = 'RESPONSE WITH HALOTHANE'
DIFF = 'DIFFERENCE OF RESPONSES';

PROC UNIVARIATE DATA=DOGS FREQ PLOT NORMAL;
VAR DIFF;
TITLE 'LET SLEEPING DOGS LIE';
RUN;

 DATA STATES;
 INFILE 'STATES1 RAW A';
 INPUT NAME $ 4-19 HIGHTEMP 25-27 LOWTEMP 29-31
        POP80 POPSQMI FARMPCT PCTURB
        @83 PCAPINC 5.;
 URB_POP = (PCTURB * POP80) / 100;
 TEMPRNGE = HIGHTEMP - LOWTEMP;
 IF  ((FARMPCT>50)&(PCAPINC>7000)) THEN RICHFARM = 'YES';
 ELSE RICHFARM = 'NO';

 FORMAT POP80 URB_POP COMMA11. FARMPCT PCTURB 6.2 PCAPINC DOLLAR8.;
 LABEL NAME = 'State name"
       HIGHTEMP = 'Highest recorded temperature'
       LOWTEMP = 'Lowest recorded temperature'
       POP80 = 'Population, 1980'
       POPSQMI = 'Population density, 1980'
       FARMPCT = '% land devoted to agriculture'
       PCTURB = '% urban pop (areas over 2,500 pop)'
       PCAPINC = '1980 per captita income'
       URB_POP = 'Urban population, 1980'
       TEMPRNGE = 'Temperature range'
       ;
 RUN;


PROC SORT DATA=STATES;
BY RICHFARM;

PROC UNIVARIATE DATA=STATES FREQ PLOT NORMAL;
```

```
VAR POP80;
BY RICHFARM;
ID NAME;
RUN;
```

7.4.1 PROC UNIVARIATE Assignment

This week we analyze aspects of our favorite dataset using PROC UNIVARIATE. Use the permanent dataset with 19 variables and 171 observations that you created in the previous assignment.

1. For each of the four difference variables created, use UNIVARIATE to examine the shapes of their distributions. Do any of them appear to be skewed? Do they appear to be reasonably normal? What is the result of the test for normality? Also, what do the sign test, the sign rank test, and the t-test tell you? Do you prefer the nonparametric tests over the t-test for any of the four variables? Why or why not?

   When the data is not normally distributed, use the non-parametric tests. When the data is normally distributed, use the t-tests. The hypotheses are $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Reject H_0 for small p-values.

2. Repeat what you have done for (1), answering the same questions, but use the ratio variables created in the previous assignment. This will let you compare the two methods from a different perspective. However, first subtract 1 from every ratio variable. What is the purpose of this?

   We subtracted 1 from each ratio variable so the test $\mu = 0$ will match with what SAS tests for.

```
OPTIONS LINESIZE = 72;

DATA RATIOS.MINUS1;
SET HW5.MISSING;

RAT1 = RATDRFFQ - 1;
RAT2 = RATTOT - 1;
RAT3 = RATALCON - 1;
RAT4 = RATCAL - 1;

LABEL RAT1 = 'RATIO OF DR/FFQ MINUS 1'
```

```
RAT2 = 'RATIO OF TOTAL FAT MINUS 1'
RAT3 = 'RATIO OF ALCOHOL MINUS 1'
RAT4 = 'RATIO OF CALORIES MINUS 1';

PROC UNIVARIATE DATA=HW5.MISSING NORMAL PLOT;
VAR DIFDRFFQ DIFTOT DIFALCON DIFCAL;
ID ID;
TITLE 'ANALYSIS OF DIFFERENCES';
RUN;

PROC UNIVARIATE DATA=RATIOS.MINUS1 NORMAL PLOT;
VAR RAT1 RAT2 RAT3 RAT4;
ID ID;
TITLE 'ANALYSIS OF RATIOS';
RUN;
```

7.5 PROC FREQ and PROC CHART

The following code will make a 2D table of frequencies:

```
PROC FREQ;
TABLES HEIGHT*GRADE;
```

The syntax of PROC CHART is:

```
PROC CHART;
VBAR <FIELD1> /GROUP = <FIELD2> TYPE = PERCENT G100;
```

Here is an example using PROC FREQ and PROC CHART:

```
PROC FREQ DATA=PRES.US ORDER = FREQ;
TABLES PARTY REGION/NOCUM;
TABLES BORNST;
RUN;

PROC CHART DATA=PRES.US(WHERE(INAUG>1866));
HBAR PARTY;
VBAR PARTY/GROUP=REGION TYPE=PERCENT;
VBAR PARTH/TYPE=MEAN SUMVAR=AGEDEATH;
VBAR AGEINQUG/MIDPOINTS=45 TO 70 BY 5 NOSPACE;
RUN;
```

7.5.1 PROC FREQ/PROC CHART Assignment

This week we analyze aspects of the FFQ/DR dataset using PROC FREQ and PROC CHART. Use the permanent dataset with 19 variables and 171 observations.

1. Make a frequency table that shows how many individuals are in each of the CONFORM, NONCONF, and SPLIT categories.

2. Make a chart that shows the same information in (1).

3. Make a chart that shows for each of the CONFORM and NONCONF categories, how many individuals have more of the 4 measurements higher using the FFQ method, and how many have more than the 4 higher using the DR method. To make this chart you need only use the two character variables.

4. Based on the charts above, does it appear that, overall for the four variables(saturated fat, total fat, alcohol, and total calories) being measured, the FFQ and DR methods are comparable? Explain.

5. Thinking back to the results from the previous two assignments, do you think the FFQ method is a suitable substitute for the DR method for any of the four variables? Explain.

```
PROC FREQ DATA=HW5.MISSING ORDER = FREQ;
TABLES CONFORM;
TITLE 'FREQUENCY TABLE OF FOOD QUESTIONNAIRE';
RUN;

PROC CHART DATA=HW5.MISSING;
HBAR CONFORM;
TITLE 'BAR CHART OF FOOD QUESTIONNAIRE';
RUN;


PROC CHART DATA=HW5.MISSING(WHERE=((CONFORM='CONFORM')OR
                                   CONFORM='NONCONFORM')));
VBAR CONFORM/GROUP = IDNT;
TITLE 'BAR CHART OF FOOD QUESTIONNAIRE BY DR AND FFQ';
RUN;
```

7.6 Combining Datasets and Handling Dates

1. Stacking — adding to the bottom of a dataset. Also called concatenation. Purpose: to create a dataset that contains all measurements from dataset 1, followed by all measurements from dataset2, and so on. Syntax: DATA —–; SET DATASET1(OPTIONS) DATASET2(OPTIONS) ....

2. Interleaving — similar to concatenation except that observations from one dataset are not all above or below those of another dataset. Instead, they are put in the new dataset in an order specified by a certain variable which we shall refer to as the "by" variable. Syntax: DATA —–; SET DATASET1 DATASET2 ...; BY VARNAME; The "by" variable must be in both datasets.

3. Matched Merge — find matching id's and add data to that line(2nd dataset overwrites the first data set). Syntax: DATA —–; MERGE DATASET1 DATASET2 ...; BY VARNAME; If there are no matching variable names in the 2 datasets, then a new column is created.

SAS provides an easy way of addressing the question: how long something took given 2 dates. When reading the data, do 1) read date values with a special date input format, 2) SAS stores the date as a number which is the number of days before or after January 1, 1960.

7.6.1 Assignment

1. The raw data files that you need for this assignment are named HW9DAT ONE and HW9DAT TWO. They are located on the account F3525605@ODUVM.CC.ODU.EDU. Use ftp to retrieve theses files. The password for this account is ORIORLE.

2. The first file contains measurements on three variables: subject's id number, subject's date of birth, and date on which the subject entered the study. The second file contains measurements on two variables: subject's id and date of first followup on the subject. Create a SAS dataset out of each of these raw data files. Be sure to use the correct in format to read the dates.

3. Merge the two datasets so that you have one SAS dataset with four variables for each subject. Now calculate the following three variables: 1) age, to the nearest .1 years, at which the subject entered the study, 2) the number of weeks, to the nearest week, between entry and first followup, and 3) the age-group to which the subject belonged at time

of entry into the study. To do the last, use these three age groups: <= 29.9 years, 30.0 - 40.9 years, >= 50.0 years. Print this dataset.

4. Use a frequency table to display how many subjects are in each age group.

```
DATA MEASURE1.DAT;
INFILE 'HW9DAT ONE A';
INPUT @1 ID 5. @7 BIRTH MMDDY8. @16 ENTRY MMDDYY8.;
LABEL ID = 'IDENTIFICATION'
BIRTH = 'DATE BORN'
ENTRY = 'DATE ENTERED STUDY';

PROC SORT; BY ID;

DATA MEASURE2.DAT;
INFILE 'HW9DAT TWO A';
INPUT @1 ID 5. @7 FOLLOW MMDDYY8.;

PROC SORT; BY ID;

LABEL FOLLOW = 'DATE OF FIRST FOLLOW-UP'
ID = 'IDENTIFICATION';

DATA MERGED.BOTH;
MERGE MEASURE1.DAT MEASURE2.DAT;
BY ID;


ENTRYAGE = (ENTRY-BIRTH)/365.25;
WEEKS = (FOLLOW-ENTRY)/7;
IF ENTRYAGE<=29.9 THEN CAT = 'UNDER 29.9 YEARS OLD';
ELSE
IF ENTRYAGE>=50 THEN CAT = 'OVER 50 YEARS OLD;
ELSE CAT = 'BETWEEN 30 AND 49.9 YEARS OLD';

LABEL ENTRYAGE = 'AGE OF ENTRY'
WEEKS = '# WEEKS BETWEEN ENTRY AND FOLLOW-UP'
CAT = 'AGE GROUP';

PROC PRINT L;
ID ID;
```

```
FORMAT BIRTH MMDDYY8. ENTRY MMDDYY8. FOLLOW MMDDYY8.
       ENTRYAGE 5.1 WEEKS 5.1;
VAR BIRTH ENTRY FOLLOW ENTRYAGE WEEKS CAT;
TITLE 'STUDY GROUP';
RUN;

PROC FREQ ORDER=FREQ;
TABLES CAT;
TITLE 'FREQUENCY TABLE BY AGE GROUP';
RUN;
```

7.7 PROC FORMAT

PROC FORMAT controls the formatting of printing on paper. PROC FORMATS are user defined. The syntax is

```
PROC FORMAT;
VALUE FMTNAME
RANGE1 = 'TEXT'
RANGE2 = 'TEXT'
...
RANGEN = 'TEXT';
```

FMTNAME is the name of you format. It can be 1–8 characters. It cannot begin with #. If the name is a format for a character variable, it must begin with $.

7.7.1 PROC FORMAT Assignment

In this assignment you will make a new SAS dataset, create formats for some of the variables, and use those formats in examining relationships among a few variables.

1. The raw data file is HSB DAT, which will be sent to your account. The data is described on the accompanying sheet.

2. Create a SAS dataset containing the 15 variables. Use the same names as shown on the handout.

3. Make formats for the following variables, so that their numeric values are replaced by the character strings shown on the handout: SES, SCTYP, HSP.

4. Using your formats, make the following two-way tables: SES vs SCTYP and SES vs HSP. In each table, make SES the row variable, and show percentages in the rows only.

5. Do you see an apparent relationship between SES and SCTYP? What is the nature of that relationship? Explain.

6. Do you see an apparent between SES and HSP? What is the nature of that relationship? Explain.

```
DATA SCHOOL;
INFILE 'HSB DAT A';
INPUT ID SEX RACE SES SCTYP HSP LOCUS CONCPT MOT CAR
RDG WRTG MATH SCI CIV;

PROC FORMAT;
VALUE SESFORM
1 = 'LOWER'
2 = 'MIDDLE'
3 = 'UPPER';

VALUE SCHFORM
1 = 'PUBLIC'
2 = 'PRIVATE';

VALUE HSPFORM
1 = 'GENERAL'
2 = 'ACADEMIC'
3 = 'VOCATIONAL';
RUN;

PROC FREQ;
TABLES SES*SCTYP SES*HSP/ NOCOL NOPERCENT;
FORMAT SCTYP SCHFORM. HSP HSPFORM. SES SESFORM.;
TITLE 'SOCIO-ECONOMIC STATUS VS SCHOOL TYPE AND HS PROGRAM';
RUN;
```

7.8 Data Analysis PROCS

PROC TTEST can be used to perform a t-test on a mean response. The null hypothesis is that $\mu = 0$.

```
PROC TTEST DATA = <DATASET NAME.;
CLASS <FIELDNAMES1>;
VAR <FIELDNAMES2>;
RUN;
```

PROC CORR performs an Spearman correlation analysis. The hypotheses are H_0 : correlation is zero, H_1 : correlation is not zero. PROC CORR is used to look at the relationship between two variables. PROC REG is used to perform a regression analysis. Type I sums of squares tells the significance of a variable given the previous ones are already in the model.

7.8.1 Assignment

```
DATA SCHOOL;
INFILE 'HSB DAT A';
INPUT ID SEX RACE SES SCTYP HSP LOCUS CONCPT
MOT CAR RDG WRTG MATH SCI CIV;

LABEL ID = 'ID NUMBER'
SEX = 'SEX'
RACE = 'RACE'
SES = 'SOCIO-ECONOMIC STATUS'
SCTYP = 'SCHOOL TYPE'
HSP = 'HIGH SCHOOL PROGRAM'
LOCUS = 'LOCUS OF CONTROL'
CONCPT = 'SELF CONCEPT'
MOT = 'MOTIVATION'
CAR = 'CAREER CHOICE'
RDG = 'READING T-SCORE'
WRTG = 'WRITING T-SCORE'
MATH = 'MATH T-SCORE'
SCI = 'SCIENCE T-SCORE'
CIV = 'CIVICS T-SCORE';

PROC FORMAT;
VALUE SESFORM
1 = 'LOWER'
2 = 'MIDDLE'
3 = 'UPPER';

VALUE SCHFORM
1 = 'PUBLIC'
```

```
2 = 'PRIVATE';

VALUE HSPFORM
1 = 'GENERAL'
2 = 'ACADEMIC'
3 = 'VOCATIONAL';

VALUE MATHFORM
LOW-25 = 'LOWER 25%'
25.0001-50 = '25%-50%'
50.0001-75 = '50%-75%'
75.0001-HIGH = 'TOP 25%';
RUN;

PROC CHART;
VBAR MATH/GROUP=SCTYP TYPE=PERCENT G100;
FORMAT SCTYP SCHFORM.;
TITLE 'BAR CHART OF SCHOOL TYPE VS MATH SCORES';
RUN;

PROC TTEST;
CLASS SCTYP;
VAR MATH;
TITLE 'T-TESTS OF SCHOOL TYPE AND MATH SCORES';
RUN;

PROC SORT; BY SCTYP;

PROC UNIVARIATE PLOT;
BY SCTYP;
VAR MATH;
FORMAT SCTYP SCHFORM.;
TITLE 'PLOT OF SCHOOL TYPE AND MATH SCORES';
RUN;

PROC CHART;
VBAR MATH/GROUP=SES SUBGROUP=SCTYP TYPE=PERCENT G100;
FORMAT SES SESFORM. SCTYP SCHFORM.;
TITLE 'BAR CHART OF SOCIO-ECONOMIC STATUS/SCHOOL
 TYPE AND MATH SCORES';
RUN;
```

```
PROC GLM;
CLASS SCTYP SES;
MODEL MATH = SCTYP SES SCTYP*SES;
MEANS SCTYP SES SCTYP*SES;
TITLE 'MODEL OF MATH SCORES AS A FUNCTION OF SOCIO-
 ECONOMIC STATUS AND SCHOOL TYPE';
RUN;
```

Chapter 8

Linear Regression

Dr. Morgan, Old Dominion University
Statistics 537, Fall 1996
Text used: Linear Statistical Models, an Applied Approach, 2nd edition, Bruce L. Bowerman and Richard T. O'Connell, Duxbury Press, 1990.

8.1 Simple Linear Regression

The simplest form of the technique known as regression is that of predicting a single variable y from a single variable x using a straight line relationship. Consider this imaginary example: y is a child's weight and x is the average of the parent's heights. We obtain n measurements on x and y. Call them (x_1, y_1), (x_2, y_2), ... (x_n, y_n). We postulate a *linear model* relating the mean of y_i to x_i.

$$y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where μ_i is the mean of y_i for all children whose parents' average height is x_i. ϵ_i is a random variable representing random variance of y_i from μ_i. $\mu_i = \beta_0 + \beta_1 x_i$ is the linear model for μ_i. β_0 is the intercept and β_1 is the slope.

Important Note: We do not know if this model is true, and in a real sense, no model is ever true. Models are mathematical approximations of reality, and we hope to find one that does a good job of approximation. We are now entertaining a simple linear model, which we will "fit(estimate)." Later, we will learn how to measure the adequacy of the model. First, we

learn how to fit the model. We will use the data $(x_1, y_1), ...$ to find estimates b_0 of β_0 and b_1 of β_1. They in turn will give

$$\widehat{y} = b_0 + b_1 x.$$

$\widehat{y}$ is the predicted value of y for a given x. To find b_0 and b_1 we us the technique of *least squares.* Here is the idea: out model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. We choose the values b_0 and b_1 that makes the deviations "small." Specifically, we will minimize the sum of squared deviations.

$$s = \sum_{i=1}^{n} \epsilon_i^2 =$$

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

β_0 moves the line up and down. β_1 changes the tilt of the line.

$$\frac{\partial s}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial s}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i.$$

Set these two equations equal to zero and solve for β_0 and β_1.

$$\left.\begin{array}{l} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{array}\right\}$$

$$\left.\begin{array}{l} \sum_{i=1}^{n} y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \end{array}\right\}$$

Then,

$$\left.\begin{array}{l} b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{array}\right\}$$

The above two equations are called *normal equations.* The solution is:

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right) \left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} =$$

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $\bar{x}$ is the average of the x's and $\bar{y}$ is the average of the y's. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

$b_0 = \bar{y} - b_1\bar{x}$. Substituting back into the model equation gives:

$$\widehat{y} = b_0 + b_1 x = \bar{y} - b_1\bar{x} + b_1 x =$$

$$\bar{y} + b_1(x - \bar{x}).$$

We now have a line. We next need to find methods for judging how well this line "fits the data." Consider the identity

$$y_i - \bar{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y}).$$

Square both sides:

$$(y_i - \bar{y})^2 =$$

$$(y_i - \widehat{y}_i)^2 + (\widehat{y}_i - \bar{y})^2 + 2(y_i - \widehat{y})(\widehat{y}_i - \bar{y}).$$

Sum both sides:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 =$$

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + (\widehat{y}_i - \bar{y})^2 + 2(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}).$$

Working with the cross product terms only:

$$2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}) =$$

$$2\sum_{i=1}^{n}(y_i - \widehat{y}_i)(b_1(x_i - \bar{x})) =$$

$$2b_1(y_i - \widehat{y}_i)(x_i - \bar{x}) =$$

$$2b_1\sum_{i=1}^{n}[y_i - \bar{y} - b_1(x_i - \bar{x})](x_i - \bar{x}) =$$

$$2b_1\left[\sum_{i=1}^{n}(y_1 - \bar{y})(x_i - \bar{x}) - b_1\sum_{i=1}^{n}(x_i - \bar{x})^2\right] =$$

$$2b_1 \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 0.$$

Hence,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 =$$

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2.$$

Page 117 in the textbook shows b_1 and b_0 derived.

$$SS(TOTAL) = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

$$SS(ERROR) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

We want $SS(ERROR)$ to be small.

$$SS(REGRESSION) = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2.$$

We want $SS(REGRESSION)$ to be large. The proportion of total variation explained by the model is

$$R^2 = \frac{SS(REGRESSION)}{SS(TOTAL)} = 1 - \frac{SS(ERROR)}{SS(TOTAL)}.$$

This quantity R^2 is used in any regression model, not just the simple linear regression.

$$R^2 = \frac{SS(Regression)}{SS(TOTAL)} =$$

$$\frac{\sum_{i=1}^{\infty}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{\infty}(y_i - \bar{y})^2} =$$

$$\frac{\sum_{i=1}^{n}[\bar{y} - b_1(x_i - \bar{x}) - \bar{y}]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} =$$

$$\frac{b_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} =$$

$$\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2 \sum_{i=1}^{n}(x_i-\bar{x})^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2 \sum_{i=1}^{n}(y_i-\bar{y})^2} =$$

$$\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}\right]^2 =$$

$$r^2,$$

where r is Pierson's correlation coefficient which is a simple coefficient of determination(see Section 5.3 of the text book).

SAS Code

```
DATA REPAIR;
INPUT N_UNITS MINUTES;
CARDS;
1 23
2 29
3 49
4 64
4 74
5 87
6 96
6 97
7 109
8 119
9 149
9 145
10 154
10 166;

PROC PLOT;
PLOT MINUTES*N_UNITS;
RUN;

PROC REG;
MODEL MINUTES = N_UNITS/P;
OUTPUT OUT=NEW P=PRED R=RESID;
RUN;

PROC PLOT;
PLOT MINUTES*N_UNITS='O' PRED*N_UNITS='*'/OVERLAY;
```

```
PLOT RESID*N_UNITS='*'/VREF=0;
RUN;
```

8.1.1 Point Estimates

The model is $y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$. The *mean response* μ_0 when $x = x_0$ is $\beta_0 + \beta_1 x_0$. An estimate of the mean response when $x = x_0$ is $\widehat{y}_0 = b_0 + b_1 x_0$. A *particular response* when $x = x_0$ is $\beta_0 + \beta_1 x_0 + \epsilon_0$. So, our prediction of the response of a new measurement at $x = x_0$ is $\widehat{y}_0 = b_0 + b_1 x_0$. This matches the main response estimate because the best estimate of ϵ_0 is zero. However, this estimate is subject to more variability than that of the mean response.

8.1.2 Homework and Answers

1. The chairman of the Accounting Department at a large university undertakes a study to relate starting salary(y) after graduation for accounting majors to grade point average(GPA) in major courses. To do this, records of seven recent accounting graduates are randomly selected.

Accounting Graduate i	GPA x_i	Starting Salary, y_i (thousands of dollars)
1	3.26	28.2
2	2.60	24.8
3	3.35	27.9
4	2.86	25.3
5	3.82	30.3
6	2.21	23.0
7	3.47	29.4

   (a) Plot y versus x. Explain why the plot suggests that the simple linear regression model having a positive slope

   $$y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

   might appropriately relate y to x.

   The plot of x versus y appears on the next page. Looking at the data points, the points do appear to fall on an approximate straight line. As the GPA increases, so does the starting salary of accounting graduates. Perhaps employers think that employees

with a higher GPA are more productive. Thus, employee's with a higher GPA are compensated more than employee's with a lower GPA. This explains the positive slope. An increase in GPA means an increase in starting salary.

(b) Discuss the meaning of the third historical population of potential starting salaries.

(c) Discuss the meaning of μ_3, the third mean starting salary. μ_3 is the mean of the population defined in part (b).

(d) Discuss the conceptual difference between μ_3 and $y_3 = \mu_3 + \epsilon_3$. What does ϵ_3 measure in this situation?

y_3 is one observation of one individual with GPA 3.35. ϵ_3 measures how far this person's salary fell from μ_3.

(e) Discuss the meaning of β_0 and β_1 in this model. Why does the interpretation of β_0 fail to make practical sense?

Again, β_0 is the Y-axis intercept. In this study, β_0 shows the starting salary of the person with a GPA of 0.00. β_0 fails to make practical sense. No data was gathered of people who did not go to college. These people's GPA would be zero. Yet, this is not reflected in this study. β_1 represents the rise in starting salary per unit rise in GPA.

(f) Calculate the least squares point estimates b_0 and b_1 of β_0 and β_1.

$$b_1 = \frac{7\sum_{i=1}^{7} x_i y_i - \left(\sum_{i=1}^{7} x_i\right)\left(\sum_{i=1}^{7} y_i\right)}{7\sum_{i=1}^{7} x_i^2 - \left(\sum_{i=1}^{7} x_i\right)^2} =$$

$$\frac{7(590.829) - (21.57)(188.9)}{7(68.3071) - (21.57)^2} = 4.752111.$$

$$b_0 = \bar{y} - b_1\bar{x} = 26.98571 - (4.752111)(3.081429) = 12.3424.$$

(g) Using the prediction equation $\hat{y}_i = b_0 + b_1 x_i$, calculate a point estimate of μ_3 and a point prediction of $y_3 = \mu_3 + \epsilon_3$.

$$\mu_3 = \beta_0 + \beta_1 x_3 = 12.34242 + (4.752111)(3.35) = 28.262.$$

$$y_3 = \mu_3 + \epsilon_3 = 28.262 + 0.0 = 28.262.$$

(h) Suppose that an accounting major will graduate with a GPA of $x_0 = 3.26$. The starting salary of this graduate may be expressed in the form $y_0 = \mu_0 + \epsilon_0$.

i. Discuss the conceptual difference between μ_0 and $y_0 = \mu_0 + \epsilon_0$.
μ_0 is the mean starting salary of all accounting majors when the GPA $x_0 = 3.26$. y_0 is the future starting salary of a future graduate. There is not an actual data point that corresponds to such a GPA for the starting salary.

ii. Is the grade point average $x_0 = 3.25$ in the experimental region?
The experimental region ranges from 2.21 to 3.82. 3.25 is in the experimental region.

iii. Using an appropriate prediction equation, calculate a point estimate of μ_0 and a point prediction of y_0.
$\mu_0 = \beta_0 + \beta_1(3.25)$.
$\hat{y}_0 = b_0 + b_1 x_0 = 12.34242 + (4.752111)(3.25) = 27.87$.

(i) Calculate SSE, s^2, and s.

$$SSE = \sum_{i=1}^{7} \epsilon_i^2 = 1.061178.$$

$$s^2 = \frac{SSE}{n-2} = \frac{1.061178}{5} = 0.21224.$$

$$s = \sqrt{s^2} = 0.461.$$

2. The no-intercept model for the simple linear regression line is

$$Y_i = \beta X_i + \epsilon_i, i = 1, 2, 3, ..., n.$$

This is sometimes appropriate when the line *must* pass through the origin $(0, 0)$. For this model,

(a) Find the least squares estimate of β. (That is, find the value b for β that minimizes the sum of squared errors for this model).

$$s^2 = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta X_i)^2.$$

$$\frac{\partial s}{\partial \beta} = -2 \sum_{i=1}^{n} (Y_i - \beta X_i) X_i = 0$$

$$\sum_{i=1}^{n} (Y_i - \beta_i X_i) X_i = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \beta X_i^2 = 0$$

$$b = \widehat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

(b) Now that you have the estimate, find its variance.

$$Var(b) = Var\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) =$$

$$\sum_{i=1}^{n} \left(\frac{X_i}{\sum_{i=1}^{n} X_i^2}\right)^2 \sigma^2 =$$

$$\frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}.$$

8.1.3 Model Assumptions and MS(E)

So far we said nothing about the probability distribution. We have *no basis* for running statistical tests of constructing confidence intervals. To perform these tasks, we require some *distribution assumptions.* The assumptions are:

1. ϵ_i is a random variable with mean zero and variance σ^2. That is $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma^2$. ϵ_i and ϵ_j are uncorrelated.
2. $E(y_i) = \beta_0 + \beta_1 x_i$, and $Var(y_i) = \sigma^2$.
3. ϵ_i has a normal distribution.

In Chapter 6 of the text book, we will learn how to check the validity of these assumptions. In Chapter 5 of the text book, we will use these to construct confidence intervals. In building confidence intervals, we will require a point estimate of the underlying variance σ^2. It can be shown that $E(SSE) = (n-2)\sigma^2$ using assumptions 1 and 2. Hence, a point estimate for σ^2 is

$$s^2 = \frac{SS(E)}{n-2} = MS(E)$$

called the *mean square error.* The *residual* is the difference between the observed value of y and the predicted value of y.

A FACT: $w_1, w_2, ..., w_k$ are random variables. $c_1, c_2, ..., c_k$ are constants. Then,

$$Var\left(\sum_{i=1}^{k} c_i w_i\right) =$$

$$\sum_{i=1}^{k} c_i^2 Var(w_i) + \sum_{i=1}^{k}\sum_{j=1}^{k} c_i c_j Cov(w_i, w_j), i \neq j.$$

If the w_i's are uncorrelated , then

$$Var\left(\sum_{i=1}^{k} c_i w_i\right) = \sum_{i=1}^{k} c_i^2 Var(w_i)$$

since the covariances are zero. If in addition, $c_i = \frac{1}{k}, \forall i$ and $Var(w_i) = \sigma^2, \forall i$, then

$$Var\left(\sum_{i=1}^{k} c_i w_i\right) =$$

$$\sum_{i=1}^{k} \left(\frac{1}{k}\right)^2 \sigma^2 = \frac{\sigma^2}{k}.$$

8.1.4 Inference in Simple Linear Regression

Inference for $\beta_.$ The estimate for β_1 is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} =$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} =$$

$$\sum_{i=1}^{n} c_i w_i$$

where $c_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}$ and $w_i = y_i$. The variance of this estimate is

$$Var(b_1) = Var\left(\sum_{i=1}^{n} c_i w_i\right) =$$

$$\sum_{i=1}^{n} c_i^2 Var(w_i) + \sum\sum Cov(w_i, w_j) =$$

$$\sum_{i=1}^{n} c_i^2 \sigma^2 + 0 = \sum_{i=1}^{n} c_i \sigma^2 =$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}\sigma^2 =$$

$$\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The standard deviation is $SD(b_1) = \sigma\sqrt{c_{11}}$. So the estimated standard deviation of b_1 is

$$s\sqrt{c_{11}} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

This result is called the *standard error* of b_1. Our $(1-\alpha)100\%$ confidence interval for β_1 is

$$b_1 \pm t_{\alpha/2}(n-2)SE(b_1) =$$

$$b_1 \pm t_{\alpha/2}(n-2)s\sqrt{c_{11}}.$$

That appears at the bottom of page 145 in the text book.

Example: The computer repair handout. $15.51 \pm 1.782(0.505)$ the degrees of freedom is 12. A 90% confidence interval is $15.51 \pm 0.90 = [14.61, 16.41]$. The test for β_1 is as follow: $H_0 : \beta_1 = \beta_1^0$ vs $H_1 : \beta_1 \neq \beta_1^0$. Reject H_0 if

$$t = b_1 = \frac{\beta_1^0}{SE(b_1)} > t_{\alpha/2}(n-2)$$

or

$$t = b_1 = \frac{\beta_1^0}{SE(b_1)} < t_{\alpha/2}(n-2).$$

The most widely used and most important application of the test is $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. Lets test β_1 of the computer repair with $\alpha = 0.05$. Then, $t_{\alpha/2}(n-2) = t_{0.025}(12) = 2.179$. Reject H_0 if the test statistic $t > 2.179$ or $t < -2.179$.

$$t = \frac{b_1}{SE(b_1)} = \frac{15.509}{0.505} = 30.711 > 2.179.$$

Therefore, reject H_0. Conclude that $\beta_1 \neq 0$. x does contribute to the prediction of y.

8.1.5 Inference for β_0

Using reasoning similar to β_1, we can derive the following quantities and expressions for b_0.

$$Var(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] = \sigma^2 c_{00}.$$

$SD(b_0) = \sigma^2\sqrt{c_{00}}$. The standard error for b_0 is $s\sqrt{c_{00}}$. A $(1-\alpha)100\%$ confidence interval for b_0 is $b_0 \pm t_{\alpha/2}(n-2)s\sqrt{c_{00}}$. The test statistic for testing $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ is

$$t = \frac{b_0}{SE(b_0)}.$$

Reject H_0 if $t > t_{\alpha/2}(n-2)$ or $t < -t_{\alpha/2}(n-2)$.

8.1.6 Inference for y

Estimate the mean value of y at a given x. The average of y at $x = x_0$ is

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

The estimate for μ_0 is

$$\mu_0 = \widehat{y}_0 = b_0 + b_1 x_0 =$$

$$\bar{y} + b_1(x_0 - \bar{x}).$$

The variance of the estimation is

$$Var(\widehat{y}_0) = Var(\bar{y} + b_1(x_0 - \bar{x})) =$$

$$Var(\bar{y}) + Var(b_1(x_0 - \bar{x})) + 2Cov(\bar{y}, b_1(x_0 - \bar{x})) =$$

$$Var(\bar{y}) + (x_0 - \bar{x})^2 Var(b_1) + 2(x_0 - \bar{x})Cov(\bar{y}, b_1) =$$

$$\frac{\sigma^2}{n} + (x_0 - \bar{x})^2\sigma^2 c_{11} =$$

$$\sigma^2 \left(\frac{1}{n} + (x_0 - \bar{x})c_{11} \right) =$$

$$\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right) =$$

$\sigma^2 h_{00}$.

From that, we get the standard deviation $\widehat{y}_0 = \sigma\sqrt{h_{00}}$ and the standard error $s\sqrt{h_{00}}$. The $(1-\alpha)100\%$ confidence interval for μ_0 is

$$\widehat{y}_0 \pm t_{\alpha/2}(n-2)s\sqrt{h_{00}} =$$

$$[\bar{y} + b_1(x_0 - \bar{x}] \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}.$$

1. The smallest standard error is when $x_0 = \bar{x}$. The further you go from the center of the data, the worse the estimate performs.
2. By changing x_0, we can produce confidence intervals for μ_0 at every possible x. This gives confidence bands. See page 169 of the text book.
3. Confidence bands are not widely used in practice.

To predict a *new* y at $x = x_0$, we have the estimate $\widehat{y}_0 = b_0 + b_1 x_0 = \bar{y} + b_1(x_0 - \bar{x})$. The standard error is $s\sqrt{1+h_{00}} =$

$$s\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}.$$

The confidence interval is

$$\widehat{y}_0 \pm t_{\alpha/2}(n-2)s\sqrt{1+h_{00}}.$$

8.1.7 Simple Coefficients of Determination and Correlation

$$SS(TOTAL) = SS(MODEL) + SS(E).$$

SS(MODEL) is also called SS(REGRESSION).

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 =$$

$$\sum_{i=1}^{n}(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

$$R^2 = \frac{SS(MODEL)}{SS(TOTAL)} = 1 - \frac{SS(E)}{SS(TOTAL)}.$$

8.1.8 An F-Test for Simple Linear Regression

Each of the sums of squares in the identity above has a degrees of freedom associated with it and these are displayed in an ANOVA table.

Source	d.f.	SS	MS
Model	1	SS(MODEL)	SS(MODEL)/1
Error	$n-2$	SS(ERROR)	SS(E)/$(n-2)$
Total	$n-1$	SS(TOTAL)	

A *mean square* is the sum of squares divided by the degrees of freedom. The *F-ratio* provides a test of $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. It is given by:

$$\frac{MS(MODEL)}{MS(E)} = F(MODEL).$$

Reject H_0 if $F(MODEL) > F_\alpha(1, n-2)$.

8.1.9 Homework and Answers

1. Problem 5.1 in the text book on pages 196-197 — do by hand and using results from HW1.

   5.1a $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. $t_{0.025}(5) = 2.571$. The test statistic is

   $$t = \frac{b_1}{SE(b_1)} = \frac{\frac{4.752111}{\sqrt{0.21224}}}{\sqrt{1.841}} = 13.996.$$

   Since $13.996 > 2.571$, reject H_0. x does contribute to y.

   5.1b $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$. The test statistic is

   $$t = \frac{12.34242}{0.461\left(\sqrt{\frac{1}{7} + \frac{(3.081429)^2}{1.841}}\right)} = 12.59.$$

   Since $12.59 > 2.571$, reject H_0. For $\alpha = 0.01$, $t_{0.005}(7) = 4.032$. Since $4.032 < 12.59$, we can reject H_0 at a higher confidence level.

   5.1c Total variance is given by

   $$\sum_{i=1}^{7}(y_i - \bar{y})^2 = 42.63.$$

Explained variance is given by

$$\sum_{i=1}^{7}(\widehat{y}_i - \bar{y})^2 = 41.5674.$$

Unexplained variance is given by

$$\sum_{i=1}^{7}(y_i - \widehat{y}_i)^2 = 1.0612.$$

Then,

$$r^2 = \frac{41.5674}{42.63} = 0.975.$$

Then, $r = 0.987$. The r^2 means that the x and y variables have a high tendency to move together in a linear fashion. r^2 is very high. In general, r^2 is the proportion of total variance of y that is explained by the simple linear regression model.

5.1d $s^2 = \frac{SS(E)}{7-2} = \frac{1.0612}{5} = 0.21224$. Then, $s = 0.461$.

5.1e $b_1 \pm t_{0.025}(5)se(b_1) =$

$$4.752111 \pm (2.571)\left(\frac{\sqrt{0.21224}}{\sqrt{1.841}}\right) =$$

$$4.75211 \pm 0.873 = (3.88, 5.63).$$

5.1f $b_0 \pm t_{0.025}(5)s\sqrt{c_{00}} =$

$$12.34242 \pm (2.571)(0.461)\sqrt{\frac{1}{7} + \frac{(3.081429)^2}{1.841}} =$$

$$12.34242 \pm 2.73 = (9.61, 15.07).$$

5.1g $\bar{y} + b_1(x_0 - \bar{x}) \pm t_{0.025}(5)s\sqrt{h_{00}} =$

$$27.787 \pm (2.571)(0.461)\sqrt{\frac{1}{7} + \frac{0.02842}{1.841}} =$$

$$27.787 \pm 0.472 = (27.315, 28.259).$$

5.1h $\widehat{y}_0 \pm t_{0.025}(5)s\sqrt{1 + h_{00}} =$

$$27.787 \pm (2.571)(0.461)\sqrt{1 + \frac{1}{7} + \frac{0.02842}{1.841}} =$$

$$27.787 \pm 1.276 = (26.511, 29.063).$$

5.1i $F(MODEL) = \frac{MS(MODEL)}{MS(E)}$. $SS(MODEL)41.5674$. $MS(MODEL) = \frac{41.5674}{1}$.

$$F(MODEL) = \frac{\frac{41.5674}{1.0612}}{5} = 195.851.$$

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. $F_{0.05}(1, 5) = 6.61$. Reject H_0.

2. Problem 5.3 on page 197 of the text book — use SAS to obtain as many of the requested calculations as you can.

5.3a $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. $t_{\alpha/2}(n-2) = t_{0.025}(28) = 2.048$. The test statistic is

$$t = \frac{b_1}{se(b_1)} = \frac{2.665214}{0.25849959} = 10.31.$$

Since $10.31 > 2.048$, reject H_0.

5.3b $b_1 \pm t_{0.025}(28)se(b_1) =$

$$2.665 \pm (2.048)(0.25849959) = 2.665 \pm 0.529 = (2.136, 3.194).$$

5.3e $\widehat{\mu}_0 = b_0 + b_1 x_0 = 7.814 + (2.665)(0.1) = 8.08$.

5.3f $\widehat{\mu}_0 \pm t_{0.025}(28)s\sqrt{h_{00}} =$

$$8.08 \pm (2.048)(0.31656)\sqrt{0.041848}.$$

Here

$$h_{00} = \frac{1}{30} + \frac{0.1 - 0.213)^2}{1.49967} = 0.041848.$$

Then,

$$8.08 \pm 0.133 = (7.947, 8.213).$$

5.3g $\widehat{y}_0 = 8.08 + (2.665)(.1) = 8.3465$.

5.3h $\widehat{y}_0 \pm t_{0.025}(28)s\sqrt{1 + h_{00}} =$

$$8.08 \pm (2.048)(0.31656)\sqrt{1.041848} =$$

$$8.08 \pm 0.662 = (7.418, 8.742).$$

5.3i $F(MODEL) = \frac{MS(MODEL)}{MS(E)} = \frac{10.65268}{0.10021} = 106.303$.

5.3j $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. $F_{0.05}(1, 28) = 4.2$. Since $106.303 > 4.2$, reject H_0.

5.3k Total variance is 13.459. Explained variance is 10.653. Unexplained variance is 2.806.

$$r^2 = \frac{10.653}{13.459} = 0.792.$$

Approximately 79% of the total variance in 30 observed samples is explained variance.

3. Prove the following for the simple linear regression model:

$$\sum_{i=1}^{n} e_i = 0.$$

$$\sum_{i=1}^{n} x_i e_i = 0.$$

$$\sum_{i=1}^{n} \widehat{y}_i e_i = 0.$$

Starting with the first one:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - b_0 - b_1 x_i =$$

$$\sum_{i=1}^{n} y_i - \bar{y} + b_1 \bar{x} - b_1 x_i =$$

$$n\bar{y} - n\bar{y} + n b_1 \bar{x} - n b_1 \bar{x} = 0.$$

Starting with the second one:

$$\sum_{i=1}^{n} x_i e_i =$$

$$\sum_{i=1}^{n} x_i [y_i - \bar{y} + b_1 \bar{x} - b_1 x_i] =$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} x_i + b_1 \bar{x} x_i - b_1 x_i^2 =$$

$$\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} yi \sum_{i=1}^{n} x_i + \frac{b_1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - b_1 \sum_{i=1}^{n} x_i^2.$$

Expand on b_1 to get zero.

8.1.10 An F-Test of Lack of Fit

The i-th residual is $\epsilon_i = y_i - \widehat{y}_i$. Let $\eta_i = E(y_i)$. If the model is correct, then $\eta_i = \mu_i = \beta_0 + \beta_1 x_1$. Rewrite ϵ_i as

$$\epsilon_i = (y_i - \widehat{y}_i) - E(y_i - \widehat{y}_i) + E(y_i - \widehat{y}_i) =$$

$$[y_i - \widehat{y}_i - E(y_i - \widehat{y}_i)] + \overbrace{[\eta_i - E(y_i)]}^{\text{biasness}}.$$

The lack-of-fit test assesses whether or not the bias terms are zero. It requires multiple observations at some x values.

Suppose we have data y_{11}, y_{12}, y_{1n_1} at $x = x_1$, y_{21}, y_{22}, y_{2n_2} measurements at $x = x_2$ and so on. Then,

$$SS(E) = \sum_{L=1}^{m}\sum_{k=1}^{n_L}(y_{Lk} - \bar{y}_L)^2 =$$

$$\overbrace{\sum_{L=1}^{m}\sum_{k=1}^{n_L}(y_{Lk} - \bar{y}_L)^2}^{\text{SS(PE)}} + \overbrace{\sum_{L=1}^{n} n_L(\bar{y}_L - \widehat{y}_L)^2}^{\text{SS(LF)}}$$

where the degrees of freedom for SS(E) are $n-2$, the degrees of freedom for SS(PE) are $n-m$ and the degrees of freedom for SS(LF) are $m-2$. Then,

$$MS(PE) = \frac{SS(PE)}{n-m},$$

$$MS(LF) = \frac{SS(LF)}{m-2},$$

$$F(LF) = \frac{MS(LF)}{MS(PE)}.$$

H_0 : the model fits versus H_1 : model is biased. Reject H_0 if $F(LF) > F_\alpha(m-2, n-m)$.

SAS Code

```
OPTIONS NODATE;
DATA REPAIR;
INPUT N_UNITS MINUTES;
LACKOFIT = N_UNITS;
CARDS;
1 23
2 29
3 49
4 64
4 74
5 87
6 96
6 97
7 109
8 119
9 149
9 145
10 154
10 166
11 162
11 174
12 180
12 176
14 179
16 193
17 193
18 195
18 210
18 198
20 205
;

PROC PLOT;
PLOT MINUTES*N_UNITS;
RUN;

PROC REG;
MODEL MINUTES = N_UNITS;
OUTPUT OUT=NEW P=PRED R=RESID;
RUN;
```

```
PROC PLOT;
PLOT MINUTES*N_UNITS='O' PRED*N_UNITS='*'/OVERLAY;
PLOT RESID*N_UNITS='*'/VREF=0;
RUN;

PROC UNIVARIATE NORMAL PLOT;
VAR RESID;
RUN;

PROC GLM DATA=REPAIR;
CLASS LACKOFIT;
MODEL MINUTES = N_UNITS LACKOFIT;
RUN;
```

The lack-of-fit test is performed in SAS by creating a variable that is identical to the independent variable, and declaring it a CLASS variable in PROC GLM. The class variable is then placed as the last term in the model statement and the regression is run. PROC GLM is another proc for fitting linear models, which we use for this application because PROC REG does not have a class statement for creating grouping variables.

8.2 Assumptions Behind Regression

The 3 assumptions behind regression make certain statements about the $\epsilon_i's$ which we wish to evaluate. We check these assumptions by inspecting the *residuals,* $\epsilon_i = y_i - \widehat{y}_i, i = 1, 2, ..., n.$ If the model(including the assumptions) is correct, then the residuals should behave in a way consistent with the model. The assumptions are:

1. Variance of the residuals is constant.
2. The measurements are uncorrelated.
3. The measurements are normally distributed.

There are two methods for assessing this behavior: 1) plots and 2) statistics for isolating unusual residuals. Plots include:

1. Plot the residuals versus x_i to verify assumption 1.
2. Plot the residuals versus $\widehat{y}_i$ to verify assumption 1.
3. Plot the residuals versus time/space ordering variable to verify assumption 2.

4. Use a scatterplot, box plot, or a stem-leaf plot to verify assumption 3.

8.2.1 Shapiro-Wilks Test of Normality

The hypotheses are H_0 : data comes from a normal distribution versus H_1 : data does not come from a normal distribution. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

when given a random sample $x_1, x_2, ..., x_n$. The numerator is the best estimate of the standard deviation based on a linear combination of the order statistics. a_i and the reject points can be found in the original paper in Biometrika(1965), volume 52. SAS uses this test for any $n \leq 2000$ data points. For $n > 2000$ SAS uses Kolmogorov's statistic.

8.2.2 Lack of Independence

Plot ϵ_i versus time/space to see if the data is ordered. Positive correlation looks at positive/negative runs of numbers. Plot the residuals against time/space variable. Look for non-random patterns.

Strategy 1: Evaluate runs of positive and negative signs. Long runs indicate positive correlation. Short runs indicate negative correlation.

Strategy 2: The Durbin-Watson test. The Durbin-Watson test looks for correlation arising from the model $E_\mu = \rho\epsilon_{\mu-1} + z_\mu$, where $z_\mu \sim N(0, \sigma^2)$ independent of the $\epsilon's$. For this model, lag-s correlation $= \rho^s, |\rho| < 1$. $H_0 : \rho = 0$ versus $H_1 : \rho < 0$ or $H_1 : \rho > 0$ or $H_1 : \rho \neq 0$. The test statistic is

$$d = \frac{\sum_{u=2}^{n}(\epsilon_u - \epsilon_{u-1})^2}{\sum_{u=2}^{n}\epsilon_u^2}.$$

See page 239 of the text book on how to determine the rejection region.

SAS Code

```
*SEE PAGE 240 OF THE TEXT BOOK;

DATA PRICEY;
```

```
INPUT T Y X;
LABEL T = 'TIME'
Y = 'DEMAND'
X = 'PRICE DIFFERENCE';
CARDS;
1 7.38 -0.05
2 8.51 0.25
...
30 9.26 0.55;

PROC PLOT;
PLOT Y*X;
RUN;

PROC REG;
MODEL Y = X/DW;
OUTPUT OUT=NEW P=PREDICTED R=RESID;
RUN;

DATA NEW1;
SET NEW;
SIGNRES = SIGN(RESID);

PROC PRINT L;
RUN;

PROC PLOT;
PLOT Y*X='O' PREDICTED*X='*'/OVERLAY;
PLOT RESID*X/VREF=0;
RUN;
```

8.2.3 Sign Testing for Checking Assumption 2

Data is in order according to some time or space variable. Examine the ordered residuals. Specifically, look at the sequence of signs they produce. Properties of the sequence of signs:

1. Few runs means positive correlation.

2. Many runs means negative correlation.

3. Moderate runs means uncorrelated.

The *sign test* test the number of runs as a way of testing for correlation. Let u be the number of runs, n_1 be the number of positive signs, and n_2 be the number of negative signs. The tables in the book report the cdf of u for given n_1, n_2. H_0 : no correlation versus H_1 : positive correlation. Reject if $P(U \leq u)$ is small. H_1 : negative correlation. Reject if $P(U \geq u)$ is small. Small is less than α or your p-value.

Example: Test for positive correlation with $n_1 = 9$, $n_2 = 7$ and $u = 5$. p-value $= P(U \leq 5) = 0.035$. Thus, reject H_0 at $\alpha = 0.05$. Conclude that there is positive correlation.

Example: Test the following sequence for negative correlation: $+, -, +, -, +, +, -, +, -, -, -, +, +, -, +, +$. Here $n_1 = 9, n_2 = 7, u = 11$. The p-value is $P(U \geq u) = 1 - P(U \leq 10) = 1 - 0.806 = 0.194$. Fail to reject H_0.

For n_1, n_2 outside the range of the tables, there is a normal approximation. $\mu =$ mean number of runs $= \frac{2n_1n_2}{n_1+n_2} + 1$. $\sigma^2 =$ variance of the number of runs $= \frac{2n_2n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$. The test statistic is

$$z = \frac{u - \mu \pm 0.5}{\sigma}.$$

Use $+0.5$ for positive correlation and -0.5 for negative correlation.

Example: Test for negative correlation when $n_1 = 12, n_2 = 18$ and $u = 17$.

$$\mu = \frac{2(12)(18)}{12+18} + 1 = 15.4.$$

$$\sigma^2 = \frac{2(12)(18)[2(12)(18) - 12 - 18)]}{(12+18)^2(12+18-1)} = 6.6538.$$

Then,

$$z = \frac{17 - 15.4 - 0.5}{\sqrt{6.6538}} = 0.43.$$

Fail to reject H_0.

8.2.4 Homework

1. Problem 6.1 on page 252 of the text book. Omit part (d). Use SAS to get these results. Write your interpretations on the printed listing.

```
* hw 4;
* Roger Goodwin;
* Stat 537;
* Due Thursday, October 3, 1996;
*;
*;
data gpasal;
 input gpa salary;
 cards;
3.26 28.2
2.60 24.8
3.35 27.9
2.86 25.3
3.82 30.3
2.21 23.0
3.47 29.4


proc reg;
model salary = gpa;
output out = new p=predictd r=resid;

proc print;

proc plot;
 plot resid*gpa;
 plot resid*predictd;

proc univariate normal plot;
var resid;

run;
```

2. Repeat (a), (b), (c), and (e) of (1) but for a simple linear regression using the data in problem 4.22(use y and x_2).

```
* hw 4;
```

```
* Roger Goodwin;
* Stat 537;
* Due Thursday, October 3, 1996;
*;
*;
data priceage;
 input age price;
 lackofit = age;
 cards;
6 4.5
4 54.9
4 47.0
3 47.0
3 70.0
3 42.5
2 48.0
4 45.5
3 44.0
3 72.0
6 65.0
3 71.0
3 72.0
3 72.0
2 73.5
2 73.0
3 73.5
3 40.3
4 45.0
4 45.0
4 72.5
2 74.0
2 73.0
;


proc reg;
model price = age;
output out = new p=predictd r=resid;

proc print;

proc plot;
```

```
 plot resid*age;
 plot resid*predictd;

proc univariate normal plot;
var resid;

proc glm data=priceage;
class lackofit;
model price=age lackofit;

run;
```

3. Use SAS to calculate the lack-of-fit test for your model in (2).

8.3 Matrix Algebra

A *matrix* is a rectangular array of numbers or symbols. **A** means matrix **A.**

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -3 \\ 2 & 1 & 7 \\ 4 & -2 & 6 \\ 0 & 0 & 3 \end{pmatrix}$$

is a 4×3 matrix.

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

is a 2×1 matrix. Matrices with 1 row or 1 column are called *row vectors* or *column vectors.* A 1×1 matrix is just an ordinary number, sometimes called a *scalar.* In general, we will discuss Matrices like the following matrix:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

a_{ij} is the element in row i, column j.

8.3.1 The Transpose of a Matrix

The *transpose* of the matrix **A** is found by interchanging rows and columns to obtain an $n \times m$ matrix denoted by $\mathbf{A}'$.

$$\mathbf{A}' = \begin{pmatrix} 1 & 2 & 4 & 0 \\ 0 & 1 & -2 & 0 \\ -3 & 7 & 6 & 3 \end{pmatrix}$$

is a 3×4 matrix.

8.3.2 Sums and Differences of Matrices

Define matrix **A** as the same as before. Define the following Matrices:

$$\mathbf{B} = \begin{pmatrix} 1 & 0 \\ -1 & 2 \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 4 & 3 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} 4 & 2 & 1 \\ -2 & -1 & -7 \\ 1 & 2 & 0 \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} 1 & 0.5 & 2 \\ -2 & 3 & 0 \\ 2 & 0 & 3 \end{pmatrix}$$

Then,

$$\mathbf{A} + \mathbf{D} = \begin{pmatrix} 5 & 2 & -2 \\ 0 & 0 & 0 \\ 1 & 0 & 10 \\ 1 & 2 & 3 \end{pmatrix}$$

And,

$$\mathbf{A} - \mathbf{D} = \begin{pmatrix} -3 & -2 & -4 \\ 4 & 2 & 14 \\ 7 & -4 & 2 \\ -1 & -2 & 3 \end{pmatrix}$$

8.3.3 Matrix Multiplication

Two Matrices can be multiplied if they are *conformable.* That is, if the number of columns of the first matrix is the same as the number of rows of the second matrix. Matrix $\mathbf{A}$ is a 4×3 matrix, matrix $\mathbf{B}$ is a 3×2 matrix, matrix $\mathbf{C}$ is a 3×3 matrix, matrix $\mathbf{D}$ is a 4×3 matrix, and matrix $\mathbf{E}$ is a 3×3 matrix. We can multiply **AB, AC,** and **AE.** But, we can not multiply **AD, BC,** etc. The element of row i column j of the product is the product of the i-th row of the first matrix with the j-th column of the second matrix.

$$\mathbf{B}'\ \mathbf{C} = \begin{pmatrix} 1 & \text{-1} & 0 \\ 0 & 2 & 1 \end{pmatrix} \begin{pmatrix} 4 & 3 & 0 \\ 2 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} \text{4-2+0} & \text{3-4+0} & \text{0+0+0} \\ \text{0+4+0} & \text{0+8+0} & \text{0+0+1} \end{pmatrix} =$$

$$\begin{pmatrix} 2 & \text{-1} & 0 \\ 4 & 8 & 1 \end{pmatrix}.$$

8.3.4 The Identity Matrix and Inverses

The identity matrix $\mathbf{I}$ of order n is an $n \times n$ matrix with 1's on the diagonal and 0's elsewhere. An example of a 3×3 identity matrix is

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Multiplication by $\mathbf{I}$ preserves any matrix. $\mathbf{AI} = \mathbf{A}$ and $\mathbf{IA} = \mathbf{A}$.

A *square matrix* is one for which the number of rows is equal to the number of columns. Some square Matrices have inverses. The *inverse* of matrix $\mathbf{A}$ is another square matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{I}$. We write $\mathbf{B} = \mathbf{A}^{-1}$.

Consider matrix $\mathbf{E}$. 1.5 times the first column plus the second column is equal to the third column. Column three is said to be *linearly dependent* on columns one and two. The columns of a given matrix are said to be linearly dependent if one of its columns can be written as a linear combination of the others. If none of the columns of a given matrix can be written as a linear combination of the other then the columns of the given matrix are said to be *linearly independent.* The maximum number of linearly independent columns of a given matrix is called the *rank* of the matrix. If the rank of the given matrix is equal to the number of columns, then the matrix is said to be of *full column rank.*

Square Matrices have inverses iff they are of full rank. A full rank square matrix implies that it is non-singular. If matrix $\mathbf{X}$ has rank r, then $\mathbf{X}'\mathbf{X}$ also has rank r. If $\mathbf{X}$ has full column rank, then $\mathbf{X}'\mathbf{X}$ is invertible.

SAS Code

Several options in PROC REG will be demonstrated. The CORRB option gives correlation estimates. The COVB option gives covariance estimates. The I option gives the inverse of $\mathbf{X}'\mathbf{X}$. The XPX option gives the model cross products. The CLM option gives confidence intervals.

```
OPTION LINESIZE = 72;

DATA REPAIR;
INPUT N_UNITS MINUTES;
N_UNITSQ = N_UNITS**2;
CARDS;
1 23
2 29
...
20 205;

PROC PLOT;
PLOT MINUTES*N_UNITS;
RUN;

PROC REG;
MODEL MINUTES = N_UNITS N_UNITSQ/CORRB COVB I XPX CLM;
OUTPUT OUT=NEW P=PRED R=RESID;
RUN;

PROC PRINT DATA=NEW;
RUN;

PROC PLOT DATA=NEW;
PLOT MINUTES*N_UNITS='O' PRED*N_UNITS='*'/OVERLAY;
PLOT RESID*N_UNITS='*'/VREF=0;
RUN;
```

8.4 Multiple Regression

In order to get a better feeling for the matrix multiplications to follow, we set up the simple linear regression in matrix terms.

$$y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, ..., n.$$

$$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1,$$

$$y_2 = \beta_0 + \beta_1 x_2 + \epsilon_2,$$

...

$$y_n = \beta_0 + \beta_1 x_n + \epsilon_n,$$

Equivalently,

$$y_1 = \beta_0 x_{10} + \beta_1 x_{11} + \epsilon_1,$$

$$y_2 = \beta_0 x_{20} + \beta_1 x_{21} + \epsilon_2,$$

...

$$y_n = \beta_0 x_{n0} + \beta_1 x_{n1} + \epsilon_n,$$

where x_{ij} is the i-th measured value for variable x_j, $i = 1, ..., n$ and $j = 0, 1$. The variable x_{i0} always takes the value 1. Let's now write the model in matrix form:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} x_{10} & x_{11} \\ x_{20} & x_{21} \\ . & . \\ . & . \\ . & . \\ x_{n0} & x_{n1} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

Then, the model can be written as $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$. If we write

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_n \end{pmatrix}$$

where $\mu_i = E(y_i)$, then $\mu = \mathbf{X}\beta$. Estimate β_0 and β_1 by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 =$$

$$\sum_{i=1}^{n}(y_i - \beta_0 x_{i0} - \beta_1 x_{i1})^2 =$$

$$\mathbf{(\ Y\ -\ X}\beta\ \mathbf{)'(\ Y\ -\ X}\beta\ \mathbf{)}.$$

Differentiating with respect to β_0 and setting to zero yields the *normal equations:*

$$nb_0 + \left(\sum_{i=1}^{n} x_{i1}\right) b_1 = \sum_{i=1}^{n} y_i,$$

$$\left(\sum_{i=1}^{n} x_{i1}\right) b_0 + \left(\sum_{i=1}^{n} x_{i1}^2\right) b_1 = \sum_{i=1}^{n} x_{i1} y_i,$$

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \end{pmatrix}.$$

Consider $\mathbf{X'X} =$

$$\begin{pmatrix} x_{10} & x_{20} & ... & x_{n0} \\ x_{11} & x_{21} & ... & x_{n1} \end{pmatrix} \begin{pmatrix} x_{10} & x_{11} \\ x_{20} & x_{21} \\ . & . \\ . & . \\ . & . \\ x_{n0} & x_{n1} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{pmatrix}.$$

The *normal equations* in matrix form are $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$. The solution to these normal equations is easy. $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, provided that $\mathbf{X}$ has full column rank of 2. The estimate of $\mu = \mathbf{X}\beta$ based on our estimate of $\mathbf{b}$ for β is

$$\widehat{\mu} = \mathbf{X}\mathbf{b} = \overbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}^{\text{call it } \mathbf{H}}\mathbf{Y} = \mathbf{H}\mathbf{Y} = \widehat{\mathbf{Y}}$$

which is the same matrix but with different variability. Next consider the sums of squares

$$SS(E) = \sum_{i=1}^{n}(y_i - \widehat{y})^2 =$$

$$(\mathbf{Y} \text{ - } \widehat{\mathbf{Y}})'(\mathbf{Y} \text{ - } \widehat{\mathbf{Y}}) =$$

$$(\mathbf{Y} \text{ - } \mathbf{H}\mathbf{Y})'(\mathbf{Y} \text{ - } \mathbf{H}\mathbf{Y}) =$$

$$[(\mathbf{I} \text{ - } \mathbf{H})\mathbf{Y}]'[(\mathbf{I} \text{ - } \mathbf{H})\mathbf{Y}] =$$

$$\mathbf{Y}'(\mathbf{I} \text{ - } \mathbf{H})'(\mathbf{I} \text{ - } \mathbf{H})\mathbf{Y}.$$

Note that

$$(\mathbf{I} \text{ - } \mathbf{H})'(\mathbf{I} \text{ - } \mathbf{H}) =$$

$$(\mathbf{I}' \text{ - } \mathbf{H}')(\mathbf{I} \text{ - } \mathbf{H}) =$$

$$(\mathbf{I} \text{ - } \mathbf{H})(\mathbf{I} \text{ - } \mathbf{H}) = \mathbf{I} \text{ - } \mathbf{H} \text{ - } \mathbf{H} + \mathbf{H}\mathbf{H};$$

Note that

$$\mathbf{H}\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} =$$

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H},$$

which means it is an idempotent matrix. Then,

$$\mathbf{I} \text{ - } \mathbf{H} \text{ - } \mathbf{H} + \mathbf{H}\mathbf{H} = \mathbf{I} \text{ - } \mathbf{H} \text{ - } \mathbf{H} + \mathbf{H} = \mathbf{I} \text{ - } \mathbf{H}.$$

The total sum of squares is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$\mathbf{Y} \text{ - } \tfrac{1}{n}\mathbf{J}_n\mathbf{Y} = (\mathbf{I} \text{ - } \tfrac{1}{n}\ \mathbf{J})\mathbf{Y}.$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = [\mathbf{Y} - \tfrac{1}{n}\mathbf{J}_n\mathbf{Y}]'[\mathbf{Y} - \tfrac{1}{n}\mathbf{J}_n\mathbf{Y}]$$

$$\mathbf{Y}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})\mathbf{Y} =$$

$$\mathbf{Y}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})(\mathbf{I} - \tfrac{1}{n}\mathbf{J})\mathbf{Y} =$$

$$\mathbf{Y}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J} - \tfrac{1}{n}\mathbf{J} + \tfrac{n}{n^2}\mathbf{J})\mathbf{Y} =$$

$$\mathbf{Y}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})\mathbf{Y} = SS(TOTAL).$$

Finally, SS(MODEL) = SS(TOTAL) - SS(E). So,

$$\mathbf{Y}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})\mathbf{Y} - \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} =$$

$$\mathbf{Y}'[(\mathbf{I} - \tfrac{1}{n}\mathbf{I}) - (\mathbf{I} - \mathbf{H})]\mathbf{Y} =$$

$$\mathbf{Y}'(\mathbf{H} - \tfrac{1}{n}\mathbf{I})\mathbf{Y} = SS(MODEL).$$

The sums are the same in the general case.

SAS Code

This is the multiple regression example using the data on page 306 of the text book. Most of this program is shown on page 346 of the text book.

```
DATA DETR;
INPUT REGION X1 X2 X4 X3 Y;
X3SQ = X3**2;
LABEL REGION = 'SALES REGION'
X1 = 'PRICE FOR FRESH DETERGENT'
X2 = 'AVERAGE INDUSTY PRICE'
X3 = 'ADVERTISING EXPENDITURE($100,000)'
X4 = 'PRICE DIFFERENCE X1-X2'
X3SQ = 'SQUARE OF X3'
Y = 'DEMAND';
CARDS;
1 3.85 3.8 -0.05 5.50 7.38
2 3.75 4.0 0.25 6.75 8.51
...
30 3.70 4.2 0.55 6.80 9.26
. . . 0.10 6.80 .;

PROC PRINT L;
```

```
TITLE 'THE DETERGENT DATA';
TITLE2 'NOTE DATA WITH MISSING VALUES FOR THE DEPENDENT VARIABLE';
TITLE3 'TIS IS NOT USED IN FITTING THE MODEL, BUT REG WILL GIVE Y-HAT';
RUN;

PROC REG DATA=DETR;
MODEL Y = X4 X3 X3SQ/P CLM CLI I XPX COVB;
OUTPUT OUT=ONE R=RESID P=YHAT;
TITLE 'THE DETERGENT DATA';
RUN;

PROC PLOT DATA=ONE;
PLOT RESID*(X4 X3 YHAT)/VREF=0;
TITLE 'RESIDUALS PLOTS FOR THE DETERGENT DATA';
RUN;

PROC UNIVARIATE NORMAL PLOT DATA=ONE;
VAR RESID;
TITLE 'CHECKING DISTRIBUTION OF THE RESIDUALS';
RUN;
```

8.5 Homework and Answers

1. Text, Chapter 7, # 11, 12, 15. Add the following to # 11:

   (a) Calculate $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and verify that $\mathbf{H}^2 = \mathbf{H}$ by direct multiplication.

   (b) Assuming that this data for a simple linear regression, calculate b_0 and b_1 using the formulas from Chapter 4, and compare to (d).

7.11a $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 12 \\ 12 & 28 \end{pmatrix}$.

7.11b $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{7}{6} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{4} \end{pmatrix}$.

7.11c $\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 65 \\ 119 \end{pmatrix}$.

7.11d $\begin{pmatrix} \frac{98}{6} \\ -\frac{11}{4} \end{pmatrix}$

7.11e $$\mathbf{H} = \begin{pmatrix} 5/12 & 2/12 & -1/12 & 2/12 & -1/12 & 5/12 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ -1/12 & 2/12 & 5/12 & 2/12 & 5/12 & -1/12 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ -1/12 & 2/12 & 5/12 & 2/12 & 5/12 & -1/12 \\ 5/12 & 2/12 & -1/12 & 2/12 & -1/12 & 5/12 \end{pmatrix}$$

7.11f $$b_1 = \frac{6(10+24+24+28+18+15) - (12)(65)}{6(28) - (12)^2} = -2.75.$$

$$b_0 = \bar{Y} - b_1\bar{X} = \frac{65}{6} + 2.75(2) = 16.333.$$

7.12a $\begin{pmatrix} -\frac{1}{12} & \frac{1}{8} \end{pmatrix}$.

7.12b $\begin{pmatrix} -\frac{1}{12} \\ \frac{1}{8} \end{pmatrix}$.

7.12c $\frac{5.5}{24}$.

7.15a $\lambda' = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}$

7.15b $\lambda' = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

7.15c $\lambda' = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & -1 \end{pmatrix}$

7.15d $\lambda' = \begin{pmatrix} 0 & 2 & 3 & 1 & 0 & 0 \end{pmatrix}$

7.15e $\lambda' = \begin{pmatrix} 0 & -1/3 & -1/3 & -1/3 & 1 & 0 \end{pmatrix}$

7.15f $\lambda' = \begin{pmatrix} 1 & 2 & 5 & 3 & 1 & 6 \end{pmatrix}$

2. Text, Chapter 8, # 4, 23, 24.

4b

Source	d.f.	SS	MS	F
Model	2	25.462472	12.731236	229.37
Error	5	0.277528	0.0555056	
Total	7	25.74		

4c $$s = \sqrt{\frac{SS(E)}{5}} = \sqrt{0.0555056} = .236.$$

$$R^2 = \frac{SS(MODEL)}{SS(TOTAL)} = \frac{25.462472}{25.74} = 0.989.$$

$R = 0.995$. The proportion of explained variance in the model to the overall variation is 0.989. This is highly significant.

4d $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$. $\alpha = 0.05$. $t_{0.025}(5) = 2.571$. The test statistic is

$$t = \left| \frac{b_0}{s\sqrt{c_{00}}} \right| = \frac{12.917034}{0.5492} = 23.52.$$

Since $23.52 > 2.571$, reject H_0. For $\alpha = 0.01$, $t_{0.005}(5) = 4.032$. Reject H_0 again. The line probably does not pass thru the origin.

4e
$$t = \left| \frac{b_1}{s\sqrt{c_{11}}} \right| = \left| \frac{-0.087064}{0.009035063} \right| = 9.636.$$

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Since $9.636 > 4.032$, reject H_0. Changes in x_1 do relate to changes in y.

For x_2 :

$$t = \left| \frac{b_2}{s\sqrt{c_{22}}} \right| = \frac{0.090221}{0.014122} = 6.389.$$

$H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. Since $6.389 > 4.032$, reject H_0. Changes in x_2 do relate to changes in y.

4f In general, the confidence interval is given by

$$b_j \pm t_{\alpha/2}(n-k)s\sqrt{c_{jj}}, j = 0, 1, 2.$$

For β_0 :

$$b_0 \pm t_{0.025}(5)s\sqrt{c_{00}} =$$

$$12.917034 \pm (2.571)(0.5492) =$$

$$12.917034 \pm 1.412 = (11.505, 14.329).$$

For β_1 :

$$b_1 \pm t_{0.025}(5)s\sqrt{c_{11}} =$$

$$-0.087064 \pm (2.571)(0.009035063) =$$

$$-0.087064 \pm 0.02323 = (-0.110, -0.064).$$

For β_2 :

$$b_2 \pm t_{0.025}(5)s\sqrt{c_{22}} =$$

$$0.090221 \pm (2.571)(0.014122) =$$

$$0.090221 \pm 0.036308 = (0.054, 0.127).$$

We can be 95% confident that β_0 is in the interval given. We can be 95% confident that β_1 is in the interval given. Finally, we can be 95% confident that β_2 is in the interval given.

4f $x_{01} = 40$ and $x_{02} = 10$.

$$y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02},$$

$$y_0 = 12.917034 - 0.087064 x_{01} + 0.090221 x_{02},$$

$$y_0 = 12.917034 - 0.087064(40) + 0.090221(10),$$

$$\widehat{y}_0 = 10.34.$$

4h $\widehat{\mu}_0 \pm t_{0.005}(5) s\sqrt{h_{00}} =$

$$\widehat{\mu}_0 \pm (4.032)(0.109411) = 10.34 \pm 0.441 =$$

$$(9.899, 10.781).$$

The population mean μ_0 is probably in the interval (9.899, 10.781) with 99% confidence.

4i $\widehat{y}_0 \pm t_{0.005}(5)\sqrt{1 + h_{00}} =$

$$s\sqrt{1 + h_{00}} = (0.236)(1.102239) = 0.26013.$$

Then

$$10.34 \pm (4.032)(0.26013) =$$

$$10.34 \pm 1.05 = (9.291, 11.39).$$

Use SAS for 8.23 and 8.24. Otherwise perform calculations with a hand calculator. The SAS code for both problems is as follow:

```
DATA CREST;
INPUT X1 X2 X3 Y;
X1SQ = X1*X1;
X2SQ = X2*X2;
X3SQ = X3*X3;
LABEL X1 = 'BUDGET'
X2 = 'RATIO'
X3 = 'PERSONAL INCOME'
X1SQ = 'X1 SQUARED'
X2SQ = 'X2 SQUARED'
X2SQ = 'X3 SQUARED';
CARDS;
16300 1.25 547600000000 1050000
...
28000 1.56 1821700000000 245000
;
```

```
*REDUCED MODEL
PROC REG DATA=CREST;
MODEL Y = X1-X3/P CLM CLI I XPX COVB;
OUTPUT OUT=NEW R=RESID P=PRED;
RUN;

PROC PLOT DATA=NEW;
PLOT RESID*(X1 X2 X3 PRED)/VREF=0;
TITLE 'RESIDUALS PLOTS FOR CREST DATA';

PROC UNIVARIATE PLOT NORMAL;
VAR RESID;
RUN;

*COMPLETE MODEL;
PROC REG DATA = CREST;
MODEL Y = X1-X3 X1SQ X2SQ X3SQ/P CLM CLI I XPX COVB;
OUTPUT OUT=NEW2 R=RESID P=PRED;
RUN;

PROC PLOT DATA=NEW2;
PLOT RESID(X1 X2 X3)/VREF=0;
RUN;

PROC UNIVARIATE PLOT NORMAL;
VAR RESID;
RUN;
```

8.6 Test #1

Most of the answers to this test can be found in the notes.

1. For the general multiple regression setting,

   (a) Write down the model, discussing the role of each matrix or vector with an explanation of the concept it represents, and stating the corresponding dimensions.

   (b) Write down the normal equations and their solutions.

   (c) Write down the assumptions 1 – 3.

2. We learned several techniques, for assessing the adequacy of the regression model and the correctness of the assumptions, that involve plotting the residuals. Briefly explain three of these plots, including sketches to show what types of behavior can be ascertained from them.

3. Here is a small dataset with measurements on two variables x and y :

y	x
1	3
1	5
2	6
3	4
4	10
4	6
6	8

   (a) You want to fit a simple linear regression of y on x. Write down the vectors $\mathbf{Y}$ and β, and the matrix $\mathbf{X}$, that would be used to write the model in matrix notation. Also find $\mathbf{X}'\mathbf{Y}$ and $\mathbf{X}'\mathbf{X}$.

   (b) Find the solution to the normal equations(you need not use the matrix formulation if you do not wish to).

   (c) Test the hypothesis $H_0 : \beta_1 = 0$ that there is no regression.

4. Dr. Morgan gave out a SAS printout and asked to identify certain parts of it. This has been intentionally omitted.

8.7 The General Linear Regression Model

Our goal is to use an arbitrary number of independent variables to predict y. Our initial model is still $y_i = \mu_i + \epsilon_i$, and we will model the mean μ_i of y_i at given values of x's with a linear function of the x's. Let x_{i1} be the i-th value of the independent variable x_1, x_{i2} be the i-th value of the independent variable x_2, and so on. Our model is

$$y_i = \mu_i + \epsilon_i = \overbrace{\beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}^{\text{model for } \mu_i} + \epsilon_i, i = 1, ..., n.$$

Define the following Matrices:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix},$$

$$\mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & ... & x_{1p} \\ x_{20} & x_{21} & x_{22} & ... & x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n0} & x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}$$

$$\mathbf{E} = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}.$$

Then the model in matrix terms is $\mathbf{Y} = \mathbf{X}\ \beta + \mathbf{E}$.

8.7.1 Special Cases

1. Simple linear regression with one independent variable. The model is $y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \epsilon_i$.

2. Quadratic regression where there are two independent variables, $p = 2$, but the second variable is the square of the first. $x_{i2} = x_{i1}^2$. Then, the model is $y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i$. Note that linear refers to the combination of independent variables.

3. p-th order polynomial relation on one independent variable. $x_{i2} = x_{i1}^2$, $x_{i3} = x_{i1}^3, ... \; x_{ip} = x_{i1}^p$. The model is $y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \cdot \cdot \cdot + \beta_p x_{i1}^p + \epsilon_i$.

4. 2-dimensional quadratic response surface. The model is $y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \epsilon_i$.

8.7.2 Assumptions, Standard Errors, and Residual Analysis

As with simple linear regression, we have 3 assumptions.

1. For any fixed values $(x_{i1}, x_{i2}, ..., x_{ip}$ of independent variables $x_1, ..., x_p$, the response y_i has the same variance σ^2.

2. Any 2 values of the response, y_i and y_i' are uncorrelated i.e. $Cov(y_i, y_i') = 0$.

3. The distribution of the population potential values for the response at any given values $x_{i1}, ..., x_{ip}$ of the independent variables $x_1, ..., x_p$ is normal i.e. $y_i \sim N(\mu_i, \sigma^2)$ where μ_i is the mean of that population.

The 3 assumptions describe the random component of our model. The fixed component is the model for μ_i where $\mu_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. We estimate μ_i by estimating the β's. The point estimate of σ^2 is

$$MS(E) = s^2 = \frac{SS(E)}{n-k},$$

where k is the number of β's which is $p+1$.

$$SS(E) = \sum_{i=1}^{n} (y_i - \widehat{y})^2 = \mathbf{Y}'(\mathbf{I} \text{ - } \mathbf{H})\mathbf{Y}.$$

Hence, a point estimate of the *standard error* is

$$s = \sqrt{\frac{SS(E)}{n-k}}.$$

The SS(E) can also be written as

$$SS(E) = \mathbf{Y}'(\mathbf{I} \text{ - } \mathbf{H})\mathbf{Y} = \mathbf{Y}'\mathbf{Y} \text{ - } \mathbf{Y}'\mathbf{H}\mathbf{Y} =$$

$$\mathbf{Y}'\mathbf{Y} \text{ - } \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} =$$

$$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}.$$

To check the assumptions, we use the same plots previously covered. The i-th residual is $\epsilon_i = y_i - \widehat{y}_i$.

$$\mathbf{E} = \begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{pmatrix} = \begin{pmatrix} y_1 - \widehat{y}_1 \\ y_2 - \widehat{y}_2 \\ . \\ . \\ . \\ y_n - \widehat{y}_n \end{pmatrix} =$$

$$\mathbf{Y - Xb} = \mathbf{Y - X(X'X)^{-1}X'Y} = \mathbf{Y - HY} = \mathbf{(I - H)Y}.$$

8.7.3 Multiple Coefficient of Determination and Correlation

Recall that SS(TOTAL) = SS(MODEL) + SS(E) where SS(TOTAL) is the sum of squares of *total variation,* SS(MODEL) is the sum of squares of *explained variation,* and SS(E) is the sum of squares of *unexplained variation.* We want to get a model that explains as much of the total variation as possible. The proportion of variation explained is

$$R^2 = \frac{SS(MODEL)}{SS(TOTAL)} = 1 - \frac{SS(E)}{SS(TOTAL)}$$

and is called *multiple coefficient of determination.* $R = \sqrt{R^2}$ is called the *multiple correlation coefficient.*

8.7.4 Overall F-Test and the Basic ANOVA Table

The overall test for the regression model tests $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ versus H_1 : at least one $\beta \neq 0$. The test statistic is

$$F(MODEL) = \frac{MS(MODEL)}{MS(E)} = \frac{\frac{SS(MODEL)}{k-1}}{\frac{SS(E)}{n-k}},$$

where k is the number of β's in the model. Reject if $F(MODEL) > F_\alpha(k-1, n-k)$. This test is usually presented in an ANOVA Table.

Source	SS	d.f.	MS	F
Model	SS(MODEL)	$k-1$	MS(MODEL)	F(MODEL)
Error	SS(E)	$n-k$	MS(E)	
Total	SS(TOTAL)	$n-1$		

8.7.5 Inference for β_j

To perform a test for β_j, or to construct a confidence interval, we need an estimate and a standard error. The estimate is b_j. It may be shown that variance of b_j is the $(j+1)$st diagonal entry of the matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Write c_{jj} for the $(j+1)$st diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. Then, $Var(b_j) = \sigma^2 c_{jj}$. Hence the standard error for b_j is

$$\sqrt{MS(E)c_{jj}} = s\sqrt{c_{jj}}.$$

We now get the confidence interval and the test. A $(1-\alpha)100\%$ confidence interval for β_j is

$$b_j \pm t_{\alpha/2}(n-k)s\sqrt{c_{jj}}.$$

The test of $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ is reject H_0 if

$$|t| = \left| \frac{b_j}{s\sqrt{c_{jj}}} \right| > t_{\alpha/2}(n-k).$$

8.7.6 Confidence Intervals and Prediction Intervals

Recall that $\widehat{y}_0 = b_0 + b_1 x_{01} + b_2 x_{02} + \cdots + b_p x_{0p} = \mathbf{b}'\mathbf{X}_0$ where

$$\mathbf{X}_0 = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ . \\ . \\ . \\ x_{0p} \end{pmatrix}.$$

1. The point estimate of $\mu_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$ is equal to the mean of the population of y's at x_0.
2. The point prediction of $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \epsilon_0$ is equal to the new value of y at this x_0.

We want a confidence interval for both (1) and (2). Consider (1). We have $\widehat{y}_0 = \mathbf{b}'\mathbf{X}_0$. It can be shown that

$$Var(\widehat{y}_0) = \sigma^2 \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0.$$

So if we write $h_{00} = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$, then $Var(\widehat{y}_0) = \sigma^2 h_{00}$, and hence the standard error is $s\sqrt{h_{00}}$. Then the $(1-\alpha)100\%$ confidence interval for μ_0 is

$$\widehat{y}_0 \pm t_{\alpha/2}(n-k)s\sqrt{h_{00}}.$$

The $(100-\alpha)100\%$ prediction interval for y_0 is

$$\widehat{y}_0 \pm t_{\alpha/2}(n-k)s\sqrt{1+h_{00}}.$$

Suppose we want to estimate the difference in the mean response at two different set of values for the independent variables, say $\mathbf{X}_{0'}, \mathbf{X}_{0''}$. The estimated means at these two vectors are

$$\widehat{y}_{0'} = \mathbf{X}'_{0'}\mathbf{b},$$

and

$$\widehat{y}_{0''} = \mathbf{X}'_{0''}\mathbf{b}.$$

So the *estimated difference* is

$$\widehat{y}_{0'} - \widehat{y}_{0''} = \mathbf{X}'_{0'}\mathbf{b} - \mathbf{X}'_{0''}\text{bf b}.$$

Let $\mathbf{X}'_0 = \mathbf{X}'_{0'} - \mathbf{X}'_{0''}$. Then the estimate of the difference in means is $\mathbf{X}'_0\mathbf{b}$. The standard error of this estimate is $s\sqrt{h_{00}}$ where $h_{00} = \mathbf{X}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$. Then, the $(1-\alpha)100\%$ confidence interval is

$$(\widehat{y}_{0'} - \widehat{y}_{0''}) \pm t_{\alpha/2}(n-k)s\sqrt{h_{00}}.$$

8.8 More on Multiple Regression

8.8.1 Interaction

See Example 9.1 on page 388 of the text book.

SAS Code

```
DATA FROZEN;
INPUT X1 X2 Y;
X3= X1*X2;
LABEL X1 = 'RADIO AND TELEVISION EXPENDITURES'
X2 = 'PRINT EXPENDITURES'
X3 = 'X1*X2'
Y = 'SALES VOLUME';
CARDS;
1 1 3.27
1 2 8.38
...
5 5 21.05
4.5 3.5 .;
```

```
PROC PLOT;
PLOT Y*X1=X2;
POLT Y*X2=X1;
RUN;

PROC REG;
MODEL Y = X1 X2;
RUN;

PROC REG;
MODEL Y = X1 X2 X3/P CLM CLI I COVB;
OUTPUT OUT=NEW R=RESID;
RUN;

PROC UNIVARIATE NORMAL PLOT;
VAR RESID;
RUN;
```

8.8.2 Testing Part of a Model

We often encounter the following problem: we have fit a multiple regression model and we want to know if we can reduce it to a simpler model. The appropriate hypothesis to test is $H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_p = 0$. Consider the following: SS(TOTAL) = SS(MODEL) + SS(E) holds for both models. Furthermore, since both models have the same y_i's, bo have the same SS(TOTAL). So,

$$SS(TOTAL) = SS(MODEL)_c + SS(E)_c,$$

where "c" means complete model which is also equal to

$$SS(TOTAL) = SS(MODEL)_r + SS(E)_r$$

where "r" means reduced model and obviously

$$SS(E)_r \geq SS(E)_c,$$

and

$$SS(MODEL)_c \geq SS(MODEL)_r.$$

Define $SS(E)_{drop} = SS(E)_r - SS(E)_c = SS(MODEL)_c - SS(MODEL)_r$ which is the reduction in unexplained variation from going from the reduced model to the complete model. The degrees of freedom of $SS(E)_{drop}$ are

$$[n - (g + 1)] - [n - (p + 1)] = p - g.$$

So,

$$MS(E)_{drop} = \frac{SS(E)_{drop}}{p-g}.$$

The test we want is based on this statistic.

$$F(x_{g+1}, x_{g+2}, ..., x_p | x_1, x_2, ..., x_g) = \frac{MS(E)_{drop}}{MS(E)_{complete}}.$$

Reject $H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_p = 0$ if $F > F_\alpha(p-g, n-k)$. The Type I SS is $SS(E)_{drop}$ for putting a variable in the model given that the variables before it are already in the model. Type I SS are order dependent. Type II SS is $SS(E)_{drop}$ for putting this variable in last given all other variables in the model statement are already in the model.

SAS Code

This is the detergent data again. We will fit an interaction term (see pages 397 — 400 in the text book) and demonstrate the use of Type I and Type III sums of squares to test for a reduced model.

```
DATA DETR;
INPUT REGION X1 X2 X4 X3 Y;
X3SQ = X3**2;
X43 = X4*X3;
CARDS;
1 3.85 3.8 -0.05 5.50 7.38
...
30 3.7 4.2 0.55 6.8 9.26
.   .   .  0.10 6.80 .
;

PROC REG DATA=DETR;
MODEL Y = X4 X3 X43 X3SQ/SS1 SS2;
TITLE 'THE FULL MODEL';
RUN;

PROC REG DATA=DETR;
MODEL Y = X4 X3;
TITLE 'THE REDUCED MODEL';
RUN;
```

8.9 Outliers and Influential Observations

Outliers are observations with un-usually large residuals (outlier with respect to y) or with x values far from the other x's(outlier with respect to x). *Influential observations* are those that play a disproportionate role in determining the regression line. Any particular measurement can be both, one or the other, or neither. We consider statistics for identifying such measurements. Here $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$.

1. The *leverage value* is h_{ii}. This is a method for identifying observations whose x values are outlying. h_{ii} is the i-th diagonal element of $\mathbf{H}$. This is called the leverage of the i-th measurement of $\mathbf{X}_i$. It can be shown that $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = k$. A leverage value is considered large if

   (a) It is substantially larger than other leverage values.

   (b) It is greater than $2\bar{h} = \frac{2k}{n}$.

   Values with high leverage values are potentially influential and bear further investigation.

2. *Studentized residual* is used to detect outliers in y. The calculation of the i-th studentized residual is

$$\frac{e_i}{s\sqrt{1-h_{ii}}}.$$

3. *Difference in fits statistic.* Let $\widehat{y}_{(i)}$ be the predicted value of y_i using the regression equation obtained from all of the data *except* the i-th measurement. $f_i = \widehat{y}_i - \widehat{y}_{(i)}$. The difference in fits statistic is f_i divided by its standard error s_{f_i}. It can be shown that the difference in fit is

$$\frac{f_i}{s_{f_i}} = \left(\frac{d_i}{s_{d_i}}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)^{\frac{1}{2}}$$

   where $d_i = y_i - \widehat{y}_{(i)}$. The will tell if the measurement is "drawing the equation to itself." Look for large difference in fits. If it is greater than 2 or greater than $2\sqrt{\frac{p}{n-p}}$ then it is an outlier. Or, look for values much larger than most others.

4. *Cook's distance* is defined as $D_i = (\mathbf{b} - \mathbf{b}^{(i)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}^{(i)})$ where $\mathbf{b}^{(i)}$ is the estimate of β with the i-th measurement deleted. It can be shown that

$$D_i = \left(\frac{e_i}{s\sqrt{h_{ii}}}\right)^2 \left(\frac{h_{ii}}{1-h_{ii}}\right)\left(\frac{1}{k}\right).$$

It combines the studentized residual with the leverage. It is very similar to the difference in fits statistic. An observation is influential if $D_i > F_{0.5}(k, n-k)$ It is best to just look for large D_i relative to the others with a stem-and-leaf plot.

5. *Difference in estimation of β_j statistic.* Let $g_j^{(i)} = b_j - b_j^{(i)}$. $s_{g_j}^{(i)}$ is the standard error of $g_j^{(i)}$. The difference in fits statistic is

$$\frac{g_j^{(i)}}{s_{g_j}^{(i)}} = \left(\frac{d_i}{s_{d_i}}\right)\left(\frac{t_{ji}}{t_j' t_j (1-h_{ii})}\right)$$

where t_{ji} is the (j,i) element of $\mathbf{T} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. t_j' is row j of $\mathbf{T}$. Absolute values greater than $\frac{2}{n}$ are nominated for further investigation.

6. The *covariance ration* corresponding to the i-th measurement is

$$CVR_i = \frac{(s_i)^{2k}}{(s)^{2k}}\left(\frac{1}{1-h_{ii}}\right)$$

where s_i^2 is s^2 calculated for the model with the i-th measurement deleted. This looks at how the variances and covariances of the b_j's change when the i-th measurement is deleted.

   (a) If $CVR_i > 1 + \frac{3k}{n}$, then eliminating the i-th observation significantly damages the precision of SS least estimates.
   (b) If $CRV_i < 1 - \frac{3k}{n}$, then it enhances precision.

SAS Code

The following data appears in "Procedures and Analysis for Staffing Standards Development: Regression Analysis Handbook" published by the Navy Manpower and Material Analysis Center(1979). The goal is to develop a workable model relating labor hours to a number of available variables. We will try to do this while also examining the data for outliers and influential points. We restrict ourselves to fitting the model for only the small to medium sized hospitals, defined as those with daily patient load below 200.

```
DATA HOSPITAL;
INPUT X1 X2 X3 X4 X5 Y;
IF (X1>199) THEN DELETE;
CARDS;
15.57 2463 479.92 18.0 4.45 566.52
...
510.22 86533 15524 371.6 6.35 18854.45
;

PROC REG DATA=HOSPITAL;
TITLE 'HOSPITAL DATA';
MODEL Y = X1 X2 X3 X4 X5; * NOTE THE LARGE P-VALUES FOR THE
    INDIVIDUAL BETAS, DESPITE THE GOOD
    RSQUARE. OBVIOUSLY THERE IS SOME
    REDUNDANT INFORMATION AMONG THE X'S.
    THIS IS A PROBLEM KNOWN AS
    MULTICOLLINEARITY, WHICH WE WILL STUDY
    IN MORE DETAIL LATER;
RUN;

PROC CORR;
VAR X1-X5; * TO STUDY LINEAR RELATIONSHIPS AMONE
 THE X'S. BASED ON THIS WE THROW OUT
 X2, X3, X4. THIS IS ALSO PART OF THE
 PROBLEM OF VARIABLE SELECTION, ALSO
 TO BE STUDIED IN MORE DETAIL LATER;
RUN;

PROC REG DATA=HOSPITAL;
MODEL Y = X1 X5/R INFLUENCE;
OUTPUT OUT=HOSPRES P=PREDICTED R=RES STUDENT=STUDRES;
RUN;

PROC UNIVARIATE NORMAL PLOT;
VAR STUDRES;
RUN;

PROC PLOT;
PLOT RES*PREDICTED RES*X1 RES*X5/VREF=0;
RUN;

*BASED ON THE ABOVE, WE DELETE TWO
```

```
OBSERVATIONS AND STUDY THE EFFECT ON
THE MODEL;

DATA HOSPITAL2;
SET HOSPITAL;
IF(_N_=2 OR _N_=12) THEN DELETE;

PROC REG DATA=HOSPITAL2;
MODEL Y = X1 X5/R INFLUENCE;
OUTPUT OUT=HOSPRES3 P=PREDICTED R=RES RSTUDENT=STUDRESS;
TITLE 'HOSPITAL DATA: X1 AND X5 ONLY WITH 2 OBS REMOVED';
RUN;

PROC PLOT;
PLOT RES*PREDICTED RES*X1 RES*X5/VREF=0;
RUN;
```

8.10 Multicollinearity(Ill Conditioning)

Recall the normal equations are

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}.$$

The solution is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Solving the equations requires that $(\mathbf{X}'\mathbf{X})$ be non-singular for which the condition is $det(\mathbf{X}'\mathbf{X}) > 0$. If $det(\mathbf{X}'\mathbf{X}) = 0$, then $\mathbf{X}'\mathbf{X}$ can not be inverted and there are *exact dependencies* among the columns of **X.** That is, among the independent variables one or more can be written as linear combinations of the other which says they are redundant and should be removed from the model. More troublesome is when dependencies hold only approximately in which the determinate is close to zero. From a computational point of view, $\mathbf{X}'\mathbf{X}$ is difficult to invert due to round off errors in very small numbers. This is no longer a significant problem with modern computers. However, there are other troubling aspects of *Multicollinearity.*

1. Extreme Multicollinearity can cause least square estimates of the β's to be far from their actual values and/or to have the wrong signs. This is because estimates become highly dependent on the particular x values obtained. Small changes in the x's make large changes in the regression equation.

2. Adding or deleting an independent variable can cause large changes in the regression coefficients.

3. Strong Multicollinearity means there is much overlapping information in the $x's$. Hence, t values of the individual $x's$ are small and is difficult to judge which $x's$ are important.

4. Standard errors tend to be very large when strong Multicollinearity is present.

How do we detect Multicollinearity? One way is to inspect the correlation matrix of the $x's$. High correlation among the variables indicate data overlap. This is a good method but not always sufficient since it just looks at redundancy among pairs of the $x's$. Other informal methods are non-significant t tests based on most of the independent variables but with a significant regression model. Another method is to look for regression coefficients with signs opposite from those expected from previous experience or theory. A formal method is provided by the *variance inflation factor.* Recall that the variance is estimated by the $(j+1)$-st diagonal obtained by $s^2(\mathbf{X}'\mathbf{X})^{-1}$. Let $x_1, x_2, ..., x_p$ be the independent variables in the model. It can be shown the $(j+1)$-st diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ is

$$c_{jj} = \frac{1}{(1-R_j^2)\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$$

where R_j^2 is the R^2 value for the model predicting x_j from $x_1, x_2, ..., x_{j-1}$, $x_{j+1}, ..., x_p$. If x_j is totally unrelated to the other independent variables, then $R_j^2 = 0$. If x_j can be perfectly predicted from the other $x's$, then $R_j^2 = 1$. The larger R_j^2, and hence the smaller $1 - R_j^2$, the worst is the Multicollinearity problem and the larger is $s^2(b_j)$. Define the variance inflation factor for b_j as

$$VIF_j = \frac{1}{1-R_j^2}.$$

As R_j^2 grows, so does the variance inflation factor. The *average variance inflation factor* is given by

$$AVIF = \frac{\sum_{j=1}^{p} VIF_j}{p}.$$

Guidelines for the use of variance inflation factors are

1. If the largest $VIF_j > 10$, then there may be a problem.

2. If AVIF is substantially greater than 1, then there may be a problem.

Partial regression plots look for overlap among the independent variables given a single independent variable.

8.10.1 Partial Leverage Plots

Consider expanding the simple linear model

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

to

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

We know that $SS(MODEL)$ using $x_2|x_1$ is equal to $SSMODEL(x_1, x_2) - SSMODEL(x_1)$. That is the sums of squares used to test $H_0 : \beta_2 = 0$ in the full model is the difference in the sums of squares model for the two models. Here we look closely at exactly what is occuring. When we fit the model $y = \beta_0 + \beta_1 x_1 + \epsilon$, we get residuals which for now we call $e_{(2)}$. That is the residuals for the model without x_2. Now, we add x_2 to the model. What can it add to our fit for y?

1. Since some of the variation in the $y's$ has already been explained by x_1, the added contribution of x_2 will be in terms of how well it can explain the as-yet unexplained variation: the residuals $e_{(2)}$.

2. Moreover, any overlapping information between x_1 and x_2 will not help for it is already in the model. Only information in x_2 that is not in x_1 will be of help at this point. What information is this? Fit the model $x_2 = \beta_0' + \beta_1' x_1 + \epsilon'$ and call the residuals from this model $e'_{(2)}$. Then, $e_{(2)}$ is the information in x_2 not found in x_1.

This suggests a plot of $e_{(2)}$ against $e'_{(2)}$. The plot is called the *partial leverage residuals plot.* If the plot shows an approximate straight line, then x_2 will be helpful in explaining y when added to the model already containing x_1. In fact, the slope of the regression of $e_{(2)}$ on $e'_{(2)}$ is the fitted value b_2 for β_2 in the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

SAS Code

This is an illustration of the use of partial regression plots employing the hospital data. Recall that we had previously identified observations 2 and 12 as possibly troublesome. Here we see where they occur in the partial regression plots, then inspect the plots again with observation 2 deleted. We also look at these plots for the full model containing all of x_1 thru x_5, which we know are highly collinear. VIF is demonstrated.

```
DATA HOSPITAL;
INPUT X1 X2 X3 X4 X5 Y;
```

```
IF (X1>199) THEN DELETE;
IF (_N_=2 | _N_=12) THEN PROBLEM='*'; ELSE PROBLEM='O';
CARDS;
15.57 2463 479.92 18.0 4.45 566.52
...
510.22 86533 15524 371.6 6.35 18854.45
;
PROC REG DATA=HOSPITAL;
ID PROBLEM;
MODEL Y=X1-X5/PARTIAL VIF;
TITLE 'HOSPITAL DATA: THE FULL MODEL';
RUN;

PROC REG DATA=HOSPITAL;
ID PROBLEM;
MODEL Y=X1 X5/PARTIAL R INFLUENCE;
TITLE 'HOSPITAL DATA: X1 AND X5 ONLY';
RUN;

DATA FEWER;
SET HOSPITAL;
IF _N_=2 THEN DELETE;

PROC REG DATA=FEWER;
IF PROBLEM;
MODEL Y = X1 X5/PARTIAL R INFLUENCE VIF;
TITLE 'HOSPITAL DATA: X1 AND X5 ONLY WITH OBS. 2 DELETED';
RUN;
```

8.11 Model Building

The *principle of parsimony* suggests models with fewer variables are easier to interpret and are less likely to fit the peculiar features of the present data too closely. Want to balance the models against "good fit:" 1) high R^2, 2) small MS(E), and 3) small prediction intervals.

8.11.1 Some Model Comparison Criteria

One possibility is to try every possible model for the set of $x's$ you have available. With p independent variables, there are 2^p possible models. This is reasonable for small p. But, for large p we need other criteria. We consider

R^2, adjusted R^2, MS(E), and C. These are closely related and will typically lead to similar conclusions.

- R^2 is a good first indicator of model fit. But it suffers from the fact that adding new variables can not decrease R^2.
- Adjusted R^2 : Let p be the number of independent variables in the model. Then adjusted R^2 is given by

$$\left(R^2 - \frac{(k-1)}{n-1}\right)\left(\frac{n-1}{n-k}\right), p = k-1.$$

  Suppose the independent variables are simply lists of random numbers with no relation to y. It can be shown that they will still explain enough variation in y to make $R^2 = \frac{k-1}{n-1}$ on average. So, subtract $\frac{k-1}{n-1}$ from R^2. This will on average make $R^2 = 0$ when the independent variables are unrelated to y. However, this over corrects when the $x's$ are related to y. A perfect fit would give $R^2 = 1$ but the subtraction makes it

$$1 - \frac{k-1}{n-1} = \frac{n-k}{n-1}.$$

  Hence, multiply by $\frac{n-k}{n-1}$ to get the final formula for adjusted R^2. It can be shown that

$$MS(E) = (1 - \text{adjusted } R^2)\left(\frac{SS(TOTAL)}{n-1}\right).$$

  So, if MS(E) decreases, then adjusted R^2 increases.

- Mallow's C_k : Let p be the number of independent variables available. Let k be the number of variables including the intercept in a particular model chosen from those available.

$$C = \frac{SS(E)}{S_p^2} - (n - 2k)$$

  where SS(E) is the sum of squares due to error for the model chosen with $k-1$ independent variables and S_p^2 is the mean square for error for the model using all p independent variables. It can be shown that if a k variable model does not suffer from lack-of-fit, then $E(C) \approx k$. Adequate models should have C close to k. Models having serious lack-of-fit will have C much bigger than k. In general, we would like

1. C_k to be small so that SS(E) is small.
2. C_k be close to k.

Note: if all p variables are used in the model, then $C_{p+1} = p + 1$. Hence, this tells you nothing.

- Press: The deleted residuals (or Press residuals) are defined by $d_i = y_i - \widehat{y}_{(i)} = \frac{e_i}{1-h_{ii}}$. The press statistic is

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 .$$

Large values of Press say the model likely has a fit that too strongly depends on a few measurements. Use Press as one method of comparing several competing models that are being examined in detail.

SAS Code

```
THIS IS THE CREST DATA FROM AN EARLIER ASSIGNMENT. WE ILLUSTRATE
ADJUSTED RSQUARE, ETC, AND SHOW ONE ASPECT OF THE STEPWISE FACILITY;

OPTION LINESIZE = 72 NODATE;

DATA CREST;
INPUT X1 X2 X3 Y;
X1SQ=X1*X1;
X2SQ=X2*X2;
X2SQ=X3*X3;
LABEL X1 = 'BUDGET'
X2 = 'RATIO'
X3 = 'PERSONAL INCOME'
Y = 'CREST SALES'
X1SQ = 'X1 SQUARED'
X2SQ = 'X2 SQUARED'
X3SQ = 'X3 SQUARED';
CARDS;
16.3 1.25 547.9 105.0
...
28.0 1.56 1821.7 245.0
;

*REDUCED MODEL;
```

```
PROC REG DATA=CREST;
MODEL Y = X1-X3/ VIF P INFLUENCE R;
TITLE 'LINEAR REGRESSION ANALYSIS WITH 3 INDEPENDENT VARIABLES';
RUN;

*COMPLETE MODEL;
PROC REG DATA=CREST;
MODEL Y = X1-X3 X1SQ X2SQ X3SQ/VIF P INFLUENCE R;
TITLE 'LINEAR REGRESSION WITH 6 INDEPENDENT VARIABLES';
RUN;

*COMPLETE MODEL;
PROC REG DATA=CREST;
MODEL Y = X1-X3 X1SQ X2SQ X3SQ/SELECTION=STEPWISE INCLUDE=3;
TITLE 'LINEAR REGRESSION ANALYSIS WITH 6 INDEPENDENT VARIABLES';
TITLE2 'USING STEPWISE WITH X1-X3 FORCED INTO THE MODEL';
```

8.11.2 Backwards, Forward, and Stepwise Selection Procedures

- Backward Elimination: operates as follow:

  1. Fit the regression with all the independent variables in the model.
  2. Calculate partial F-test for every variable in the model.
  3. Let F_L be the lowest of the partial F's. If the p-value for $F_L > \alpha_{stay}$ then remove the corresponding variable for the model and return to step 2 with one less variable. If p-value for $F_L \leq \alpha_{stay}$ then adopt the current model and no more variables are removed. The default in SAS is $\alpha_{stay} = 0.10$.

- Forward Selection: Start with no variables in the model. Add variables one-by-one until none meet the criteria for entry.

  1. Start with no variables in the model.
  2. Examine all variables not in the model and choose the one with the largest partial correlation. Calculate the partial F for the variable and label it F_U.
  3. If the p-value for $F_U < \alpha_{entry}$ then add the corresponding variable to the model. Go to step 2. If the p-value for $F_U \geq \alpha_{entry}$ then do not add this or any other variable. Stop here. In SAS, the default for $\alpha_{entry} = 0.5$.

Forward selection is not very popular. Useful for explaining stepwise selection. Something important can be left out of the model due to overlap of information.

- Stepwise Regression: Stepwise regression is a blend of forward and backward selection. It begins with forward selection by adding variables one at a time. But, after a variable is entered, all others in the model are then tested to see if any can be thrown out. Once this is done, those not in the model are examined for entry, etc, etc... This procedure requires two significant levels: α_{entry} and α_{stay}. The defaults in SAS are $\alpha_{entry} = \alpha stay = 0.15$. Always keep $\alpha_{stay} \geq \alpha_{entry}$. Otherwise, variables added will be immediately removed.

SAS Code

```
CHOOSING A MODEL FOR THE HOSPITAL DATA. THIS EXAMPLE WILL DEMONSTRATE
IMPLEMENTATION OF THE VARIOUS VARIABLE SELECTION TECHNIQUES IN SAS;

OPTION NODATE;

DATA HOSPITAL;
INPUT X1-X5 Y;
IF (X1 > 199) THEN DELETE;
LABEL X1 = 'AVERAGE DAILY PATIENT LOAD'
X2 = 'MONTHLY X-RAY EXPOSURES'
X3 = 'MONTHLY OCCUPIED BED DAYS'
X4 = 'ELIGIBLE POP IN AREA(DIVIDED BY 1000)'
X5 = 'MONTHLY LABOR HOURS';
CARDS;
15.57 2463 479.92 18.0 4.45 566.52
...
510.22 86533 15524.00 371.6 6.35 18854.45
;

PROC REG DATA=HOSPITAL;
MODEL Y=X1-X5/SELECTION = FORWARD;
TITLE 'HOSPITAL DATA: FORWARD SELECTION';
RUN;


PROC REG DATA=HOSPITAL;
MODEL Y=X1-X5/SELECTION = BACKWARD;
```

```
TITLE 'HOSPITAL DATA: BACKWARD SELECTION';
RUN;


PROC REG DATA=HOSPITAL;
MODEL Y=X1-X5/SELECTION = STEPWISE INCLUDE=1 SLE=.10 SLS=.20;
TITLE 'HOSPITAL DATA: STEPWISE SELECTION';
RUN;


PROC REG DATA=HOSPITAL;
MODEL Y=X1-X5/SELECTION = RSQUARE CP MSE ADJRSQ BEST=10 STOP=5;
TITLE 'HOSPITAL DATA: RSQUARE SELECTION';
RUN;
```

Some of the options used are as follow: INCLUDE=1 means include the first variable in the model automatically, SLE=.10 is α_{entry}, SLS=.20 is α_{stay}, BEST=10 means only show the best 10 models with highest R^2, and STOP=5 means don't show any models with more than five variables in it.

8.11.3 Transforming X and Y to Get a Linear Fit

Polynomials are by no means the only families of models available to the regression worker. Here are some others:

$$y = \beta_0 + \beta_1 \left(\frac{1}{x_1}\right) + \beta_2 \left(\frac{1}{x_2}\right) + \epsilon,$$

$$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \epsilon,$$

$$y = \beta_0 + \beta_1 \sqrt{x_1} + \beta_2 \sqrt{x_2} + \epsilon,$$

or

$$\ln y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

SAS Code

```
ANALYSIS USING A 2ND-ORDER RESPONSE SURFACE. THIS EXAMPLE IS ADAPTED
FROM NASA TECHNICAL MEMO 81556(1980): CYCLES TILL FAILURE OF
SILVER-ZINC CELLS WITH COMPETING FAILURE MODES-PRELIMINARY DATA
ANALYSIS. RESEARCHERS STUDIED THE EFFECT OF CHARGE RATE AND TEMPERATURE
ON A NEW TYPE OF POWER CELL. THE VARIABLES WERE
X1: CHARGE RATE(.6, 1.0, AND 1.4 AMPERES)
```

```
X2: TEMPERATURE(10, 20, 30C)
Y:  LIFE OF CELL MEASURED IN TERMS OF THE NUMBER OF DISCHARGE-CHARGE
    CYCLES BEFORE FAILURE.
NOT KNOWING THE NATURE OF THE RESPONSE FUNCTION, WE START WITH THE FULL
2-ND ORDER MODEL WHICH SEEMS TO OVERFIT THE DATA. NEXT STEP IS TO
DECIDE HOW TO PARE DOWN THAT MODEL. THE TEST STATEMENT IS ALSO
DEMONSTRATED;

DATA POWER;
INPUT X1 X2 Y;
X1SQ=X1**2; X2SQ=X2**2; X12=X1*X2;
CARDS;
0.6 10 50
1.0 10 86
1.4 10 49
0.6 20 88
1.0 20 157
1.0 20 131
1.0 20 184
1.4 20 109
0.6 20 179
1.0 30 235
1.4 30 224
;

PROC REG DATA=POWER;
MODEL Y = X1 X2 X1SQ X2SQ X12/R INFLUENCE;
LINONLY: TEST X1SQ=0 X2SQ=0 X12=0;
TITLE 'POWER CELLS EXAMPLE';
RUN;

PROC REG DATA=POWER;
MODEL Y = X1 X2 X1SQ X2SQ /R INFLUENCE;
LINONLY: TEST X1SQ=0 X2SQ=0;
RUN;

PROC REG DATA=POWER;
MODEL Y = X1 X2 X1SQ  X12/R INFLUENCE;
LINONLY: TEST X1SQ=0 X12=0;
RUN;

PROC REG DATA=POWER;
```

```
MODEL Y = X1 X2 X1SQ /R INFLUENCE;
OUTPUT OUT=NEW1 R=RESID STUDENT=STUDRES;
RUN;

PROC PLOT DATA=NEW1;
PLOT RESID*X1 RESID*X2;
RUN;

PROC UNIVARIATE PLOT NORMAL;
VAR STUDRES;
RUN;
```

SAS Code

```
THIS EXAMPLE IS FROM BOWERMAN AND O'CONNELL(PAGES 640-643). THE STATE
DEPARTMENT OF TAXATION WISHES TO INVESTIGATE THE TIME TO COMPLETE FORM
ST 1040 AVG AND ITS RELATIONSHIP TO THE
FILER'S EXPERIENCE IN FILLING IT OUT. WE EXPECT TIME TO DECAY WITH
EXPERIENCE, AND EXPECT AN ASYMPTOTIC BEHAVIOR TOWARDS A LOWEST BOUND.
NINE PEOPLE FOR WHOM INCOME AVERAGING IS ADVANTAGEOUS ARE RANDOMLY
SELECTED. WE INVESTIGATE TWO POSSIBLE MODELS FOR THE RELATIONSHIP;

DATA COMPLETE;
INPUT Y X @@;
XINV=1/X; LNX=LOG(X);
LNY=LOG(Y);

CARDS;
80 1 47 8 37 4 28 16 89 1 58 2 20 12 19 5 33 3;

PROC PLOT;
PLOT Y*X;
RUN;

PROC REG;
MODEL Y=XINV/R INFLUENCE;
OUTPUT OUT=NEW1 P=PREDICTD R=RES STUDENT=STUDRES;
TITLE 'MODEL 1';
RUN;

PROC UNIVARIATE NORMAL PLOT;
VAR STUDRES;
```

```
RUN;

PROC PLOT;
PLOT Y*X='O' PREDICTD*X='*'/OVERLAY;
PLOT Y*XINV='O' PREDICTD*XINV='*'/OVERLAY;
RUN;

PROC REG DATA=COMPLETE;
MODEL LNY=LNX;
OUTPUT OUT=NEW2 P=PREDICTD R=RES STUDENT=STUDRES;
TITLE 'MODEL 2';
RUN;

PROC UNIVARIATE NORMAL PLOT;
VAR STUDRES;
RUN;

PROC PLOT;
PLOT LNY*X='O' PREDICTD*X='*'/OVERLAY;
PLOT LNY*LNX='O' PREDICTD*LNX='*'/OVERLAY;
RUN;
```

8.12 Homework and Answers

Text, Chapter 10, # 5.

Text, Chapter 11, # 1, 7, 18. In # 18, you can use all of the techniques learned so far in this class. You are limited, however, to obtaining a final model that has at least 22 degrees of freedom for error. I will summarize the models chosen and show them to the class when the assignment is returned.

10.5: It seems that model # 2 has more multicollinearity. The p-values for most of the variables are large, thus indicating non-significance. We know that the t-statistic is calculated with

$$t = \frac{b_j}{s\sqrt{c_{jj}}}.$$

If c_{jj} which is related to multicollinearity is large(indicating much overlap of data), then the t statistic will be small which will give the impression that a $\beta_j = 0$ and thus the independent variable is not important. All of the variance inflation factors for the second model

are much greater than 10. Thus, we can see that multicollinearity is seriously influencing the least squares point estimates.

11.1: $\bar{R}^2$ is the adjusted R^2 for the number of independent variables.

$$\bar{R}^2 = \left(R^2 - \frac{k-1}{n-1}\right)\left(\frac{n-1}{n-k}\right),$$

where n is the number of observations and $k = p + 1$. Then,

$$\bar{R}^2 = \left(0.8628 - \frac{5-1}{18-1}\right)\left(\frac{18-1}{18-5}\right) = 0.8206.$$

(b): The model with the smallest MS(E) is the 4-th model. It is also the model with the highest R^2 value. Additionally, the prediction interval is smallest with model 4.

Just by looking at the p-values, it does appear that model 3 is best in the table. But we know that multicollinearity could be influencing the p-values in model 4, thus making the p-values insignificant.

11.7: The usual residuals were calculated by taking the difference of the observed dependent variable and the predicted dependent variable. $e_i = y_i - \widehat{y}_i$, where $\widehat{y}_i = 2585.52 + 1.2324x_{i3} - 530.933x_{i5}$ where x_3 is the monthly occupied beds and x_5 is the average length of a patients stay in days.

The press residuals were calculated with the following equation:

$$d_i = \frac{e_i}{1 - h_{ii}},$$

where $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ and e_i is the usual residual. We know that

$$d_{15} = \frac{e_{15}}{1 - h_{15}}.$$

Substituting in values

$$-1448.866 = \frac{-469.07}{1 - h_{15}}.$$

Solve for h_{15} to get 0.6763.

8.13 Remedies for Non-Constant Error Variance

Non-constancy of variances comes chiefly in one of two ways:

1. Var(y) is a function of some independent variable of x.
2. Var(y) is a function of E(y).

Correspondingly, there are two chief avenues for remedy:

1. A weighted analysis: divide the model by a function of x.
2. Apply a transformation: to y, then fit the model for the transformed y.

8.13.1 Weighted Analysis

If the residuals plot of e_i versus x_{ij} shows a fan pattern, then the Var(y) is a function of x_{ij}. We might suspect $\text{Var}(y_i) = \sigma_i^2$, where $\sigma_i = x_{ij}^c \sigma$ for some constant c. Consider the transformed model,

$$\frac{y_i}{x_{ij}^c} = \beta_0 \left(\frac{1}{x_{ij}^c}\right) + \beta_1 \left(\frac{x_{i1}}{x_{ij}^c}\right) + \beta_2 \left(\frac{x_{i2}}{x_{ij}^c}\right) + \cdots + \beta_p \left(\frac{x_{ip}}{x_{ij}^c}\right) + \frac{\epsilon_i}{x_{ij}^c}.$$

Let $y_i^* = \frac{y_i}{x_{ij}^c}$ for the dependent variable in this transformed model. Then,

$$\frac{Var(y_i^*)}{x_{ij}^{2c}} = \frac{1}{x_{ij}^{2c}}\sigma_i = \frac{x_{ij}^{2c}\sigma^2}{x_{ij}^{2c}} = \sigma^2.$$

The transformed model does not have the unequal variances problem. This is called the *weighted analysis.* The weights here are

$$w_i = \frac{1}{x_{ij}^{2c}}.$$

In general, weights are proportional to the reciprocal of the variance. The weighted analysis multiplies the model by $\sqrt{w_i}$.

SAS Code

Results of estimation and prediction using the weighted analysis: our estimate of μ_0 and our prediction of y_0 are

$$\widehat{y}_0 = b_0 + b_1 x_{01} + b_2 x_{02} + \cdots + b_p x_{0p}.$$

Our confidence interval for μ_0 is

$$\widehat{y}_0 \pm t_{\alpha/2}(n-k)s\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

Our prediction interval for y_0 is

$$\widehat{y}_0 \pm t_{\alpha/2}(n-k)s\sqrt{\mathbf{X}_{0j}^{2c} + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

Then, $\sigma_0^2 = \sigma^2 x_{0j}^{2c}$.

```
OPTION NODATE; LINESIZE=72; PS=30;

IN A STUDY OF 27 INDUSTRIAL ESTABLISHMENTS OF VARYING SIZES, THE NUMBER
OF SUPERVISED WORKERS(X) AND THE NUMBER OF SUPERVISORS(Y) WERE
RECORDED. IT WAS DECIDED TO STUDY THE RELATIONSHIP BETWEEN THE TWO
VARIABLES, AND AS A START A SIMPLE LINEAR MODEL WAS POSTULATED.;
*CHATTERJEE AND PRICE, 1991;

DATA SUPER;
INPUT X Y;
LABEL X='NUMBER OF WORKERS' Y='NUMBER OF SUPERVISORS';
CARDS;
294 30
247 32
267 37
358 44
423 47
311 49
450 56
534 62
438 68
697 78
688 80
630 84
709 88
627 97
615 100
```

```
999 109
1022 114
1015 117
700 106
850 128
980 130
1025 160
1021 97
1200 180
1250 112
1500 210
1650 135
750 .
;

PROC REG;
MODEL Y=X/XPX I CLM CLI;
OUTPUT OUT=NEW R=RESID P=PRED;
TITLE 'ORDINARY LEAST SQUARES ANALYSIS';
RUN;


PROC UNIVARIATE PLOT NORMAL;
VAR RESID;
RUN;

PROC PLOT;
PLOT Y*X='O' PRED*X='*'/OVERLAY;
PLOT RESID*X='*'/VREF=0;
RUN;

* THERE ARE TWO WAYS OT GO ABOUT A WEIGHTED ANALYSIS IN SAS. THE FAR
EASIER WAY IS SHOWN HERE -- THE OTHER WAY IS SHOWN IN THE TEXT BOOK.
CREATE A VARIABLE IN YOUR DATASET WHICH CONTAINS THE WEIGHTS YOU WANT TO
USE. THEN JUST RUN PROC REG AS USUAL BUT SPECIFY THE WEIGHTING VARIABLE
IN A 'WEIGHT' STATEMENT. SAS THEN DOES EVERY THING FOR YOU. ONE CAUTION-
IF YOU OUTPUT THE RESIDUALS TO PLOT, YOU WILL HAVE TO WEIGHT THEM
YOURSELF(SEE BELOW).

THE ALTERNATIVE METHOD, SHOWN IN OUR TEXT, IS TO TRANSFORM ALL OF THE
VARIABLES YOURSELF, AND THEN RUN THE TRANSFORMED MODEL WITH PROC REG.
NOT ONLY IS THIS A LOT OF EXTRA WORK, BUT THE CONFIDENCE INTERVALS
```

```
PRODUCED BY CLM AND CLI WILL BE FOR WEIGHTED UNITS, NOT THE ORGININAL
UNITS. I DO NOT RECOMMEND THAT APPROACH! ;

DATA SUPER1;
SET SUPER;
XSQ = X**2; XSQINV=1/XSQ;

PROC REG DATA=SUPER1;
MODEL Y=X/XPX I CLM CLI R INFULENCE;
WEIGHT XSQINV;
OUTPUT OUT=NEW1 R=RESID1 P=PRED1;
TITLE 'WEIGHTED LEAST SQUARES ANALYSIS';
RUN;

DATA NEW1;
SET NEW1;
WTRESID=RESID1/X;

PROC UNIVARIATE PLOT NORMAL DATA=NEW1;
VAR WTRESID;
RUN;

PROC PLOT DATA=NEW1;
PLOT Y*X='O' PRED1*X='*'/OVERLAY;
PLOT WTRESID*X='*'/VREF=0;
RUN;
```

8.13.2 Transformations of Y to Stabilize Variance

1. Try fixing everything else first. This may fix the non-normality problem.

2. If all else fails, try a transformation on y. The more popular one is the Box-Cox family of transformations:

$$y \rightarrow \frac{y^{\lambda} - 1}{\lambda}$$

for some λ.

SAS Code

```
OPTION NODATE; LS=72; PS=30;
```

```
* THIS IS EXAMPLE 13.5(PP 656-660);

DATA TEL;
INFILE 'BOC TAB134 A';
INPUT HOUR NORDERS @@;
Y=SQRT(NORDERS);
HOURSQ = HOUR**2;
LABEL HOUR = 'HOUR OF BUSINESS DAY'
HOURSQ = 'SQUARE OF HOUR'
NORDERS = 'NUMBER OF ORDERS'
Y = 'SQUARE ROOT OF NUMBER OF ORDERS';

PROC SORT;
BY HOUR;
RUN;

PROC UNIVARIATE PLOT;
BY HOUR; VAR NORDERS;
OUTPUT OUT=NEW1 MEAN=HOURMEAN VAR=HOURVAR STD=HOURSTD;
TITLE 'THE UNTRANSFORMED DATA';
RUN;

PROC PRINT DATA=NEW1;
RUN;

PROC REG DATA=TEL;
MODEL NORDERS = HOUR HOURSQ;
OUTPUT OUT=NEW1A STUDENT=STUDRES;
RUN;

PROC UNIVARIATE NORMAL PLOT DATA=NEW1A;
VAR STUDRES;
RUN;

PROC UNIVARIATE PLOT DATA=TEL;
BY HOUR; VAR Y;
OUTPUT OUT=NEW2 MEAN=HOURMEAN VAR=HOURVAR STD=HOURSTD;
TITLE 'DATA WITH TRANSFORMATION ON NUMBER OF ORDERS;
RUN;

PROC PRINT DATA=NEW2;
```

```
RUN;

PROC REG DATA=TEL;
MODEL Y = HOUR HOURSQ;
OUTPUT OUT=NEW2A STUDENT=STUDRES;
RUN;

PROC UNIVARIATE NORMAL PLOT DATA=NEW2A;
VAR STUDRES;
RUN;
```

8.14 Dummy Variables(Indicators)

Usually the independent variables in a regression model cover some continuous range. But, this need not always be the case. Consider the following example.

Example: 13 turkeys are measured for x as the age in weeks and y as the weight in pounds. 4 turkeys are from Georgia, 4 turkeys are from Virginia and 5 turkeys are from Wisconsin. Let

$$z_1 = \begin{cases} 1 & \text{if turkey from GA} \\ 0 & \text{if not} \end{cases}$$

$$z_2 = \begin{cases} 1 & \text{if turkey from VA} \\ 0 & \text{if not} \end{cases}$$

If $z1 = z2 = 0$, then the turkey must be from WI. In general, to represent a categorical variable with s levels, you need $s-1$ dummy variables. The model for turkeys is $y = \beta_0+\beta_1 x+\beta_2 z_1+\beta_3 z_2+\epsilon$. The model for GA turkeys is $y = \beta_0 + \beta_1 x + \beta_2 + \epsilon = (\beta_0 + \beta_2) + \beta_1 x + \epsilon$. The model for VA turkeys is $y = (\beta_0 + \beta_3) + \beta_1 x + \epsilon$. The model for WI turkeys is $y = \beta_0 + \beta_1 x + \epsilon$. Can we get a model that does not assume the slope is the same for each state? Yes: $y = \beta_0 + \beta_1 x + z_1(\delta_0 + \delta_1 x) + z_2(\lambda_0 + \lambda_1 x) + \epsilon$. For WI turkeys: $y = \beta_0 + \beta_1 x + \epsilon$. For GA turkeys: $y = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)x + \epsilon$. For VA turkeys: $y = (\beta_0 + \lambda_0) + (\beta_1 + \lambda_1)x + \epsilon$. These three models imply the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 z_1 x + \beta_5 z_2 x + \epsilon.$$

Let's test to see if VA turkeys and GA turkeys have the same model:

$$H_0 : \beta_2 = \beta_3$$

or

$$H_0 : \beta_2 - \beta_3 = 0.$$

The general setup is $\mathbf{Y = X}\ \beta + \epsilon$. **Let** $\lambda' = (\lambda_0, \lambda_1, \lambda_2, \cdots, \lambda_p)$ **be a vector of constants. Look at testing and estimation of** $\lambda'\beta =$

$$\lambda_0\beta_0 + \lambda_1\beta_1 + \cdots + \lambda_p\beta_p.$$

For the turkey example:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Here, $\lambda' = (0, 0, 1, -1)$. **Here are the basic results:**

1. **The estimates of** $\lambda'\beta$ **are** $\lambda'\mathbf{b}$.
2. **The standard error of the estimate is** $s_{\lambda'\mathbf{b}} = s\sqrt{\lambda'(\mathbf{X}'\mathbf{X})^{-1}\lambda}$.
3. **The test statistic for** $H_0 : \lambda'\beta = 0$ **is**

   $$T = \frac{\lambda'\mathbf{X}}{s\sqrt{\lambda'(\ \mathbf{X}'\mathbf{X})^{-1}\lambda}}.$$

   Reject H_0 **if** $|T| > t_{\alpha/2}(n-k)$.
4. **A** $(1-\alpha)100\%$ **confidence interval for** $\lambda'\beta$ **is**

   $$\lambda'\mathbf{b} \pm t_{\alpha/2}(n-k)s\sqrt{\lambda'(\ \mathbf{X}'\mathbf{X})^{-1}\lambda}.$$

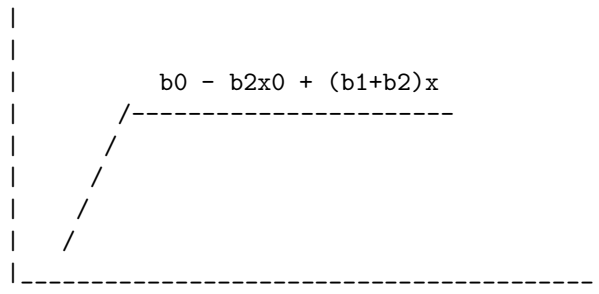
SAS Code

```
USE OF INDICATOR VARIABLES. THIS EXAMPLE IS ON PAGE 243 OF DRAPER AND
SMITH. 13 TURKEYS HAVE BEEN MEASURED FOR THEIR HEIGHT, AGE AND STATE
OF ORIGIN;

DATA GOBBLER;
INPUT X Y  ORIGIN $ Z1 Z2;
Z1X = X1*X; Z2X = Z2*X;
LABEL X = 'AGE'
Y = 'WEIGHT'
ORGIN = 'STATE OF ORIGIN';
```

```
CARDS;
28 13.3 G 1 0
20 8.9 G 1 0
32 15.1 G 1 0
22 10.4 G 1 0
29 13.1 V 0 1
27 12.4 V 0 1
28 13.2 V 0 1
26 11.8 V 0 1
21 11.5 W 0 0
27 14.2 W 0 0
29 15.4 W 0 0
23 13.1 W 0 0
25 13.8 W 0 0
;

PROC REG;
MODEL Y = X Z1 Z2;
OUTPUT OUT=NEW P=PREDICTD;
NOSTATE: TEST Z1=0, Z2=0;
TITLE 'PARALLEL LINES MODEL';
RUN;

PROC PLOT;
PLOT Y*X=ORIGIN PREDICTD*X='*'/OVERLAY;
RUN;

PROC REG DATA=GOBBLER;
MODEL Y = X Z1 Z2 Z1X Z2X;
OUTPUT OUT=NEW1 P=PREDICTD;
SAMEINT: TEST Z1=0, Z2=0;
SAMESLOP: TEST Z1X=0, Z2X=0;
TITLE 'SEPARATE LINES(OR INTERACTION) MODEL';
RUN;

PROC PLOT;
PLOT Y*X=ORIGIN PREDICTD*X='*'/OVERLAY;
RUN;
```

8.14.1 Piece-wise Linear Regression

Another important use of dummy variables is for *piece-wise linear regression.*



We want a different slope to kick-in at $x = x_0$. The model is $y = \beta_0 + \beta_1 x + \beta_2(x - x_0)z + \epsilon$, where

$$z = \begin{cases} 1 & \text{if } x > x_0 \\ 0 & \text{if } x \leq x_0 \end{cases}$$

SAS Code

```
USE OF INDICATORS TO FIT A PIECEWISE LINEAR REGRESSION. THE DATA
FIRST APPEARED IN THE NEW YORK TIMES(SEPT 28, 1975). WE ARE TRYING
TO MODEL LIFE EXPECTANCY AS A FUNCTION OF PER-CAPITA INCOME;

OPTION NODATE;

DATA LIFEINC;
INPUT OBSNO COUNTRY $ LIFE INC;
LABEL COUNTRY = 'COUNTRY'
LIFE = 'LIFE EXPECTANCY'
INC = 'PER-CAPITA INCOME;
DROP OBSNO;

CARDS;
1 AUSTRALIA 71.0 3426
...
101 ZAIRE 38.8 118
;

PROC PLOT;
PLOT LIFE*INC;
```

```
RUN;

DATA LIFEINC1;
SET LIFEINC;
IF INC>1100 THEN Z1=1; ELSE Z1=0;
Z1INC=Z1*(INC-1100);
INCSQ=INC**2;

PROC REG;
MODEL LIFE=INC;
TITLE 'SIMPLE LINEAR REGRESSION';
RUN;

PROC REG;
MODEL LIFE=INC INCSQ;
TITLE 'QUADRATIC MODEL';
RUN;

PROC REG;
MODEL LIFE=INC Z1INC/R INFLUENCE;
OUTPUT OUT=NEW P=PREDICTD R=RESI STUDENT=STUDRESI COOKD=COOKS
   H=LEVERAGE DFFITS=DIFFITS;
ID COUNTRY;
TITLE 'PIECEWISE LINEAR MODEL';
RUN;

PROC PLOT;
PLOT LIFE*INC='O' PREDICTD*INC='*'/OVERLAY;
RUN;

PROC UNIVARIATE PLOT NORMAL;
VAR STUDRESI COOKS LEVERAGE;
ID COUNTRY;
RUN;
```

8.15 Homework and Answers

Text, Chapter 12, # 1,2,3,4(a,c),5(a,b,c,d,e),26,27,28.

12.1: $R^2 = 1 - \frac{SS(E)}{SS(TOTAL)} = 1 - \frac{45}{500} = 0.91$. 91% of the variation in the model is explained by the independent variables and the dummy

variables.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ versus at least one $\beta \neq 0$.

$$F(MODEL) = \frac{MS(MODEL)}{MS(E)} = \frac{\frac{SS(MODEL)}{4}}{\frac{SS(E)}{25-5}} =$$

$$\frac{\frac{500-45}{4}}{\frac{45}{20}} = \frac{113.75}{2.25} = 50.56.$$

$F_{0.05}(4, 20) = 2.87$. Since $50.56 > 2.87$, reject H_0. At least one β does not equal zero.

12.2: $\mu_{[aC]} = \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3(0) + \beta_4(1),$

$\mu - [aA] = \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3(0) + \beta_4(0).$

$\mu_{[aC]} - \mu_{[aA]} = \beta_4.$

$\mu_{[aB]} = \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3(1) + \beta_4(0).$

$\mu_{[aB]} - \mu_{[aA]} = \beta_3.$

$\mu_{[aC]} - \mu_{[aB]} = \beta_4 - \beta_3.$

$\beta_4 = \mu_{[aC]} - \mu_{[aA]}$ measures the effects on mean sales of video recorders when changing from advertiser A to advertiser C. $\beta_3 = \mu_{[aB]} - \mu_{[aA]}$ measures the effects on mean sales of video recorders when changing from advertiser A to advertiser B. $\beta_4 - \beta_3 = \mu_{[aC]} - \mu_{[aB]}$ measures the effects on mean sales of video recorders when changing from advertiser B to advertiser C.

12.3: $H_0 : \mu_{[aC]} = \mu_{[aB]} = \mu_{[aA]}$ versus at least one μ is not equal.

The complete model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 D_{iB} + \beta_4 D_{iC}.$$

The reduced model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

$$F(D_{iB}, D_{iC}|x_{i1}, x_{i1}^2) = \frac{MS_{drop}}{MS(E)_{complete}} =$$

$$\frac{\frac{SS_{drop}}{p-g}}{\frac{SS(E)_{comp}}{n-k}}.$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} =$$

$$\begin{pmatrix} 0.02 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 0.05 \end{pmatrix} \begin{pmatrix} 1000 \\ 300 \\ 50 \end{pmatrix} = \begin{pmatrix} 20 \\ 3 \\ 2.5 \end{pmatrix}.$$

For the reduced model,

$$SS(E)_{drop} = 21500 - \begin{pmatrix} 20 & 3 & 2.5 \end{pmatrix} \begin{pmatrix} 1000 \\ 300 \\ 50 \end{pmatrix} = 475.$$

For the complete model $SS(E)_{drop} = 475 - 45 = 430$. Then,

$$\frac{\frac{SS_{drop}}{4-2}}{\frac{SS(E)_{comp}}{25-5}} \frac{\frac{430}{2}}{\frac{45}{20}} = 95.56.$$

$F_{0.05}(2, 20) = 3.49$. Since $95.56 > 3.49$, reject H_0. At least 2 of the μ's are different or at least one $\beta \neq 0$. The partial F test says that the mean difference between agencies A and C and agencies A and B are significant and should not be removed from the model for at least one of them or both of them. In all practical sense, different advertising agencies do influence the mean sales of video recorders.

12.4: $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$.

$$t = \frac{b_3}{s\sqrt{c_{33}}} = \frac{3}{1.5\sqrt{0.05}} = 8.94.$$

$t_{0.025}(20) = 2.086$. Since $8.94 > 2.086$, reject H_0. $\beta_3 \neq 3$. There does exist a significant difference between advertising thru agency B and agency A.

$$b_3 \pm t_{0.025}(25)s\sqrt{c_{33}}$$

$$3 \pm 2.086(1.5)\sqrt{0.05}$$

$$(2.3, 3.7).$$

This interval makes Panasound 95% confident that the effect of changing from agency A to agency B will increase sales between 2300 units and 3700 units.

12.5: $\lambda = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}$.

$$\begin{pmatrix} 0 & 0 & 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} 0.02 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0 \\ 0 & 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0 & 0.10 \end{pmatrix} \times$$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix} = 0.15.$$

$H_0 : \beta_4 - \beta_3 = 0$ versus $H_1 : \beta_4 - \beta_3 \neq 0$.

$$t = \frac{b_4 - b_3}{s\sqrt{\lambda'(\mathbf{X}'\mathbf{X})^{-1}\lambda}} = \frac{5-3}{1.5\sqrt{0.15}} = 3.44.$$

$t_{0.025}(2) = 2.086$. Since $3.44 > 2.086$, reject H_0. It can be concluded that the effects of advertising agencies B and C do differ.

12.27: $D_{iB} = \begin{cases} 1 & \text{if gas type B.} \\ 0 & \text{otherwise.} \end{cases}$

$$D_{iC} = \begin{cases} 1 & \text{if gas type C.} \\ 0 & \text{otherwise.} \end{cases}$$

If $D_{iB} = D_{iC} = 0$, then the gas type is A.

The variable x_2(additive) plotted against y resembles an upside-down curve. We know that there is interaction between gas type and gas additive when plotted with the dependent variable. Thus, the two terms $D_{iB}x_{i2}$ and $D_{iC}x_{i2}$ should be in the model. The other three terms D_{iB}, D_{iC}, and x_{i2} are the usual independent variables that help predict gas mileage.

The y vector and the $\mathbf{X}$ matrix are

$$y = \begin{pmatrix} 28.0 \\ 28.6 \\ 27.4 \\ 33.3 \\ 34.5 \\ 33.0 \\ 32.0 \\ 35.6 \\ 34.4 \\ 35.0 \\ 34.0 \\ 33.3 \\ 34.7 \\ 33.5 \\ 32.3 \\ 33.4 \\ 33.0 \\ 32.0 \\ 29.6 \\ 30.6 \\ 28.6 \\ 29.8 \end{pmatrix}, X = \begin{pmatrix} x_0 & D_{iB} & D_{iC} & x_2 & D_{ib}x_2 & D_{iC}x_2 & D_{iB}x_2^2 & D_{iC}x_2^2 & \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 2 & 4 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 4 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 & 4 & 2 & 0 & 4 & 0 \\ 1 & 0 & 1 & 2 & 4 & 0 & 2 & 0 & 4 \\ 1 & 0 & 1 & 2 & 4 & 0 & 2 & 0 & 4 \\ 1 & 1 & 0 & 3 & 9 & 3 & 0 & 9 & 0 \\ 1 & 1 & 0 & 3 & 9 & 3 & 0 & 9 & 0 \\ 1 & 0 & 1 & 3 & 9 & 0 & 3 & 0 & 9 \\ 1 & 0 & 1 & 3 & 9 & 0 & 3 & 0 & 9 \end{pmatrix}.$$

12.28

$$\mu_{Bx_2} = \beta_0 + \beta_1 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \beta_5 x_{i2} + \beta_7 x_{i2}^2 + \epsilon_i.$$

$$\mu_{Ax_2} = \beta_0 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \epsilon_i.$$

$$\mu_{Cx_2} = \beta_0 + \beta_2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2 + \beta_6 x_{i2} + \beta_8 x_{i2}^2 + \epsilon_i.$$

$$\mu_{Bx_2} - \mu_{Ax_2} = \beta_1 + \beta_5 x_{i2} + \beta_7 x_{i2}^2.$$

$$\mu_{Cx_2} - \mu_{Ax_2} = \beta_2 + \beta_6 x_{i2} + \beta_8 x_{i2}^2.$$

$$\mu_{Cx_2} - \mu_{Bx_2} = \beta_2 - \beta_1 + \beta_6 x_{i2} - \beta_5 x_{i2} + \beta_8 x_{i2}^2 - \beta_7 x_{i2}^2.$$

$$\frac{\mu_{Cx_2} + \mu_{Bx_2}}{2} - \mu_{Ax_2} = \frac{\beta_1 + \beta_2 + \beta_5 x_{i2} + \beta_6 x_{i2} + \beta_7 x_{i2}^2 + \beta_8 x_{i2}^2}{2}.$$

$\mu_{Bx_2} - \mu_{Ax_2}$ measures the effects on gas mileage when changing from unleaded gas type A to type B at a given amount of x_2. $\mu_{Cx_2} - \mu_{Ax_2}$ measures the effects on gas mileage when changing from unleaded gas type A to type C at a given amount of x_2. $\mu_{Cx_2} - \mu_{Bx_2}$ measures the effect on gas mileage when changing from unleaded gas type B to type C at a given amount of x_2. $\frac{\mu_{Cx_2} - \mu_{Bx_2}}{2} - \mu_{Ax_2}$ measures the effect on gas mileage when changing from unleaded gas type A to the average of C and B.

8.16 1-Factor Experiments

One version of the model statement is $y_{lk} = \mu_l + \epsilon_{lk}$, where μ_l is the average response at the l-th level, and ϵ_{lk} is the error. The usual assumptions for ϵ_{lk} are normality, independence and constancy. Alternate way of writing the model is $y_{lk} = \mu + \tau_l + \epsilon_{lk}$. If a 1-factor model has v numeric variables, then it is equivalent to a $v-1$ degree polynomial model.

In the caffeine example in the SAS Code section, $v = 3$. There are 3 levels of caffeine. The first test is $H_0 : \mu_1 = \mu_2 = \mu_3$, versus H_1 : at least one mean is not equal. If we reject the overall test, we need to look closely at where the differences are. There are two techniques:

1. Multiple comparisons: test all pairs $\mu_l - \mu_{l'}$, which controls use of one which has the smallest coefficient.

2. Estimate and test the more general linear function

   $$\sum_{l=1}^{v} c_l \mu_l,$$

   where $\sum c_l = 0$. These are called *contrasts.* The estimation of a contrast is

   $$\sum_{l=1}^{v} c_l \bar{y}_l,$$

   where $\bar{y}_l$ is the sample mean at level l. The *standard error* is

   $$s\sqrt{\sum_{l=1}^{v} \frac{c_l^2}{n_l}}.$$

   Hence the confidence interval is

   $$\sum_{l=1}^{v} c_l \bar{y}_l \pm t_{\alpha/2}(\sum n_l - v) s \sqrt{\sum_{l=1}^{v} \frac{c_l^2}{n_l}}.$$

   The test statistic is

   $$t = \frac{\sum c_l \mu_l}{s\sqrt{\sum \frac{c_l^2}{n_l}}}$$

   Reject H_0 if $|t| > t_{\alpha/2}(\sum n_l - v)$.

SAS Code

```
*THREE DIFFERENT RUNS OF THE SAME 1-WAY ANOVA DATA;
*SEE HANDOUT FOR AN EXPLANATION OF THE EXPERIMENT THAT
PRODUCED THESE DATA;

OPTIONS NODATE;

DATA TABLE9_1;
INPUT CAFFEINE TAPS @@;
IF CAFFEINE=1 THEN D1=1; ELSE D1=0; *D1 AND D2 ARE ONE POSSIBLE SET;
IF CAFFEINE=2 THEN D2=1; ELSE D2=0; * OF INDICATOR VARIABLES;
CAFFSQ=CAFFEINE**2; *CAFFSQ WILL BE USED IN A QUADRATIC REG;
CARDS;
1 242 1 245 1 244 1 248 1 247 1 248 1 242 1 244 1 246 1 242
...
;

PROC PRINT;
TITLE 'THE DATA AND THE INDICATOR VARIABLES';
RUN;

PROC SORT; BY CAFFEINE;
PROC UNIVARIATE PLOT;
VAR TAPS; BY CAFFEINE;
TITLE 'VISUAL DISPLAY SHOWING EACH LEVEL OF THE EXPERIMENTAL VALUE';
RUN;

PROC REG;
MODEL TAPS = D1 D2;
TITLE 'ONE-WAY ANOVA WITH INDICATORS Z1 AND Z2';

PROC REG;
MODEL TAPS = CAFFEINE CAFFSQ/SS1 SS2;
TITLE 'THE QUADRATIC REGRESSION MODEL';
RUN;

PROC GLM;
MODEL TAPS = CAFFEINE;
MEANS CAFFEINE/BON TUKEY SCHEFFE;
OUTPUT OUT=NEW R=RESI COOKD=COOKS;
TITLE 'THE ONE-WAY ANOVA USING GLM';
```

```
RUN;

PROC UNIVARIATE PLOT NORMAL;
VAR RESI;
TITLE 'RESIDUALS FOR ENTIRE DATA SET';
RUN;
```

SAS Code

```
*DEMONSTRATION OF THE CONTRAST STATEMENT IN PROC GLM;
*SHOWING THE SAME TESTS WITH THE REG REGRESSION;
* THIS IS THE SAME DATA AS IN THE PREVIOUS SECTION;

OPTIONS NODATE;

DATA TABLE9_1;
INPUT CAFFEINE TAPS @@;
IF CAFFEINE=1 THEN D1=1; ELSE D1=0; *D1 AND D2 ARE ONE POSSIBLE SET;
IF CAFFEINE=2 THEN D2=1; ELSE D2=0; * OF INDICATOR VARIABLES;
CARDS;
1 242 1 245 1 244 1 248 1 247 1 248 1 242 1 244 1 246 1 242
...
;

PROC GLM;
CLASS CAFFEINE;
MODEL TAPS = CAFFEINE;
MEANS CAFFEINE/LSD;
CONTRAST '0MG VS 200MG' CAFFEINE 1 0 -1;
CONTRAST '100MG VS AVG OF OTHERS' CAFFEINE 1 -2 1;
CONTRAST '100MG VS 200MG' CAFFEINE 0 -1 1;
TUTLE 'ANALYSIS OF COFFEE DATA';
RUN;

PROC REG;
MODEL TAPS = D1 D2;
CF0VS200: TEST D1=0;
AVGVS100: TEST D1-2*D2=0;
CF100200: TEST -D2=0;
RUN;
```

SAS Code

```
* ANOTHER 1-WAY ANOVA EXAMPLE. THE GROUPING VARIABLE IS
NON-NUMERIC, SO A POLYNOMIAL MODEL DOESN'T HAVE ANY
INTERPRETATION OR MAKE SENSE LIKE IN EXAMPLE 19. BUT THIS
DOESN'T KEEP US FROM DEFINING DUMMY VARIABLES THAT WILL
RUN THE ANOVA AS A REGRESSION, WHICH IS DONE IN THE PROC
REG BELOW. NEVERTHELESS, WE PREFER TO RUN THE ANOVA WITH
PROC GLM, WHICH IS SHOWN AFTER THE REG. THE GLM APPROACH
IS EASIER TO CODE AND THE RESULTS ARE EASIER TO INTERPRET.
THE REG APPROACH IS SOLELY TO HELP US UNDERSTAND THAT
ANALYSIS OF VARIANCE IS A SPECIAL CASE OF REGRESSION,
EVEN THOUGH WE DON'T USUALLY THINK OF IT IN THAT WAY;

* THIS EXAMPLE IS TAKEN FROM BOWERMAN & OCONNELL (PAGE 786).
TO COMPARE THE DURABILITY OF FOUR DIFFERENT BRANDS OF GOLF
BALLS, FIVE OF EACH BRAND ARE SELECTED AND PLACED INTO A
MACHINE THAT EXERTS THE FORCE OF A 250 YARD DRIVE. THE
NUMBER OF SIMULATED DRIVES NEEDED TO BREAK OR CHIP EACH
BALL IS RECORDED;

OPTION NODATE;

DATA HACKER;
INPUT BRAND $ DRIVES @ @;
IF BRAND='ALPHA' THEN Z1=1; ELSE Z1=0;
IF BRAND='BEST' THEN Z2=1; ELSE Z2=0;
IF BRAND='CENTURY' THEN Z3=1; ELSE Z3=0;
CARDS;
ALPHA 281 ALPHA 220 ALPHA 274 ALPHA 242 ALPHA 251
...
;

PROC SORT; BY BRAND;
RUN;

PROC UNIVARIATE PLOT;
VAR DRIVES;
BY BRAND;
TITLE 'VISUAL DISPLAY OF DATA ON EACH BRAND';
RUN;
```

```
PROC REG;
MODEL DRIVES= Z1 Z2 Z3;
CONTRAST1: TEST Z1+Z2+Z3=0;
CONTRAST2: TEST Z1-Z2+Z3=0;
TITLE 'ONE-WAY ANOVA WITH INDICATORS';
RUN;

PROC GLM;
CLASS BRAND;
MODEL DRIVES=BRANDS;
MEANS BRAND/LSD;
CONTRAST 'DIVOT VS OTHERS' BRAND -1 -1 -1 3;
CONTRAST 'ALPHA/CENT VS BEST/DIV' BRAND 1 -1 1 -1;
ESTIMATE 'ALPHA/CENT VS BEST/DIV' BRAND 1 -1 1 -1;
ESTIMATE 'DIVOT MEAN' INTERCEPT 1 BRAND 0 0 0 1;
OUTPUT OUT=NEW R=RESI COOKD=COOKS;
TITLE 'THE ONE-WAY ANOVA USING PROC GLM';
RUN;

PROC SORT; BY BRAND;
RUN;

PROC UNIVARIATE PLOT NORMAL;
VAR RESI;
TITLE 'RESIDUALS FOR ENTIRE DATA SET';
RUN;

PROC UNIVARIATE PLOT;
VAR COOKS;
TITLE 'COOK'S D FOR ENTIRE DATA SET';
RUN;
```

8.17 2-Factor Experiments

In a 2-factor experiment, we have two factors to be assessed for their effects on a response variable. The model is $y_{ijk} = \mu_{ij} + \epsilon_{ijk}$, where y_{ijk} is the k-th response using level i of factor 1 and with level j of factor 2. μ_{ij} is the mean response when using the combination of level i of factor 1 and level j of factor 2. ϵ_{ijk} is the error with the usuall assumptions. An alternative way of writing the model statement is $y_{ijk} = \mu + \alpha_i + \gamma_j + \theta_{ij} + \epsilon_{ijk}$, where μ is the overall mean, α_i is the effect of level i of factor 1, γ_j is the effect

of level j of factor 2, θ_{ij} is the interaction of the i-th level of factor 1 with the j-th level of factor 2.

The primary questions are:

1. Do the factors interact?
2. How does changing the level of factor 1 affect the response?
3. How does changing the level of factor 2 affect the response?

Consider the example in the text book: consider changing from level 1 to leve 2 of shelf height. If width is equal to regular, the response is $\mu_{21}-\mu_{11} = (\mu+\alpha_2+\gamma_1+\theta_{21})-(\mu+\alpha_1+\gamma_1+\theta_{11}) = (\alpha_2-\alpha_1)+(\theta_{21}+\theta_{11})$. If the width is wide, then the change in response is $\mu_{22}-\mu_{12} = (\mu+\alpha_2+\gamma_2+\theta_{22})-(\mu+\alpha_1+\gamma_2+\theta_{12}) = (\alpha_2-\alpha_1)+(\theta_{22}-\theta_{12})$. If no interaction, then the answer is the same whether we look at wide or regular shelves. But, if the factors interact, the answer for wide shelves is different than that for regular shelves.

SAS Code

```
* EXAMPLE OF A 2-FACTOR EXPERIMENT. DEMAND FOR A PRODUCT IS STUDIED AS
A FUNCTION OF SHELF DISPLAY HEIGHT(AT THREE LEVELS: BOTTOM, MIDDLE,
AND TOP) AND DISPLAY WIDTH(AT TWO LEVELS: REGULAR AND WIDE). THREE
MEASUREMENTS HAVE BEEN OBTAINED AT EACH FACTOR COMBINATION. THIS DATA
FORMS THE MAIN EXAMPLE IN CHAPTER 15 OF OUR TEXT;

OPTION NODATE LS=72 PS=40;

DATA SALES;
INPUT A B Y @@;
IF A=1 THEN DA1=1; ELSE DA1=0;
IF A=2 THEN DA2=1; ELSE DA2=0;
IF B=1 THEN DB1=1; ELSE DB1=0;
DAB11=DA1*DB1; DAB21=DA2*DB1;
LABEL A = 'DISPLAY HEIGHT'
B = 'DISPLAY WIDTH'
Y = 'MONTHLY DEMAND';
CARDS;
1 1 58.2 1 1 53.7 1 1 55.8
...
;

PROC FORMAT;
```

```
VALUE DW 1='BOTTOM' 2='MIDDLE' 2='TOP';
VALUE DH 1='REGULAR' 2='WIDE';
RUN;

* LET'S FIRST MAKE A MEANS PLOT FOR THE SIX COMBINATIONS - THIS IS
AN IMPORTANT GRAPHICAL TOOL IN ANALYZING TWO-FACTOR EXPERIMENTS;

PROC MEANS NOPRINT DATA=SALES;
CLASS A B;
VAR Y;
OUTPUT OUT=MEANDAT MEAN=SALESAVG;
RUN;

PROC PRINT DATA=MEANDAT;
RUN;

DATA MEANDAT;
SET MEANDAT; IF (_N_>6.5);

PROC PLOT DATA=MEANDAT;
FORMAT A DW. B DH.;
PLOT SALESAVG*A=B;
PLOT SALESAVG*B=A;
TITLE 'MEANS PLOTS FOR SHELF DATA';
RUN;

PROC REG DATA=SALES;
MODEL Y=DA1 DA2 DB1 DAB11 DAB21;;
HITETEST: TEST DA1+.5*DB1+.5DAB11-DA2-.5DB1-.5DAB21=0,
         DA2+.5DB1+.5DAB21-.5DB1=0;
WIDTEST: TEST 3*DB1+DAB11+DAB21=0;
INTTEST: TEST DAB11=0, DAB21=0;
TITLE1 'SHELF DISPLAY DATA';
TITLE2 'THE REGRESSION APPROACH';

* THE TESTS SHOWN IN PROC REG ARE THE STANDARD TESTS FOR A 2-FACTOR
ANOVA. THEY ARE A BIT MESSY TO WRITE IN TERMS OF THE INDICATORS. ONE
ADVANTAGE OF THE ANOVA APPROACH WITH PROC GLM IS THAT THESE TESTS
ARE GENERATED AUTOMATICALLY FOR YOU;

PROC GLM DATA=SALES;
FORMAT A DW. B DH.;
```

```
CLASS A B;
MODEL Y=A B A*B;
OUTPUT OUT=NEW2 R=RESI COOKD=COOKS;
TITLE2 'THE ANOVA APPROACH';
RUN;

PROC UNIVARIATE PLOT NORMAL;
VAR RESI COOKS;
TITLE 'RESIDUAL ANALYSIS';
RUN;

* SINCE THE INTERACTION TERM IS HIGHLY INSIGNIFICANT, WE MAY WANT TO
REMOVE IT FROM THE MODEL. THIS WILL RESULT IN MORE D.F. FOR THE
ERROR TERM, AND IS USUALLY CALLED POOLING THE INTERACTION WITH ERROR.
WITH THE INTERACTION TERM REMOVED FROM THE MODEL, WE EFFECTIVELY
ANALYZE EACH FACTOR AS IF THE OTHER WAS NOT PRESENT;


PROC GLM DATA=SALES;
FORMAT A DW. B DH.;
CLASS A B;
MODEL Y=A B;
CONTRAST 'TOP VS OTHERS' A -1 -1 2;
ESTIMATE 'WIDE SHELF MEAN' INTERCEPT 1 B 0 1;
MEANS A B / TUKEY BON;
MEANS A / TUKEY BON CLDIFF; * THIS OPTION EXPRESSES THE COMPARISONS IN
TERMS OF CONFIDENCE INTERVALS;
TITLE 'SHELF DISPLAY DATA WITH INTERACTION TERM POOLED TO ERROR';
RUN;

* ONE CONTRAST STATEMENT AND ONE ESTIMATE STATEMENT HAVE BEEN SHOWN
AS EXAMPLES - MANY OTHERS COULD BE DONE. SINCE THERE IS NO INTERACTION
IN THE MODEL, TESTS AND CONFIDENCE INTERVALS INVOLVING LEVELS OF A DO
NOT INVOLVE LEVELS OF B, AND VICE-VERSA;
```

SAS Code

```
* ANOTHER EXAMPLE OF A 2-WAY CROSS-CLASSIFICATION. THIS DATA, FROM THE
JOURNAL OF QUALITY TECHNOLOGY(1969) VIA DRAPER AND SMITH(1981),
CONCERNS A PROBLEM WITH PRODUCTION RATES IN A CATALYST PLANT. AFTER
EXTENSIVE DISCUSSION IN THE RESEARCH UNIT, IT WAS DECIDED TO FOCUS
THE INVESTIGATION ON FOUR REAGENTS AND THREE CATALYSTS;
```

```
OPTION NODATE; LS=72 PS=40;

DATA JQT;
INPUT REAGENT $ CATALYST PRODRATE @@;
IF REAGENT='A' THEN Z1=1; ELSE Z1=0;
IF REAGENT='B' THEN Z2=1; ELSE Z2=0;
IF REAGENT='C' THEN Z3=1; ELSE Z3=0;
IF CATALYST=1 THEN W1=1; ELSE W1=0;
IF CATALYST=2 THEN W2=1; ELSE W2=0;
ZW11=Z1*W1; ZW12=Z1*W2;
ZW21=Z2*W1; ZW22=X2*W2;
ZW31=Z3*W1; ZW32=Z3*W2;
CARDS;
A 1 4 A 1 6 A 2 11 A 2 7 A 3 5 A 3 9
...
;

* FIRST, THE MEANS PLOT;

PROC MEANS NOPRINT;
CLASS REAGENT CATALYST;
VAR PRODRATE;
OUTPUT OUT=MEANDAT MEAN=PRATEAVG;
RUN;

DATA MEANDAT;
SET MEANDAT; IF (_N_>8.5);


PROC PLOT DATA=MEANDAT;
PLOT PRATEAVG*REAGENT=CATALYST;
PLOT PRATEAVG*CATALYST=REAGENT;
TITLE 'MEANS PLOTS FOR PRODUCTION DATA';
RUN;

PROC REG DATA=JQT;
MODEL PRODRATE = Z1 Z2 Z3 W1 W2 ZW11 ZW12 ZW21 ZW22 ZW31 ZW32;
REAT: TEST 3*Z1+ZW11+ZW12, 3*Z2-3*Z1+ZW21+ZW22-ZW11-ZW12,
   3*Z3-3*Z2+ZW31+ZW32-ZW21-ZW22;
INTERACT: TEST ZW11, ZW12, ZW21, ZW22, ZW31, ZW32;
TITLE 'THE REGRESSION APPROACH';
```

```
RUN;

PROC GLM DATA=JQT;
CLASS CATALYST REAGENT;
MODEL PRODRATE=REAGENT CATALYST REAGENT*CATALYST;
TITLE 'THE ANOVA APPROACH';
RUN;

* SINCE THE INTERACTION IS PLAYING A STRONG ROLE, ONE OPTION IS TO
EXAMINE THE 4*3=12 TREATMENT COMBINATIONS AS LEVELS OF A SINGLE
FACTOR, WHICH CAN BE DONE WITH A 1-FACTOR ANOVA;

DATA NEW;
SET JQT;
IF REAGENT='A' THEN DO;
IF CATALYST=1 THEN TREAT='A1';
ELSE IF CATALYST=2 THEN TREAT='A2';
ELSE TREAT='A3';
END;

IF REAGENT='B' THEN DO;
IF CATALYST=1 THEN TREAT='B1';
ELSE IF CATALYST=2 THEN TREAT='B2';
ELSE TREAT='B3';
END;

IF REAGENT='C' THEN DO;
IF CATALYST=1 THEN TREAT='C1';
ELSE IF CATALYST=2 THEN TREAT='C2';
ELSE TREAT='C3';
END;

IF REAGENT='D' THEN DO;
IF CATALYST=1 THEN TREAT='D1';
ELSE IF CATALYST=2 THEN TREAT='D2';
ELSE TREAT='D3';
END;

KEEP PRODRATE TREAT;

PROC PRINT;
RUN;
```

```
PROC GLM DATA=NEW;
CLASS TREAT;
MODEL PRODRATE=TREAT;
MEANS TREAT/TUKEY BON;
OUTPUT OUT=NEW1 R=RESI COOKD=COOKS;
TITLE 'ANALYZING DATA AS 1-WAY CLASSIFICATION DUE TO INTERACTION';
RUN;

PROC UNIVARIATE PLOT NORMAL;
VAR RESI COOKS;
TITLE 'RESIDUAL ANALYSIS';
RUN;
```

8.18 Homework and Answers

1. We were asked to solve Problem 14.46 in the text book. We were asked to analyze the data using `PROC GLM` and `PROC REG` for a 1-factor experiment. In addition, we were asked to run an overall test to see if there are any differences in the factor levels, to use a multiple comparisons technique to investigate any differences in the factor levels, to calculate a few contrasts in means, and finally to verify the assumptions of the model.

   We are to predict aptitude test scores based on 5 levels of college degree type. The 5 levels are: Business, Science, Liberal Arts, Fine Arts, and Engineering. The 5 levels can be represented with the following dummy variables:

$$D1 = \begin{cases} 1, & \text{if degree type is Business} \\ 0, & \text{if degree type is not Business} \end{cases}$$

$$D2 = \begin{cases} 1, & \text{if degree type is Science} \\ 0, & \text{if degree type is not Science} \end{cases}$$

$$D3 = \begin{cases} 1, & \text{if degree type is Liberal Arts} \\ 0, & \text{if degree type is not Liberal Arts} \end{cases}$$

$$D4 = \begin{cases} 1, & \text{if degree type is Fine Arts} \\ 0, & \text{if degree type is not Fine Arts} \end{cases}$$

$$D5 = \begin{cases} 1, & \text{if not Business, Science, Liberal Arts, or Fine Arts} \\ 0, & \text{otherwise} \end{cases}$$

Note that it only takes 4 dummy variables to model the 5 independent variables. This is because if `D1, D2, D3,` and `D4` are all equal to zero, then it can be deduced that the degree type is Engineering.

The model statement produced by `PROC REG` appears as follow:

$$TESTS = 80.6 - 27.88D1 - 5.63D2 - 27D3 - 40.13D4.$$

Both `PROC REG` and `PROC GLM` produced similar results.

Approximately 72% of the variation in the model can be explained by the 5 levels of college degree type(or 4 dummy variables). It appears that `D2` (Science) does not contribute much information to the model. This could be that there is overlapping information with the Engineering variable. We could look at the variance inflation factor to determine whether or not this is the case. But, we were not asked to do this. The overall F-test that tested whether any of the coefficients were not equal to zero had a p-value of 0.0001. Thus we reject the null hypothesis that all of the coefficients are equal to zero. We can conclude that one or more are not equal to zero. A multiple comparisons test was run to determine which factor levels differed from each other. The results using Tukey's method shows that there is a significant difference between both Engineering and Science test scores together, and Liberal Arts, Business and Fine Arts test scores together. That is to mean that there is no significant difference between Engineering and Science test scores, and that there is no significant difference among Liberal Arts, Business, and Fine Arts test scores.

Further tests were made concerning the differences in means. The means of each dummy variable are as follow:

$$\mu_{D1} = \beta_0 + \beta_1 D_1.$$

$$\mu_{D2} = \beta_0 + \beta_2 D_2.$$

$$\mu_{D3} = \beta_0 + \beta_3 D3.$$

$$\mu_{D4} = \beta_0 + \beta_4 D4.$$

$$\mu_{D5} = \beta_0.$$

Even though `D5` for Engineering is not explicitly stated in the model, it is implied to be the y intercept. The following is a list of specific tests performed:

(a) We tested for the difference in average test scores for Science and Engineering. This can be written as:

$$\mu_{D2} - \mu_{D5} = \beta_0 + \beta_2 D2 - \beta_0 = \beta_2 D2 = 0.$$

The result of this test shows that there is no significant difference in average test scores for Science and Engineering students at the 95% confidence level. This is the same result obtained by Tukey's method.

(b) We tested for the difference between business and the average of the two arts majors. This can be written as:

$$\mu_{D1} - \frac{\mu_{D3} + \mu_{D4}}{2} =$$

$$\beta_0 + \beta_1 D1 - \frac{1}{2}\beta_0 - \frac{1}{2}\beta_3 D3 - \frac{1}{2}\beta_0 - \frac{1}{2}\beta_4 D4 =$$

$$\beta_1 D1 - \frac{1}{2}\beta_3 D3 - \frac{1}{2}\beta_4 D4 = 0.$$

The result of this test shows that there is no significant difference in average test scores for the difference between Business majors and the average of the two arts majors at the 95% confidence level.

(c) We tested for the difference between the average of Science and Engineering majors and the average of the other three majors. This can be written as:

$$\frac{\mu_{D2} + \mu_{D5}}{2} - \frac{\mu_{D1} + \mu_{D3} + \mu_{D4}}{3} =$$

$$\frac{\beta_0 + \beta_2 D2 + \beta_0}{2} - \frac{\beta_0 + \beta_1 D1 + \beta_0 + \beta_3 D3 + \beta_0 + \beta_4 D4}{3} =$$

$$\frac{1}{2}\beta_2 D2 - \frac{1}{3}\beta_1 D1 - \frac{1}{3}\beta_3 D3 - \frac{1}{3}\beta_4 D4 = 0.$$

The result of this test shows that there is a significant difference between the average of Science and Engineering majors and the average of the other three majors at the 95% confidence level.

Finally, the assumptions of the model were verified using plots. The residuals were plotted against each of the independent variables. The results show a random pattern for all the variables. Thus constant variance holds true. A stem-and-leaf plot was used to verify the Normality assumption. The residuals appear to be Normally distributed. Thus the assumptions of the model have not been violated.

2. In problem 15.6, we were asked to model the effects of time of day and positioning of advertisements on tele-marketing response. This is a two factored experiment where Factor 1 is the time of day and Factor 2 is positioning of advertisements. Factor 1 has 3 levels: 10:00 morning (rerun of the Honeymooners), 4:00 afternoon(rerun of MASH), and 9:00 evening(first run of Cheers). Factor 2 has 4 levels: on the hour, on the half-hour, early in the program, and late in the program. We were to model this experiment using `PROC GLM.`

   The first thing we needed to determine was whether or not interaction between the two factors were significant. That is because our interpretation of the model would be different if interaction is significant. As it turns out, in this experiment interaction is not significant. This was demonstrated in two ways: 1) the p-value for the mean response of interaction was very large, and 2) the plots of the means vs each independent variable showed parallel patterns. The interaction term was removed from the model and `PROC GLM` was run again.

   Both means of the independent variables appear to be significant. 99% of the variation is explained by the model. Using Tukey's method to determine which levels of the two factors differed, it was determined that for the positioning of advertisements, early in the program differed significantly from late in the program, from on the hour and from on the half-hour. Late in the program differed significantly from on the hour and on the half-hour. However, on the hour and on the half-hour did not differ significantly from each other. Again using Tukey's method on the time of tele-marketing, all three levels differed significantly from each other. That is to say that the mean response of 9:00 evening, 4:00 afternoon, and 10:00 morning differed significantly at the 95% level.

   `PROC UNIVARIATE` was run on Cook's D statistic. It showed that observation 20 may be an outlier.

   To test the constant variance assumption, plots of the residuals and factors were made. Both plots(one for each factor) showed random patterns. Thus, constance variance does hold true. `PROC UNIVARIATE` was run on the residuals. A Normally distributed pattern appeared

in the stem and leaf diagram. Thus, the Normality assumption holds true.

8.19 Logistic Regression

The reference book used in this section is *Applied Logistic Regression,* Homser and Lemeshow, Wiley(1989).

Logistic regression is regression with a binary response variable. An example would be letting y be the presence of coronary heart disease and the independent variables being age, cholesterol, weight, etc. In general, logistic regression deals with regression models for binary response variables. The independent variables can be continuous, categorical, or both as usual. The distinguishing characteristic is that y has only 2 values and thus is not normal. Note that direct prediction of y is useless. Since y can only be 0 or 1, and both are possible at any x, what we will do is model, like in ordinary regression, the expected value of y.

$$E(y|x) = P(y = 1|x)1 + P(y = 0|x)0 = P(y = 1|x) = \Pi(x),$$

which is the probability of having a coronary disease in the example. The models we will use are *intrinsically linear:* as stated, they do not appear to be linear, but become so with a suitable transformation. NOTE: Invert 0 and 1, you would get the model for people without heart disease. The *logistic model* fits s-shaped curves. s-shaped curves

$$\Pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

This can be transformed to a linear model as follow:

$$1 - \Pi(x) = \frac{1}{e^{\beta_0+\beta_1 x}},$$

$$\Rightarrow \frac{\Pi(x)}{1 - \Pi(x)} = e^{\beta_0+\beta_1 x},$$

$$\Rightarrow \ln\left(\frac{\Pi(x)}{1 - \Pi(x)}\right) = \beta_0 + \beta_1 x.$$

How about the error terms? With $\Pi(x)$, as given above, the model is $y = \Pi(x) + \epsilon$. But, unlike in ordinary regression, ϵ is not normal, since y has only 2 possible values.

$$y = 1 \Rightarrow \epsilon = 1 - \Pi(x),$$

and

$$y = 0 \Rightarrow \epsilon = -\Pi(x),$$

8.19.1 Fitting the Logistic Regression Model

In ordinary regression, we fit the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ by least sum of squares. With the normal distribution, the joint density for $y_1, y_2, ..., y_n$ is

$$f(y_1, y_2, ..., y_n) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right)$$

which for purposes of estimation is called the *likelihood.* Maximizing the likelihood function is the same as minimizing SS(E). The technique is called *maximum likelihood estimation* say to estimate β_0 and β_1 by maximizing the likelihood function. In this course, we have presented regression as an estimation technique based on minimizing SS(E) — the underlying reason for doing that is that it *maximizes the likelihood.* Maximum likelihood is the fundamental principle of estimation, and is what we shall use for the logistic regression model.

For our simple logistic regression model, the parameters are $\beta = (\beta_0, \beta_1)$. The likelihood function is

$$L = \prod_{i=1}^{n} \Pi(x_i)^{y_i} [1 - \Pi(x_i)]^{1-y_i}.$$

The *log likelihood* is

$$\log L = \sum_{i=1}^{n} \left(y_i \ln[\Pi(x_i)] + (1 - y_i) \ln[1 - \Pi(x_i)] \right).$$

We want to find the values b_0, and b_1 of β_0 and β_1 which maximizes L. So, evaluate

$$\frac{\partial L}{\partial \beta_0}$$

and

$$\frac{\partial L}{\partial \beta_1},$$

which yields

$$\sum_{i=1}^{n} (y_i - \Pi(x_i)) = 0$$

and

$$\sum_{i=1}^{n} x_i [y_i - \Pi(x_i)] = 0.$$

There is no explicit analytic solution to these equations. Solutions are found computationally by a computer algorithm. Note that a consequence of the first equation is

$$\sum_{i=1}^{n} \frac{y_i}{n} = \sum_{i=1}^{n} \frac{\widehat{\Pi}(x_i)}{n}.$$

SAS Code

```
* THE AGE AND PRESENCE OR ABSENCE OF CORONARY HEART DISEASE, WAS
MEASURED FOR 100 SUBJECTS PARTICIPATING IN A STUDY REPORTED BY
HOSMER AND LEMESHOW(1989). AFTER DETERMINING HOW TO PLOT THE DATA
WE WILL FIT A LOGISTIC REGRESSION OF CHD ON AGE;

OPTION NODATE;

DATA HEARTY;
INPUT AGE CHD @@;
LABEL AGE = 'AGE OF SUBJECT'
CHD = 'CORONARY HEART DISEASE STATUS';
IF AGE<30 THEN AGEGRUP=25;
ELSE IF AGE<35 THEN AGEGRUP=32;
ELSE IF AGE<40 THEN AGEGRUP=37;
ELSE IF AGE<45 THEN AGEGRUP=42;
ELSE IF AGE<50 THEN AGEGRUP=47;
ELSE IF AGE<55 THEN AGEGRUP=52;
ELSE IF AGE<60 THEN AGEGRUP=57;
ELSE AGEGRUP=65;

CARDS;
20 0 23 0 24 0 25 1 26 0 26 0 28 0 28 0 29 0
...
;

* THE VARIABLE AGEGRUP CONTAINS THE MIDPOINTS FOR THE AGE GROUPS.
IT WILL BE USED TO GIVE A PLOT THAT IS MUCH MORE INFORMATIVE THAN
THE RAW DATA PLOT;

PROC PLOT;
PLOT CHD*AGE;
TITLE 'THE RAW DATA PLOT: NOT VERY INFORMATIVE';
RUN;
```

```
PROC MEANS;
CLASS AGEGRUP;
VAR CHD;
OUTPUT OUT=NEW MEAN=CHDPROP;
TITLE 'USING THE MEANS PROCEDURE TO CALCULATE GROUP PROPORTIONS';
RUN;

PROC PRINT DATA=NEW;
TITLE 'OUTPUT DATASET CREATED BY THE MEANS PROCEDURE';
RUN;

DATA NEW;
SET NEW;
IF _N_=1 THEN DELETE;
DROP _TYPE_ _FREQ_;

PROC PLOT DATA=NEW;
PLOT CHDPROP*AGEGRUP='*';
TITLE 'PLOT OF GROUP PROPORTIONS AGAINST GROUP MIDPOINTS';
TITLE2 'NOTICE THE S-SHAPED PATTERN';
RUN;

PROC LOGISTIC DATA=HEARTY DESCENDING;
MODEL CHD=AGE;
TITLE 'LOGISTIC REGRESSION FOR CHD STATUS: AGE IS THE INDEPENDENT VAR';
RUN;

* NOTE: USE THE DESCENDING OPTION TO MODEL THE 1'S AND NOT THE 0'S;
```

8.19.2 Testing for Significance of the Coefficients

As in ordinary regression, when we fit the model, we would like to test independent variables for significance. In ordinary regression, this can be done with t of F tests which come from calculating the change in SS(E) for models with or without variables in question.

In logistic regression the idea is similar, but we do not work with sums of squares. Instead, we work with the likelihood function. We use the change in the log likelihood for the models with and without the variables. For a model with one independent variable, this is $L(\beta_0) - L(\beta_0, \beta_1) =$

$\ln l(\beta_0) - \ln l(\beta_0, \beta_1)$. For theoretical reasons we multiply this by -2 to get:

$$G = -2 \ln l(\beta_0) - (-2) \ln l(\beta_0, \beta_1).$$

Under the hypothesis $\beta_1 = 0$, the distribution of G is approximately $\chi^2(1)$.

There are two other commonly used tests of significance for the independent variables:

1. The Wald test — The Wald test is analogous to the t-test in ordinary linear regression. It compares the MLE $\widehat{\beta}_1$ of the slope to an estimate of its standard error, but then squares the ratio to obtain a χ^2 with 1 degree of freedom. The Wald Chi-square is

$$\left(\frac{\widehat{\beta}_1}{s.e.(\widehat{\beta}_1)} \right)^2 = \frac{\widehat{\beta}_1^2}{s^2 \beta_1}.$$

   The $s.e.(\widehat{\beta}_1)$ calculation is based on an asymptotic expression. For small samples the likelihood ratio test above is preferable.

2. The Score test — The score test comes from the second likelihood equation:

$$\sum_{i=1}^{n} x_i [y_i - \Pi(x_i)] = 0.$$

   If $H_0 : \beta_i = 0$ is true, then $\widetilde{\Pi}(x_i) = \bar{y}$ and so the left hand side of this equation becomes

$$\sum_{i=1}^{n} x_i (y_i - \bar{y})$$

   whose standard error is

$$\sqrt{\bar{y}(1 - \bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

   The score test compares the value to its standard error. Squaring gives an approximate Chi-square. The score Chi-square is

$$\frac{\left[\sum_{i=1}^{n} x_i (y_i - \bar{y})^2 \right]^2}{\bar{y}(1 - \bar{y}) \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

8.19.3 The Multiple Logistic Regression Model

Just as in ordinary regression, we can have multiple independent variables in a logistic regression model. If the independent variables are $x_1, x_2, ..., x_p$, then the multiple logistic regression model is

$$\ln\left(\frac{\Pi(x_i)}{1-\Pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

or

$$\Pi(x_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}.$$

The x's can be continuous or categorical. Hence, we can have ANOVA-like models, regression-like models, and combinations of the two. For x's that are categorical, we have to create dummy variables just as we did for regression formulations of 1-factor and 2-factor experiments.

We wish to use the mle technique to estimate the β's. There will be $p+1$ likelihood equations obtained by differentiating the log likelihood with respect to each of the β's. After simplification, the equations become

$$\sum_{i=1}^{n} y_i - \Pi(x_i) = 0,$$

$$\sum_{i=1}^{n} x_{ij}[y_i - \Pi(x_i)] = 0, i = 1, 2, ..., p.$$

These are solved computationally to get estimates $\widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_p$ and hence $\widehat{\Pi}(x_i)$.

Tests for significance:

1. The overall test for regression $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ based on

   $$G = -2\ln[l(\beta_0)] - [-2\ln l(\beta_0, \beta_1, ..., \beta_p)]$$

   is called the *likelihood ratio test statistic.* Reject H_0 if $G > \chi^2_\alpha(p)$.

2. A score test can be done. Stating the form of this test requires much complicated notation. So, we skip.

3. Wald tests for individual variables are as before. Calculate

   $$\frac{(\widehat{\beta}_j)^2}{s^2(\widehat{\beta}_j)}$$

   and compare to $\chi^2_\alpha(1)$.

4. A general likelihood ratio test for throwing out a subset of variables, say $\beta_{k+1}, \beta_{k+1}, ..., \beta_p$ is

$$G = -2 \ln l(\beta_0, \beta_1, ..., \beta_k) - [-2 \ln l(\beta_0, \beta_1, ..., \beta_p)].$$

$H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$. Reject H_0 if $G > \chi^2_\alpha(p-k)$.

SAS Code

```
FIRST EXAMPLE OF A MULTIPLE LOGISTIC REGRESSION. THE DATA IS FROM
A STUDY ON LOW INFANT BIRTH WEIGHTS. DETAILS ARE ON THE ACCOMPANYING
HANDOUT. HERE WE ILLUSTRATE MULTIPLE LOGISTIC REGRESSION BY FITTING
TWO DIFFERENT MODELS USING A FEW OF THE INDEPENDENT VARIABLES THAT
ARE AVAILABLE;

OPTION NODATE;

DATA LOWBIRTH;
INFILE 'BWT RAW A';
INPUT ID LOW AGE LWT RACT SMOKE PTL HT UI FTV BWT;
IF RACE=2 THEN RACE_1=1; ELSE RACE_1=0;
IF RACE=3 THEN RACE_2=1; ELSE RACE_2=0;

PROC LOGISTIC DESCENDING;
MODEL LOW=AGE LWT RACE_1 RACE_2 FTV;
TITLE 'THE FULL MODEL';
RUN;

PROC LOGISTIC DESCENDING;
MODEL LOW=LWT RACE_1 RACE_2/COVB;
TITLE 'THE REDUCED MODEL';
RUN;

* TO ASSESS THE JOINT CONTRIBUTION OF AGE AND FTV, CALCULATE(BY HAND)
THE DIFFERENCE IN THE -2 LOG L STATISTICS FOR THE TWO MODELS.
```

8.19.4 Lack-of-Fit

We investigate 3 basic methods for evaluating the fit of a logistic regression model. There is no R^2.

1. Evaluation relative to an expanded model. Suppose our current model has independent variables $x_1, x_2, ..., x_k$. Available but not in the model

are $x_{k+1}, ..., x_p$. We may judge our model adequate if we are unable to improve it with the addition of other variables.

In SAS, this can be done in 2 ways: 1) use PROC LOGISTIC to fit each model. The current model and then the model with $x_1, x_2, ..., x_p$. Then, calculate by hand

$$G = -2 \ln l(\beta_0, \beta_1, ..., \beta_k) - [-2 \ln l(\beta_0, \beta_1, ..., \beta_p)]$$

and compare to $\chi^2(p-k)$. We studied this in the previous section. 2) Use of SELECTION=FORWARD in a model that INCLUDES $x_1, x_2, ..., x_k$ automatically. SAS generates a score test for comparing the model $x_1, x_2, ..., x_k$ only to the model with all available variables. There are no calculations by hand here.

2. Hosmer-Lemeshow Lack-of-Fit — Having fit the model, proceed as follow:

   (a) Form groups of observations based on decile of fitted probabilities, ie the observations with lowest 10% predicted probability in group 1, second lowest 10% group, etc to get 10 groups.

   (b) For each group, count how many of the observations have $y = 1$ and how many observations have $y = 0$.

   (c) For each group, calculate the expected number of observations with $y = 1$ and the expected number of observations with $y = 0$. For $y = 1$, just add the predicted probabilities.

   (d) Use a chi^2 test with $t-2$ degrees of freedom to compare observed to expected. Here t is the number of groups.

   SAS implements this with the LACKFIT option. Depending on the number of observations, it may produce $t < 10$ groups.

3. Statistics for Comparing Models — The two statistics introduced here provide a way to compare different models for the same data while accounting for the number of independent variables p and the number of observations N.

   (a) AIC(Akaike's Information Criterion) is $-2 \log L + 2(p + 2)$.

   (b) SC(Schwartz's Criterion) is $-2 \log L + (P + 2) \log N$.

   For either one, smaller values are more desirable.

SAS Code

```
THE CHD DATA. MODELS ARE RUN, ALONG WITH THE LACKFIT OPTION;

OPTION NODATE;

DATA HEARTY;
INPUT AGE CHD @@;
LABEL AGE = 'AGE OF SUBJECT'
CHD = 'CORONARY HEART DISEASE STATUS';
IF AGE<30 THEN AGEGRUP=25;
ELSE IF AGE<35 THEN AGEGRUP=32;
ELSE IF AGE<40 THEN AGEGRUP=37;
ELSE IF AGE<45 THEN AGEGRUP=42;
ELSE IF AGE<50 THEN AGEGRUP=47;
ELSE IF AGE<55 THEN AGEGRUP=52;
ELSE IF AGE<60 THEN AGEGRUP=57;
ELSE AGEGRUP=65;

IF AGEGRUP=25 THEN Z1=1; ELSE Z1=0;
IF AGEGRUP=32 THEN Z2=1; ELSE Z2=0;
IF AGEGRUP=37 THEN Z3=1; ELSE Z3=0;
IF AGEGRUP=42 THEN Z4=1; ELSE Z4=0;
IF AGEGRUP=47 THEN Z5=1; ELSE Z5=0;
IF AGEGRUP=52 THEN Z6=1; ELSE Z6=0;
IF AGEGRUP=57 THEN Z7=1; ELSE Z7=0;

AGE2=AGE**2; AGE3=AGE**3; AGE4=AGE**4;
AGE5=AGE**5; AGE6=AGE**6; AGE7=AGE**7;
AGEGRUP2=AGEGRUP**2; AGEGRUP3=AGEGRUP**3; AGEGRUP4=AGEGRUP**4;
AGEGRUP5=AGEGRUP**5; AGEGRUP6=AGEGRUP**6; AGEGRUP7=AGEGRUP**7;

CARDS;
20 0 23 0 24 0 25 1 26 0 26 0 28 0 28 0 29 0
...
;

PROC LOGISTIC DATA=HEARTH DESCENDING;
MODEL CHD=AGE/LACKFIT;
TITLE 'LOGISTIC REGRESSION WITH AGE AS INDEPENDENT VARIABLE';
RUN;
```

```
PROC LOGISTIC DATA=HEARTY DESCENDING;
MODEL CHD=AGE AGE2 AGE3 AGE4/SELECTION=FORWARD INCLUDE=1 DETAILS;
TITLE 'USING FORWARD SELECTION TO EVALUATE RESIDUAL SCORE STATISTICS';
TITLE2 'EVALUATES LINEAR MODEL IN CONTEXT OF A 4TH DEGREE POLYNOMIAL';
RUN;

PROC LOGISTIC DATA=HEARTY DESCENDING;
MODEL CHD=Z1 Z2 Z3 Z4 Z5 Z6 Z7/LACKFIT;
TITLE 'SATURATED MODEL USING INDICATORS FOR GROUPED AGES';
TITLE2 'THIS IS LIKE A 1-FACTOR ANOVA IN CHPT 14 OF BOWERMAN-OCONNELL';
TITLE3 'BUT NOW THE DEPENDENT VARIABLE IS 0-1';
RUN;

PROC LOGISTIC DATA=HEARTY DESCENDING;
MODEL CHD=AGEGRUP AGEGRUP2 AGEGRUP3 AGEGRUP4 AGEGRUP5 AGEGRUP6 AGEGRUP7/LACKFIT;
TITLE 'SATURATED POLYNOMIAL MODEL FOR GROUPED AGES';
RUN;

* THESE LAST TWO PROC LOGISTICS DEMONSTRATE THAT, LIKE IN ORDINARY
REGRESSION, IF GROUPS ARE DEFINED BY A NUMERIC VARIABLE, WITH I VALUES,
THEN FITTING AN I-1 DEGREE POLYNOMIAL IS EQUIVALENT TO USING I-1
INDICATORS. IN THE LOGISTIC REGRESSION SETTING, A MODEL WHICH CONTAINS
A PARAMETER FOR EVERY LEVEL OF THE GROUPING VARIABLE IS SAID TO BE
"SATURATED.";
```

8.19.5 Interpreting the Regression Coefficients

Interpreting the coefficient involves 2 issues:

1. Determining the functional relationship between the dependent variable and the independent variable.

2. Appropriately defining the unit of change for the independent variable.

For instance, with the simple logistic regression with only one independent variable

$$g(x) = \ln\left(\frac{\Pi(x)}{1-\Pi(x)}\right) = \beta_0 + \beta_1 x$$

where $g(x)$ is the logit transform. Then $\beta_1 = g(x+1) - g(x)$ is the change in the logit for unit change in x.

1. Categorical Dictonomous Independent Variables — A *dictonomous variable* only has two categories.

   Example: $age < 55$, $age \geq 55$.

   With two categories, we need 1 dummy variable with values 0 and 1. So,

   $$g(1) = \ln\left(\frac{\Pi(1)}{1-\Pi(1)}\right),$$

   $$g(0) = \ln\left(\frac{\Pi(0)}{1-\Pi(0)}\right).$$

   $$\beta_1 = g(1) - g(0) = \ln\left(\frac{\frac{\Pi(1)}{1-\Pi(1)}}{\frac{\Pi(0)}{1-\Pi(0)}}\right)$$

   which is called the *log odds ratio.* The *odds ratio* is given by

   $$e^{\beta_1} = \frac{\frac{\Pi(1)}{1-\Pi(1)}}{\frac{\Pi(0)}{1-\Pi(0)}}.$$

   The confidence interval for the odds ratio is $e^{\widehat{\beta} \pm z_{\alpha/2} s.e.(\widehat{\beta})}$. The group in the denominator is called the *reference group.*

2. Polytomized Independent Variable — When an independent variable is categorical with more than 2 groups, it is said to be *polytomious.* If there are k categories, we need $k-1$ indicators(dummy variables) for the logistic model. Our model is

   $$g(z) = \ln\left(\frac{\Pi(z)}{1-\Pi(z)}\right) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_{k-1} z_{k-1}$$

   where $z_1, z_2, ..., z_{k-1}$ are indicators defined in the usual way. The group that does not have an indicator is called the *reference group.* We seek meaning to $\beta_1, \beta_2, ..., \beta_{k-1}$. Consider $\mathbf{Z} = (z_1, z_2, ..., z_{k-1})$ Then, $z_1 = (1, 0, ..., 0)$. $z_0 = (0, 0, ..., 0)$. $g(z_1) - g(z_0) = (\beta_0 + \beta_1) - \beta_0 = \beta_1$. $beta_1$ is the log odds ratio for group 1 relative to group k. β_2 is the log odds ratio for group 2 relative to group k. Thus, e^{β_1} is the odds ratio for group 1, etc. The odds ratio for comparing group j to group j' is $e^{\beta_j - \beta'_j}$.

3. Continuous Independent Random Variable — The equation for the logit is

$$g(x) = \ln\left(\frac{\Pi(x)}{1-\Pi(x)}\right) = \beta_0 + \beta_1 x.$$

From that

$$\beta_1 - g(x+1) - g(x) = \ln\left(\frac{\frac{\Pi(x+1)}{1-\Pi(x+1)}}{\frac{\Pi(x)}{1-\Pi(x)}}\right)$$

which is the log odds ratio for comparing individuals separated by 1 unit of x. The estimate is $e^{\widehat{\beta}_1}$. The confidence interval is $e^{\widehat{\beta}_1 \pm z_{\alpha/2} s.e.(\widehat{\beta}_1)}$. CAUTION: Does a 1 unit change in x have meaning? Sometimes yes and sometimes no.

4. Multivariate Case(No Interaction) — Take the low birth rate data. Consider the model with LWD(x_1) and age(x_2). The model is

$$\ln\left(\frac{\Pi(x_1,x_2)}{1-\Pi(x_1,x_2)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = g(x).$$

Then, $g(1,x_2) - g(0,x_2) = (\beta_0 + \beta_1 + \beta_2 x_2) - (\beta_0 + \beta_2 x_2) = \beta_1$.

$$\beta_1 = \ln\left(\frac{\frac{\Pi(1,x_2)}{1-\Pi(1,x_2)}}{\frac{\Pi(0,x_2)}{1-\Pi(0,x_2)}}\right)$$

is the odds ratio for comparing the low-last weight group to the high last weight group adjusting(accounting) for age. x_2 must be kept fixed. Vary x_1 by 1 unit.

More generally, for a model with independent variables $x_1, ..., x_p$

$$Y(x) = \ln\left(\frac{\Pi(x)}{1-\Pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

β_i is the log odds ratio for a unit change in x_i holding the other $x's$ fixed.

5. Multivariate Case with Interaction — What distinguishes the interaction case from the no-interaction case is no interaction can change the variable of interest while keeping other variables fixed. In the interaction model, changing the variable of interest changes other variables, also.

SAS Code

```
MODEL-FITTING WITH THE LOW BIRTHWEIGHT DATA;


OPTION NODATE;

DATA LOWBIRTH;
INFILE 'BWT RAW A';
INPUT ID LOW AGE LWT RACT SMOKE PTL HT UI FTV BWT;
IF RACE=2 THEN RACE_1=1; ELSE RACE_1=0;
IF RACE=3 THEN RACE_2=1; ELSE RACE_2=0;

IF LWT<110 THEN LWD=1; ELSE LWD=0; *DICHOTOMIZING;
IF PTL>0.5 THEN PTD=1; ELSE PTD=0; *DICHOTOMIZING;

* THE FOLLOWING ARE POSSIBLE 2-WAY INTERACTION TERMS;

R1LWD=RACE_1*LWD; R1SMOKE=RACE_1*SMOKE; R1PTD=RACE_1*PTD;
R1HT=RACE_1*HT;
R2LWD=RACE_2*LWD; R2SMOKE=RACE_2*SMOKE; R2PTD=RACE_2*PTD;
R2HT=RACE_2*HT;

LWDSMOKE=LWD*SMOKE; LWDPTD=LWD*PTD; LWDHT=LWD*HT;
SMOKEPTD=SMOKE*PTD; SMOKEHT=SMOKE*HT; PTDHT=PTD*HT;

* WE WILL EXAMINE TWO-WAY TABLES TO ASSESS THE MARGINAL BENEFIT OF
EACH VARIABLE. THIS IS ANALOGOUS TO INITIAL PLOTS OF Y VERSUS X
IN ORDINARY REGRESSION;

PROC FREQ;
TABLES LWD*LOW RACE*LOW SMOKE*LOW PTD*LOW HT*LOW
UI*LOW FTV*LOW/NOCOL NOPERCENT CHISQ;
TITLE 'TWO-WAY TABLES: DEPENDENT AGAINST EACH PREDICTOR';
RUN;

PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 AGE LWD SMOKE PTD HT UI FTV/INCLUDE=2
SELECTION=STEPWISE SLENTRY=.15 LSSTAY=.20;
TITLE 'STEPWISE SELECTION WITH RACE INCLUDED';
RUN;
```

```
PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 AGE LWD SMOKE PTD HT UI FTV/
SELECTION=BACKWARD SLSTAY=.20;

TITLE 'BACKWARD SELECTION';
RUN;

PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 LWD SMOKE PTD HT
R1LWD R1SMOKE R1PTD R1HT
R2LWD R2SMOKE R2PTD R2HT
LWDSMOKE LWDPTD LWDHT
SMOKEPTD SMOKEHT PTDHT/INCLUDE=6
SELECTION=FORWARD;
TITLE 'TRYING INTERACTION TERMS/CALCULATION OF RESIDUAL SCORE STATISTIC';
RUN;


PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 LWD SMOKE PTD HT
R1LWD R1SMOKE R1PTD R1HT
R2LWD R2SMOKE R2PTD R2HT
LWDSMOKE LWDPTD LWDHT
SMOKEPTD SMOKEHT PTDHT/INCLUDE=6
SELECTION=STEPWISE SLSTAY=.20 SLENTRY=.15;
TITLE 'STEPWISE INVESTIGATION OF TWO-WAY INTERACTION TERMS';
RUN;

PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 LWD SMOKE PTD HT
R1LWD R2LWD R1SMOKE R2SMOKE LWDHT/INCLUDE=6
SELECTION=FORWARD DETAILS;
TITLE 'RESIDUAL SCORE STATISTIC FOR A SUBSET OF INTERACTIONS';
RUN;

PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 LWD SMOKE PTD HT
R1LWD R2LWD;
TITLE 'RACE-LWD INTERACTION ONLY';
RUN;
```

```
PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 LWD SMOKE PTD HT
R1SMOKE R2SMOKE;
TITLE 'RACE-SMOKE INTERACTION ONLY';
RUN;

PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 LWD SMOKE PTD HT
LWDHT;
TITLE 'LWD-HT INTERACTION ONLY';
RUN;

PROC LOGISTIC DESCENDING;
MODEL LOW=RACE_1 RACE_2 LWD SMOKE PTD HT /LACKFIT;
TITLE 'THE MAIN EFFECTS MODEL';
RUN;
```

Chapter 9

Clinical Trials

Dr. Lee, Old Dominion University

STAT 540, Spring 1997

Text used: Friedman, L. M., C. D. Furberg, and D. L. DeMets, *Fundamentals of Clinical Trials, 3-rd edition,* 1996

9.1 Outline

1. Definition and types of trials.
2. Clinical trial phases.
3. The study protocol.
4. Essentials of good clinical trial design.
5. Brief history of clinical trials.
6. Defining the study population.
7. Four major study designs: their strengths and weaknesses.
8. Methods of randomization: simple, blocked, stratified.
9. Baseline assessment.
10. Blinding or masking of treatment assignments.
11. Monitoring compliance.

12. Exclusions, withdrawals, losses: which subjects should be included in the analysis of trial results?

A clinical trial is a prospective study of the effects of one or more test treatments and a control treatment on human subjects. It is generally accepted that a randomized controlled trial is the most reliable method of conducting clinical research. The subjects are enrolled, treated, and followed forward in time; however, they need not all enter the trial simultaneously, and in fact, patients often enter the trial at different times. The outcome measure may be death, a non-fatal clinical event or a laboratory test. The time period of the trial may be short or long, depending on the outcome measure. In the phases, the historical controls include 1) animal studies, 2) healthy volunteers, and 3) borderline individual case studies. Under this definition, studies that use a historical control group do not qualify as a clinical trial. Experiments on healthy human volunteers are borderline in that they provide only indirect evidence of the effect of a treatment. An individual case study, wherein one individual's pattern of treatment and response is reported as an interesting occurrence, also does not constitute a clinical trial. This is because other patients with the same condition will almost certainly show varied responses to the same treatment. That is, the experience of one patient does not adequately enable an inference to be made about the effect of treating future patients in the same way. One of the main problems in conducting a trial is getting a large enough group of patients on different treatments to make reliable treatment comparisons. There are various ways of classifying clinical trials. One way is by the type of treatment. The great majority of clinical trials are concerned with evaluation of drug therapy. However, clinical trials may also be concerned with surgical procedures, radiation treatments, and different forms of medical advice (e.g. diet and exercise). The treatments used in clinical trials are often called interventions and are classified as follows:

1. Prophylactic — a treatment aimed at preventing the development of a disease or condition (e.g. a new vaccine).

2. Therapeutic — a treatment aimed at curing or healing a disease or condition (e.g. drugs, devices, procedures).

3. Diagnostic — identifying the sign or symptoms by which a disease is known (e.g. an AIDS test).

9.1.1 Clinical Trial Phases

The process of running clinical trials to qualify a new drug for sale to the public has been classified by the FDA into the following three phases:

1. Phase I Trial (Early) — Determine whether there is a dosage regimen not overly toxic and suitable for further study of therapeutic effect. The sample size is usually 4-10 subjects and lasts no more than 20-80 days. In this design, select healthy volunteers who are given a new drug; observations are made; then a washout period; the subject is then given the next drug in the same manner; order of receiving the drugs is randomized.

2. Phase II Trial (Efficacy) — Estimate the effectiveness of the regimen after passing a preliminary trial. The sample size is usually 20-80 subjects; usually no more than 150-200 subjects; In this design, the features are controlled; double-blinded; monitored; and randomized.

3. Phase III Trial (Effectiveness) — Compare the effectiveness of the new treatment with a standard or control treatment. The sample size is usually several thousand subjects.

Efficacy refers to what the treatment will accomplish in an ideal setting where all participants are eligible to receive the treatment and comply perfectly with the assigned treatment and treatment schedule. Effectiveness refers to what the treatment will accomplish in actual practice. For example, if 60% of the patients respond favorably to a new treatment under ideal conditions, then the treatment efficacy is a 60% success rate. Our text (Chapter 14) emphasizes the importance of analyzing the trial results in ways that reflect the treatment effectiveness.

9.1.2 The Study Protocol

A written protocol must document the purpose, design, and conduct of a clinical trial. It must be specific with respect with respect to which patients are to be included and excluded from the trial, the treatments to be tested, the outcome measures, etc. The following list is a brief outline of a protocol document.

1. Background of the study.

2. Objectives

   (a) Primary question and response variable.
   (b) Secondary questions and response variables.
   (c) Sub-group hypotheses.

3. Design of the study.
   (a) Study population.
      i. Inclusion and exclusion criteria.
      ii. Sample size estimates.
   (b) Enrollment of subjects.
      i. Informed consent.
      ii. Assessment of eligibility.
      iii. Baseline examination.
      iv. Treatment allocation.
   (c) Intervention (i.e. treatment).
      i. Description and schedule.
      ii. Measures of compliance — want this to be high.
   (d) Follow-up visit description and schedule.
   (e) Ascertainment of response variables.
      i. Training.
      ii. Data collection.
      iii. Data monitoring and quality control.
      iv. Data analysis.
   (f) Organization.
      i. Participating investigators.
      ii. Study administration.
         A. Committees and sub-committees.
         B. Policy and data monitoring committee.

9.1.3 Essentials of Good Clinical Trial Design

This section comes from Wooding, Planning Pharmaceutical Clinical Trials, 1994. There are many features of a good trial design that are important, but there are three that should always be present, except in rare cases.

1. The use of concurrent controls (a placebo, a standard treatment, or both).

2. Double-blinding.

3. Randomization.

A trial lacking any of these features is likely to be rejected by the FDA and will be unconvincing to other informed critics such as journal readers. The use of before and after tests, that omit concurrent controls and use only baseline data, is not an acceptable substitute for concurrent controls. One reason for this is that baseline responses may change from one time period to another, even in the absence of a treatment effect.

9.1.4 The History of Clinical Trials

The modern era of properly designed clinical trials begins about 1950. Before this time, most trials did not involve a control group or randomization.

By 1950, or shortly thereafter, most of the ground rules for conducting clinical trials on a scientific basis had been established. The concept of multiple investigators, different study sites, and a common study protocol emerged in the 1930's. The first trial with a properly randomized control group (1948) studied streptomycin as a treatment of pulmonary tuberculosis. The U.S. polio vaccine trials began in 1953 and involved thousands of volunteers and a placebo group. One of the first multi-center, randomized trials in the U.S. was the University Group Diabetes Program (1961 - 1974). It involved seven clinics, four treatment groups, and a sample size over 1,000 patients. In 1969, the FDA began requiring evidence from a randomized controlled trial to gain approval for marketing a new drug. It was only in the mid 1970's that properly done clinical trials began to be widespread in the pharmaceutical industry. Any clinical trial requires a clear definition of which patients are eligible for inclusion and also a list of exclusion criteria to supplement the main definition of the disease. The study population is a subset of the general population defined by the inclusion-exclusion criteria. Usually, subjects are not randomly chosen from the study population. Judgments about whether the study results can be generalized from the subjects actually in the trial to the study population can be made from defined medical conditions and by observing explanatory variables such as age, sex, elevated blood pressure, etc. The impact of the eligibility criteria on the recruitment of patients should be considered when deciding on these criteria. Excessive restrictions implies extreme difficulty in recruiting an adequate number of participants.

9.1.5 General Criteria for Inclusion and Exclusion

1. Include those patients most likely to benefit from the treatments being studied (those that have the disease being studied and those that are likely to respond to treatment). Knowing the mechanism of action of a treatment enables an investigator to identify those subjects most likely to respond to treatment. For example, type of bacteria and site of an infection enables determining whether a new antibiotic may be effective in treating a urinary infection. If the mechanism of action of a treatment is unclear or if there is uncertainty about the stage of the disease at which a treatment may be most beneficial, then the subjects that are most likely to respond to the treatment cannot easily be selected. An unknown mechanism of action implies a) a more heterogeneous study population, b) a reduced chance of detecting a treatment difference.

2. Include those subjects for which there is a high likelihood of detecting a treatment difference. If the primary response variable is continuous (e.g. blood pressure, serum cholesterol level, etc) change is easier to detect if the initial level is extreme. For example, one would expect a more pronounced drop in blood pressure in a subject with an initial systolic blood pressure of 135mm of Hg than in a subject with an initial BP of 125mm of Hg. Subjects with frequent cardiac arrhythmic episodes are more likely to respond to an anti-arrhythmic agent than subjects with only one brief episode.

3. Exclude any subject for which a treatment is known to be harmful. Pregnant women are often excluded from drug trials unless the primary question concerns pregnancy. People with a history of gastric bleeding are usually excluded from trials of almost any anti-inflammatory drug. These exclusions apply only before the enrollment of a subject in a trial; participants my develop conditions during a trial which would have excluded them had the conditions been present earlier. In these circumstances, a participant may be removed from a treatment but kept in the trial for the purpose of analysis.

4. Exclude subjects at high risk of developing conditions that preclude observing the event of interest. People with cancer or severe kidney disorders are usually excluded from trials in which the primary question concerns heart disease, as they may die or withdraw from the study before the primary response variable can be observed. When there is a competing risk, the ability to assess a treatment effect is decreased and biased results may be obtained for the primary study

question if the treatment in some way has a harmful or beneficial effect on the co-existing condition.

5. Include only those subjects who are likely to comply with the assigned treatment and treatment schedule. Non-compliance reduces the opportunity to observe the effect of a treatment. Unfortunately, there is no way to guarantee that a given subject will comply.

9.1.6 Example of Selection Criteria

This example of selection criteria of patients comes from the study of chemotherapy trial for advanced colorectal cancer (Pocock, 1983).

1. Patients must have histologically confirmed metastic or locally recurrent carcinoma of the colon of the rectum.
2. The tumor must be beyond hope of surgical eradication.
3. The tumor masses must be measurable by physical examination or chest x-ray.
4. No previous chemotherapy treatment for the disease.
5. Expected survival of at least 90 days and absence of severe malnutrition, nausea, and vomiting.
6. If the white blood cell count is too low, then the patient can not tolerate chemotherapy. The white blood cell count should be greater than 4,000 per mm^3. The platelet count should be greater than 100,000 per mm^3. The haemoglobin should be greater than 10 grams per 100 ml. The creatinine should be less-than 1.5 mg per 100 ml.

9.1.7 Baseline Assessment

Baseline refers to the status of a patient before he is randomized to a treatment group. Baseline data may be collected by interview, questionnaire, physical exam, and laboratory tests. These include personal characteristics (e.g. age, sex) and clinical history (e.g. duration of illness and previous therapy); includes measurable parameters that may change during the trial (e.g. blood pressure, tumor size, Hamilton rating of depression). The following lists the uses of baseline measurements.

1. To determine patient eligibility.

2. To characterize the population to which the results of the trial may be generalized.

3. To identify subgroups of patients for which a treatment may have a beneficial or harmful effect.

4. To stratify the patients at the time of randomization or during the analysis of trial results.

Problem 1: A patient's baseline status may change before the initiation of treatment. By definition, a baseline measurement is any measurement taken at or before randomization. Any change in a baseline measurement that occurs after randomization is usually regarded as an outcome of the trial and is, therefore, not a baseline measurement. A patient's baseline status may change between the time of randomization and the time that the treatment begins. For example, a patient's diastolic blood pressure may drop from 95 mm of Hg to 85 mm of Hg. Such a change may dilute the results of the trial and thus decrease the chance of detecting a treatment difference. Another example, a new antihypertensive agent would not produce as large of a decrease in dbp as it would have in the baseline value where higher. Possible solution: wait until the latest possible moment to randomize each patient.

Problem 2: Observer variation. Some clinicians and nurses record consistently higher or lower blood pressure values than others. Possible solutions: use training sessions; take repeated measurements on each subject using two or three different observers and record the average.

Problem 3: Objective assessment. Example: three cardiologist were asked to interpret the same ECGs for 1,252 patients suspected of heart disease. They disagreed in 132 cases on whether the patient had an infarct (Pocock, 1983). Possible solutions: use an objective classification of ECGs; use a panel of two or three observers to resolve the differences.

9.1.8 Four Major Study Designs

1. The randomized controlled design. Patients are randomly allocated to treatment and control groups.

2. Non-randomized concurrent controls. One group is given a standard treatment at approximately the same time as another group receiving the new treatment. For example, the survival time is compared for

patients in each of two different hospitals where one hospital is using a new surgical technique and the other is using traditional care.

3. Historical controls. A new treatment is used on a current set of patients and the results are compared to those obtained at a different point in time for a "comparable" group of patients that received a standard treatment.

4. Cross-over design. Patients are randomized to 2 groups. Patients assigned to one group first receive treatment A and then later receive treatment B. Patients in the other group receive B followed by A. This design uses small sample sizes because of "within" variability.

The following table lists the strengths and weaknesses of the four study designs.

Study Design	Strengths	Weaknesses
Randomized Control Design	Removes potential bias when randomly allocating subjects to the two groups. Tends to produce comparable groups since biases average out over the two groups.	Investigators as well as patients may be reluctant to place a patient in the (perceived) inferior treatment group. For rare diseases it may be difficult to find an adequate number of patients for a randomized control study.
Non-randomized concurrent controls	The design is easier to use since one selects the treatment group and tries to match characteristics with a control group.	The treatment and control group may not be directly comparable. The burden of proof falls on the investigator to show that appropriate matching has been done (very difficult to do).
Historical Controls	No patient is denied access to a new therapy. Roughly half as many patients are needed as in a randomized controlled study.	It is difficult, if not impossible to obtain a control group comparable to the treatment group in all ways except for the new treatment regimen. Historical records may be difficult to obtain; population characteristics (e.g. diet) may change over time.
The Cross-Over Design	Requires fewer patients to achieve the same power as a similar parallel groups design.	The effect of the treatment may carry over into the second period. Thus, results of the trial are considered more difficult to interpret than for a parallel groups design.

9.1.9 Methods of Randomization

1. Fixed allocation randomization. Fixed allocation procedures assign the study subjects (usually equally) to the treatment groups with fixed probabilities. The allocation probability does not change during the course of the trial. Some researchers advocate allocation procedures that are not 50:50. For example, Peto, et. al. (1976) recommend that sometimes a 2:1 allocation ratio, treatment-to-control be used. The rationale is that more information may be desired about a new treatment than about a control group. However, the general view is that 50:50 is more in keeping with the attitude of indifference toward which of the two groups a patient will be assigned. PROBLEM: Suppose the recruitment goal of a trial is 200 patients which are to be randomly and equally allocated to treatments A and B . Only about 30 of the patients have been identified and the remaining 170 will be recruited at a rate of about 4 per week. As a member of a design committee, you have been asked to develop a randomization schedule that ensures the following: a) a balanced assignment in case the trial terminates earlier than scheduled, and b) clinic personnel nor patients will be able to predict the next assignment (otherwise, patients with a good prognosis might be deliberately placed on a particular treatment and thus bias the results).

2. Simple Randomization. When a patient becomes eligible to be randomized, he may be assigned to a treatment by the toss of a coin or by a selection from a random number table. The main disadvantage of simple randomization is that a large imbalance may occur, especially in small trials. For example, with $N = 20$ subjects to be randomized, the chance of a 12:8 split or worse is about 0.50. With $N = 100$ subjects to be randomized, the chance of a 60:40 split or worse is about 0.05. Although imbalances are unlikely in large trials, there could be a serious imbalance if the trial ends earlier than scheduled.

3. Blocked randomization. Blocked randomization is used to avoid an imbalance in the number of subjects assigned to each group. Blocked randomization guarantees that at no time during the entry of patients into a trial will the imbalance be large. If equal allocation is used and subjects are randomly assigned to groups A and B, then for each block of even size (e.g. 4, 6, 8) exactly one-half of the subjects will be assigned to A and the other half to B. Within blocks the order in which treatments A and B are assigned to individual patients is randomized. This process is repeated for consecutive blocks of subjects until all subjects have been randomized. The smaller the

block size the more likely it becomes that an investigator may discover blocking pattern and delay enrolling the next patient until, in his own judgment the "right" patient comes along. **Notation:** Let k be the number of treatment groups, and b be the block size (i.e. number of patients included in each block). The block size is greater than 2. The block size must be at least as large as k.

Example 1: Develop a randomization schedule for assigning the first 4 patients to one of two treatments (say A and B) using equal allocation and a block size of 4. SOLUTION: Number the patients according to the time order in which they become available (1, 2, 3, 4). Note that there are 6 possible permutations of AABB, $\frac{4!}{2!2!}$. Make a list of these 6 permutations and number them 1-6. Select a number from the random number table that falls in the range 1-6. The result identifies the particular order (i.e. permutation) in which the treatments are assigned to the 4 patients.

Example 2: For the next block of 4 patients a second number is selected from the random number table to randomly allocate the next 4 patients to the 2 treatments. Assuming that the entries from the random number table are 2 and 5, give the treatment assignment of the first 8 patients. What is the maximum imbalance that could occur if the trial ends earlier than scheduled?

Example 3: Repeat Example 1 using a block of size 6.

Example 4: Repeat Example 3 using an allocation ratio 2:1, A:B. Use a block of size 6. What is the maximum imbalance that could occur?

4. Stratified Randomization. *Stratification* refers to grouping the subjects on the basis of baseline measurements. Grouping done at or before randomization is called *pre-stratification.* Grouping done after randomization (usually at the analysis stage) is called *post-stratification.* Blocking, as described earlier, is a form of stratification with patients being grouped by time of entry into a trial. Variables with error prone classifications or subjective interpretations should not be used to form strata. Pre-stratification is primarily used in small to moderate sized trials where it is more likely that an imbalance may occur. For example, a large number of good prognosis patients may, by chance, be assigned to treatment #1 and a much smaller number to treatment

#2. The main purpose of pre-stratification is to ensure comparable treatment groups. Pre-stratification is usually unnecessary in large trials because an imbalance is unlikely to occur. It is generally recommended that pre-stratification be used in multi-center trials where patients are grouped naturally by center or clinic. Stratified randomization proceeds in much the same way as described earlier. Blocking within strata is done with patients identified by time of entry within strata.

Example 5: This is an example of a stratified randomization with a block size of 4.

Age	Sex	Systolic Blood Pressure
(1) 40-49	(1) Male	(1) Greater than 130
(2) 50-59	(2) Female	(2) Less than or equal to 130
(3) 60-69		

The number of strata is $3 \times 2 \times 2 = 12$. So, number the strata 1 through 12. Subjects within stratum #1 become available in the time order 1, 2, ...

9.1.10 Blinding or Masking of Treatment Assignment

Blinding refers to the condition where knowledge of the treatment assignment is withheld from some individual or group of individuals. The purpose of blinding is to improve the objectivity of the data collection, reporting, and analysis process. It is important to use blinding when the outcome measures are subject to interpretation errors. Laboratory tests should be performed by personnel who are blinded to the treatment assignment. The only exception is those cases where the treatment assignment is needed to determine the test to be performed. ECGs, x-rays, and other photographs should be read by individuals blinded to the treatment assignment. Blinding should not be used if the study participant assumes a measurable risk in order to achieve or maintain blinding. Blinding is feasible only when it is possible to administer all treatments in an identical fashion and the clinic personnel do not need to know the treatment to care for the patient receiving it. A *single blind trial* is one in which only the participant does not know his treatment assignment. A *double blind trial* is one where neither the participant nor the investigator responsible for following the participant know the identity of the treatment assignment. A *triple blind trial*

is one where neither the participant, nor the investigator, nor the committee responsible for monitoring response variables know the identity of the treatment assignment. Blinding is often used in drug trials and is difficult to use in trials where the treatment is invasive (e.g. surgery).

9.1.11 Monitoring Compliance

The term *compliance* refers to the extent to which a participant adheres to the assigned treatment, treatment schedule, and follow-up schedule of clinic visits e.g. takes medication at the prescribed dosage and frequency. Reasons for noncompliance include side effects, unwillingness to change behavior, not understanding instructions, condition deteriorates, etc. Noncompliance dilutes the treatment effect and may have a major effect on the power of the trial. A low level of compliance may be caused by inept planning and execution of the trial. Steps that can be taken before enrollment to minimize compliance problems include:

1. Exclude people addicted to drugs or alcohol, those who live too far from the clinic, and those likely to move before termination of the trial.
2. Exclude those that have concomitant diseases and are taking other medications.
3. Ensure that each participant is clearly instructed on the purpose of the trial and what is expected of him.
4. When feasible, before randomization use a run-in period to identify poor compliers and exclude them from the trial. A run-in period is also useful for detecting drug intolerance. The number of participants eliminated by the run-in period is usually small (5-10%). In 1988, a total of 26 trials used a run-in phase.

Steps that can be taken after randomization to maintain compliance include the following list.

1. Use special pill dispensers that help the participant keep track of when he has taken his medication.
2. Use counselors, physicians, and brochures to inform and encourage the participant to comply with his assigned treatment.
3. Have clinic staff remind the participant of upcoming visits.

Even if all of the above steps have been taken, we must still monitor and assess the level of compliance. The study protocol usually specifies that the level of compliance be monitored. Assessment of compliance level is needed to publish the study results.

Problem: As a member of a monitoring committee you have been asked to indicate the frequency and type of data to be collected to monitor patient compliance with the treatment and treatment schedule. A goal of the trial is a compliance rate of at least 90%. The committee wants to track the compliance rate in the two treatment groups and also monitor any trends in the compliance rates. In drug trials, pill or capsule count is the easiest and most commonly used measure of compliance. Laboratory results (e.g. blood and urine tests) are sometimes used but usually indicate only what has happened in the preceding one or two days.

Example : Consider the tablet compliance rate in the aspirin myocardial infarction trial.

	Ave. # of tablets Prescribed per Day		Ave. # of tablets Taken per Day		Ave. Percentage Compliance	
Visit	Group A	Group B	Group A	Group B	Group A	Group B
1	1.95	1.94	1.91	1.89	97.9	97.4
2	1.93	1.90	1.82	1.78	94.3	93.7
3	1.92	1.89	1.83	1.80	95.3	95.2

Example: Consider (a different trial) of aspirin compliance as indicated by salicylate level in urine tests. The following table of data was collected.

	Percentage of Participants Showing Positive Tests	
Visit	A (aspirin)	B (placebo)
1	88.5	3.0
2	86.1	4.4
3	86.8	4.2

9.1.12 Exclusions, Withdrawals, and Losses

Even if a trial is carefully planned, some participants may be lost to follow-up, not comply with the study protocol (i.e. deviate from the4 assigned treatment or treatment schedule), or perhaps not even have met the entrance requirements. Some investigators prefer to remove from the analysis those participants who do not meet the inclusion criteria or who do not follow the study protocol perfectly. Others believe that once a subject is

randomized, he should always be followed and included in the analysis in the group to which he was randomly assigned. The latter point of view is widely accepted and is called an *analysis by intent to treat.*

Exclusions

Exclusions are people who are screened as potential trial participants, but who do not meet the entrance criteria and are therefore not randomized. Individual physicians may have, for certain of their patients, a definition preference for one or the other of the trial treatments. When this happens, the subject cannot ethically be admitted to the trial. He must be excluded from the trial and be given the treatment thought best for him, even if there is little objective basis for this preference. The trial protocol should contain specific instructions against randomizing patients who:

1. Are unlikely to tolerate one or more of the treatments.
2. Are extremely young or extremely old for the disease.
3. Seem unlikely to cooperate.
4. Live so far away that regular treatment may prove difficult.
5. Have a disease likely to take an abnormal course.

Exclusions, whether for whimsical or serious reasons, do not bias the treatment-control comparisons. They do influence the interpretation of th trial results by limiting the study population to which the results can be generalized.

Withdrawals

Withdrawals are randomized participants who are deliberately excluded from the analysis of trial results. The most common reasons for withdrawals are 1) the patient is found ineligible after he has been randomized (e.g. by miss diagnosis), 2) the patient does not comply with the assigned treatment or treatment schedule. The decision not to comply with the treatment or treatment schedule may be made by the participant, his primary care physician, or a trial investigator. Noncompliance may occur because of adverse effects of the treatment, loss of participant interest, changes in the condition of the participant, and a variety of other reasons. To minimize the number of randomized participants who are later found ineligible, Peto, et. al. (1976) suggest the following: 1) wait until the latest possible moment to

randomize a patient, and 2) do not randomize a patient unless a diagnosis is unequivocal. Since withdrawals, for any reason, can bias the treatment-control comparisons, the ideal policy is to accept withdrawals under any circumstances. This uses the intent to treat philosophy. This policy may not be satisfactory if diagnosis of the disease is difficult. In such cases, Peto, et. al. (1976) suggest randomizing some or all of the patients whose diagnosis is doubtful and then withdraw any patient who is later proved to have the wrong disease.

Losses

Losses are randomized patients who are lost to follow-up. Losses may occur because the therapy was not successful or because it has been completely successful. There is no completely satisfactory way to account for such losses so do not let it happen. Patients who move away from the tral centers where they were admitted should still be followed when survival time is the main response variable. Losses can bias the treatment-control comparisons so every effort should be made to minimize losses.

9.1.13 Treatment Efficacy and Effectiveness

Some investigators argue that non compliers should be withdrawn from the analysis because otherwise the trial is not a fair test of a new treatment. That is, the treatment-control difference may be less that what it would have been if all patients had complied with the assigned treatment. To adjust for compliance, subgrouping is sometimes used to analyze treatment efficacy. However, many investigators believe that subgrouping on the basis of compliance level, withdrawing non compliers, or including them only up to the date at which they fail to comply with the assigned treatment is seriously wrong. Compliance level is an outcome of the trial and is therefore not considered to be an appropriate basis for subgrouping.

Example: Consider the coronary artery bypass surgery trial (1979). It was a randomized trial that compared bypass surgery to medical therapy in terms of the effect on mortality. After randomization, some of the patients were switched from medical treatment to surgery or from surgery to medical treatment. Critics of the trial argued that when the trial was started, the surgical techniques were still evolving. Thus, surgical mortality in the study did not reflect what occurred in actual practice at the end of this long-term trial. In addition, there were wide differences in surgical mortality between the cooperation clinics, which may have been related to the experience of the surgeons. The following table shows the two year mortal-

ity versus compliance levels.

	Treatment Group			
	Medical		Surgery	
	Received Medicine	Switched to Surgery	Received Surgery	Switched to Medicine
Died	27	2	15	6
Survived 2 yrs	296	48	354	20
Totals	323	50	369	26

Analysis	Medical	Surgery	z_{obs}	p-value
Intent to Treat	29/373 (7.8%)	21/395 (5.3%)	1.38	0.170
Efficacy	27/323 (8.4%)	15/369 (4.1 %)	2.37	0.018

9.2 Background and Review

Use an external reference to find the pdf, mean, and variance of a random variable that has a hyper geometric distribution. The population size $n = n_1 + n_2$, has n_1 (S)success' and n_2 (F)failures. The sample size is s. Let X be the number of successes in a sample of size s taken without replacement on this population, $x_1, x_2, ..., x_s$.

FDA evidence requires:

- Safety.
- Efficacy — how effective if a treatment is taken at regular dosage.
- Effectiveness — the highest level phase of a trial.

Example:
AABB (1)
ABAB (2)
BAAB (3)
BABA (4)
BBAA (5)
ABBA (6)
Block size = 4.
Most imbalance is 2As, and 4Bs if trial stops early.

Example:

Take the same sample, but with the block size = 6. Then,

$$\frac{6!}{3!3!} = 20 \text{ cases.}$$

Example: Take AAAABB. Then, there are $\frac{6!}{4!2!} = 15$ cases.

The time at which each subject is randomized is a defining moment in the conduct of a clinical trial. The *survival time* is the studied outcome. Fifty or more subjects are needed in a clinical trial. Suppose n_1 and n_2 are fixed by the allocation ratio. The problem is to determine n. The sample size and power involve α, β, and γ_1. n is the number of subjects to be randomized (not the number of subjects to recruit before randomization). Some subjects are lost to followup. Some subjects are non-compliance patients (non-adherence or protocol deviations).

Let x_i be independent random variables and let $W = \sum a_i x_i + c$. Then, $E(W) = \sum a_i E(x_i) + c$, and $Var(W) = \sum a_i^2 Var(x_i)$. Let z have a standard normal distribution. Then z^2 has a chi-square distribution with 1 degree of freedom. Consider the Rao-Blackwell variance decomposition. Let the pair (x, y) have a bivariate distribution. Then $Var(y) = Var[E(y|x)] + E[Var(y|x)]$. Consider the univariate Cramer δ theorem (i.e. the δ-method). Suppose that $\sqrt{n}(x_n - \mu) \to N(0, \sigma_1^2)$. Let $g(x)$ be any function differentiable at $x = \mu$. Then $\sqrt{n}[g(x_n) - g(\mu)] \to N(0, \sigma_1^2)$ where $\sigma_1^2 = [g'(\mu)]^2\sigma^2$. That is, $g(x_n)$ has approximately a normal distribution with a mean $g(\mu)$ and variance $\frac{[g'(\mu)]^2\sigma^2}{n}$.

The multivariate δ Theorem. Let $\underline{x}_n' = (x_{1n}, x_{2n}, ..., x_{kn})$ be a k-dimensional vector of random variables such that $\sqrt{n}(\underline{x}_n - \underline{\mu}) \to N(0, \Sigma)$ where Σ is a $k \times k$ covariance matrix. Let $g(x)$ be a differentiable function at $\underline{x} = \underline{\mu}$. Then $\sqrt{n}[g(\underline{x}_n) - g(\underline{\mu})] \to N(0, \sigma^2)$ where $\sigma^2 = \underline{a}'\Sigma\underline{a}$,

9.3 Large Sample Tests

Consider any statistic S_n that has approximately, if n is large, a normal distribution. Let H_0 and H_1 denote the null and alternative hypotheses. Use the notation:

Notation	Interpretation
$\mu = E(S_n)$	General expected value of S_n.
μ_0	Null expected value of S_n.
$\mu_1 = E_{h_1}(S_n)$	Expected value of S_n computed under the assumption that a particular H_1 is true.
$\sigma^2_{on} = Var_{H_0}(S_n)$	Null variance of S_n.
$\sigma^2_{1H} = Var_{H_1}(S_n)$	Variance of S_n computed under the assumption that a particular H_1 is true.
$\delta = \mu_1 - \mu_0$	Difference is expected values of S_n under H_0 and H_1.

Note that $\mu_0 = 0$. There are three cases:

1. Right sided alternatives:

   $$H_0 : \mu \leq \mu_0.$$

   $$H_1 : \mu > \mu_0.$$

   $$\alpha = \text{ P(making a Type I error).}$$

   The test statistic is

   $$z = \frac{S_n - \mu_0}{\widehat{\sigma}_{0n}}$$

   Reject H_0 if $z_{\text{observed}} \geq z_\alpha$.

2. Left sided alternatives:

   $$H_0 : \mu \geq \mu_0.$$

   $$H_1 : \mu < \mu_0.$$

   The test statistic is

   $$z = \frac{S_n - \mu_0}{\widehat{\sigma}_{0n}}$$

   Reject H_0 if $z_{\text{observed}} \leq z_\alpha$.

3. Two-sided alternatives:

$H_0 : \mu = \mu_0.$

$H_1 : \mu \neq \mu_0.$

$\alpha =$ P(making a Type I error).

The test statistic is

$$z = \frac{S_n - \mu_0}{\widehat{\sigma}_{0n}}$$

Reject H_0 if $z_{\text{observed}} \leq z_{\alpha/2}$ or if $z_{\text{observed}} \geq z_{\alpha/2}$

The sample size n needed so a two-sided test with significance level of α has the power $1 - \beta$ at a particular alternative μ_1(or $\delta_1 = \mu_1 - \mu_0$) must satisfy the following equation: $|\delta_1| = z_{\alpha/2}\sigma_{0n} + z_\beta \sigma_{1n}$ is called the *sample size power equation.*

Notes:

1. For a one-sided test, $\frac{\alpha}{2}$ must be replaced by α.
2. The power is greater than 0.50 iff $z_\beta > 0$.

The proof for two-sided alternatives:

$H_0 : \mu = \mu_0.$

$H_1 : \mu \neq \mu_0.$

Reject H_0 if $z_{\text{obs}} \leq -z_{\alpha/2}$ or if $z_{\text{obs}} \geq z_{\alpha/2}$ or if

$$\frac{S_n - \mu_0}{\sigma_{0n}} \leq -z_{\alpha/2}$$

or

$$\frac{S_n - \mu_0}{\sigma_{0n}} > z_{\alpha/2}$$

Equivalently, reject H_0 if

$$S_n \leq \mu_0 - z_{\alpha/2}\sigma_{0n}$$

or

$$S_n \geq \mu_0 + z_{\alpha/2}\sigma_{0n}.$$

The power at μ_1 is

$$P(\text{reject } H_0 |\ H_1 \text{ is true or } \mu = \mu_1) =$$

$$P_{\mu_1}(S_n \leq \mu_0 - z_{\alpha/2}\sigma_{0n} \text{ or } z_n \geq \mu_o + z_{\alpha/2}\sigma_{0n}) =$$

$$P_{\mu_1}(S_n \leq \mu_0 - z_{\alpha/2}\sigma_{0n}) + P_{\mu_1}(S_n \geq \mu_0 + z_{\alpha/2}\sigma_{0n}) =$$

$$P_{\mu_1}\left(\frac{S_n - \mu_1}{\sigma_{1n}} \leq \frac{\mu_0 - \mu_1 - z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}}\right),$$

$$1 - \beta = P\left(z \leq \frac{\mu_0 - \mu_1 - z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}}\right) + P\left(z \leq \frac{\mu_0 - \mu_1 + z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}}\right) =$$

$$P(z \leq a - c) + P(z \geq a + c)$$

where

$$a = \frac{\mu_0 - \mu_1}{\sigma_{1n}},$$

and

$$c = \frac{z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}},$$

- CASE 1: $a > 0$. If $a > 0$, then $P(z > a + c) \approx 0$. Choose $z_\beta > 0$. Then,

$$a - c = z_\beta =$$

$$\frac{\mu_0 - \mu_1}{\sigma_{1n}} - \frac{z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}} = z_\beta,$$

$$\frac{\mu_0 - \mu_1}{\sigma_{1n}} = z_\beta + \frac{z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}},$$

$$\mu_0 - \mu_1 = z_\beta\sigma_{1n} + z_{\alpha/2}\sigma_{0n},$$

$$|\mu_0 - \mu_1| = z_\beta\sigma_{1n} + z_{\alpha/2}\sigma_{0n},$$

which is $|\delta_1|$.

- CASE 2: $a < 0$.

  $$a < 0 \Rightarrow P(z < a - c) \approx 0.$$

  So, the power

  $$P(\mu_1) = P(z > a + c) = 1 - \beta,$$

  $$\Rightarrow a + c = -z_\beta,$$

  $$\frac{\mu_0 - \mu_1}{\sigma_{1n}} + \frac{z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}} = -z_\beta,$$

  $$\frac{\mu_0 - \mu_1}{\sigma_{1n}} = -z_\beta - \frac{z_{\alpha/2}\sigma_{0n}}{\sigma_{1n}} =$$

  $$\mu_0 - \mu_1 = -z_\beta\sigma_{1n} - z_{\alpha/2}\sigma_{0n} - (\mu_0 - \mu_1) =$$

  $$z_\beta\sigma_{1n} + z_{\alpha/2}\sigma_{0n} =$$

  $$|\mu_0 - \mu_1| = z_\beta\sigma_{1n} + z_{\alpha/2}\sigma_{0n}.$$

9.4 General Approach to Sample Size Determination

A parameter δ is chosen to represent the contrast between responses on the two treatments. Testing $\delta = 0$ under the null hypothesis H_0 tests for no difference. A significance level (usually 2-sided) of H_0 will be based on some statistic S_n, and a significance level of α. The sample size is $n = n_1 + n_2$ is chosen to ensure that the test has a specified power of $1 - \beta$ for a specified alternative δ_1. Allowance must be made for a hypothetical rate at which patients are lost to follow-up. Advantages of the sample size calculations include:

- They make investigators aware of the consequences of their choice of trial size.
- They may prevent implementing trials that are too small to be of any real value.

Difficulties with the sample size calculations include:

- The choices of α and β are somewhat arbitrary (though usually $\alpha = 0.05$ and $\beta \leq 0.10$) and different choices of α and β affect the value of n.
- The alternative δ_1 statistic is difficult to interpret. It may represent any of the following: a) the smallest clinically worth while difference, b) a difference worth detecting, and c) a difference thought likely to occur. Thus the value δ_1 must be chosen subjectively and agreement between investigators may not be easy to achieve.
- The sample size usually depends on other parameters (e.g. response rate in the control group and variance) about which there is some uncertainty.
- Information concerning the last difficulty listed can be obtained from: a) a pilot trial, b) an earlier Phase I or II study, and c) the literature concerning a similar study.

Conclusions: The sample size calculations should be regarded in a flexible way as providing guidance concerning the size of a study rather than a rigid prescription. They are a required part of the study protocol.

9.4.1 Sample Size and Power Calculations

Example 1: Suppose a trial is to be conducted to study the effectiveness of a new cholesterol lowering drug. The difference in serum cholesterol at baseline and six months after randomization is to be observed for each subject.

Group	Sample Size	Mean Change in Cholesterol Level	Sample Mean Difference	Variance
1 (drug)	n_1	v_1	$\bar{x}_1$	σ^2
2 (placebo)	n_2	v_2	$\bar{x}_2$	σ^2

Assume that subjects are allocated equally to the two groups. The hypotheses are $H_0 : v_1 = v_2$, versus $H_1 : v_1 > v_2$.

1. Determine $n = n_1 + n_2$ so that a 5% confidence level test has a power of 0.90 for detecting the difference $\delta = v_1 - v_2 = 10$ when $\bar{v} = 20$.
2. Determine the power of the 5% test at $\delta = 10$ if only $n = 100$ patients are recruited.

3. What power is attained if only $n = 36$ subjects are recruited?
4. What difference δ can be detected with a power of 0.90 and $n = 100$?
5. What sample size n is needed to detect the difference $\delta = 10$ with a power of 0.90 when using a two-tailed test with $\alpha = 0.05$?

Example 2: A trial is to be conducted to study two different ways of administering chemotherapy as a treatment for small cell lung cancer.

Treatment	Sample Size	Success
1 (new)	n_1	p_1
2 (control)	n_2	p_2

Previous studies indicate that the success rate for treatment 2 is $p_2 = 0.15$. Determine the total number n of participants that must be randomized so that a 5% level test of $H_0 : p_1 = p_2$ versus $H_1 : p_1 > p_2$ has a power of 0.90 for detecting an increase $\delta = p_1 - p_2$ of $\delta = 0.20$.

Example 3: Consider a repeated measures trial for comparing slopes (text book, page 113). The assumptions are

1. In the control group, the response variable decreases at a rate of 80 units per year ($\theta_1 = 80$).
2. A 25% reduction in this rate is expected for a new treatment (i.e. $\theta_2 = 60$).
3. Other studies indicate $\sigma_\epsilon = 150$ and $\sigma_\theta = 63$.
4. Subjects are to be allocated equally to the two treatment groups.

Determine $n = n_1 + n_2$ so that a 5% level test of $\delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \theta_1 - \theta_2$ has a power of 0.90 for the following cases:

1. A 3 year study with 4 visits per year. $D = 3, k = 13$, one visit at baseline.
2. A 4 year study with 4 visits per year. $D = 4, k = 17$.

Example 4: This example can be found in the text book on page 108. Consider an eye study in which one eye is treated by laser surgery and the other eye by a standard therapy. The left and right eyes are randomly allocated to the two treatments. For each subject the data consists of two responses.

1. Vision improves (success) or does not improve (failure) in the eye receiving treatment 1.

2. Vision improves or does not improve in the eye receiving treatment 2.

π_i is the success rate for treatment $i = 1, 2$. $\pi_1 = 0.80, \pi_2 = 0.60, \delta = \pi_1 - \pi_2$. The discordant rate f is $f = 0.50$. Determine n so that a 5% level, two-sided test of $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ has a power of 0.90 against the alternative $\delta = 0.20$.

9.5 General Sample Size and the Power Equation

$|\delta| = z_{\alpha/2}\sigma_{0n} + z_\beta \sigma_{1n}$.

Comparing proportions with independent samples gives the following table:

Group	Sample Size	Population Proportion	Sample Proportion
1	n_1	p_1	$\widehat{p}_1$
2	n_2	p_2	$\widehat{p}_2$

$S_n = \widehat{p}_1 - \widehat{p}_2$,

$\mu = E(S_n) = p_1 - p_2$,

$\sigma_n^2 = Var(S_n) = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$,

$H_0 : p_1 - p_2 = 0$.

$H_1 : p_1 - p_2 > 0$.

$\mu_0 = E_{H_0}(S_n) = 0$,

$\mu_1 = E_{H_1}(S_n) = p_1 - p_2$,

$\delta = \mu_1 - \mu_0 = p_1 - p_2$,

$\sigma_{0n}^2 = Var_{H_0}(S_n) = \bar{p}(1-\bar{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$

$$\sigma_{1n}^2 = Var_{H_1}(S_n) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

where

$$\bar{p} = \frac{p_1 + p_2}{2}$$

The sample size of the power equation is

$$|\delta| = z_\alpha \sigma_{0n} + z_\beta \sigma_{1n},$$

$$|\delta| = z_\alpha \sqrt{\bar{p}(1-\bar{p})}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} + z_\beta \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Substitute

$$Q_i = \frac{n_i}{n}$$

or

$$n_i = Q_i n, \text{ for } i = 1, 2.$$

Then,

$$|\delta| = \frac{z_\alpha \sqrt{\bar{p}(1-\bar{p})}\sqrt{\frac{1}{Q_1} + \frac{1}{Q_2}} + z_\beta \sqrt{\frac{p_1(1-p_1)}{Q_1} + \frac{p_2(1-p_2)}{Q_2}}}{\sqrt{n}}.$$

$$z = \frac{\text{statistic} - E(\text{statistic})}{\text{std deviation}},$$

$$E(z) = \frac{1}{\sigma}x + \left(-\frac{\mu}{\sigma}\right) =$$

$$\frac{1}{\sigma}E(x) + \left(-\frac{\mu}{\sigma}\right) =$$

$$\frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0.$$

$$Var(z) = \left(\frac{1}{\sigma}\right)^2 Var(x) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

$$\widehat{p} = \frac{y}{n} = \frac{1}{n}y,$$

$$E(\widehat{p}) = \frac{1}{n}E(y) = \frac{1}{n}np = p,$$

$$Var(\widehat{p}) = \left(\frac{1}{n}\right)^2 Var(y) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n},$$

$$z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$z' = \frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}$$

Note that $\widehat{p} =$ sample mean

$$\mu = E(x) = 0(1-p) + 1p = p,$$

$$\sigma^2 = E(x^2) - [E(x)]^2,$$

$$E(x^2) = 0^2(1-p) + 1^2 p = p,$$

$$\Rightarrow \sigma^2 = p - p^2 = p(1-p),$$

$$Var(\widehat{p}) = \frac{p(1-p)}{n}$$

Suppose that $\bar{x}_1$, and $\bar{x}_2$ are independent.

$$S_n = \bar{x}_1 - \bar{x}_2 = (1)\bar{x}_1 + (-1)\bar{x}_2,$$

$$E(S_n) = 1E(\bar{x}_1) + (-1)E(\bar{x}_x) = 1\mu_1 + (-1)\mu_2 = \mu_1 - \mu_2,$$

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z' = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2} = \frac{y_1 + y_2}{n_1 + n_2}.$$

The Q_i's are the allocation amounts. $Q_i = \frac{n_i}{n}$.

Example:

$$z = \frac{\widehat{p}_1 - \widehat{p}_2 - 0}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_1} + \frac{\widehat{p}(1-\widehat{p})}{n_2}}}$$

Reject H_0 if $z_{\text{obs}} \geq z_\alpha = 1.645$. $p_2 = 0.15$.

$$\delta = p_1 - p_2 = p_1 - 0.15 \Rightarrow p_1 = 0.35,$$

because $\delta = 0.20$. The power is

$$1 - \beta = 0.90 \Rightarrow \beta = 0.10,$$

$$z_{0.10} = 1.28; Q_1 = 0.60; Q_2 = 0.40;$$

$$\bar{p} = \frac{p_1 + p_2}{2} = \frac{0.15 + 0.35}{2} = 0.25.$$

Then,

$$0.20 =$$

$$\frac{1.645\sqrt{0.25(0.75)}\sqrt{\frac{1}{0.60} + \frac{1}{0.40}} + 1.28\sqrt{\frac{0.35(0.65)}{0.60} + \frac{0.15(0.85)}{0.40}}}{\sqrt{n}}$$

Then,

$$\sqrt{n} =$$

$$\frac{1.645\sqrt{0.25(0.75)}\sqrt{\frac{1}{0.60} + \frac{1}{0.40}} + 1.28\sqrt{\frac{0.35(0.65)}{0.60} + \frac{0.15(0.85)}{0.40}}}{0.20} = 12.62,$$

$n = 159.2 \Rightarrow n = 160, n_1 = Q_1 n = 0.60(160) = 96, n_2 = Q_2 n = 0.40(160) = 64$. What about the non-compliance people? Reference page 107 of the text book.

Definition: A *drop-out* is any participant randomized to the treatment group who does not adhere to the assigned treatment or treatment schedule. A *drop-in* is any participant randomized to the control group who begins to use the study treatment. The text book, page 108, uses the following notation: R_0 = drop-out rate, and R_I = drop-in rate. n = unadjusted sample size, n^* = adjusted sample size = $\frac{n}{(1-R_0-R_I)^2}$.

Example: Suppose $R_0 = 0.20, R_I = 0.05$, and $n = 200$. Then, $n^* = \frac{200}{(1-0.20-0.05)^2} = \frac{200}{(0.75)^2} = 356$. Another example appears on page 108 of the text book using p_I^*, p_C^* from p_I, and p_C.

9.5.1 Further Background

In all that follows we will primarily use the following:

1. Let X be any random variable and let a, b be any constants. Then, $E(ax+b) = aE(x) + b$ and $Var(ax+b) = a^2 Var(x)$.

2. The standardized form of any random variable X is $z = \frac{x-\mu}{\sigma}$. As a consequence of (1), any standardized random variable z has mean, $E(z) = 0$, and $Var(z) = 1$.

3. Let $x_1, x_2, ..., x_n$ be any independent random variables and let $a_1, a_2, ..., a_n$ be any constants. Then,

$$E\left(\sum_{i=1}^{n} a_i x_i\right) \sum_{i=1}^{n} a_i E(x_i),$$

$$Var\left(\sum_{i=1}^{n} a_i x_i\right) \sum_{i=1}^{n} a_i^2 Var(x_i),$$

4. An important special case of (3) is

$$E(a_1 x_1 + a_2 x_2) = a_1 E(x_1) + a_2 E(x_2),$$

and

$$Var(a_1 x_1 + a_2 x_2) = a_1^2 Var(x_1) + a_2^2 Var(x_2).$$

Random Sample from a Single Population

Let $x_1, x_2, ..., x_n$ be iid random variables with mean μ and variance σ^2. That is, $E(x_i) = \mu$, and $Var(x_i) = \sigma^2$. The sample mean $\bar{x}$ is a linear function of independent random variables.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Using (3), we have

$$E(\bar{x}) = \frac{1}{n}E(x_1) + \frac{1}{n}E(x_2) + \cdots + \frac{1}{n}E(x_n) =$$

$$\frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu =$$

$$n\left(\frac{1}{n}\mu\right) = \mu.$$

$$Var(\bar{x}) = \left(\frac{1}{n}\right)^2 Var(x_1) + \left(\frac{1}{n}\right)^2 Var(x_2) + \cdots + \left(\frac{1}{n}\right)^2 Var(x_n) =$$

$$\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2 =$$

$$n\left(\frac{\sigma^2}{n^2}\right) = \frac{\sigma^2}{n}.$$

Standardized Form of $\bar{X}$

Using (2) we write,

$$z = \frac{\bar{x} - E(\bar{x})}{\sqrt{Var(\bar{x})}}$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}.$$

We know that z has a distribution with mean zero and variance 1.

Modified Standardized Form o $\bar{X}$

The standardized form of $\bar{x}$ can not be used for inference about a population mean μ unless the value of σ is known. Since usually the value of σ is not known, we replace σ by its estimator, namely, the sample standard deviation, $s = \sqrt{s^2}$. The result is the modified standardized form

$$z' = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

The Central Limit Theorem

The Central Limit Theorem implies that if n is large ($n > 30$) then both z and z' have approximately standard normal distributions. If $x_1, x_2, ..., x_n$ is a random sample from a normal distribution, then

a. z has exactly a standard normal distribution for every n.

b. z' has exactly a $t-$distribution for every n.

In case (b), the modified standardized form z' is usually labeled T to denote that it has a $t-$distribution. That is, we write

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

since if the sample arises from a normal distribution, then T has a $t-$distribution with degrees of freedom $n - 1$.

In clinical trials the sample size is usually much larger than 30 so we use the normal approximation to the distribution of z'.

Special Case: Binary Response Variables

As before assume that we have iid random variables $x_1, x_2, ..., x_n$ where the response of each subject is binary:

$$x_i = \begin{cases} 1, & \text{if subject \# i responds favorably to treatment.} \\ 0, & \text{otherwise.} \end{cases}$$

It is clear that each x_i then has the following probability distribution:

x	0	1
$f(x)$	$1-p$	p

where $p = P(x_i = 1)$ is the probability that subject i responds favorable to treatment. Note that $x_1 + x_2 + \cdots + x_n$ is the total number of successes occurring in n iid trials, call it y. We know y has a binomial distribution with pdf

$$f(y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}, y = 0, 1, 2, ..., n.$$

The mean and variance of a binomial random variable y are $E(y) = np$, and $Var(y) = np(1-p)$. We usually estimate the parameter p by the sample proportion $\widehat{p} = \frac{y}{n}$. At this point we should use (1) to get the mean, variance, and standardized form of $\widehat{p}$

Since iid binary random variables are a special case of general iid random variables, we should be able to apply earlier results obtained for the general case. In particular note that

$$\widehat{p} = \frac{y}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Thus $\widehat{p}$ is actually a sample mean and we know that

$$\left.\begin{array}{l} E(\widehat{p}) = \mu \\ Var(\widehat{p}) = \frac{\sigma^2}{n} \end{array}\right\}$$

where μ is the population mean and σ^2 is the population variance. μ and σ^2 can be calculated from the pdf

x	0	1
$f(x)$	$1-p$	p

Independent Samples from 2 Populations

Population	Population Mean	Population Variance	Sample Size	Sample Mean	Sample Variance
#1	μ_1	σ_1^2	n_1	$\bar{x}_1$	s_1^2
#2	μ_2	σ_2^2	n_2	$\bar{x}_2$	s_2^2

Since the samples are take independently, $\bar{x}_1$ and $\bar{x}_2$ are independent random variables. Usually we want to estimate $\mu_1 - \mu_2$ and for this reason we are interested in the statistic $\bar{x}_1 - \bar{x}_2$. Note that $\bar{x}_1 - \bar{x}_2$ is a linear function of $\bar{x}_1$ and $\bar{x}_2$. That is

$$\bar{x}_1 - \bar{x}_2 = (1)\bar{x}_1 + (-1)\bar{x}_2.$$

Using our theorems on expected value and variance of linear functions of independent random variables, we have

$$E(\bar{x}_1 - \bar{x}_2) = (1)E(\bar{x}_1) + (-1)E(\bar{x}_2) =$$

$$(1)\mu_1 + (-1)\mu_2 = \mu_1 - \mu_2.$$

$$Var(\bar{x}_1 - \bar{x}_2) = Var[(1)\bar{x}_1 + (-1)\bar{x}_2] =$$

$$(1)^2 Var(\bar{x}_1) + (-1)^2 Var(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Let z and z' be the standardized form and the modified standardized forms of $\bar{x}_1 - \bar{x}_2$. The Central Limit Theorem implies that if n_1 and n_2 are large, both greater than 30, then z and z' have approximately standard normal distributions. Often we are willing to assume that the two populations have a common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$. Then the standardized form of $\bar{x}_1 - \bar{x}_2$ is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Since the value of σ is usually unknown we estimate σ^2 by forming a weighted average of s_1^2 and s_2^2.

$$s_p = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

When the two populations can be assumed to have a common variance, the modified standardized form of $\bar{x}_1 - \bar{x}_2$ is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

As before, the Central Limit Theorem implies that z' has approximately a standard normal distribution if $n_i \geq 30, i = 1, 2$.

Alternatively, if the independent samples arise from normal distributions, then z' has exactly a $t-$distribution with $n_1 + n_2 - 2$ degrees of freedom and is usually labeled T rather than z'.

Independent Samples and a Binary Response

Population	Population Proportion	Sample Size	Number Giving a Favorable Response	Sample Proportion
#1	p_1	n_1	y_1	$\widehat{p}_1 = y_1/n_1$
#2	p_2	n_2	y_2	$\widehat{p}_2 = y_2/n_2$

Clearly $\widehat{p}_1 - \widehat{p}_2$ is the difference of two sample means. So this case is very similar to the one just described (except that there is no possibility that binary random variables can have normal distributions. So there is no $T-$statistic). Now determine the mean, variance and standardized form of $\widehat{p}_1 - \widehat{p}_2$.

$$E(\widehat{p}_1 - \widehat{p}_2) = E[(1)\widehat{p}_1 + (-1)\widehat{p}_2] =$$

$$(1)E(\widehat{p}_1) + (-1)E(\widehat{p}_2) = p_1 - p_2.$$

$$Var(\widehat{p}_1 - \widehat{p}_2) = Var[(1)\widehat{p}_1 + (-1)\widehat{p}_2] =$$

$$(1)^2 Var(\widehat{p}_1) + (-1)^2 Var(\widehat{p}_2) =$$

$$Var(\widehat{p}_1) + Var(\widehat{p}_2) = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}.$$

The standardized form is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

The modified standardized form is

$$z' = \frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}$$

According to the Central Limit Theorem, z and z' have approximate standard normal distributions.

Confidence Limits for $p_1 - p_2$

To obtain $100(1-\alpha)\%$ confidence limits, choose the limit $z_{\alpha/2}$ from the tabled normal distribution so that

$$P\left(-z_{\alpha/2} < Z' < z_{\alpha/2}\right) = 1 - \alpha.$$

Then,

$$P\left(-z_{\alpha/2} < \frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Algebraic manipulation gives $1 - \alpha =$

$$P\left(\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}\right.$$

$$< p_1 - p_2 <$$

$$\left.\widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}\right)$$

Thus, the upper and lower confidence limits are

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

Testing the Hypothesis

The standardized form under H_0 is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2 - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

and the modified standardized form under H_0 is

$$z' = \frac{\widehat{p}_1 - \widehat{p}_2 - 0}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}}$$

which is the proper test statistic.

9.6 Comparing Slopes with Repeated Measures

A trial is sometimes designed to investigate changes over time in some characteristic that is measured repeatedly for each subject. The characteristic may be blood pressure, bone mass, cholesterol level, lung volume, etc. A model sometimes used in this setting is the simple linear random effects model with follow-up time being an explanatory variable. The model is similar to the ordinary straight line model except that the slope and intercept are permitted to vary between different subjects. Thus, the slope and intercept are modeled as random quantities and, for this reason, the model is called a *linear random effects model.*

Let k be the number of planned follow-up visits and $x_1, x_2, ..., x_k$ be the follow-up times measured from the date of randomization. In the Linear Random Effects model, y_{ij} is response of subject i at follow-up time x_j :

$$y_{ij} = \gamma_i + \theta_i x_j + \epsilon_{ij}, i = 1, 2, ..., n; j = 1, 2, ..., k,$$

where γ_i is the subject specific intercept, θ_i is the subject specific slope, and ϵ_{ij} is the random error term. The model assumptions are:

1. (γ_i, θ_i) are iid random vectors that have a bivariate normal distribution with mean vector (γ, θ) and covariance matrix, $Var(\gamma_i) = \sigma^2_\gamma, Var(\theta_i) = \sigma^2_\theta$, and $Cov(\gamma, \theta_i) = \sigma_{\gamma\theta}$.

2. ϵ_{ij} are iid and have a normal distribution with mean zero and variance σ^2_ϵ.

3. (γ_i, θ_i) and ϵ_{ij} are independent collections of random variables.

Note: The model implies that

1. Repeated observations on the same subject are dependent random variables.

2. Observations on different subjects are independent.

9.6.1 Estimators of Subject Specific Slopes

Let L_i be the number of visits completed by subject i. We assume there are no missed interim visits but that subject i may not complete all k visits. Thus, $L_i = k$ only if subject i completes all k visits. The data for subject i are $(x_j, y_{ij}), j = 1, 2, 3, ..., L_i; i = 1, 2, ..., n$. The least squares estimator of the subject specific slope θ_i is

$$\widehat{\theta}_i = \sum_{j=1}^{L_i} \frac{(x_j - \bar{x}_i)y_{ij}}{S_i},$$

where

$$S_i = \sum_{j=1}^{L_i} (x_j - \bar{x}_i)^2,$$

and

$$\bar{x}_i = \frac{x_1 + x_2 + \cdots + \bar{x}_{L_i}}{L_i}.$$

Expected Value and Variance of Slope Estimator for a Single Subject

Standard results for the straight line regression model with fixed slope and intercept give the following:

$$E(\widehat{\theta}_i|\theta_i) = \theta_i$$

$$Var(\widehat{\theta}_i|\theta_i) = \frac{\sigma_\epsilon^2}{S_i}$$

To derive the unconditional mean and variance, recall that

$$E[E(Y|X)] = E(Y),$$

and

$$Var(Y) = Var[E(Y|X)] + E(Var(Y|X)].$$

Applying this to the two equations above gives:

$$E(\widehat{\theta}_i) = E(\theta_i) = \theta,$$

and

$$v_i = Var(\widehat{\theta}_i) = Var[E(\widehat{\theta}_i|\theta_i)] + E[Var(\widehat{\theta}_i|\theta_i)] =$$

$$Var(\theta_i) + E\left[\frac{\sigma_\epsilon^2}{S_i}\right] =$$

$$\sigma_\theta^2 + \frac{\sigma_\epsilon^2}{S_i} =$$

$$\sigma_\theta^2[1 + R/S_i]$$

where

$$R = \frac{\sigma_\epsilon^2}{\sigma_\theta^2}.$$

Estimate of the Population Mean Slope

Since observations on different subjects are assumed independent, we have, for n subjects, n different estimators of θ. Namely, $\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_n$ are all estimators of θ. How should we combine the subject specific slope estimators to get a single estimate of θ? The maximum likelihood estimator of θ is the following weighted linear combination of $\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_n$.

$$\widehat{\theta} = \frac{\sum_{i=1}^n v_i^{-1}\widehat{\theta}_i}{\sum_{i=1}^n v_i^{-1}}$$

where $v_i = \sigma_\theta^2[1 + R/S_i], i = 1, 2, ..., n$.

Expected Value and Variance of Slope Estimation

Note that $\widehat{\theta}$ is a linear function of $\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_n$. Thus,

$$E(\widehat{\theta}) = E\left[\frac{\sum_{i=1}^n v_i^{-1}\widehat{\theta}_i}{\sum_{i=1}^n v_i^{-1}}\right] =$$

$$\frac{\sum_{i=1}^n v_i^{-1}E(\widehat{\theta}_i)}{\sum_{i=1}^n v_i^{-1}} =$$

$$\theta\frac{\sum_{i=1}^n v_i^{-1}}{\sum_{i=1}^n v_i^{-1}} = \theta.$$

Since $\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_n$ are independent, we have

$$Var(\widehat{\theta}) = \frac{\sum_{i=1}^n (v_i^{-1})^2 Var(\widehat{\theta}_i)}{\left(\sum_{=1}^n v_i^{-1}\right)^2},$$

$$(v_i^{-1})^2 Var(\widehat{\theta}_i) = \frac{1}{v_i^2} Var(\widehat{\theta}_i) = \frac{1}{v_i},$$

since by definition $v_i = Var(\widehat{\theta}_i)$. Thus,

$$Var(\widehat{\theta}) = \frac{\sum_{i=1}^{n} v_i^{-1}}{\left(\sum_{i=1}^{n} v_i^{-1}\right)^2} = \frac{1}{\sum_{i=1}^{n} v_i^{-1}}$$

Comparing Slopes for Two Treatment Groups

Group	Population Mean Slope	Sample Size
1	θ_1	n_1
2	θ_2	n_2

Some other notation: L_{1i} is the number of visits completed by subject i in group 1. L_{2i} is the number of visits completed by subject i in group 2.

$$S_{1i} = \sum_{j=1}^{L_{1i}} (x_j - \bar{x}_{1i})^2$$

$$S_{2i} = \sum_{j=1}^{L_{2i}} (x_j - \bar{x}_{2i})^2$$

The parameters $\sigma_\theta^2, \sigma_\gamma^2, \sigma_{\theta\gamma}, \sigma_\epsilon^2$ are assumed to be identical for the two groups. $\widehat{\theta}_1$ is the slope estimator for group 1. $\widehat{\theta}_2$ is the slope estimator for group 2. Observations in the two groups are assumed to be independent. Thus, $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent random variables.

Mean and Variance of the Standardized form of $\widehat{\theta}_1 - \widehat{\theta}_2$

$$E[\widehat{\theta}_1 - \widehat{\theta}_2] = E(\widehat{\theta}_1) - E(\widehat{\theta}_2) = \theta_1 - \theta_2.$$

$$Var(\widehat{\theta}_1 - \widehat{\theta}_2) = (1)^2 Var(\widehat{\theta}_1) + (-1)^2 Var(\widehat{\theta}_2) =$$

$$Var(\widehat{\theta}_1) + Var(\widehat{\theta}_2) =$$

$$\left(\sum_{i=1}^{n_1} v_{1i}^{-1}\right)^{-1} + \left(\sum_{i=1}^{n_2} v_{2i}^{-1}\right)^{-1} =$$

where $v_{1i} = \sigma_\theta^2[1 + R/S_{1i}]$, and $v_{1i} = \sigma_\theta^2[1 + R/S_{2i}]$

Standardized Form of $\widehat{\theta}_1 - \widehat{\theta}_2$

$$Z = \frac{\widehat{\theta}_1 - \widehat{\theta}_2 - (\theta_1 - \theta_2)}{\sqrt{\left(\sum_{i=1}^{n_1} v_{1i}^{-1}\right)^{-1} + \left(\sum_{i=1}^{n_2} v_{2i}^{-1}\right)^{-1}}}$$

Sample Size and Power (Ideal Conditions)

We assume that each subject will complete all follow-up visits. Then, $L_{1i} = L_{2i} = k$, and

$$S_{1i} = S_{2i} = \sum_{j=1}^{k} (x_j - \bar{x})^2.$$

Let

$$S = \sum_{j=1}^{k} (x_j - \bar{x})^2.$$

Then,

$$v_{1i} = v_{2i} = \sigma_\theta^2 [1 + R/S].$$

We have $\delta = \theta_1 - \theta_2$.

$$Var(\widehat{\theta}_1 - \widehat{\theta}_2) = \left(\sum_{i=1}^{n_1} v_{1i}^{-1}\right)^{-1} + \left(\sum_{i=1}^{n_2} v_{2i}^{-1}\right)^{-1} =$$

$$\sigma_\theta^2 [1 + R/S] \left[\frac{1}{n_1} + \frac{1}{n_2}\right].$$

Sample Size — Power Equation

The hypotheses are

$$H_0 : \delta = 0,$$

$$H_1 : \delta > 0,$$

where $\delta = \theta_1 - \theta_2$, and $\widehat{\delta} = \widehat{\theta}_1 - \widehat{\theta}_2$.

$$\mu_1 = E_{H_1}(\widehat{\delta}) = \delta_1,$$

$$\mu_0 = E_{H_0}(\widehat{\delta}) = 0.$$

$$\sigma_{0n}^2 = \sigma_{1n}^2 = Var(\widehat{\theta}_1 - \widehat{\theta}_2) = \sigma_\theta^2[1 + R/S]\left[\frac{1}{n_1} + \frac{1}{n_2}\right] =$$

$$\frac{\sigma_\theta^2[1 + R/S]}{n}\left[\frac{1}{Q_1} + \frac{1}{Q_2}\right].$$

$$|\mu_1 - \mu_0| = z_\alpha \sigma_{0n} + z_\beta \sigma_{1n}$$

or

$$|\delta_1| = \frac{(z_\alpha + z_\beta)\sigma_\theta\sqrt{1 + R/S}\sqrt{\frac{1}{Q_1} + \frac{1}{Q_2}}}{\sqrt{n}}$$

$$n = \frac{(z_\alpha + z_\beta)^2\sigma_\theta^2(1 + R/S)\left[\frac{1}{Q_1} + \frac{1}{Q_2}\right]}{(\delta_1)^2}$$

A Simple Formula for S

It is common to make simplifying assumptions when determining sample size. We assume the follow-up times are equally spaced but may begin at an arbitrary point in time. That is,

$$x_1 = b, x_2 = b + a, x_3 = b + 2a, ..., x_k = b + (k - 1)a.$$

Since

$$S = \sum_{i=1}^{k}(x_i - \bar{x})^2$$

does not depend on b, we can set b equal to any convenient value to get a simple form of S. Let $b = a$. Then, $x_i = ai, i = 1, 2, ..., k$ and

$$\sum_{i=1}^{k} x_i = a\sum_{i=1}^{k} i = \frac{ak(k+1)}{2},$$

$$\sum_{i=1}^{k} x_i^2 = a^2\sum_{i=1}^{k} i^2 = \frac{a^2k(k+1)(2k+1)}{6}.$$

Thus,

$$S = \sum_{i=1}^{2} x_i^2 - \frac{\left(\sum_{i=1}^{k} x_i\right)^2}{k} =$$

$$\frac{a^2k(k+1)(2k+1)}{6} - \frac{a^2k(k+1)}{4} = \frac{a^2k(k^2-1)}{12}.$$

Let D equal the duration of the follow-up. $D = x_k - x_1 = ka - a = a(k-1)$. Then,

$$S = \frac{a^2k(k-1)^2(k+1)}{12(k-1)} = \frac{D^2k(k+1)}{12(k-1)}.$$

Our sample size formula becomes

$$n = \frac{(z_\alpha + z_\beta)^2\sigma_\theta^2(1+R/S)\left[\frac{1}{Q_1}+\frac{1}{Q_2}\right]}{(\delta_1)^2} =$$

$$\frac{(z_\alpha + z_\beta)^2\sigma_\theta^2\left[1+\frac{12R(k-1)}{D^2k(k+1)}\right]}{(\delta_1)^2}.$$

The choice of k and D depend on:

(a) How long we can afford to follow participants.

(b) How many times a participant may be willing to visit a clinic.

Example: Text book, page 113. Assumptions:

1. In the control group, the response variable decreases at a rate of 80 units per year (i.e. $\theta_1 = 80$).
2. A 25% reduction in this rate is expected in the treatment group ($\theta_2 = 60$).
3. Other studies indicate that $\sigma_\epsilon = 150$ and $\sigma_\theta = 63$.
4. Equal allocation: $Q_1 = Q_2 = \frac{1}{2}$.

Determine the sample size n needed so a 5% level test has a power of 0.90 for

a. a 3 year study with 4 visits per year (i.e. $D = 3, k = 13$, one visit is baseline).

b. a 4 year study with 4 visits per year (i.e. $D = 4, k = 17$).

Note that $k \geq 2$ due to slope calculations. The solution is as follow:

$H_0 : \delta = 0,$

$H_A \delta \neq 0.$

$\sigma_\epsilon = 150,$

$\sigma_\theta = 63,$

$R = \frac{\sigma_\epsilon}{\sigma_\theta}^2 = 5.67,$

Let $\alpha = 0.05$. Then $z_{\alpha/2} = 1.96$. If power is set at 0.90, Then $Z_{2,3} = 1.28$.

$\delta_1 = \theta_1 - \theta_2 = 80 - 60 = 20.$

1. $D = 3, k = 13.$

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \left[1 + \frac{12R(k-1)}{D^2 k(k+1)}\right] \left[\frac{1}{Q_1} + \frac{1}{Q_2}\right] =$$

$$4\frac{(1.96 + 1.28)^2 (63)^2}{(20)^2} \left[1 + \frac{12(5.67)(12)}{(3)^2(13)(14))}\right] = 625.$$

2. $D = 4$, $k = 17.$

$$4\frac{(1.96 + 1.28)^2 (63)^2}{(20)^2} \left[1 + \frac{12(5.67)(16)}{(4)^2(17)(18))}\right] = 510.$$

9.6.2 Paired Binary Response

Occasionally, trials are conducted by using matched pairs of subjects or some form of natural pairing.

Example: Consider an eye study in which one eye is treated for loss of visual acuity by laser surgery and the other eye is treated by standard therapy. The left and right eyes are randomly allocated in the two treatments. For each subject, the data consists of two responses:

a. Vision improves (success) or does not improve (failure) in the eye receiving treatment 1.

b. Vision improves (success) or does not improve (failure) in the eye receiving treatment 2.

Main Parameters of Interest

Let π_i, $i = 1, 2$ denote the success rate for treatment $i = 1, 2$. Then the main parameter is

$$\delta = \pi_1 - \pi_2$$

is the difference in success rates of the two treatments. In the following, we discuss the basis for McNemar's test for comparing π_1 and π_2. We then derive the sample size, the power equation. We will see that McNemar's test is in some ways similar to the paired $t-$test and in other ways similar to a test for comparing proportions with independent samples.

Notation

Let u_i be the response of subject i to treatment 1.

$$u_i = \begin{cases} 1, \text{ if success occurs on treatment 1} \\ 0, \text{ otherwise} \end{cases}$$

Let v_i be the response of subject i to treatment 2.

$$v_i = \begin{cases} 1, \text{ if success occurs on treatment 2} \\ 0, \text{ otherwise} \end{cases}$$

Assumptions

(u_i, v_i) are independent and identically distributed pairs with the following probability distribution:

(u,v)	(0,0)	(0,1)	(1,0)	(1,1)
$f(u,v)$	$\pi_{0,0}$	$\pi_{0,1}$	$\pi_{1,0}$	$\pi_{1,1}$

Recall that π_i is the success rate of treatment i.

$$\pi_1 = P(u = 1) = P(u = 1, v = 0) + P(u = 1, v = 1) = \pi_{1,0} + \pi_{1,1}.$$

Similarly,

$$\pi_2 = P(v = 1) = P(u = 0, v = 1) + P(u = 1, v = 1) = \pi_{0,1} + \pi_{1,1}.$$

Note that

$$\delta = \pi_1 - \pi_2 = (\pi_{1,0} - \pi_{1,1}) - (\pi_{0,1} - \pi_{1,1}) = \pi_{1,0} - \pi_{0,1}.$$

The outcomes (0,1) and (1,0) are called *discordant responses.* Thus, δ is the difference in success rates which is the same as the difference in discordant response rate.

Analysis Based on Within Subject Differences

Let $w_i = u_i - v_i$, $i = 1, 2, ..., n$. The w_i are iid with the following probability distribution:

w	-1	0	1
$g(w)$	$\pi_{0,1}$	$(\pi_{0,0} + \pi_{1,1})$	$\pi_{1,0}$

That is,

$g(-1) = P(w = -1) = P(u - v = -1) = P(u = 0 \text{ and } v = 1) = \pi_{0,1}.$

$g(0) = P(w = 0) = P(u - v = 0) =$

$P(u = 0 \text{ and } v = 0) + P(u = 1 \text{ and } v = 1) = \pi_{0,0} + \pi_{1,1}.$

Similarly, $g(1) = \pi_{1,0}$.

$E(w_i) = -1(\pi_{0,1}) + 0(\pi_{0,0} + \pi_{1,1}) + 1(\pi_{1,0}) =$

$\pi_{1,0} - \pi_{0,1} = \delta.$

$Var(w_i) = E(w_i^2) - [E(w_i)]^2,$

$E(w_i^2) = (-1)^2\pi_{0,1} + (0)^2(\pi_{0,0} - \pi_{1,1}) + (1)^2\pi_{1,0} = \pi_{1,0} + \pi_{0,1}.$

$Var(w_i) = (\pi_{1,0} + \pi_{0,1}) - \delta^2 = f - \delta^2$

where $f = \pi_{1,0} + \pi_{0,1}$ is the discordant response rate. Let

$$\bar{w} = \sum_{i=1}^{n} \frac{w_i}{n}.$$

Then $E(\bar{w}) = \delta$ and $Var(w_i) = \frac{f-\delta^2}{n}$. For large n, the following quantity has approximately a standard normal distribution:

$$Z = \frac{\bar{w} - \delta}{\sqrt{\frac{f-\delta^2}{n}}}.$$

An Alternative For of $\bar{w}$

Let $Y_{i,j}$ be the number of subjects for which the response is the pair (i,j). $(i,i) = (0,0),(0,1),(1,0),(1,1)$. Then $(Y_{0,0}, Y_{1,0}, Y_{0,1}, Y_{1,1})$ has a multinomial distribution and each $Y_{i,j}$ has a marginal binomial distribution $\sum_{i,j} Y_{i,j} = n$. $\pi_{i,j}$ is the sample proportion of the subjects giving response $(i,j) = \frac{Y_{i,j}}{n}$. To relate these quantities to $\bar{w}$, note that $\sum_{i=1}^{n} u_i$ is the number of successes in treatment 1 which is $Y_{1,0} + Y_{1,1}$.

$$\sum_{i=1}^{n} \frac{u_i}{n} = \frac{Y_{1,0} + Y_{1,1}}{n} = \hat{\pi}_{1,0} + \hat{\pi}_{1,1}$$

Similarly, $\sum_{i=1}^{n} v_i$ is the number of successes in treatment 2 which is $Y_{0,1} + Y_{1,1}$.

$$\sum_{i=1}^{n} \frac{v_i}{n} = \frac{Y_{0,1} + Y_{1,1}}{n} = \hat{\pi}_{0,1} + \hat{\pi}_{1,1}.$$

Thus,

$$\bar{w} = \sum_{i=1}^{n} \frac{(u_i - v_i)}{n} = \frac{\sum_{i=1}^{n} u_i}{n} - \frac{\sum_{i=1}^{n} v_i}{n} = \hat{\pi}_1 - \hat{\pi}_2 =$$

$$(\hat{\pi}_{1,0} + \hat{\pi}_{1,1}) - (\hat{\pi}_{0,1} + \hat{\pi}_{1,1}) = (\hat{\pi}_{1,0} - \hat{\pi}_{0,1})$$

which is the sample difference in discordant response rates.

McNemar's Test

Consider the hypothesis $H_0 : \delta = 0$ versus $H_A : \delta \neq 0$ where $\delta = \pi_1 - \pi_2 = \pi_{1,0} - \pi_{0,1}$. Let π denote the common value of $\pi_{1,0}$ and $\pi_{0,1}$. An estimate of π is

$$\bar{p} = \frac{1}{2}\left(\hat{\pi}_{1,0} + \hat{\pi}_{0,1}\right).$$

$$E_{H_0}(\bar{w}) = \delta = 0.$$

$$Var_{H_0}(\bar{w}) = \frac{f - \delta^2}{n} = \frac{2\pi}{n}.$$

with $\delta = 0$ and $f = \pi_{1,0} + \pi_{0,1} = 2\pi$. The test statistic is

$$Z = \frac{\bar{w} - 0}{\sqrt{\frac{2\bar{p}}{n}}}.$$

This is equivalent to

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{2\bar{p}}{n}}}.$$

Z^2 is the McNemar's statistic, which has approximately a chi-square distribution with degrees of freedom equal to 1. Reject H_0 if $z_{\text{obs}} \leq -z_{\alpha/2}$ or if $z_{\text{obs}} \geq z_{\alpha/2}$ which is equivalent to, reject H_0 if $z^2 \geq z_\alpha$.

Sample Size-Power Equation

Recall that

$$|\delta| = z_{\alpha/2}\sigma_{0n} + z_\beta \sigma_{1n}$$

where $\delta = E(\bar{w})$ and

$$\sigma_{0n}^2 = Var_{H_0}(\bar{w}) = \frac{2\pi}{n}$$

and

$$\sigma_{1n}^2 = Var_{H_A}(\bar{w}) = \frac{f - \delta^2}{n}.$$

Substituting gives,

$$|\delta| = z_{\alpha/2}\frac{\sqrt{2\pi}}{\sqrt{n}} + z_\beta \frac{\sqrt{f - \delta^2}}{\sqrt{n}}$$

or

$$\sqrt{n} = \frac{z_{\alpha/2}\sqrt{2\pi} + z_\beta\sqrt{f - \delta^2}}{|\delta|}.$$

2π is usually replaced by the discordant response rate f. That is,

$$\sqrt{n} = \frac{z_{\alpha/2}\sqrt{f} + z_\beta\sqrt{f - \delta^2}}{|\delta|}.$$

Example: Consider the eye study described earlier. Let π_1 be the success rate on treatment 1 equal to 0.80. Let π_2 be the success rate on treatment 2 equal to 0.60. Let f be the discordant rate equal to 0.50 which is equal to $\pi_{1,0} + \pi_{0,1}$. For a 2-sided test, let $\alpha = 0.05$, and the power be 0.90. Then,

$$\delta_1 = \pi_1 - \pi_2 = 0.20,$$

$$\sqrt{n} = \frac{1.96\sqrt{0.50} + 1.28\sqrt{0.50 - (0.20)^2}}{0.20} = 11.27 \rightarrow n = 128.$$

9.7 Simple Linear Regression

Take the simple linear regression model with constant slope and intercept. The model statement is

$$y_i = \gamma + \theta_i + \epsilon_i, i = 1, 2, ..., n,$$

where ϵ_i is iid $N(0, \sigma^2)$. The sums of squares are

$$s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i.$$

The *slope estimator* is

$$\widehat{\theta} = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{s_{xx}}.$$

The *mean and variance estimators* are

$$E(\widehat{\theta}) = \theta,$$

$$Var(\widehat{\theta}) = \frac{\sigma_\epsilon^2}{s_{xx}}$$

The *standardized* of $\widehat{\theta}$ is

$$z = \frac{\widehat{\theta} - \theta}{\sigma_\epsilon / s_{xx}}.$$

The notation above is used in one of the sequential monitoring articles. Also note,

$$(x_i - \bar{x})(y_i - \bar{y}) = (x_i - \bar{x}) y_i - (x_i - \bar{x}) \bar{y},$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^{n} x_i - \bar{x} =$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) y_i - 0,$$

$$E(\widehat{\theta}) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})E(y_i)}{s_{xx}},$$

$$E(y_i) = \delta + \theta x_i,$$

$$Var(y_i) = Var(\epsilon_i) = \sigma_\epsilon^2.$$

$$\Rightarrow E(\widehat{\theta}) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(\delta + \theta x_i)}{s_{xx}} =$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})\delta + \sum_{i=1}^{n}(x_i - \bar{x})\theta x_i}{s_{xx}} =$$

$$\frac{\delta \sum_{i=1}^{n}(x_i - \bar{x}) + \theta \sum_{i=1}^{n}(x_i - \bar{x})x_i}{s_{xx}} =$$

$$0 + \frac{\theta \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{s_{xx}} = \theta \frac{s_{xx}}{s_{xx}} = \theta \Rightarrow \text{unbiased.}$$

Now find the variance of $\widehat{\theta}$.

$$\widehat{\theta} = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{s_{xx}}\right] y_i,$$

$$Var(\widehat{\theta}) = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{s_{xx}}\right]^2 Var(y_i) =$$

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{s_{xx}^2}\sigma_\epsilon^2 = \sigma_\epsilon^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{s_{xx}^2} =$$

$$\sigma_\epsilon^2 \frac{s_{xx}}{s_{xx}^2} = \frac{\sigma_\epsilon^2}{s_{xx}} \Rightarrow \text{ unbiased.}$$

9.7.1 Model for Clinical Trials

The model is

$$y_{ij} = \delta_i + \theta_i x_j + \epsilon_{ij}, j = 1, 2, .., k; i = 1, 2, ..., n.$$

k is the number of visits and n is the number of subjects. The model for subject 1 is

$$y_{1j} = \delta_1 + \theta_1 x_j + \epsilon_{1j}, j = 1, 2, .., k.$$

For the estimator of $\widehat{\theta}_1$ see page 25 of the handout. The model for subject 2 is

$$y_{2j} = \delta_2 + \theta_2 x_j + \epsilon_{2j}, j = 1, 2, .., k.$$

For the estimator of $\widehat{\theta}_2$ see page 25 of the handout.

$$Var(\widehat{\theta}_i) = Var\left[E(\widehat{\theta}_i|\theta_i) + E(Var(\widehat{\theta}_i|\theta_i)\right] =$$

$$Var(\widehat{\theta}_i) + E\left[\frac{\sigma^2_\epsilon}{s_i}\right] =$$

$$\sigma^2_\theta + \frac{\sigma^2_\epsilon}{s_i}.$$

R (used in calculating the sample size) is

$$R = \frac{\sigma^2_\epsilon}{\sigma^2_\theta} = \frac{\text{within subject variabilities}}{\text{between subject variability}}.$$

9.8 Comparing Slopes for 2 Treatment Groups

Group	Population Mean Slope	Sample Size
1	θ_1	n_1
2	θ_2	n_2

Reference page 29 of the text book.

$$E(\widehat{\delta}) = E[\widehat{\theta}_1 - \widehat{\theta}_2] = E(\widehat{\theta}_1) - E(\widehat{\theta}_2) = \theta_1 - \theta_2,$$

$$Var(\widehat{\theta}_1 - \widehat{\theta}_2) = Var(\widehat{\theta}_1) + Var(\widehat{\theta}_2) =$$

$$\frac{1}{\sum_{i=1}^{n_1} v_{1i}^{-1}} + \frac{1}{\sum_{i=1}^{n_2} v_{2i}^{-1}}$$

$$v_{i1}^{-1} = \sigma^2_\theta[1 + R/S_{1i}],$$

$$v_{i2}^{-1} = \sigma^2_\theta[1 + R/S_{2i}],$$

Reference page 30 of the text book. Pages 28-30 are in error. On page 33 of the text book:

$L_{1i} = k; L_{2i} = k$ (no missed visits).

If $x_1 = 0$, then the calculation on page 33 of the text book is correct. On page 40, the response $(0, 0)$ and $(11, 1)$ do not say anything about the difference between treatment 1 and treatment 2.

$$H_0 : \Pi_1 = \Pi_2.$$

$$\delta = \Pi_1 - \Pi_2 = (\Pi_{10} - \Pi_{11}) - (\Pi_{01} - \Pi_{11}) = \Pi_{10} - \Pi_{01}$$

On page 42 of the text,

$$E(w) = \sum_{\forall w} wg(w) = \cdots,$$

$$E(w^2) = \sum_{\forall w} w^2 g(w) = \cdots,$$

Another form of H_0 is $H_0 : \delta = 0$. The test statistic under H_0 is

$$z = \frac{\bar{w} - 0}{\sqrt{\frac{f}{n}}}$$

$$z = \frac{\bar{w} - 0}{\sqrt{\frac{2\bar{p}}{n}}}$$

Continuing on to page 46 of the text book,

$$\mu_0 = E_{H_0}(\bar{w}),$$

$$\mu_1 = E_{H_1}(\bar{w}),$$

$$\delta = \mu_1 - \mu_0,$$

$$\sigma_{0n}^2 = Var_{H_0}(\bar{w}),$$

$$\sigma_{1n}^2 = Var_{H_1}(\bar{w}),$$

With $\mu_0 = 0$, and $\delta = 0$,

$$\mu_1 = \Pi_1 - \Pi_2,$$

$$\delta = \Pi_1 - \Pi_2,$$

$$\sigma_{0n}^2 = \frac{2\Pi}{n},$$

$$\sigma_{1n}^2 = \frac{f - \delta^2}{n}.$$

9.9 Homework and Answers

1. Fizz Laboratories, a pharmaceutical company, has developed a new pain relief medication(drug #1) for patients suffering from arthritis. The new medication is to be compared with a commonly marketed medication (drug #2). Equal numbers of subjects will be allocated to the 2 groups. Each subject will be treated and asked one hour later to rate the medication as either "complete relief" or "less than complete relief." Based upon the literature, an investigator believes that the standard drug will produce about 40% positive responses (i.e. $p_2 = 0.40$). Answer the following in terms of a 5% level test of $H_0 : p_1 = p_2$, versus $H_1 : p_1 \neq p_2$.

   a. What sample size ($n = n_1 + n_2$) is needed so the test has a power of 0.90 for detecting the difference $\delta_1 = 0.20$? ($\delta = p_1 - p_2$). **Answer:** $\delta_1 = p_1 - p_2 = p_1 - 0.40 = 0.20 \Rightarrow p_1 = 0.60$.

$$\bar{p} = \frac{p_1 + p_2}{2} = \frac{0.60 + 0.40}{2} = 0.50.$$

$$\sigma^2_{0n} = Var_{H_0}(S_n) = \bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

$$\sigma^2_{1n} = Var_{H_1}(S_n) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} =$$

$$\frac{0.60(0.40)}{n_1} + \frac{0.40(0.60)}{n_2}.$$

$\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$, and the power is $1 - \beta \Rightarrow \beta = 0.10$. Using the equation:

$$|\delta_1| = z_{\alpha/2}\sigma_{0n} + z_\beta \sigma_{1n},$$

$$0.20 = z_{\alpha/2}\sqrt{(0.25)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} + z_\beta\sqrt{\frac{0.24}{n_1} + \frac{0.24}{n_2}}$$

$n_1 = n_2, \alpha_{0.025} = \pm 1.96, \alpha_{0.10} = 1.282.$

$$0.20 = 1.96\sqrt{(0.25)\left(\frac{1}{n_1} + \frac{1}{n_1}\right)} + 1.282\sqrt{\frac{0.24}{n_1} + \frac{0.24}{n_1}}$$

$$0.20 = \frac{1.38593}{\sqrt{n_1}} + \frac{0.8882}{\sqrt{n_1}},$$

$$0.20\sqrt{n_1} = 1.38593 + 0.8882,$$

$$0.20\sqrt{n_1} = 2.2741,$$

$$\sqrt{n_1} = 11.3706,$$

$$n_1 = 129.29 \Rightarrow n_2 = 129.29 \Rightarrow n = n_1 + n_2 = 260.$$

b. If only $n = 200$ subjects can be recruited what will be the power of the test for detecting the difference $\delta_1 = 0.20$? **Answer:** $n = 200$

$$0.20 = 1.96\sqrt{(0.25)\left(\frac{1}{100} + \frac{1}{100}\right)} + z_\beta\sqrt{\frac{0.24}{100} + \frac{0.24}{100}}$$

$$0.20 = 0.138593 + z_\beta(0.06928),$$

$$0.061407 = z_\beta(0.06928),$$

$$z_\beta = 0.8864,$$

The power is

$$1 - \beta \approx 0.816.$$

2. Consider a trial that involves repeated measures on $n = n_1 + n_2$ subjects over a follow-up period of length D with K equally spaced follow-up times, Assume:

   i. Each subject completes all follow-up visits.

   ii. Subjects are allocated equally to the two groups.

   The difference $\widehat{\theta}_1 - \widehat{\theta}_2$ in slope estimators has variance

$$Var(\widehat{\theta}_1 - \widehat{\theta}_2) = \frac{4\sigma_\theta^2}{n}\left[1 + \frac{12R(k-1)}{D^2k(k+1)}\right]$$

   The *amount of information* available for estimating $\theta_1 - \theta_2$ is

$$I = \frac{1}{Var(\widehat{\theta}_1 - \widehat{\theta}_2)} = \frac{n}{4\sigma_\theta^2\left[1 + \frac{12R(k-1)}{D^2k(k+1)}\right]}, k \geq 2.$$

   The information varies with D, k, R and takes its maximum attainable value when k or D tend to ∞ (or $R = 0$). The maximum attainable value of I is $I_{\max} = n/4\sigma_\theta^2$. The ratio of I to $I_{\max}$ is

$$\frac{I}{I_{\max}} = \frac{1}{1 + \frac{12R(k-1)}{D^2k(k+1)}}$$

a. Construct a table showing how $(I/I_{\max}) \times 100$ varies as a function of k and R for a trial of duration $D = 3$ years. (Show the table for $k = 2, 4, 6, 8, 10$ and $R = 0.20, 1.0, 5.0$). **Answer:**

k	R	$I/I_{\max}$
2	0.2	95.74468
2	1	81.8181
2	5	47.36842
4	0.2	96.15385
4	1	83.333
4	5	50
6	0.2	96.92308
6	1	86.30137
6	5	55.75221
8	0.2	97.47292
8	1	88.52459
8	5	60.67416
10	0.2	97.86477
10	1	90.16393
10	5	64.70588

b. Using the table obtained in part (a), what general conclusion can be drawn concerning how the choice of k may depend on R. **Answer:** When k is constant and R increase, the ratio $I/I_{\max}$ decreases. As k increases and R is constant, $I/I_{\max}$ increases. In general choose k such that $R < k$ to maximize $I/I_{\max}$.

3. Let Y_1, Y_2 be *independent* binomial random variables with the following probability distributions:

$$f_i(y) = \left(\begin{array}{c} n_i \\ y \end{array} \right) p_i^y (1-p_i)^{n_i - y}, y = 0, 1, 2, ..., n_i; i = 1, 2.$$

Let $\widehat{p}_i = \frac{y_i}{n_i}$ denote the sample proportions for $i = 1, 2$. Assuming n_1 and n_2 are both large, $\widehat{p}_i$ has approximately a normal distribution with mean p_i and variance $\frac{p_i(1-p_i)}{n_i}$. If p_i is the probability that a subject in group i experiences a certain event during a particular time period, then $\theta = \frac{p_1}{p_2}$ is called the *relative risk* of experiencing that event.

a. Let $\widehat{\theta} = \frac{\widehat{p_1}}{\widehat{p_2}}$ denote an estimator of θ. Use the δ method to determine the mean and variance of the limiting normal distri-

bution of $\ln(\widehat{\theta})$, where $\ln(\cdot)$ denotes natural logarithm. (Note: $\ln(\widehat{\theta}) = \ln(\widehat{p}_1) - \ln(\widehat{p}_2)$).

b. Use the result obtained in part (a) to give a general formula for $100(1-\alpha)\%$ confidence limits for $\ln(\theta)$. Then invert the limits to give a formula for confidence limits on θ.

4. Let S_n denote a general statistic that has approximately, if n is large, a normal distribution with mean $\mu = E(S_n)$ and variance $\sigma_n^2 = Var(S_n)$. Consider testing $H_0 : \mu = u_0$ versus the one-sided alternative $H_1 : \mu > \mu_0$. The test statistic is $z = \frac{S_N - \mu_0}{\sigma_{0n}}$ where $\mu_0 = E_{H_0}(S_n)$ and $\sigma_{0n}^2 = Var_{H_0}(S_n)$. If H_0 is true, then z has approximately a standard normal distribution. Thus, H_0 will be rejected at significance level α if $z_{\text{obs}} \geq z_\alpha$. Beginning with the definition of power, show that the sample size needed so the test has power $1--\beta$ for detecting the difference $\delta_1 = \mu_1 - \mu_0$ must satisfy the following equation:

$$|\delta_1| = z_\alpha \sigma_{0n} + z_\beta \sigma_{1n}$$

where $\sigma_{1n}^2 = Var_{H_1}(S_n)$.

9.10 Dummy Variables

Corrections to the handout given. Listed are the dummy variables for Example 1:

$$V_{1,1,i} = \begin{cases} 1, & \text{if } (x_i, y_i) = (1,1) \\ 0, & \text{otherwise} \end{cases}$$

$$V_{1,2,i} = \begin{cases} 1, & \text{if } (x_i, y_i) = (1,0) \\ 0, & \text{otherwise} \end{cases}$$

$$V_{2,1,i} = \begin{cases} 1, & \text{if } (x_i, y_i) = (0,1) \\ 0, & \text{otherwise} \end{cases}$$

$$V_{2,2,i} = \begin{cases} 1, & \text{if } (x_i, y_i) = (0,0) \\ 0, & \text{otherwise} \end{cases}$$

$$n_{1,1} = \sum_{i=1}^{n} V_{1,1,i}.$$

$$n_{1,2} = \sum_{i=1}^{n} V_{1,2,i}.$$

$$n_{2,1} = \sum_{i=1}^{n} V_{2,1,i}.$$

$$n_{2,2} = \sum_{i=1}^{n} V_{2,2,i}.$$

where i indexes the number of successes in n trials. Marginally (individually), it is clear $n_{1,1}$, $n_{1,2}$, $n_{2,1}$, $n_{2,2}$ have binomial distributions. $n_{1,1}$, $n_{1,2}$, $n_{2,1}$, and $n_{2,2}$ are independent.

9.11 2×2 Frequency Tables

2×2 frequency tables arise in a number of settings:

- Paired binary data (studied earlier).
- Analysis of survival rates.
- Comparing proportions across strata.

Our discussion concerns a parameter called *the odds ratio.* The logarithm of the odds ratio is closely related to the logistic transform, which has an important role in the analysis of binary data.

Outline:

1. Two distinct sampling models.
2. Different forms of the chi-square statistic.
3. Relationship to the standardized difference of two sample proportions.
4. The odds ratio: definition, interpretation, estimation, and confidence limits.
5. Fisher's exact test.
6. Comparing proportions across strata: the Mantel-Haenszel statistic.

2×2 frequency tables arise when sampling a single population or when independently sampling two populations.

Example: A sample of n individuals is selected from a certain population and, for each individual, a pair of response variables (x, y) is observed.

		Tumor Regression	
		Yes	No
Toxicity	Yes	a	b
	No	c	d

Example: In a sample of $n = n_1 + n_2$ subjects, suppose n_1 of them are randomly allocated to drug A and n_2 to drug B. A single response variable is observed for each subject.

		Nausea		
		Yes	No	
Drug	A	a	b	n_1
	B	c	d	n_2

The interpretation of the above two examples is different, but the statistics are calculated the same way.

Definition If the row *or* column totals are fixed through sampling, the *sampling model* is two independent binomial distributions. If neither the row nor the column totals are fixed through sampling, the *sampling model* is the multinomial distribution.

Notation:

		Column		
		1	2	
Row	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$

and the corresponding parameters

π_{11}	π_{12}	$\pi_{1\cdot}$
π_{21}	π_{22}	$\pi_{2\cdot}$
$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	

With the multinomial sampling model, $\sum \pi_{i,j} = 1$ and $n_{11} + n_{12} + n_{21} + n_{22} = n$. With the binomial sampling model, $\pi_{11} + \pi_{12} = 1$, $\pi_{21} + \pi_{22} = 1$, and $n_{11} + n_{12} = n_{1\cdot}$, $n_{21} + n_{22} = n_{2\cdot}$ where $n_{1\cdot}$ and $n_{2\cdot}$ are sample sizes.

9.11.1 Binomial Sampling Model

n_{11} has a binomial distribution

$$f(n_{11}) = \left(\begin{array}{c} n_{1\cdot} \\ n_{11} \end{array} \right) (\pi_{11})^{n_{11}} (1 - \pi_{11})^{n_{1\cdot} - n_{11}}, n_{11} = 0, 1, 2, ..., n_{1\cdot}.$$

n_{21} has a binomial distribution

$$g(n_{21}) = \left(\begin{array}{c} n_{2\cdot} \\ n_{21} \end{array} \right) (\pi_{21})^{n_{21}} (1 - \pi_{21})^{n_{2\cdot} - n_{21}}, n_{21} = 0, 1, 2, ..., n_{2\cdot}.$$

where n_{11} and n_{21} are independent random variables.

9.11.2 Multinomial Sampling Model

For the multinomial model, the joint distribution is

$$f(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!} (\pi_{11})^{n_{11}} (\pi_{12})^{n_{12}} (\pi_{21})^{n_{21}} (\pi_{22})^{n_{22}}$$

where $n_{11} + n_{12} + n_{21} + n_{22} = n$, and $\sum \pi_{ij} = 1$.

9.11.3 Chi-Square Statistic

The standard chi-square statistic for testing homogeneity ($H_0 : \pi_{11} = \pi_{21}$) in two independent binomial populations or for testing independence ($H_0 : \pi_{11}\pi_{22} = \pi_{21}\pi_{12}$) in a multinomial population is

$$\chi^2 = \frac{(n_{11} n_{22} - n_{12} n_{21})^2 n}{n_{\cdot 1} n_{\cdot 2} n_{1\cdot} n_{2\cdot}}$$

where if $n \to \infty$ or if $n_1, n_2 \to \infty$, then χ^2 has a limiting chi-square distribution with degrees of freedom of 1. Note that the numerator is a cross product difference and the denominator is the product of the marginal totals.

9.11.4 Continuity Correction

The continuity correction for the chi-square is

$$\chi^2 = \frac{(|n_{11} n_{22} - n_{12} n_{21}| - \frac{n}{2})^2 n}{n_{\cdot 1} n_{\cdot 2} n_{1\cdot} n_{2\cdot}}.$$

9.11.5 Summary of the Chi-Square Statistics

The following summary applies to the chi-square statistics for 2×2 tables.

1. $$\chi^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2 n}{n_{.1}n_{.2}n_{1.}n_{2.}}$$

2. The typical goodness-of-fit:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

where $\hat{E}_{11} = \frac{n_{.1}n_{1.}}{n}$, $\hat{E}_{12} = \frac{n_{.2}n_{1.}}{n}$, $\hat{E}_{21} = \frac{n_{.1}n_{2.}}{n}$, and $\hat{E}_{22} = \frac{n_{.2}n_{2.}}{n}$.

3. $$\chi^2 = \frac{[n_{11} - E_{H_0}(n_{11})]^2}{Var_{H_0}(n_{11})}$$

where

$$E_{H_0} = n_{1.}\frac{n_{.1}}{n},$$

$$Var_{H_0}(n_{11}) = \frac{n_{.1}n_{.2}n_{1.}n_{2.}}{n^2(n-1)}.$$

4. $$\chi^2 = \frac{[w(\hat{p}_1 - \hat{p}_2)]^2}{\frac{nw\hat{p}(1-\hat{p})}{n-1}}$$

where

$$w = \frac{n_{1.}n_{2.}}{n},$$

$$\hat{p}_1 = \frac{n_{11}}{n_{1.}},$$

$$\hat{p}_2 = \frac{n_{21}}{n_{2.}},$$

$$\hat{p} = \frac{n_{11} + n_{21}}{n},$$

Note that (1) and (2) are equivalent. (3) and (4) are equivalent, but slightly different from (1) and (2). (3) and (4) are closely related to the Mantel-Haenszel statistic that we will study later. It can be shown that, in (4),

$$\chi^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2(n-1)}{n_{.1}n_{.2}n_{1.}n_{2.}}.$$

Thus, (3) and (4) differ from (1) and (2) only by n and $n-1$ in the numerator. If n is large, the all four equations are nearly identical. Later we will show that the difference between (1) or (2) and (3) or (4) is, for large samples, the difference between a conditional test and an unconditional test.

9.11.6 Relationship to the Standardized Differences

This section will cover the relationship of the chi-squares to the standardized differences of two sample proportions. In the binomial case, consider $H_0 : \pi_{11} = \pi_{21}$. Let

$$\hat{p}_1 = \frac{n_{11}}{n_{1\cdot}},$$

$$\hat{p}_2 = \frac{n_{21}}{n_{2\cdot}},$$

and let

$$\hat{p} = \frac{n_{11} + n_{21}}{n},$$

be the estimate of a common value of π_{11} and π_{21}. The standardized difference is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{1\cdot}} + \frac{\hat{p}(1-\hat{p})}{n_{2\cdot}}}}.$$

It has a approximate standard normal distribution under H_0. Thus,

$$Z^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}(1-\hat{p})[\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}]}$$

has an approximate chi-square distribution with 1 degrees of freedom. It is easy to show that

$$Z^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

is the statistic found in (1) listed earlier.

9.11.7 The Odds Ratio

Let p denote the probability of getting a head when tossing a coin. The *odds* of getting a head, as opposed to not getting a head, is $\frac{p}{1-p}$. Thus, if $p = 0.40$, the odds of getting a head is $\frac{0.40}{0.60} = \frac{4}{6}$ or 4:6. Consider a 2×2

table with parameters

π_{11}	π_{12}
π_{21}	π_{22}

In row 1, the odds that the event in column 1 occurs is

$$\frac{\pi_{11}}{\pi_{12}} = \frac{\pi_{11}}{1-\pi_{11}}.$$

In row 2, the odds that the event in column 1 occurs is

$$\frac{\pi_{21}}{\pi_{22}} = \frac{\pi_{21}}{1-\pi_{21}}.$$

Definition: The odds ratio, ψ is row 1 odds divided by row 2 odds. So,

$$\psi = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$$

If the rows of the table represent independent binomial samples, then, $\pi_{11} + \pi_{12} = 1$ and $\pi_{21} + \pi_{22} = 1$ or $\pi_{12} = 1 - \pi_{11}$ and $\pi_{22} = 1 - \pi_{21}$. In this case, the odds ratio is

$$\psi = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}} = \frac{\frac{\pi_{11}}{1-\pi_{11}}}{\frac{\pi_{21}}{1-\pi_{21}}}.$$

If π_{11} and π_{21} are both small, the

$$\psi = \frac{\pi_{11}}{\pi_{21}}$$

where the later quantity is called the *relative risk* of the event in column 1.

9.11.8 Odds Ratio Estimate

$\psi = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$. $\hat{\psi}$ is an estimate of ψ is $\hat{\psi} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{21}\hat{\pi}_{12}}$. $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$ is the sample proportion in row i and column j. Thus, $\hat{\psi} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$.

Example: [Physician's Health Study (Meinert, 1986, page 314)] Take two independent binomials. The following table comes from a double blinded study with physicians as the participants.

	Myocardial Infarction		
	Attack	No Attack	
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

The relative odds of a heart attack (placebo versus aspirin) is

$$\hat{\psi} = \frac{189(10,933)}{104(10,845)} = 1.83.$$

Thus, the odds of an attack in the placebo group is 1.83 times the odds in the aspirin group.

9.11.9 Properties of the Odds Ratio

Recall that ψ is the row 1 odds divided by the row 2 odds.

1. In general, $0 < \psi < \infty$.
2. When the row 1 odds is greater than the row 2 odds, $1 < \psi < \infty$.
3. When the row 1 odds is less than the row 2 odds, $0 < \psi < 1$.
4. ψ is estimated in the same way in the binomial and multinomial sampling models.
5. In the binomial model, ψ is a measure of *departure from homogeneous* rows. $\psi = \frac{\pi_{11}(1-\pi_{21})}{\pi_{21}(1-\pi_{11})}$.
   (a) $\psi = 1$ implies $\pi_{11} = \pi_{21}$.
   (b) $\psi < 1$ implies $\pi_{11} < \pi_{21}$.
   (c) $\psi > 1$ implies $\pi_{11} > \pi_{21}$.
6. In multinomial sampling, ψ is a *measure of association.*
   (a) $\psi = 1$ implies independent row and column categories.
   (b) $\psi < 1$ implies negative association.
   (c) $\psi > 1$ implies positive association.

   Those are measures of dependence and independence.

Proof of item (6): Let (x, y) have the following bivariate distribution.

		y = 1	y = 0	
x	1	π_{11}	π_{12}	$\pi_{1\cdot}$
	0	π_{21}	π_{22}	$\pi_{2\cdot}$
		$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	

The marginal of x is

x	0	1
$f_1(x)$	$\pi_{2\cdot}$	$\pi_{1\cdot}$

The marginal of y is

y	0	1
$f_2(y)$	$\pi_{\cdot 2}$	$\pi_{\cdot 1}$

The expectations of x and y are as follow:

$$E(x) = \pi_{1\cdot} = \pi_{11} + \pi_{12},$$

$$E(y) = \pi_{\cdot 1} = \pi_{11} + \pi_{21},$$

Thus the covariance is

$$Cov(x, y) = \pi_{11} - \pi_{1\cdot}\pi_{\cdot 1} = \pi_{11} - (\pi_{11} + \pi_{12})(\pi_{11} + \pi_{21}) =$$

$$\pi_{11} - \pi_{11}^2 - \pi_{11}\pi_{12} - \pi_{12}\pi_{11} - \pi_{12}\pi_{21} =$$

$$\pi_{11}(1 - \pi_{11} - \pi_{21} - \pi_{12}) - \pi_{12}\pi_{21} =$$

$$\pi_{11}\pi_{22} - \pi_{12}\pi_{21} =$$

$$\pi_{12}\pi_{21}\left[\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} - 1\right] =$$

$$\pi_{12}\pi_{21}[\psi - 1].$$

Assume that $0 < \pi_{ij} < 1$ for all (i, j). Then,

$$Cov(x, y) = 0 \Rightarrow \psi = 1.$$

$$Cov(x, y) > 0 \Rightarrow \psi > 1.$$

$$Cov(x, y) < 0 \Rightarrow \psi < 1.$$

All that remains to be shown is that $Cov(x, y) = 0 \Rightarrow x, y$ are independent. By a well known property of independence of random variables, we have, x, y are independent implies $Cov(x, y) = 0$. We need only show the converse. Note that

$$E(x) = P(x = 1) = \pi_1.$$

$$E(y) = P(y = 1) = \pi_{\cdot 1}$$

$$E(xy) = P(x = 1 \text{ and } y = 1) = \pi_{11}.$$

Thus,

$$Cov(x, y) = E(xy) - E(x)E(y), \forall x, y \Rightarrow$$

$$P(x = 1 \text{ and } y = 1) - P(x = 1)P(y = 1)$$

and $Cov(x, y) = 0 \Rightarrow P(x = 1 \text{ and } y = 1) = P(x = 1)P(y = 1)$. We omit showing the other cases, although they follow by repeatedly using the law of complements.

9.11.10 Large Sample Confidence Limits

This section contains the derivation for confidence limits for 2×2 tables for the odds ratio. The following derivations only apply to a 2×2 table for either the multinomial or binomial sampling schemes. The maximum likelihood estimator of ψ is

$$\hat{\psi} = \frac{n_{11}n_{22}}{n_{21}n_{12}}.$$

The $\log \hat{\psi}$ has approximately a normal distribution with mean $\ln \psi$ and a standard deviation of

$$\hat{\sigma}_{\log \hat{\psi}} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

as $n \to \infty$. Approximate $100(1 - \alpha)\%$ confidence limits for $\ln \psi$ are

$$\ln \hat{\psi} \pm z_{\alpha/2}\hat{\sigma}_{\ln \hat{\psi}}.$$

or

$$P\left(-z_{\alpha/2} < \frac{\ln \hat{\psi} - \ln \psi}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}} < z_{\alpha/2}\right)$$

Any n_{ij} that is small or close to zero should be replaced by $n_{ij} + 0.50$.

Example: [The Physician's Health Study] For very small p and large n, a Poisson distribution approximation will work, but a Normal distribution approximation will not work. From a previous section $\psi = 1.83$. $\log \hat{\psi} = 0.6043$ and

$$\hat{\sigma}_{\log \hat{\psi}} = \sqrt{\frac{1}{189} + \frac{1}{104} + \frac{1}{10,845} + \frac{1}{10,933}} = 0.12284.$$

The 95% confidence limits for $\ln \psi$ are

$$0.6043 \pm 1.96(0.12284)$$

$$0.6043 \pm 0.2408$$

$$0.3635 < \ln \psi < 0.8451$$

or

$$1.44 \leq \psi \leq 2.33.$$

The odds of an heart attack in the placebo group relative to the aspirin group is estimated to lie between 1.44 and 2.33.

9.11.11 Basis for the Confidence Limit

The log odds ratio is a function of the sample proportions $\hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{21}$, and $\hat{\pi}_{22}$ which in turn are functions of the cell counts n_{11}, n_{12}, n_{21}, and n_{22}, which either have a multinomial distribution or arise from independent binomial distributions. Since $(\hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{21}, \hat{\pi}_{22})$ has a limiting multivariate normal distribution, the $\delta-$method can be used to determine the limiting normal distribution of

$$\log \hat{\psi} = \log \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{21}\hat{\pi}_{12}}.$$

9.12 Homework and Answers

1. Suppose each subject in a clinical trial can be classified by two factors, each of which is described by a vector (x, y) of random variables that has the following joint distribution.

y / x	1	0
1	π_{11}	π_{12}
0	π_{21}	π_{22}

Now consider a sequence $(x_i, y_i), i = 1, 2, ..., n$ of iid random vectors corresponding to observing the response of n subjects. It will be convenient to introduce the following binary random variables.

$$V_{i11} = \begin{cases} 1, & \text{if } (x_i, y_i) = (1, 1). \\ 0, & \text{otherwise.} \end{cases}$$

$$V_{i12} = \begin{cases} 1, & \text{if } (x_i, y_i) = (1, 0). \\ 0, & \text{otherwise.} \end{cases}$$

$$V_{i21} = \begin{cases} 1, & \text{if } (x_i, y_i) = (0, 1). \\ 0, & \text{otherwise.} \end{cases}$$

$$V_{i22} = \begin{cases} 1, & \text{if } (x_i, y_i) = (0, 0). \\ 0, & \text{otherwise.} \end{cases}$$

Let $V_i' = (V_{i11}, V_{i12}, V_{i21}, V_{i22}), i = 1, 2, ..., n$ and note that $V_1, V_2, ..., V_n$ are iid random vectors.

(a) The covariance matrix of $V_i' = (V_{i11}, V_{i12}, V_{i21}, V_{i22})$ is

$$\Sigma = \begin{bmatrix} Var(V_{11}) & Cov(V_{11}, V_{12}) & Cov(V_{11}, V_{21}) & Cov(V_{11}, V_{22}) \\ Cov(V_{12}, V_{11}) & Var(V_{12}) & Cov(V_{12}, V_{21}) & Cov(V_{12}, V_{22}) \\ Cov(V_{21}, V_{11}) & Cov(V_{21}, V_{12}) & Var(V_{21}) & Cov(V_{21}, V_{22}) \\ Cov(V_{22}, V_{11}) & Cov(V_{22}, V_{12}) & Cov(V_{22}, V_{21}) & Var(V_{22}) \end{bmatrix}$$

where the entries have a certain order to facilitate solving Part (b) below. For example, $Cov(V_{11}, V_{12}) = E(V_{11}V_{12}) - E(V_{11})E(V_{12})$. But, $V_{11}V_{12} = 0$ because a single subject's response can not simultaneously fall into category $(1, 1)$ and $(1, 0)$. Thus, $Cov(V_{11}, V_{12}) = 0 - E(V_{11})E(V_{12}) = -E(V_{11})E(V_{12})$, $E(V_{11}) = P(V_{11} = 1) = \pi_{11}, E(V_{12}) = P(V_{12} = 1) = \pi_{12} \Rightarrow Cov(V_{11}, V_{12}) = -\pi_{11}\pi_{12}$. Complete the derivation of all entries of Σ. Solution: $E(V_{11}) = \pi_{11}, E(V_{12}) = \pi_{12}, E(V_{21}) = \pi_{21}, E(V_{22}) = \pi_{22}, E(V_{11}^2) = E(V_{11}) = \pi_{11}$, etc. $Var(V_{11}) = E(V_{11}^2) - [E(V_{11})]^2 = \pi_{11} - \pi_{11}^2 = \pi_{11}(1 - \pi_{11})$. Similarly, $Var(V_{12}) = \pi_{12}(1 - \pi_{12}), Var(V_{21}) = \pi_{21}(1 - \pi_{21})$, and $Var(V_{22}) = \pi_{22}(1 - \pi_{22})$. $Cov(V_{11}, V_{12}) = E(V_{11}V_{12}) - E(V_{11})E(V_{12}) = 0 - \pi_{11}\pi_{12} = -\pi_{11}\pi_{12}$, etc. Thus, the vector

$$\underline{V} = \begin{bmatrix} V_{11} \\ V_{12} \\ V_{21} \\ V_{22} \end{bmatrix}$$

has a covariance matrix

$$\Sigma = (\sigma_{ij}) =$$

$$\begin{bmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{12} & -\pi_{11}\pi_{21} & -\pi_{11}\pi_{22} \\ -\pi_{12}\pi_{11} & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{21} & -\pi_{12}\pi_{22} \\ -\pi_{21}\pi_{11} & -\pi_{21}\pi_{12} & \pi_{21}(1-\pi_{21}) & -\pi_{21}\pi_{22} \\ -\pi_{22}\pi_{11} & -\pi_{22}\pi_{12} & -\pi_{22}\pi_{21} & \pi_{22}(1-\pi_{22}) \end{bmatrix}$$

(b) Let the vector of sample proportions $\widehat{\pi}_{ij} = n_{ij}/n$ be written in the form

$$\underline{\widehat{\pi}}_i = \begin{bmatrix} \widehat{\pi}_{11} \\ \widehat{\pi}_{12} \\ \widehat{\pi}_{21} \\ \widehat{\pi}_{22} \end{bmatrix}$$

which corresponds to the order used in deriving the covariance matrix in Part (a). Recall that the log of the odds ratio $\psi = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$ can be written as a function of $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$, say $\phi = \phi(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ in the following way.

$$\phi = \ln(\psi) = \ln(\pi_{11}) + \ln(\pi_{22}) - \ln(\pi_{12}) - \ln(\pi_{21}).$$

The δ method theorem states that

$$\widehat{\phi} = \ln(\widehat{\pi}_{11}) + \ln(\widehat{\pi}_{22}) - \ln(\widehat{\pi}_{12}) - \ln(\widehat{\pi}_{21}).$$

has approximately a normal distribution as $n \to \infty$ with mean of ϕ and variance

$$Var(\widehat{\phi}) = \frac{1}{n}\underline{a}'\Sigma\underline{a}$$

where

$$a_1 = \frac{\partial\phi}{\partial\pi_{11}},$$

$$a_2 = \frac{\partial\phi}{\partial\pi_{12}},$$

$$a_3 = \frac{\partial\phi}{\partial\pi_{21}},$$

$$a_4 = \frac{\partial \phi}{\partial \pi_{22}},$$

Simplify $\underline{a}'\Sigma\underline{a}$ to show that

$$Var(\widehat{\phi}) = \frac{1}{n}\left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right).$$

Solution:

$$a_1 = \frac{\partial \phi}{\partial \pi_{11}} = \frac{1}{\pi_{11}},$$

$$a_2 = \frac{\partial \phi}{\partial \pi_{12}} = -\frac{1}{\pi_{12}},$$

$$a_3 = \frac{\partial \phi}{\partial \pi_{21}} = -\frac{1}{\pi_{21}},$$

$$a_4 = \frac{\partial \phi}{\partial \pi_{22}} = \frac{1}{\pi_{22}}.$$

$\underline{a}'\Sigma\underline{a} =$

$$[a_1, a_2, a_3, a_4] \begin{bmatrix} a_1\sigma_{11} + a_2\sigma_{12} + a_3\sigma_{13} + a_4\sigma_{14} \\ a_1\sigma_{21} + a_2\sigma_{22} + a_3\sigma_{23} + a_4\sigma_{24} \\ a_1\sigma_{31} + a_2\sigma_{32} + a_3\sigma_{33} + a_4\sigma_{34} \\ a_1\sigma_{41} + a_2\sigma_{42} + a_3\sigma_{43} + a_4\sigma_{44} \end{bmatrix} =$$

$$[a_1, a_2, a_3, a_4] \begin{bmatrix} (1-\pi_{11}) + \pi_{11} + \pi_{11} - \pi_{11} \\ -\pi_{12} - (1-\pi_{12}) + \pi_{12} - \pi_{12} \\ -\pi_{21} + \pi_{21} - (1-\pi_{21}) - \pi_{21} \\ -\pi_{22} + \pi_{22} + \pi_{22} + (1-\pi_{22}) \end{bmatrix} =$$

$$\left[\frac{1}{\pi_{11}}, -\frac{1}{\pi_{12}}, -\frac{1}{\pi_{21}}, \frac{1}{\pi_{22}}\right] \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} = \sum\sum \frac{1}{\pi_{ij}}.$$

Thus, the asymptotic variance of $\log\widehat{\psi}$ is $\sigma^2 = \frac{1}{n}\sum \frac{1}{\pi_{ij}}$ which is estimated by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum \frac{1}{\widehat{\pi}_{ij}} = \frac{1}{n}\sum \frac{n}{n_{ij}} = \sum \frac{1}{n_{ij}}.$$

2. The following data are from a study that compared two local anesthetics used in dental surgery to relive pain. The table below shows the numbers x_{ij} of subjects who reported no pain during surgery.

Kind of		Anesthetic # 1	Anesthetic # 2	
Surgery	n	x	n	x
1(Periodontic)	23	13	24	16
2(Endontic)	29	20	31	27

Assume that $\{x_{ij}\}$ are independent and have binomial distributions with sample sizes n_{ij} (indicated in the table) and with success probabilities θ_{ij}. That is, θ_{ij} is the probability that a subject in row i, column j suffers no pain. The two anesthetics represent different treatments while the kind of surgery is a factor with two levels. Recall that the logistic transform is

$$\lambda_{ij} = \log\left(\frac{\theta_{ij}}{1-\theta_{ij}}\right).$$

(a) Let $\Delta_1 = \lambda_{11} - \lambda_{12}$ denote the log odds ratio for row 1 and let $\Delta_2 = \lambda_{21} - \lambda_{22}$ denote the log odds ratio for row 2. The model described above implies *that all observations in row 1 are independent of all observations in row 2.* Use this fact and our previous results on confidence limits for a single log odds ratio to give a general formula for confidence limits for $\Delta_1 - \Delta_2$. You will need to think of the data in rows 1 and 2 as consisting of two 2×2 tables.

	S	F
1	$n_{11}^{(1)}$	$n_{12}^{(1)}$
2	$n_{21}^{(1)}$	$n_{22}^{(1)}$

1	$n_{11}^{(2)}$	$n_{12}^{(2)}$
2	$n_{21}^{(2)}$	$n_{22}^{(2)}$

Solution:

$$\widehat{\Delta}_1 = \log\left(\frac{n_{11}^{(1)} n_{22}^{(1)}}{n_{21}^{(1)} n_{12}^{(1)}}\right)$$

Earlier we showed that $\widehat{\Delta}_1$ has a mean of

$$\Delta_1 = \log\left(\frac{\theta_{11}^{(1)}(1-\theta_{21}^{(1)})}{\theta_{21}^{(1)}(1-\theta_{11}^{(1)})}\right).$$

$$Var[\widehat{\Delta}_1] = \frac{1}{n_{11}^{(1)}} + \frac{1}{n_{12}^{(1)}} + \frac{1}{n_{21}^{(1)}} + \frac{1}{n_{22}^{(1)}}.$$

Similarly for the 2×2 table in row 2,

$$\widehat{\Delta}_2 = \log\left(\frac{n_{11}^{(2)} n_{22}^{(2)}}{n_{21}^{(2)} n_{12}^{(2)}}\right)$$

has a mean

$$\Delta_2 = \log\left(\frac{\theta_{11}^{(2)}(1-\theta_{21}^{(2)})}{\theta_{21}^{(2)}(1-\theta_{11}^{(2)})}\right).$$

$$Var[\widehat{\Delta}_2] = \frac{1}{n_{11}^{(2)}} + \frac{1}{n_{12}^{(2)}} + \frac{1}{n_{21}^{(2)}} + \frac{1}{n_{22}^{(2)}}.$$

Since $\widehat{\Delta}_1, \widehat{\Delta}_2$ are independent, $\widehat{\Delta}_1 - \widehat{\Delta}_2$ is asymptotically normal with a mean of $\Delta_1 - \Delta_2$ and a variance of

$$Var[\widehat{\Delta}_1 - \widehat{\Delta}_2] = \sum \frac{1}{n_{ij}^{(1)}} + \sum \frac{1}{n_{ij}^{(2)}}.$$

Thus, a general $100(1-\alpha)\%$ confidence interval for $\Delta_1 - \Delta_2$ is

$$\widehat{\Delta}_1 - \widehat{\Delta}_2 \pm \sqrt{\sum \frac{1}{n_{ij}^{(1)}} + \sum \frac{1}{n_{ij}^{(2)}}}.$$

(b) Use the given data to obtain a 95% confidence limit for $\Delta_1 - \Delta_2$.
Solution: For the given data, we form the following 2×2 tables:

For Row 1:

	S	F	
1	13	10	23
2	16	8	24
	29	18	47

$$\widehat{\Delta}_1 = \log\left(\frac{13(8)}{16(10)}\right) = -0.43,$$

$$e_1 = E_{H_0}(n_{11}^{(1)}) = \frac{29(23)}{47} = 14.19,$$

$$v_1 = Var_{H_0}(n_{11}^{(1)}) = \frac{29(18)(23)(24)}{(47)^2(46)} = 2.84.$$

For Row 2:

	S	F	
1	20	9	29
2	27	4	31
	47	13	60

$$\widehat{\Delta}_2 = \log\left(\frac{20(4)}{27(9)}\right) = -1.11,$$

$$e_2 = E_{H_0}(n_{11}^{(2)}) = \frac{47(29)}{60} = 22.72,$$

$$v_2 = Var_{H_0}(n_{11}^{(2)}) = \frac{47(13)(29)(31)}{(60)^2(59)} = 2.59.$$

Since the 95% confidence limits for $\Delta_1 - \Delta_2$ are

$$-0.43 - (-1.11) \pm$$

$$1.96\sqrt{\left(\frac{1}{13} + \frac{1}{10} + \frac{1}{16} + \frac{1}{8}\right) + \left(\frac{1}{20} + \frac{1}{9} + \frac{1}{27} + \frac{1}{4}\right)} =$$

$$0.68 \pm 1.96\sqrt{0.813},$$

$$-1.09 < \Delta_1 - \Delta_2 < 2.45.$$

(c) On the basis of the confidence limits in Part (b), is there any evidence of interaction between treatments and the kind of surgery? (Explain how you are using the confidence limits to arrive at a conclusion). Solution: Since zero is included between these limits, we have no evidence against the assumption $\Delta_1 - \Delta_2 = 0$. Conclusion: No evidence of interaction.

(d) Use the Mantel-Hanszel statistic to test, at significance level $\alpha = 0.05$ the hypothesis of no treatment difference (Show what you are doing). Solution: $H_0 : \Delta = 0$, versus $H_1 : \Delta \neq 0$, $\alpha = 0.05$ where Δ is the common log odds ratio of rows 1 and 2. The test statistic is

$$\chi^2 = \frac{(|S - E| - 0.50)^2}{V}$$

where

$$S = n_{11}^{(1)} + n_{11}^{(2)},$$

$$E = E_{H_0}(S),$$

$$V = Var_{H_0}(S).$$

Reject H_0 if $\chi^2_{obs} \geq 3.84$. Conclusion:

$$S = 13 + 20 = 33,$$

$$E = e_1 + e_2 = 14.19 + 22.72 = 36.91,$$

$$V = v_1 + v_2 = 2.84 + 2.59 = 5.43,$$

$$\chi^2_{obs} = \frac{(|33 - 36.91| - 0.50)^2}{5.43} = 2.14$$

Thus, do not reject the null hypothesis.

3. Let X, Y have independent binomial distributions,

$$P(X = x) = \left(\begin{array}{c} n_1 \\ x \end{array} \right) (\theta_1)^x (1 - \theta_1)^{n_1 - x}, x = 0, 1, ..., n_1,$$

$$P(Y = y) = \left(\begin{array}{c} n_2 \\ y \end{array} \right) (\theta_2)^y (1 - \theta_2)^{n_2 - y}, y = 0, 1, ..., n_2,$$

where $n_1 + n_2 = 7$. Consider the problem of testing $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 < \theta_2$ at the significance level $\alpha = 0.02$. When using the Fisher exact test, the rejection region is left tailed and consists of rejecting H_0 whenever the conditional p-value is less-than or equal to 0.02 where the p-value is equal to

$$P_{H_0}(X \leq x_{obs} | S = s),$$

where x_{obs} is the observed value of x and s is the observed value of $S = x + y$.

(a) Use the attached table to determine the p-value when $x_{obs} = 0$ and $s = 6$. Solution:

$$p - value = P_{H_0}(X \leq 0 | s = 6) = \frac{7}{3003}.$$

(b) Note that $x = 0$ and $s = 6$ corresponds to the (x, y) pair $x = 0$, and $y = 6$ (because $y = s - x$). Use the attached table to determine the set, denoted by $W_{0.02}$, of all (x, y) pairs for which H_0 will be rejected at $\alpha = 0.02$ (i.e. all (x, y) pairs for which the p-value is less than or equal to 0.02). The set $W_{0.02}$ contains 6 pairs. Solution:

$$W_{0.02} = \{(0, 5), (0, 6), (0, 7), (1, 6), (1, 7), (2, 7)\}$$

These outcomes occur when $s = 5, 6, 7, 8, 9$ and result in rejecting H_0 at the level $\alpha = 0.02$.

(c) The set $W_{0.02}$ is the rejection region of an $\alpha = 0.02$ level test. Thus the power of the test is the probability, calculated under the assumption of specific values of $\theta_1 < \theta_2$, that (x, y) takes a value in the set $W_{0.02}$. Determine the power of the test at the specific alternative $\theta_1 = 0.20$ and $\theta_2 = 0.80$. Hint: The power is

$$\sum_{(x,y) \in W_{0.02}} P(X = x, Y = y).$$

The terms of this sum are easy to calculate.

$$\phi(0.20, 0.80) =$$

$$P_{0.20}(X = 0)P_{0.80}(Y = 5) + P_{0.20}(X = 0)P_{0.80}(Y = 6) +$$

$$P_{0.20}(X = 0)P_{0.80}(Y = 7) + P_{0.20}(X = 1)P_{0.80}(Y = 6) +$$

$$P_{0.20}(X = 1)P_{0.80}(Y = 7) + P_{0.20}(X = 2)P_{0.80}(Y = 7).$$

$$P_{0.20}(X = 0) = \left(\begin{array}{c} 7 \\ 0 \end{array} \right) (0.20)^0 (0.80)^7 = 0.2097,$$

$$P_{0.20}(X=1) = \left(\begin{array}{c} 7 \\ 1 \end{array} \right) (0.20)^1(0.80)^6 = 0.3670,$$

$$P_{0.20}(X=2) = \left(\begin{array}{c} 7 \\ 2 \end{array} \right) (0.20)^2(0.80)^5 = 0.2753,$$

$$P_{0.80}(X=5) = \left(\begin{array}{c} 7 \\ 5 \end{array} \right) (0.80)^5(0.20)^2 = 0.2753,$$

$$P_{0.80}(X=6) = \left(\begin{array}{c} 7 \\ 6 \end{array} \right) (0.80)^6(0.20)^1 = 0.3670,$$

$$P_{0.80}(X=7) = \left(\begin{array}{c} 7 \\ 7 \end{array} \right) (0.80)^7(0.20)^0 = 0.2097.$$

Then,

$$\phi(0.20, 0.80) = 0.2097(0.2753) + (0.2097)(0.3670) +$$
$$(0.2097)(0.2097) + 0.3670(0.3670) + 0.3670(0.2097) +$$
$$(0.2753)(0.2097) = 0.4480.$$

4. The formula given in class for large sample confidence limits for the log odds ratio, $\log \psi$ is

$$\log \widehat{\psi} \pm z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}}.$$

This formula, however, was derived under the assumption that n_{11}, n_{21}, n_{12}, n_{22} has a multinomial distribution and $n \to \infty$. Assume that n_{11} and n_{21} have independent binomial distributions:

$$f_1(n_{11}) = \left(\begin{array}{c} n_{1\cdot} \\ n_{11} \end{array} \right) (\pi_{11})^{n_{11}} (1-\pi_{11})^{n_{1\cdot} - n_{11}}, n_{11} = 0, 1, 2, ..., n_{1\cdot}$$

$$f_2(n_{21}) = \left(\begin{array}{c} n_{2\cdot} \\ n_{21} \end{array} \right) (\pi_{21})^{n_{21}} (1-\pi_{21})^{n_{2\cdot} - n_{21}}, n_{21} = 0, 1, 2, ..., n_{2\cdot}$$

The log odds ratio

$$\phi = \ln(\psi) = \ln \left[\frac{\pi_{11}}{1-\pi_{11}} \right] - \ln \left[\frac{\pi_{21}}{1-\pi_{21}} \right]$$

is estimated by

$$\widehat{\phi} = \ln(\widehat{\psi}) = \ln\left[\frac{\widehat{\pi}_{11}}{1-\widehat{\pi}_{11}}\right] - \ln\left[\frac{\widehat{\pi}_{21}}{1-\widehat{\pi}_{21}}\right]$$

which is the difference of two independent random variables.

(a) Use the δ method to get the mean and variance of the approximate normal distribution of $\widehat{\phi}$. Solution: Let

$$\phi(x) = \ln\left(\frac{x}{1-x}\right).$$

Then,

$$\phi'(x) = \frac{1}{x(1-x)}.$$

$$Var[\phi(\widehat{\pi}_{11})] = [\phi'(\widehat{\pi}_{11})]^2 \frac{\pi_{11}(1-\pi_{11})}{n_{1\cdot}} =$$

$$\left[\frac{1}{\pi_{11}(1-\pi_{11})}\right]^2 \frac{\pi_{11}(1-\pi_{11})}{n_{1\cdot}} =$$

$$\frac{1}{n_{1\cdot}\pi_{11}(1-\pi_{11})}.$$

$$Var[\phi(\widehat{\pi}_{21})] = [\phi'(\widehat{\pi}_{21})]^2 \frac{\pi_{21}(1-\pi_{21})}{n_{2\cdot}} =$$

$$\left[\frac{1}{\pi_{21}(1-\pi_{21})}\right]^2 \frac{\pi_{21}(1-\pi_{21})}{n_{2\cdot}} =$$

$$\frac{1}{n_{2\cdot}\pi_{21}(1-\pi_{21})}.$$

Thus,

$$\ln(\widehat{\psi}) = \ln\left[\frac{\widehat{\pi}_{11}}{1-\widehat{\pi}_{11}}\right] - \ln\left[\frac{\widehat{\pi}_{21}}{1-\widehat{\pi}_{21}}\right]$$

is the difference of independent random variables and has a variance

$$\sigma^2 = Var[\ln \widehat{\psi}] = \frac{1}{n_{1\cdot}\pi_{11}(1-\pi_{11})} + \frac{1}{n_{2\cdot}\pi_{21}(1-\pi_{21})}$$

which is estimated by

$$\widehat{\sigma}^2 = \frac{1}{n_{1\cdot}\left(\frac{n_{11}}{n_{1\cdot}}\right)\left(\frac{n_{12}}{n_{1\cdot}}\right)} + \frac{1}{n_{2\cdot}\left(\frac{n_{21}}{n_{2\cdot}}\right)\left(\frac{n_{22}}{n_{2\cdot}}\right)} =$$

$$\frac{n_{1\cdot}}{n_{11}n_{12}} + \frac{n_{2\cdot}}{n_{21}n_{22}} = \frac{(n_{11}+n_{12})}{n_{11}n_{12}} + \frac{(n_{21}+n_{22})}{n_{21}n_{22}} =$$

$$\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}$$

(b) Show that $\log \widehat{\psi}$ still applies in the case of independent binomial samples. Solution: The $100(1-\alpha)\%$ confidence limits for $\log \psi$ are

$$\log \widehat{\psi} \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}}$$

which is the same as before.

9.13 Comparing Two Population Proportions

From now on, we assume binomial sampling in which subjects are randomly allocated to two treatment groups.

		Response S	Response F	
Trt	1	x	$n_1 - x$	n_1
	2	y	$n_2 - y$	n_2

where x is the number of successes among the n_1 subjects allocated to treatment 1, and y is the number of successes among the n_2 subjects allocated to treatment 2. x and y have independent binomial distributions. The marginal distributions are

$$f_1(x) = \left(\begin{array}{c} n_1 \\ x \end{array} \right) \theta_1^x (1-\theta_1)^{n_1-x}, x = 0, 1, 2, ..., n_1.$$

$$f_2(y) = \left(\begin{array}{c} n_2 \\ y \end{array} \right) \theta_2^y (1-\theta_2)^{n_2-y}, y = 0, 1, 2, ..., n_2.$$

9.13.1 Odds Ratio

The odds ratio is

$$\psi = \frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_2}{1-\theta_2}}$$

where the numerator is the odds of success in row 1 and the denominator is the odds of success in row 2.

9.13.2 Log Odds Ratio

The log odds ratio is given by

$$\Delta = \log \psi = \log \frac{\theta_1}{1-\theta_1} - \log \frac{\theta_2}{1-\theta_2}$$

where Δ can be interpreted as a parameter indicating a treatment effect in the sense that

$\Delta = 0 \Rightarrow \theta_1 = \theta_2 \Rightarrow$ No treatment effect.

$\Delta > 0 \Rightarrow \theta_1 > \theta_2 \Rightarrow$ Treatment 1 produces a higher success rate than 2.

$\Delta < 0 \Rightarrow \theta_1 < \theta_2 \Rightarrow$ Treatment 2 produces a higher success rate than 1.

9.13.3 The Likelihood Function

The likelihood function is

$$h(x, y) = \left(\begin{array}{c} n_1 \\ x \end{array} \right) \theta_1^x (1-\theta_1)^{n_1-x} \left(\begin{array}{c} n_2 \\ y \end{array} \right) \theta_2^y (1-\theta_2)^{n_2-y} =$$

$$\left(\begin{array}{c} n_1 \\ x \end{array} \right) \left(\begin{array}{c} n_2 \\ y \end{array} \right) (\frac{\theta_1}{1-\theta_1})^x (\frac{\theta_2}{1-\theta_2})^y (1-\theta_1)^{n_1} (1-\theta_2)^{n_2}.$$

9.13.4 Re parameterization

Re parameterizing the likelihood function can be done as follow. Let $\Delta = \lambda_1 - \lambda_2$ where, by the logistic transform,

$$\lambda_1 = \log \frac{\theta_1}{1-\theta_1} \Rightarrow \frac{\theta_1}{1-\theta_1} = e^{\lambda_1} \text{ and } 1-\theta_1 = \frac{1}{1+e^{\lambda_1}},$$

$$\lambda_2 = \log \frac{\theta_2}{1-\theta_2} \Rightarrow \frac{\theta_2}{1-\theta_2} = e^{\lambda_2} \text{ and } 1-\theta_2 = \frac{1}{1+e^{\lambda_2}}$$

9.13.5 Re parameterizing the Likelihood Function

The joint likelihood function of x, y is

$$h(x,y) = f_1(x)f_2(y) = \frac{\left(\begin{array}{c} n_1 \\ x \end{array}\right)\left(\begin{array}{c} n_2 \\ y \end{array}\right) e^{x\lambda_1 + y\lambda_2}}{(1+e^{\lambda_1})^{n_1}(1+e^{\lambda_2})^{n_2}}.$$

Let $\phi = \lambda_2$ and $\Delta = \lambda_1 - \lambda_2$. Substituting $\lambda_2 = \phi$ and $\lambda_1 = \phi + \Delta$ gives

$$h(x,y) = \frac{\left(\begin{array}{c} n_1 \\ x \end{array}\right)\left(\begin{array}{c} n_2 \\ y \end{array}\right) e^{x\Delta + (x+y)\phi}}{(1+e^{\phi+\Delta})^{n_1}(1+e^{\phi})^{n_2}}.$$

The parameter space $\{(\theta_1, \theta_2) : 0 < \theta_i < 1, i = 1, 2\}$ is mapped one-to-one to the new parameter space $\{(\phi, \Delta) : -\infty < \phi < \infty, -\infty < \Delta < \infty\}$ where Δ is the treatment effect on a logistic scale, and ϕ is a nuisance parameter.

9.13.6 Fisher's Exact Test

Fisher's test is based on the conditional distribution of x given $S = s$ where $S = x + y$. The joint distribution of (x, y) is

$$h(x,y) = \frac{\left(\begin{array}{c} n_1 \\ x \end{array}\right)\left(\begin{array}{c} n_2 \\ y \end{array}\right) e^{x\Delta + (x+y)\phi}}{(1+e^{\phi+\Delta})^{n_1}(1+e^{\phi})^{n_2}}.$$

The marginal distribution of S is

$$g(s) = P(S = s) = \sum_{\{(x,y):x+y=s\}} h(x,y) =$$

$$\sum_{x=a}^{b} \frac{\left(\begin{array}{c} n_1 \\ x \end{array}\right)\left(\begin{array}{c} n_2 \\ s-x \end{array}\right) e^{x\Delta + \phi s}}{(1+e^{\phi+\Delta})^{n_1}(1+e^{\phi})^{n_2}}$$

when s is fixed at its observed value. The range of x is $a \leq x \leq b$ where $a = \max(0, s - n_2)$, and $b = \min(n_1, s)$. To see this, note that $x \leq s$ and $x \leq n_1$ implies that $x \leq \min(s, n_1) = b$. Also, $s - x = y$. So, $y = s - x \leq n_2$ implies that $x \geq s - n_2$ and $x \geq 0$ which implies that $x \geq \max(0, s - n_2) = a$. We have $a \leq x \leq b$ where $a = \max(0, s - n_2)$ and $b = \min(s, n_1)$.

9.13.7 Conditional Distribution of $x|S = s$

$$h(x|s) = \frac{h(x, s-x)}{g(s)}, a \leq x \leq b,$$

$$\frac{\left(\begin{array}{c} n_1 \\ x \end{array} \right) \left(\begin{array}{c} n_2 \\ s-x \end{array} \right) e^{x\Delta}}{\sum_{u=a}^{b} \left(\begin{array}{c} n_1 \\ u \end{array} \right) \left(\begin{array}{c} n_2 \\ s-u \end{array} \right) e^{u\Delta}}, a \leq x \leq b$$

Note that $h(x|s)$ does not depend on ϕ. This distribution is called a *generalized hyper geometric* distribution. When $\Delta = 0$, the denominator is

$$\sum_{u=a}^{b} \left(\begin{array}{c} n_1 \\ u \end{array} \right) \left(\begin{array}{c} n_2 \\ s-u \end{array} \right) = \left(\begin{array}{c} n_1 + n_2 \\ s \end{array} \right), a \leq x \leq b,$$

which follows from a combinatorial identity.

9.13.8 The Null Distribution

Under $H_0 : \Delta = 0$, the distribution of x given $S = s$ is

$$h_0(x|s) = \frac{\left(\begin{array}{c} n_1 \\ x \end{array} \right) \left(\begin{array}{c} n_2 \\ s-x \end{array} \right)}{\left(\begin{array}{c} n_1 + n_2 \\ s \end{array} \right)}, a \leq x \leq b,$$

which is the ordinary hyper geometric distribution. That is, under H_0, x is distributed like the total number of successes occurring in a sample of size s taken *without replacement* from a population containing n_1 objects labeled S and n_2 objects labeled F.

9.13.9 Null Mean and Variance

$$E_{H_0}(x|s) = s\frac{n_1}{n},$$

$$Var_{H_0}(x|s) = \frac{n-s}{n-1}(s)\left(\frac{n_1}{n}\right)\left(\frac{n_2}{n}\right)$$

where $\frac{n-s}{n-1}$ is called the finite population correction factor. Note the similarity and differences between this mean and variance in comparison to the mean and variance of a binomial random variable.

9.13.10 Relationship to a 2×2 Table

Since we will often need to calculate the null mean and variance in the context of a 2×2 table, we now note the following relationship:

$$E_{H_0}(x|s) = s\frac{n_{\cdot 1}}{n},$$

and

$$Var_{H_0}(x|s) = \frac{n-s}{n-1}(s)\left(\frac{n_{\cdot 1}}{n}\right)\left(\frac{n_{\cdot 2}}{n}\right).$$

Then,

$$Var_{H_0}(x|s) = \frac{(n-s)(s)n_{\cdot 1}n_{\cdot 2}}{n^2(n-1)}$$

which is the product of the marginal totals divided by $n^2(n-1)$.

9.13.11 The Likelihood Ratio Test

$H_0 : \Delta = 0$, versus $H_A : \Delta = \Delta_1$. The test statistic is

$$\frac{h(x|s,\Delta_1)}{h_0(x|s)} = \frac{\dfrac{\binom{n_1}{x}\binom{n_2}{s-x}e^{x\Delta_1}}{\sum_{u=a}^{b}\binom{n_1}{u}\binom{n_2}{s-u}e^{u\Delta_1}}}{\dfrac{\binom{n_1}{x}\binom{n_2}{s-x}}{\binom{n_1+n_2}{s}}} =$$

$$c_s(\Delta_1)e^{x\Delta_1}.$$

The ratio increases in x if $\Delta_1 > 0 \Rightarrow$ reject H_0 if $x \geq c_1$. The ratio decreases in x if $\Delta_1 < 0 \Rightarrow$ reject H_0 if $x \leq c_2$. Thus, the alternative hypotheses are

$H_A : \Delta > 0 \Rightarrow$ reject H_0 if $x \geq c_1$.

$H_A : \Delta < 0 \Rightarrow$ reject H_0 if $x \leq c_2$.

$H_A : \Delta \neq 0 \Rightarrow$ reject H_0 if $x \geq c_1$ or if $x \leq c_2$.

9.13.12 Large Sample Conditional Tests

As n increases and s also increases,

$$Z = \frac{X - E_{H_0}(x|s)}{\sqrt{Var_{H_0}(x|s)}}$$

converges to a standard normal distribution.

Example:

		Tumor Regression		
		Yes	No	
Trt	1	14	36	50
	2	8	42	50

Here, $s = 14 + 8 = 22$ where $x = 14$ is the number of successes observed for Fisher's test. Let θ_i be the population proportion of subjects that experience tumor regression under treatment $i, i = 1, 2$. The hypotheses are $H_0 : \theta_1 = \theta_2$ versus $H_A : \theta_1 > \theta_2$. These hypotheses are the same as $H_0 : \Delta = 0$ versus $H_0 : \Delta > 0$ where

$$\Delta = \log\left[\frac{\theta_1(1-\theta_1)}{\theta_2(1-\theta_2)}\right]$$

1. Use the standard test for comparing two population proportions. Solution: The test statistic is

$$Z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Reject H_0 if $Z_{obs} \geq 1.65 = z_\alpha$. So,

$$\hat{\theta}_1 = \frac{14}{50} = 0.28,$$

$$\hat{\theta}_2 = \frac{8}{50} = 0.16,$$

$$\hat{\theta} = \frac{14+8}{50+50} = 0.22,$$

$$Z_{obs} = \frac{0.28 - 0.16}{\sqrt{0.22(0.78)\left(\frac{1}{50} + \frac{1}{50}\right)}} = 1.45,$$

thus, do not reject H_0.

2. Use Fisher's exact test and the normal approximation for comparing two population proportions. Solution: The test statistic is

$$Z = \frac{X - E_{H_0}(x|s)}{\sqrt{Var_{H_0}(x|s)}}.$$

Reject H_0 if $Z_{obs} \geq 1.65 = Z_\alpha$. Conclusion, $x = 14$,

$$E_{H_0}(x|s) = s\left(\frac{n_1}{n}\right) = 22\left(\frac{50}{100}\right) = 11,$$

$$Var_{H_0}(x|s) = \frac{n-s}{n-1}(s)\left(\frac{n_1}{n}\right)\left(\frac{n_2}{n}\right) =$$

$$\frac{100-22}{99}(22)\left(\frac{50}{100}\right)\left(\frac{50}{100}\right),$$

$$\sqrt{Var_{H_0}(x|s)} = 2.08.$$

$Z_{obs} = \frac{14-11}{2.08} = 1.44$; Thus, do not reject H_0.

It would be surprising if these tests did not agree; squaring each test statistic yields one of the chi-square statistics mentioned earlier.

9.13.13 Small Sample Case of Fisher's Exact Test

Small sample yield tests have low power; and therefore, we only briefly discuss this case.

		Tumor Regression		
		Yes	No	
Trt	1	1	4	5
	2	4	1	5

$H_0 : \theta_1 = \theta_2$ versus $H_A : \theta_1 < \theta_2$. Reject H_0 if $x_{obs} \leq c$. $x_{obs} = 1$. The p-value is given by

$$P_{H_0}(x \leq 1|s = 5) = h_0(x|s = 5) = \frac{\left(\begin{array}{c} n_1 \\ x \end{array} \right) \left(\begin{array}{c} n_2 \\ s - x \end{array} \right)}{\left(\begin{array}{c} n_1 + n_2 \\ s \end{array} \right)} =$$

$$\frac{\left(\begin{array}{c} 5 \\ x \end{array} \right) \left(\begin{array}{c} 5 \\ s - x \end{array} \right)}{\left(\begin{array}{c} 10 \\ 5 \end{array} \right)}, x = 0, 1, 2, ..., 5.$$

The the p-value is

$$\sum_{x=0}^{1} h_0(x|s = 5) =$$

$$\frac{\left(\begin{array}{c} 5 \\ 0 \end{array} \right) \left(\begin{array}{c} 5 \\ 5 \end{array} \right)}{\left(\begin{array}{c} 10 \\ 5 \end{array} \right)} + \frac{\left(\begin{array}{c} 5 \\ 1 \end{array} \right) \left(\begin{array}{c} 5 \\ 4 \end{array} \right)}{\left(\begin{array}{c} 10 \\ 5 \end{array} \right)} = \frac{26}{252} = 0.103.$$

Since the p-value is greater than 0.05, do not reject H_0. At *the nominal significance level* of 0.05, a test of $H_0 : \psi = 1$ versus $H_A : \psi < 1$ rejects H_0 only for the (x, y) pairs $w_{0.05} = \{(0, 4), (0, 5), (1, 5)\}$.

9.14 Example of Grouping by an Explanatory Variable

Question: Since the choice of strata (i.e. blocks) is somewhat arbitrary and can be selected in such a way as to affect the conclusion concerning a treatment difference, isn't it always best to use pre-stratification? Answer:

1. Peto, et. al. (1976, 1977) recommend post-stratification in large trials.

2. Our text book (page 69) points out that post-stratification is nearly as efficient (i.e. little loss in power) as pre-stratification.

3. It is often possible to select specific covariates out of a larger set to achieve a desired result (text book, bottom of page 301). For this reason, the process of selecting covariates to be used in an adjustment should be specified in the study protocol and adhered to in the primary analysis (text book, page 302).

Consider the following calculations.

Clinic	Group 1	Group 2	Log Odds Ratio
1	$x_{11} = 4M$	$x_{12} = 8M$	$\widehat{\Delta}_1 = \widehat{\lambda}_{11} - \widehat{\lambda}_{12} = -0.188$
	$n_{11} = 7M$	$n_{12} = 13M$	
	$\widehat{\theta}_{11} = \frac{4M}{7M} = 0.57$	$\widehat{\theta}_{12} = \frac{8M}{13M} = 0.62$	
	$\widehat{\lambda}_{11} = 0.282$	$\widehat{\lambda}_{12} = 0.470$	
2	$x_{21} = 2M$	$x_{22} = 12M$	$\widehat{\Delta}_2 = \widehat{\lambda}_{21} - \widehat{\lambda}_{22} = -0.164$
	$n_{21} = 5M$	$n_{22} = 27M$	
	$\widehat{\theta}_{21} = \frac{2M}{5M} = 0.40$	$\widehat{\theta}_{22} = \frac{12M}{27M} = 0.44$	
	$\widehat{\lambda}_{21} = -0.405$	$\widehat{\lambda}_{22} = -0.241$	

where the log odds ratio is given by $\widehat{\lambda}_{ij} = \log\left[\frac{\widehat{\theta}_{ij}}{1-\widehat{\theta}_{ij}}\right]$. Notes: 1) $\widehat{\theta}_{11} < \widehat{\theta}_{12}$ implies a treatment effect exists. $\widehat{\theta}_{21} < \widehat{\theta}_{22}$ implies that within clinic effect exists. 2) $\widehat{\theta}_1 = \frac{4M+2M}{7M+5M} = 0.50$ and $\widehat{\theta} = \frac{8M+12M}{13M+27M} = 0.50$ implies that if we do not stratify by clinics, we will not be able to detect a treatment difference.

Summary of the Logistic Model for Comparing Proportions in a 2 Factor Experiment

The layout of the data is given in the following table.

Row	Trt 1	Trt 2	
1	x_{11}	x_{12}	$\Delta_1 = \lambda_{11} - \lambda_{12}$
	n_{11}	n_{12}	
	θ_{11}	θ_{12}	
2	x_{21}	x_{22}	$\Delta_2 = \lambda_{21} - \lambda_{22}$
	n_{21}	n_{22}	
	θ_{21}	θ_{22}	
$\vdots k$	x_{k1}	x_{k2}	$\Delta_k = \lambda_{k1} - \lambda_{k2}$
	n_{k1}	n_{k2}	
	θ_{k1}	θ_{k2}	

The general model is x_{ij} independent, binomials, with sample sizes of n_{ij}, and with parameters θ_{ij}. The logistic transform is $\lambda_{ij} = \log\left[\frac{\theta_{ij}}{1-\theta_{ij}}\right]$. The treatment effects are measured by $\Delta_i = \lambda_{i1} - \lambda_{i2}$ which is the log odds ratio for row i. The full model is given by $\lambda_{i2} = \alpha_i$ where α_i is the row effects, and $\lambda_{i1} = \alpha_i + \Delta_i$ where Δ_i is the treatment effects. The model with no row by treatment interaction is $\lambda_{1i} = \alpha_i + \Delta$, and $\lambda_{2i} = \alpha_i$. The row effects α_i are nuisance parameters; the main parameter of interest is Δ. The likelihood function is

$$L(\alpha_1, \alpha_2, ..., \alpha_k, \Delta) = \prod_{i=1}^{k} \frac{\left(\begin{array}{c} n_{i1} \\ x_{i1} \end{array}\right)\left(\begin{array}{c} n_{i2} \\ x_{i2} \end{array}\right) e^{x_{i1}\Delta + \alpha_i s_i}}{(1+e^{\alpha_i+\Delta})^{n_{i1}}(1+e^{\alpha_i})^{n_{i2}}}$$

The sufficient statistic for the likelihood function L is $(W, s_1, s_2, ..., s_k)$ where $W = \sum_{i=1}^{k} x_{i1}$, $s_1 = x_{11} + x_{12}$, $s_2 = x_{21} + x_{22}$, ..., $s_k = x_{k1} + x_{k2}$. The s's are the total number of successes.

The Distribution of W Conditional on $S_1 = s_1, S_2 = s_2, ..., S_k = s_k$

W is distributed conditionally as a sum of independent random variables $x_{i1}, i = 1, 2, ..., k$ each having the following distributions.

$$f_i(x|\Delta, s_i) = P(x_{i1} = x | S_i = s_i) =$$

$$\frac{\left(\begin{array}{c} n_{i1} \\ x \end{array}\right)\left(\begin{array}{c} n_{i2} \\ s_i - x \end{array}\right) e^{\Delta x}}{\sum_{u=a_i}^{b_i}\left(\begin{array}{c} n_{i1} \\ u \end{array}\right)\left(\begin{array}{c} n_{i2} \\ s_i - u \end{array}\right) e^{\Delta x u}}, a_i \leq x \leq b_i,$$

where $a_i = \max(0, s_i - n_{i2})$ and $b_i = \min(s_i, n_{i1})$ We know this from our previous study of Fisher's exact test.

Null Distribution of W Given $S_1 = s_1, S_2 = s_2, ..., S_k = s_k$

Under the null hypothesis $H_0 : \Delta = 0$, W is distributed conditionally as a sum of independent random variables $x_{i1}, i = 1, 2, ..., k$ each having a hyper geometric distribution,

$$f_i(x|0, s_i) = \frac{\left(\begin{array}{c} n_{i1} \\ x \end{array}\right)\left(\begin{array}{c} n_{i2} \\ s_i - x \end{array}\right)}{\left(\begin{array}{c} n_{i1} + n_{i2} \\ s_i \end{array}\right)}, a_i \leq x \leq b_i.$$

Thus, $E_{H_0}(x_{i1}|s_i) = s_i\left(\frac{n_{i1}}{n_i}\right)$ where $n_i = n_{i1} + n_{i2}$.

$$v_i = Var_{H_0}(x_i|s_i) = \frac{n_{i1}n_{i2}s_i(n_i - s_i)}{n_i^2(n_i - 1)}.$$

The Mantel-Hanszel statistic is

$$\chi_1^2 = \frac{[|\sum_{i=1}^k x_{i1} - \sum_{i=1}^k E_{H_0}(x_{i1})| - 0.50]^2}{\sum_{i=1}^k v_i}.$$

Equivalent Form of the Mantel-Hanszel Statistic

Let $w_i = \frac{n_{i1}n_{i2}}{n_{i1}+n_{i2}}, i = 1, 2, ..., k$ $p_{i1} = \frac{x_{i1}}{n_{i1}}, p_{i2} = \frac{x_{i2}}{n_{i2}}, n_i = n_{i1} + n_{i2}, \bar{p}_i = \frac{x_{i1}+x_{i2}}{n_{i1}+n_{i2}}$. $\bar{p}_i$ is a pooled estimate of a common value of θ_{i1} and θ_{i2} under H_0. Then the Mantel-Hanszel statistic is also given by

$$\chi_1^2 = \frac{(|\sum_{i=1}^k w_i(p_{i1} - p_{i2})| - 0.50)^2}{\sum_{i=1}^k \frac{n_i w_i \bar{p}_i(1-\bar{p}_i)}{n_i - 1}}.$$

This form of the statistic is appealing because of its similarity to the statistic used to compare two independent binomial proportions.

Rule of Five

Mantel and Hanszel suggested the following rule for deciding when the sample sizes (or number of groups) is large enough to adequately approximate the null distribution of their χ^2 statistic by the chi-square distribution.

$$\sum_{i=1}^{k} E_i - \sum_{i=1}^{k} L_i > 5$$

and

$$\sum_{i=1}^{k} H_i - \sum_{i=1}^{k} E_i > 5$$

where $E_i = n_{i1}\bar{p}_i, M_i = n_i\bar{p}_i, L_i = \max(0, M_i - n_{i2}), H_i = \min(n_{i1}, M_i)$.

Example: To study a new drug for hypertension (Fleiss, 1986), a total of $n = 41$ patients were available who had recently experienced a stroke. Of these, 16 were given the new drug and the remaining 25 served as a control. The patients were grouped by age. x_{ij} is the number of patients not experiencing a new stroke during a certain recovery period.

Age Strata	1 (Drug)	2 (Control)
1	$x_{11} = 4$	$x_{12} = 0$
	$n_{11} = 4$	$n_{12} = 1$
2	$x_{21} = 7$	$x_{22} = 3$
	$n_{21} = 11$	$n_{22} = 11$
3	$x_{31} = 1$	$x_{32} = 4$
	$n_{31} = 1$	$n_{32} = 13$

The analysis can be done in terms of 2×2 tables.

Strtm	Trt	No Strk	Strk		x_{i1}	M_{i1}	V_i
1	1	4	0	4	4	$\frac{4(4)}{5} = 3.2$	$\frac{4(1)(4)(1)}{5^2(4)} = 0.16$
	2	0	1	1			
		4	1	5			
2	1	7	4	11	7	$\frac{10(11)}{22} = 5.0$	$\frac{10(12)(11)(11)}{(22)^2(21)} = 1.43$
	2	3	8	11			
		10	12	22			
3	1	1	0	1	1	$\frac{5(1)}{14} = 0.36$	$\frac{5(9)(1)(13)}{(14)^2(13)} = 0.23$
	2	4	9	13			
		5	9	14			

$\sum x_{i1} = 4 + 7 + 1 = 12, \sum M_{i1} = 8.56, \sum V_i = 1.82$. The hypotheses are $H_0 : \Delta = 0$, versus $H_1 : \Delta \neq 0$. $\chi^2_{obs} = \frac{(|\sum x_{i1} - \sum M_{i1}| - 0.50)^2}{\sum V_i} = 4.75$, $\alpha = 0.05$. Since $\chi^2_{obs} > 3.841$, reject the null hypothesis. The Mantel-Hanszel statistic is not appropriate if there is evidence of interaction. In particular, the direction of the treatment effect should tend to be constant across strata.

Check of the Adequacy of the χ^2 Approximation

Strm	Trt 1	Trt 2	n_i	$\bar{p}_i$	E_i	M_i	L_i	H_i
1	$x_{11} = 4$	$x_{12} = 0$						
	$n_{11} = 4$	$n_{12} = 1$	5	0.80	3.2	4.0	3	4
2	$x_{21} = 7$	$x_{22} = 3$						
	$n_{21} = 11$	$n_{22} = 11$	22	0.45	4.95	9.9	0	9.9
3	$x_{31} = 1$	$x_{32} = 4$						
	$n_{31} = 1$	$n_{32} = 13$	14	0.36	0.36	5.04	0	1.0

$\sum E_i = 8.51, \sum L_i = 3, \sum E_i - \sum L_i = 8.51 - 3 = 5.51 > 5$. $\sum H_i = 14.9, \sum H_i - \sum E_i = 6.39 > 5$. The sample sizes barely satisfy the rule of five.

9.15 Summary of Peto, et. al. (1976) Part I

Section 3. Numbers of patients required.

- The ability of a trial to distinguish between two treatments depend upon how many patients die rather than on the number of patients entered.
- Clinical trials about the influence of treatment on time to death should rarely be undertaken unless: a) there is some home that the death rate can be halved, or b) the trial will be able to continue until at least 100 patients have died.

Section 4. What treatment schedules should be compared?

- The question to be answered by a clinical trial should be the most important question the investigators can think of.
- A lesser study of an important question is usually of more value than an excellent study of a trivial question.
- Many trials yield null results and it is a mark of a good trial design that a null result, if it occurs, will be of interest.
- If you are trying out a new drug, give the biggest dose of it you safely can so nobody can say, if you get a negative result, that if only you had given more, it would have worked.
- A drug trial is always a trial of the drug in the particular dose and manner given, not a trial of the drug per say.
- A question is more likely to be successfully answered by a clinical trial if it can be answered by comparing just two alternative treatments and no more, those treatments being as markedly different as possible.
- The most common reason for deviations from a treatment schedule is treatment toxicity, necessitating that less than a specified dose be given or that courses of treatment be delayed.
- Specification of treatment schedules should therefore include details of what to do if undue toxicity emerges (i.e. details concerning the flexibility permitted in the treatment schedule).

Section 11. Treatment allocation.

- Balanced randomization at the latest possible time is recommended, with no stratification.
- Balanced randomization means that randomization is performed in such a way that approximately equal numbers of patients would be equally allocated to each of the treatment groups if the trial were to end earlier than scheduled.
- One reason for waiting until the latest possible moment to randomize each patient is so that almost immediately after randomization the patients in different groups will start to receive the different treatments.
- In large trials, there is no need for randomization to be stratified by some prognostic variable.
- Instead, groups of patients can be formed at the analysis stage using those features (e.g. age or disease stage) which are eventually found to be really relevant to prognosis.
- This is called post (or retrospective) stratification.
- The patients within each stratum are then compared with each other and the results combined over different strata to give an overall p-value for the effect of treatment adjusted for the grouping variable.
- The only advantage gained by stratification at entry is that reasonable balance between the numbers on each treatment will automatically be achieved and a wasteful situation where almost all patients happen to get the same treatment is avoided.
- This advantage, however, is an illusion unless the trial is very small.

Section 13. Exclusions, withdrawals, losses. This topic was discussed extensively during the first week of class.

Section 14. When to analyze and publish your results.

1. Early analysis of a trial can be misleading if a temporary difference causes the trial to be aborted so that large numbers of patients never accumulate.

2. Most statistical tests applied to clinical trial data are based on the assumption, usually false, that the decision to stop and publish has been taken independently of the current results.

3. However, it is not uncommon to examine the data say every 6 months, and if there is an apparent difference, a more formal analysis is then undertaken, the trial then stopped and the results published if a positive result is obtained.

4. Suppose the nominal significance level is 0.05 and we look at the data on 5 different occasions to determine whether the results are yet significant at the 0.05 level.

5. Then, the actual significance level is actually about 0.15 rather than the 0.05 that is claimed.

6. For this reason, many published p-values should be doubled or tripled.

7. Simple rule: Avoid any analysis or brief inspection until dozens of deaths have accumulated for it is trials first looked at when very small that are most likely to be misleading.

Section 15. Ethical considerations.

- To avoid having trials grind to a halt before obtaining statistical significance, it may be necessary to keep the treating physicians ignorant of the current state of the treatment comparison and only allow access to the trial results by the steering committee.

- If a developing trend has already been appreciated by the treating physician before his last patient is randomized, how can allocation to the inferior treatment be justified?

- A continuation of this argument suggests that serious consideration of each patient's welfare will lead to policies that prevent any clinical trial from producing a clear answer.

- However, an ethical imperative exists which is frequently ignored that we must, if we can, discover how patients can be treated most effectively; thus policies against randomization are detrimental to the very people they are intended to help.

9.16 Homework and Answers

1. A trial is conducted at two different centers to study the difference in mortality for two treatment groups. A total of 22 subjects at center 1 are randomly and equally allocated to the two treatment groups. Similarly, 20 subjects at center 2 are randomly and equally allocated to the two treatment groups. Use the data on the attached pages to do the following:

   (a) Use the combined set of trial times from both centers to construct the Kaplan-Meier estimate of the survival function for the group receiving treatment 1. Solution:

   The combined set of treatment #1 trial times:

Trial Time	# at Risk	# of Deaths	Interval Probability of Death	Interval Probability of Survival	Survival Probability $\widehat{F}(t)$
2.6	21	1	0.048	0.952	0.95
3.9	20	1	0.050	0.950	0.90
4.3	19	1	0.053	0.947	0.86
4.8	18	1	0.056	0.944	0.81
5.4	17	2	0.118	0.882	0.71
5.4					
6.9	15	1	0.067	0.933	0.67
7.8^+	14	0			
7.9	13	1	0.077	0.923	0.61
8.1^+	12	0			
8.2	11	1	0.091	0.909	0.56
8.3	10	1	0.100	0.900	0.50
10.5^+	9	0			
11.0^+	8	0			
11.2	7	1	0.143	0.857	0.43
12.2	6	1	0.167	0.833	0.36
12.3^+	5	0			
13.8^+	4	0			
14.8	3	1	0.333	0.667	0.24
16.0^+	2	0			
16.2^+	1	0			

The following table contains the combined trial times at Center #1.

Trial		Treatment Group 1		Treatment Group 2	
Time	Trt	# at Risk	# of Deaths	# at Risk	# of Deaths
3.9	1	11	1	11	0
5.4	1	10	1	11	0
6.9	1	9	1	11	0
7.7	2	8	0	11	1
7.8^+	1				
7.9	2	7	1	10	1
7.9	1				
8.2	1	6	1	9	1
8.2	2				
8.3	1	5	1	8	0
10.5	24	0	8	1	
10.5^+	1				
11.0^+	1				
12.2	2	2	0	7	1
12.5	2	2	0	6	1
14.8	1	2	1	5	0
16.0^+	1				
16.6	2	0	0	5	1
16.9^+	2				
17.1^+	2				
18.1^+	2				
19.5^+	2				

The following table contains the combined trial times at Center #2.

Trial		Treatment Group 1		Treatment Group 2	
Time	Trt	# at Risk	# of Deaths	# at Risk	# of Deaths
2.6	1	10	1	10	0
4.3	1	9	1	10	0
4.8	1	8	1	10	0
5.4	1	7	1	10	0
7.7	2	6	0	10	1
7.8	2	6	0	9	1
8.1	2	6	0	8	1
8.1^+	1				
8.2	2	5	0	7	1
10.1	2	5	0	6	1
11.2	1	5	1	5	0
12.2	1	4	1	5	0
12.3^+	1				
13.8^+	1				
14.1	2	1	0	5	1
16.2^+	1				
16.9	2	0	0	4	1
17.3^+	2				
22.1	2				
23.9^+	2				

(b) Use the adjusted (for centers) log rank statistic to test the hypothesis of no difference in survival functions for the two treatment groups. Show how you are making the calculations. Solution:

The following sets of 2×2 tables contain the adjust log rank statistic for the trial times at Center #1.

		D	S		d_{1i}	$e_i = E_{H_0}(d_{1i})$	$v_i = Var_{H_0}(d_{1i})$
$t_9 = 12.2$	1	0	2	2	0	0.22	$\frac{1(8)(2)(7)}{(9)^2(8)} = 0.17$
	2	1	6	7			
		1	8	9			
$t_{10} = 12.5$	1	0	2	2	0	0.25	$\frac{1(7)(2)(6)}{(8)^2(7)} = 0.19$
	2	1	5	6			
		1	7	8			
$t_{11} = 14.8$	1	1	1	2	1	0.29	$\frac{1(6)(2)(5)}{(7)^2(6)} = 0.20$
	2	0	5	5			
		1	6	7			
$t_{12} = 16.6$	1	0	0	0	0	0.00	$\frac{1(4)(0)(5)}{(5)^2(4)} = 0.00$
	2	1	5	6			
		1	4	5			

$$D_1 = \sum d_{1i} = 7,$$

$$E_1 = \sum e_i = 4.94,$$

$$V_1 = \sum v_i = 2.9.$$

The following sets of 2×2 tables contain the adjust log rank statistic for the trial times at Center #2.

		D	S		d_{1i}	$e_i = E_{H_0}(d_{1i})$	$v_i = Var_{H_0}(d_{1i})$
$t_1 = 2.6$	1	1	9	10	1	0.50	$\frac{1(19)(10)(10)}{(20)^2(19)} = 0.25$
	2	0	10	10			
		1	19	20			
$t_2 = 4.3$	1	1	8	9	1	0.47	$\frac{1(18)(9)(10)}{(19)^2(18)} = 0.25$
	2	0	10	10			
		1	18	19			
$t_3 = 4.8$	1	1	7	8	1	0.44	$\frac{1(17)(8)(10)}{(18)^2(17)} = 0.25$
	2	0	10	10			
		1	17	18			
$t_4 = 5.4$	1	1	6	7	1	0.41	$\frac{1(16)(7)(10)}{(16)^2(15)} = 0.24$
	2	0	10	10			
		1	16	17			
$t_5 = 7.7$	1	0	6	6	0	0.38	$\frac{1(15)(6)(10)}{(16)^2(15)} = 0.23$
	2	1	9	10			
		1	15	16			
$t_6 = 7.8$	1	0	6	6	0	0.40	$\frac{1(14)(6)(9)}{(15)^2(14)} = 0.24$
	2	1	8	9			
		1	14	15			
$t_7 = 8.1$	1	0	6	6	0	0.43	$\frac{1(13)(6)(8)}{(14)^2(13)} = 0.24$
	2	1	7	8			
		1	13	14			
$t_8 = 8.2$	1	0	5	5	0	0.42	$\frac{1(11)(5)(7)}{(12)^2(11)} = 0.24$
	2	1	6	7			
		1	11	12			
$t_9 = 10.1$	1	0	5	5	0	0.45	$\frac{1(10)(5)(6)}{(11)^2(10)} = 0.25$
	2	1	5	6			
		1	10	11			
$t_{10} = 11.2$	1	1	4	5	1	0.50	$\frac{1(9)(5)(5)}{(10)^2(9)} = 0.25$
	2	0	5	5			
		1	9	10			

$$D_2 = \sum d_{1i} = 6,$$

$$E_2 = \sum e_i = 5.01,$$

$$V_2 = \sum v_i = 2.83.$$

Then, combining the statistics,

$$D = D_1 + D_2 = 7 + 6 = 13,$$

$$E = E_1 + E_2 = 4.94 + 5.01 = 9.95,$$

$$V = V_1 + V_2 = 2.9 + 2.83 = 5.73.$$

H_0 : the survival distributions at Centers 1 and 2 are equal, versus H_1 : the survival distributions are not equal. The test statistic is

$$\chi^2 = \frac{(|D - E| - 0.50)^2}{V}.$$

Reject H_0 if $\chi^2_{obs} \geq 3.84$.

$$\chi^2_{obs} = \frac{(|13 - 9.95| - 0.50)^2}{5.73} = 1.13.$$

Thus, do not reject H_0.

(c) Use only the data from center 1 to give a 95% confidence limit for the relative hazard rate $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$. Solution: The confidence limits for $\ln \phi$ have the form

$$\frac{D_1 - E_1}{V} \pm \frac{z_{\alpha/2}}{\sqrt{V}}$$

where

$$D_1 = 7, E_1 = 4.94, V = 2.9, z_{\alpha/2} = 1.96.$$

So,

$$\frac{7 - 4.94}{2.9} \pm \frac{1.96}{\sqrt{2.9}}$$

or

$$0.71 \pm 1.15.$$

The lower limit is -0.44 and the upper limit is 1.86. The 95% confidence limits for ϕ are: 1) the lower limit $e^{-0.44} = 0.64$, and 2) the upper limit $e^{1.86} = 0.42$.

2. Consider a maximum duration trial of length $T = 4$ years. Assume the following:

   - All subjects enter the trial at the same point in time.
   - There are no losses to follow-up other than those subjects that are alive when the trial ends.
   - Survival times for the two groups have distributions with proportional hazard rates:

     $$\bar{F}_1(t) = [\bar{F}_2(t)]^{\phi}$$

     where $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$.
   - Subjects are randomized equally to treatment 1 (a new drug) and treatment 2 (the standard drug).

   (a) How many deaths must be observed before the trial ends so a 5% level test of $H_0 : \phi = 1$ versus $H_1 : \phi \neq 1$ has a power of 0.90 at the particular alternative $\phi_1 = 0.60$? Solution:

   $$d = 4\frac{(z_{\alpha/2} + z_{\beta})^2}{[\log \phi_1]^2} =$$

   $$\frac{41.99}{0.2609} = 160.90$$

   (b) Assuming 48% of the patients receiving the standard drug will die during the 4 year period, determine the total number of subjects needed so the 5% level test has a power of 0.90. Solution:

   $$p_2 = 0.48,$$

   $$p_1 = 1 - (1 - p_2)^{\phi} = 1 - (1 - 0.48)^{0.60} = 0.32.$$

   $$p = Q_1 p_1 + Q_2 p_2 = \frac{p_1 + p_2}{2} = \frac{0.32 + 0.48}{2} = 0.40,$$

   $$n = \frac{d}{p} = \frac{160.90}{0.40} = 402.25.$$

   So, $n = 403$.

3. In Exercise # 2, suppose the subjects can be grouped as follows: a) those entering after the starting date, and b) those recruited before the starting date. Let n_A and n_B denote the numbers of subjects in each of these groups. Let us replace assumption (1) in Exercise # 2 by the following:

   - Group B participants are all randomized on the trial starting date and have a maximum exposure period of length of 4.0 years.
   - Group A participants enter the trial at a uniform rate over a period of length $R = 2.0$ years.
   - The survival times of subjects in the two treatment groups have exponential distributions with hazard rates λ_1 and λ_2.

   If only $n_B = 300$ participants have been recruited before the trial begins, how many group A participants must be recruited during an initial 2 year period so the 5% level test in Exercise # 2 has a power of 0.90? Solution:

   $$T = 4.0, R = 2.0, \phi_1 = 0.60, p_2 = 0.48, \alpha = 0.05.$$

   $$n_A p_A + n_B p_B = d$$

   where

   $$n_B = 300, d = 160.9, p_B = 0.40.$$

   $$p_A = \frac{p_1 + p_2}{2}$$

   where

   $$p_i = 1 - \frac{[e^{-\lambda_i(T-R)} - e^{-\lambda_i T}]}{\lambda_i R}, i = 1, 2$$

   $$\lambda_1 = \phi_1 \lambda_2.$$

   $$0.48 = p_2 = F_2(t) = 1 - \bar{F}_2(t) \Rightarrow \bar{F}_2(t) = 0.52 \Rightarrow$$

   $$e^{-\lambda_2 T} = 0.52 \Rightarrow$$

$$\lambda_2 = \frac{-\log 0.52}{T} = \frac{-\log 0.52}{4} \Rightarrow \lambda_2 = 0.1635.$$

$$\lambda_1 = \phi_1 \lambda_2 = 0.60(0.1635) = 0.0981.$$

$$p_1 = 1 - \frac{[e^{-(0.0981)(4-2)} - e^{-(0.0981)(4)}]}{(0.0981)(2)} = 0.2538,$$

$$p_2 = 1 - \frac{[e^{-(0.1635)(4-2)} - e^{-(0.1635)(4)}]}{(0.1635)(2)} = 0.3849.$$

$$p_A = \frac{p_1 + p_2}{2} = \frac{0.2538 + 0.3848}{2} = 0.3194.$$

$$n_A p_A + n_B p_B = d \Rightarrow$$

$$n_A(0.3194) + 300(0.40) = 160.9 \Rightarrow n_A = 128.05.$$

9.17 Comparing Proportions Across Strata

- The subjects in a clinical trial may be grouped at the time of randomization or a the analysis stage.
- When grouping occurs at the analysis stage, it is done to take advantage of explanatory information, which may improve the power of a test for detecting a treatment difference.
- It is appropriate to group the subjects using baseline measurements, but it is not appropriate to group on the basis of outcomes (e.g. level of compliance).

Pre stratification refers to grouping data at the design state. *Post stratification* refers to grouping data after the experiment has been conducted. It's accepted practice to group data by baseline measurements, but not outcomes of the experiment.

Refer to page 31 of the Peto paper for this example. x_{ij} is the number of successes observed for the n_{ij} subjects in row i, and column j. θ_{ij} is the probability of success for the subjects in row i, and column j. Ignore the grouping by clinics. Then,

$$\hat{\theta}_1 = \frac{(4+2)m}{(7+5)m} = \frac{6m}{12m} = 0.5,$$

$$\hat{\theta}_2 = \frac{(8+12)m}{(13+27)m} = \frac{20m}{40m} = 0.5,$$

Cannot tell the difference of the two treatments, ignoring clinics. Also, in the Peto article,

$$\lambda_{11} = \log\left(\frac{\theta_{11}}{1-\theta_{11}}\right) = \log\left(\frac{0.57}{1-0.57}\right) = 0.282$$

Suppose we have [Peto, page 32]

	Treatment	
Row	1	2
1	μ_{11}	μ_{12}
2	μ_{21}	μ_{22}
3	μ_{31}	μ_{32}

Interaction occurs between the different rows (i.e. the $\alpha_i's$ in the handout). $\Delta_1 = \Delta_2 = \cdots = \Delta_k$ in the no interaction concept. $w = \sum_{i=1}^{k} x_{i1} \sim$ hyper geometric distribution.

$$E_{H_0}(x_{i1}) = s_i\left(\frac{n_{i1}}{n_i}\right),$$

$$Var_{H_0}(x_{i1}) = s_i\left(\frac{n_{i1}}{n_i}\right)$$

which are all know numbers. Then standardizing,

$$Z = \frac{\left(\sum_{i=1}^{k} x_{i1}\right) - \sum_{i=1}^{k} E_{H_0}(x_{i1})}{\sqrt{\sum_{i=1}^{k} v_i}}$$

9.18 Censoring

Censoring may occur in a number of ways:

1. The trial terminates before the patient dies.
2. A patient dies from a cause unrelated to the disease being studied.
3. A patient moves away from the clinic.

All *censored* death times are viewed as losses to follow-up. The *survival time* x of a patient is the time from randomization until death occurs. Of course, survival time would not be observed if censoring occurs. If c is the censoring time of a subject, then the subject's trial time is $\min(x, c)$ (which ever one occurs first). It is assumed x and c are independent. In some trials, events other than death my be the primary outcome of interest. For example, the primary outcome my be the occurrence of:

- First stroke.
- Disease recurrence.
- Transplant rejection.
- Etc.

We will refer to death as the primary outcome of interest.

9.18.1 Censored Survival Times

Most trials have a maximum duration that may be defined as the termination date minus the starting date. Usually, patient entry is staggered over a time period of length $R \leq T$. Patients that enter the trial near the end of the recruitment period are more likely to be active when the trial terminates than are patients that enter early. If a patient does not die during the study period, then his exposure time (T) becomes his censoring time.

9.18.2 Survival Function and Hazard Rate

Using the notation, x as the survival time, which is a non-negative random variable. Let $f(x)$ be the pdf of x. $F(t)$ or $F(x)$ is the cdf of x. $\bar{F}(t)$ or $\bar{F}(x)$ is the survival function of x. $\lambda(x)$ or $\lambda(t)$ is the hazard rate. Using the assumption that the distribution of x is continuous with some pdf,

$$\bar{F}(t) = P(x > t) = 1 - P(x \leq t) = 1 - F(t).$$

$$P(x > t) = \int_t^\infty f(x)dx.$$

Two notes:

1. $\bar{F}(t) = 1 - F(t)$.

2. Since x is a non-negative random variable with a distribution of $f(x)$, the pdf $f(x) = 0, \forall x < 0$.

A typical survival function has $F(t)$ plotted on the y-axis, and t plotted on the x-axis where $\bar{F}(t) = P(x > 0) = 1$. The survival function of any continuous distribution must satisfy the following:

1. $\bar{F}(0) = 1$.
2. $\bar{F}(t)$ is non-increasing in $t \geq 0$.
3. $\lim_{t\to\infty} \bar{F}(t) = 0$.

$\bar{F}(t)$ represents the proportion of a large population of patients that survive at least t time units. The *hazard rate* of a distribution with pdf $f(x)$ and a survival function $\bar{F}(t)$ is $\lambda(x) = \frac{f(x)}{\bar{F}(t)}, x \geq 0$. $\lambda(x) \geq 0$ is always true. The hazard rate uniquely determines the distribution of the survival time.

$$\int_0^t \lambda(x)dx = \int_0^t \frac{f(x)}{\bar{F}(x)}dx = \int_0^t \frac{1}{\bar{F}(x)}\partial x$$

because $\frac{dF(x)}{dx} = f(x)$. So, $\partial F(x) = f(x)dx$. Integrate by substitution,

$$u = \bar{F}(x),$$

$$-dF(x) = -du,$$

$$d\bar{F}(x) = -dF(X),$$

and substitute back,

$$\int_1^{\bar{F}(x)} -\frac{1}{u}du,$$

when $x = 0$, and $u = 1$. The new limits of integration are

$$\int_{\bar{F}(x)}^1 \frac{1}{u}du = \ln u\Big|_{\bar{F}(x)}^1 = 0 - \ln \bar{F}(x) = -\ln \bar{F}(x)$$

when $x = t$, and $\bar{F}(x) = u$.

$$\Rightarrow \int_0^t \lambda(x)dx = -\ln \bar{F}(x).$$

Raise to the power of e,

$$e^{-\int_0^t \lambda(x)dx} = \bar{F}(x).$$

What functions $\lambda(x)$ can be hazard rate functions?

$$\bar{F}(\infty) = \lim_{t\to\infty} \bar{F}(x) = 0$$

$$\Rightarrow 0 = e^{-\int_0^\infty \lambda(x)dx} = \frac{1}{e^{\int_0^\infty \lambda(x)dx}}$$

$$\Rightarrow \lim_{t\to\infty} \int_0^t \lambda(x)dx = \infty.$$

Example: $\lambda(x) = \frac{1}{x}, x \geq 0$ cannot be used as a hazard rate function. $\bar{F}(x)$ must have a finite area for all t.

Example:

$$\lambda(x) = \begin{cases} e^{\beta x}, & x \geq 0. \\ 0, & x < 0. \end{cases}$$

Select β first. This function can be used as a hazard rate function.

9.19 The Hazard Rate Function

The *hazard rate function* is defined as

$$\lambda(x) = \frac{f(x)}{\bar{F}(x)},$$

where

$$\bar{F}(x) = P(x > t) = \int_t^\infty f(x)dx = e^{-\int_0^\infty \lambda(x)dx}.$$

$\lambda(x)$ must satisfy the following properties:

1. $\lambda(x)$ must be non-negative.
2. $\lim_{t\to\infty} \int_0^t \lambda(x)dx = \infty$.
3. $\int_0^t \lambda(x)dx$ must be finite for $t \geq 0$.

Conversely, any function $\lambda(x)$ that satisfies (1) thru (3) uniquely determines a *survival function* thru the relation

$$\bar{F}(x) = e^{-\int_0^t \lambda(x)dx}.$$

We want $\lim_{t\to\infty} \bar{F}(x) = 0$, and $\frac{1}{e^{\int_0^t \lambda(x)dx}} \to 0$.

Suppose that $\int_0^t \lambda(x)dx = \infty \Rightarrow \bar{F}(x) = 0$ for some $t > 0 \Rightarrow \bar{F}(x) = 0$ for every $t > 0 \Rightarrow$ no one survives.

Example: $\lambda(x) = \frac{1}{x}, x > 0$. Plot x on the x-axis and $\lambda(x)$ on the y-axis.

$$\int_0^t \lambda(x)dx = \int_0^t \frac{1}{x}dx = \ln x\Big|_0^t = \ln t - \lim_{x\to\infty} \ln x \to \infty \Rightarrow \bar{F}(x) = 0$$

for every $t > 0$.

Example: Let x have an exponential distribution with pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0. \\ 0, & \text{otherwise.} \end{cases}$$

where λ is a positive parameter. Determine (a) the survival function, and (b) the hazard rate. For (a):

$$\bar{F}(t) = \int_t^\infty f(x)dx =$$

$$\begin{cases} 1, & \text{if } t < 0. \\ e^{-\lambda t}, & \text{if } t \geq 0. \end{cases}$$

because,

$$\int_t^\infty f(x)dx = \int_t^\infty \lambda e^{-\lambda x}dx =$$

$$-e^{-\lambda x}\Big|_t^\infty = 0 - (-e^{-\lambda t}) = e^{-\lambda t}, t \geq 0.$$

For part (b):

$$\lambda(x) = \frac{f(x)}{\bar{F}(x)} = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \begin{cases} \lambda, & \text{if } x \geq 0. \\ 0, & \text{if } x > 0. \end{cases}$$

Example: Let x have the pdf of a Weibull distribution $\lambda\beta x^{\beta-1}e^{-\lambda x^{\beta}}, x \geq 0$, where the parameters $\lambda, \beta > 0$.

$$\bar{F}(t) = \int_t^{\infty} f(x)dx = \int_t^{\infty} \lambda\beta x^{\beta-1}e^{-\lambda x^{\beta}}dx,$$

$$\frac{d(\lambda x^{\beta})}{dx} = \lambda\beta x^{\beta-1}.$$

Using integration by parts, let $u = \lambda x^{\beta}$, and $du = \lambda\beta x^{\beta-1}dx$. Then,

$$\int_{\lambda t^{\beta}}^{\infty} e^{-u}du = -e^{-u}\Big|_{\lambda t^{\beta}}^{\infty} = e^{-\lambda t^{\beta}} = e^{-\lambda t^{\beta}}, t \geq 0.$$

To find the hazard function,

$$\lambda(x) = \frac{f(x)}{\bar{F}(x)} = \frac{\lambda\beta x^{\beta-1}e^{-\lambda x^{\beta}}}{e^{-\lambda t^{\beta}}} = \lambda\beta x^{\beta-1}$$

which still depends on whether or not $\beta > 1$ or if $\beta < 1$.

9.20 Estimates for $\bar{F}(t)$

To estimate $\bar{F}(t)$, there are two methods in Chapter 14 of the text book.

1. The life table method.
2. The Kaplan-Meier estimator.

Consider the empirical cdf

$$F_n(t) = \frac{\text{number of } x_i \leq t}{n} \rightarrow^D F(t) = P(x \leq t).$$

Without censoring, the two methods reduce to the theoretical cdf.

9.20.1 Life Table Method

The life table method is one of the oldest methods of estimating a survival function $\bar{F}(t)$.

- Can be used when the data is grouped into intervals and the exact failure and censoring times are not known. For example, a death may only be known to occur during a certain month.
- Requires a fairly large number of observations.

- Assumes that subjects lost to follow-up (i.e. censored) are not exposed to the risk of dying for the entire duration of the grouping interval.
- Adjusts the count of the number of individuals exposed to the risk of dying by the effective number at risk by subtracting the number of losses divided by 2, from the number of individuals alive at the beginning of the interval.

Notation:

- There are $0 = t_0 < t_1 < \cdots < t_k$ boundaries of the grouping intervals.
- d_i is the number of deaths occuring in the interval $[t_{i-1}, t_i)$.
- r_i is the number of individuals alive (at risk) just before interval t_{i-1}.
- l_i is the number of losses (censored observations) during interval $[t_{i-1}, t_i)$.
- r'_i is the effective number of individuals at risk at time $t_{i-1} = r_i - \frac{l_i}{2}$.

Let $p_i = P(x \geq t_i | x \geq t_{i-1})$ and $p_0 = P(x \geq 0) = 1$. Some notes:

1. $$p_i = \frac{\bar{F}(t)}{\bar{F}(t_{i-1})}$$

2. $$\bar{F}(t_j) = \bar{F}(0) \times \frac{\bar{F}(t_1)}{\bar{F}(0)} \times \frac{\bar{F}(t_2)}{\bar{F}(t_1)} \times \cdots \times \frac{\bar{F}(t_j)}{\bar{F}(t_{j-1})} =$$

   $$p_0 \times p_1 \times p_2 \times \cdots \times p_j.$$

3. $$1 - p_i = 1 - \frac{\bar{F}(t_i)}{\bar{F}(t_{i-1})} = \frac{\bar{F}(t_{i-1})}{\bar{F}(t_{i-1})} - \frac{\bar{F}(t_i)}{\bar{F}(t_{i-1})} =$$

   $$\frac{\bar{F}(t_{i-1}) - \bar{F}(t_i)}{\bar{F}(t_{i-1})} = q_i.$$

4. A natural and intuitive estimate of q_i is $q_i = \frac{d_i}{r'_i}$. Thus, the life table estimate of $\bar{F}(t)$ is

   $$\prod_{i=1}^{j} \left[1 - \frac{d_i}{r'_i}\right]$$

(a) If $A = (x \geq t_i)$ and $B = (x \geq t_{i-1})$ the $A \subseteq B$.

(b) $P(A|B) = \frac{P(A \cap B)}{P(B)} =$

$$\frac{P(x \geq t_i)}{P(x \geq t_{i-1})}$$

(c) $$q_i = \frac{P(t_{i-1} \leq x \leq t_i)}{P(x \geq t_{i-1})} = \frac{\text{dying in } [t_{i-1}, t_i]}{\text{surviving past } t_{i-1}}$$

Note that $p_0 = 1$ is always true.

$$\hat{q}_i = \frac{d_i}{r'_i}$$

$$\Rightarrow 1 - p_i = q_i,$$

$$\Rightarrow p_i = 1 - q_i,$$

$$\Rightarrow \hat{\bar{F}}(t_j) = \hat{p}_1 \times \hat{p}_2 \times \cdots \times \hat{p}_j =$$

$$(1 - \hat{q}_1)(1 - \hat{q}_2) \cdots (1 - \hat{q}_j).$$

Since $\hat{p}_i = 1 - \hat{q}_i = 1 - \frac{d_i}{r'_i}$,

$$\Rightarrow \left(1 - \frac{d_1}{r'_1}\right)\left(1 - \frac{d_2}{r'_2}\right) \cdots \left(1 - \frac{d_j}{r'_j}\right) = \hat{\bar{F}}(t_j).$$

Example: The survival and censoring times of $n = 356$ subjects are grouped into one year intervals as shown in the four columns below.

Interval	r_i	d_i	l_i
0-1	356	60	0
1-2	296	47	1
2-3	248	29	5
3-4	214	24	45
4-5	145	11	63
5-6	71	4	57

Note that $q_1 = \frac{60}{356} = 0.1685$. The following table of probabilities is calculated in a similar way.

Interval	r_i'	Death $\hat{q}_i$	Survival $\hat{p}_i$	Survival $\hat{\bar{F}}(t_i)$
0-1	356	0.1685	0.8315	0.8315
1-2	295.5	0.1591	0.8409	0.6992
2-3	245.5	0.1181	0.8819	0.6166
3-4	191.5	0.1253	0.8747	0.5394
4-5	113.5	0.0969	0.9031	0.4871
5-6	42.5	0.0941	0.9059	0.4413

9.21 Proportional Hazard Rate Model

Consider two treatment groups. Let $\bar{F}_i(x)$ be the survival function for subjects in group $i = 1, 2$. Let $\lambda_i(x)$ be the hazard rate for group i. Then,

$$\bar{F}_i(t) = e^{-\int_0^t \lambda_i(x)dx}, i = 1, 2.$$

The proportional hazard rate model is defined by the requirement that

$$\frac{\lambda_1(x)}{\lambda_2(x)} = \phi, \forall x \geq 0, \text{ where } \phi > 0.$$

The model implies that $\lambda_1(x) = \phi\lambda_2(x), \forall x$, which implies that

$$\int_0^t \lambda_1(x)dx = \phi \int_0^t \lambda_2(x)dx$$

which implies that

$$e^{-\int_0^t \lambda_1(x)dx} = e^{-\phi \int_0^t \lambda_2(x)dx} = \bar{F}_1(t) = [\bar{F}_2(t)]^\phi$$

The model is robust in the sense that it does not specify a particular form of $\bar{F}_1(t)$ and $\bar{F}_2(t)$. However, the model does specify how the survival functions are related, namely $\bar{F}_1(t) = [\bar{F}_2(t)]^\phi, \forall t$. Similarly, the model does not specify a particular form of the hazard rates, but does specify how they are related.

Example: Which of the following pairs of hazard rates are proportional?

a. $\lambda_1(x) = \alpha_j\lambda_2(x) = \beta$.

b. $\lambda_1(x) = \alpha x^2; \lambda_2(x) = \beta x^2$.

b. $\lambda_1(x) = \alpha x^2; \lambda_2(x) = \beta x^3$.

Answers (a) and (b) only are true. Answer (c) is not a proportional constant. We say that

1. x is *stochastically* larger than y, if,
$$P(X > t) \geq P(Y > t), \forall t$$
with strict inequality for some t.

2. x is stochastically equal to y, if, X and Y have identical distributions.

3. x is stochastically smaller then y, if,
$$P(X > t) < P(Y > t), \forall t$$
with strict inequality for some t.

Recall that the proportional hazard rate, $\bar{F}_1(t) = [\bar{F}_2(t)]^\phi$, implies the following properties:

1. When $\phi = 1 \Leftrightarrow \bar{F}_1(t) = \bar{F}_2(t), \forall t$ where $x \sim F_1$ and $y \sim F_2$.

2. When $\phi > 1 \Leftrightarrow \bar{F}_1(t) < \bar{F}_2(t), \forall t$.

3. When $\phi < 1 \Leftrightarrow \bar{F}_1(t) > \bar{F}_2(t), \forall t$.

Recall from the literature that $\lambda_i(x)$ are the death rates, $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$ is critical to the comparison of the two treatments, and when $\phi > 1$, it implies that treatment 2 is superior, and when $\phi < 1$, it implies that treatment 1 is superior. The log rank statistic is the same as the Mantel-Hanszel statistic applied to a series of 2×2 tables formed at each of the death times. The notation is as follows. n is the total number of trial participants. n_i is the number of trial participants allocated to treatments $i = 1, 2$. $n = n_1 + n_2$. The procedure for calculating the log rank statistic is as follow:

1. For the combined ordered set of n trials times.

2. Place censored values after the uncensored values if they are tied.

3. At each death time t_i, form the following 2×2 table:

	D	S	
1	d_{1i}	—	r_{1i}
2	d_{2i}	—	r_{2i}
	d_i	$r_i - d_i$	r_i

4. Determine the following: $e_{1i} = E_{H_0}(d_{1i}) = d_i \frac{r_{1i}}{r_i}$, $V_i = Var_{H_0}(d_{1i}) = \frac{d_i(r_i - d_i)r_{1i}r_{2i}}{r_i^2(r_i - 1)}$, $D_1 = \sum d_{1i}$, $E_1 = \sum e_{1i}$, $V = \sum V_i$.

where d_{1i} is the number of deaths in group 1 at time t_i. d_{2i} is the number of deaths in group 2 at time t_{2i}. $d_i = d_{1i} + d_{2i}$. r_{1i} is equal to the number of subjects alive (at risk) in group 1 just prior to time t_i. r_{2i} is equal to the number of subjects alive (at risk) in group 2 just prior to time t_i. The *log rank test* assumes that the proportional hazard rate model is the true model. In this setting, the null and alternative hypotheses are the following: H_o means there is no difference in survival distributions, and H_1 means that the two distributions differ. Alternatively stated: $H_0 : \phi = 1$, $H_1 : \phi \neq 1$.

The test statistic is

$$\chi^2(1) = \frac{(|D_1 - E_1| - 0.50)^2}{V}.$$

Reject H_0 if $\chi^2_{obs} \geq c$. The null distribution is approximately χ^2 with 1 degree of freedom. An example can be found in Peto (1977). See the handout in the next few sections.

9.21.1 The Kaplan-Meier Estimator

The Kaplan-Meier estimator is the preferred method of estimating a survival function from data obtained in a clinical trial. The actual values of the death and censoring times are known. So, there is no need to approximate the number of individuals at risk at the beginning of the intervals. The Kaplan-Meier estimator is also called the *product limit estimator.* We later show that the Kaplan-Meier estimator can be derived as a maximum likelihood estimator. The definition of the Kaplan-Meier estimator is as follow:

1. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered failure (death) times.
2. The set $\{t_1, t_2, \cdots, t_k\}$ does *not* include censoring times.
3. If a failure time is tied with a censoring time, then the censored value must be placed *after* the uncensored value.
4. The failure times $t_1, t_2, \cdots, t_k$ are viewed as defining a series of grouping intervals $[t_i, t_{i+1}], i = 0, 1, ..., k$ where $t_0 = 0, t_{k+1} = \infty$ and t_0 is not a death time.
5. d_i is the number of subjects that die at time t_i.
6. m_i is the number of censored survival times in the interval $[t_i, t_{i-1})$.
7. $t_{i1}, t_{i2}, ..., t_{im_i}$ are the censored survival times in the interval $[t_i, t_{i+1})$.
8. r_i is the number of subjects at risk (still alive) just prior to t_i.
9. $r_i = (d_i + m_i) + (d_{i+1} + m_{i+1}) + \cdots + (d_k + m_k)$ which is the total number of individuals whose death or censoring time is greater or equal to t_i.
10. If $\bar{F}(t)$ is continuous, then all of the $d_i = 1$ with probability equal to 1, that is, ties do not occur.
11. However, ties will occur if survival times are recorded in discrete time unites (e.g. days).
12. Therefore, any estimate of $\bar{F}(t)$ must allow for ties. The *Kaplan-Meier estimator* is

$$\hat{\bar{F}}(t) = \prod_{\{i:t_i \leq t\}} \left(1 - \frac{d_i}{r_i}\right)$$

which is the probability of surviving to time t.

Some notes:

- $\hat{\bar{F}}(t)$ estimates $P(x > t)$.
- The set $\{i : t_i \leq t\}$ is understood to include the case $t_0 = 0$ and the number of deaths is $d_0 = 0$.
- Thus, if $0 \leq t < t_1$, then $\hat{\bar{F}}(t) = 1 - \frac{d_0}{r_0} = 1 - 0 = 1$.

- If $t > t_k$, then

$$\hat{\bar{F}}(t) = \prod_{i=1}^{k} \left(1 - \frac{d_i}{r_i}\right).$$

Thus

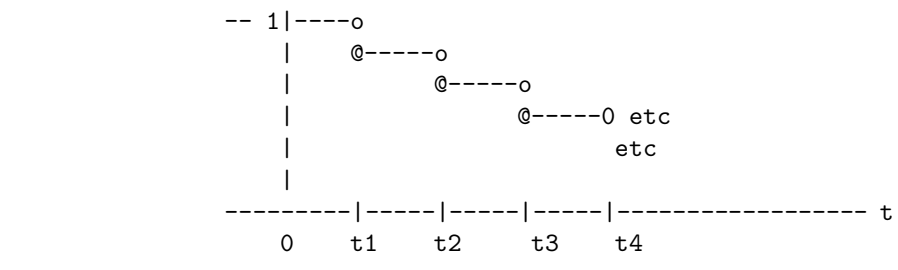
$$\lim_{t\to\infty} \hat{\bar{F}}(t) = \prod_{i=1}^{k} \left(1 - \frac{d_i}{r_i}\right) = 0,$$

which is a constant, iff $d_k = r_k = \{$only one person alive at $t_k\}$, iff $m_k = 0$. This is reasonable because if one or more failure times are censored after time t_k, then there is some evidence of a positive probability of surviving past time t_k.

- $\hat{\bar{F}}(t)$ is a non-increasing, right continuous step-function.

$$\hat{\bar{F}}(t) = \begin{cases} 1, & \text{if } 0 \le t < t_1 \\ 1 - \frac{d_1}{r_1}, & \text{if } t_1 \le t < t_2 \\ \left(1 - \frac{d_1}{r_1}\right)\left(1 - \frac{d_2}{r_2}\right), & \text{if } t_2 \le t < t_3 \\ . & . \\ . & . \\ . & . \\ \prod_{i=1}^{k}\left(1 - \frac{d_i}{r_i}\right) & \text{if } t_k \le t \end{cases}$$

The graph would look like:



Example: The ordered remission durations (in months) and censoring times of $n = 10$ subjects are the following: 3.0, 4.0^+, 5.7^+, 6.5, 6.5, 8.4^+,

10.0, 10.0^+, 12.0, 15.0 where the + represents censored times. The Kaplan-Meier estimator of survival gives the following table:

Remission	r_i	d_i	q_i	p_i	$\hat{\bar{F}}(t)$
3.0	10	1	0.1	0.9	0.9
4.0^+	9	0	0.0	1	0.9
5.7^+	8	0	0.0	1	0.9
6.5	7	2	0.285	0.714	0.643
6.5					0.643
8.4^+	5	0	0.0	1	0.482
10.0	4	1	0.25	0.75	0.482
10.0^+	—	0	0.0	1	0.482
12.0	2	1	0.50	0.50	0.241
15.0	1	1	1.0	0	0.0

There is no need to calculate the Kaplan-Meier estimator at censored survival times because $\hat{\bar{F}}(t)$ is constant over such values.

9.21.2 Examples Illustrating the Kaplan-Meier Estimator and the Log Rank Test

Here's an outline for this section.

- Estimate of survival functions for two treatment groups.
- Log-log plot for checking the proportional hazard rate assumption.
- Kaplan-Meier estimator for the combined set of trial times.
- Use of the log rank test for comparing two treatment groups.
- The adjusted log rank statistic.

9.21.3 Example (Peto, et. al., 1977)

The following example comes from the Peto article for calculating the Kaplin-Meier Estimator for treatment group 1. Note that one line in the table is used for tied ordered trial times.

Ordered Trial Times	Number at Risk r_i	Number of Deaths d_i	Interval of Death $\hat{q}_i$	Probability of Survival $\hat{p}_i$	Survival Function $\bar{\hat{F}}(t_i)$
8	12	2	0.1667	0.8333	0.8333
8					
52	10	1	0.1000	0.9000	0.7500
63	9	2	0.2222	0.7778	0.5833
63					
220	7	1	0.1429	0.8571	0.5000
365^+					
852^+					
1296^+					
1328^+					
1460^+					
1976^+					

The next table shows the calculations for treatment group 2.

Ordered Trial Times	Number at Risk r_i	Number of Deaths d_i	Interval of Death $\hat{q}_i$	Probability of Survival $\hat{p}_i$	Survival Function $\bar{\hat{F}}(t_i)$
13	13	1	0.0769	0.9231	0.9231
18	12	1	0.0833	0.9167	0.8462
23	11	1	0.0909	0.9091	0.7693
70	10	1	0.1000	0.9000	0.6924
76	9	1	0.1111	0.8889	0.6154
180	8	1	0.1250	0.8750	0.5385
195	7	1	0.1429	0.8571	0.4616
210	6	1	0.1667	0.8333	0.3846
632	5	1	0.2000	0.8000	0.3077
700	4	1	0.2500	0.7500	0.2308
1296	3	1	0.3333	0.6667	0.1539
1990^+					
2240^+					

Log-Log Plot for Checking the Proportional Hazard Rate Assumption

Group 1			Group 2		
Ordered Death Times	$\bar{\hat{F}}_1(t_i)$	$-\log-\log\bar{\hat{F}}_1(t_i)$	Ordered Death Times	$\bar{\hat{F}}_2(t_i)$	$-\log-\log\bar{\hat{F}}_2(t_i)$
0	1.0000	undefined	0	1.0000	undefined
8	0.8333	1.70	13	0.9231	2.53
52	0.7500	1.25	18	0.8462	1.79
63	0.5833	0.62	23	0.7693	1.34
220	0.5000	0.37	70	0.6924	1.00
			76	0.6154	0.72
			180	0.5385	0.48
			195	0.4616	0.26
			210	0.3846	0.05
			632	0.3077	-0.16
			700	0.2308	-0.38
			1296	0.1539	-0.63

Plotting $-\log-\log\bar{\hat{F}}(t_i)$ on the y-axis and ordered death times on the x-axis, the two plots do criss-cross each other. Thus, there is a possible violation of the proportional hazard rate assumption.

9.21.4 Example Based on the Combined Data

The following table shows the calculations for the Kaplan-Meier Estimator based on the combined data for groups 1 and 2 of Peto, et. al., 1977.

Ordered Trial Times	Treatment Group	Number at Risk r_i	Number of Deaths d_i	Interval of Death $\hat{q}_i$	Probability of Survival $\hat{p}_i$	Survival Function $\hat{\bar{F}}(t_i)$
8	1	25	2	0.080	0.920	0.920
8	1					
13	2	23	1	0.043	0.957	0.880
18	2	22	1	0.045	0.955	0.840
23	2	21	1	0.048	0.952	0.800
52	1	20	1	0.050	0.950	0.760
63	1	19	2	0.105	0.895	0.680
63	1					
70	2	17	1	0.059	0.941	0.640
76	2	16	1	0.063	0.938	0.600
180	2	15	1	0.067	0.933	0.560
195	2	14	1	0.071	0.929	0.520
210	2	13	1	0.077	0.923	0.480
220	1	12	1	0.083	0.917	0.440
365^+	1	11				
632	2	10	1	0.100	0.900	0.396
700	2	9	1	0.111	0.889	0.352
852^+	1	8				
1296	2	7	1	0.143	0.857	0.302
1296^+	1					
1328^+	1	5				
1460^+	1	4				
1976^+	1	3				
1990^+	2	2				
2240^+	2	1				

9.21.5 Hypothesis Testing

This section shows the calculations using the Log Rank statistic to test the hypothesis of no difference in the survival distributions. The following table contains 15 death times. Thus, there will be 15 2×2 tables.

t_i					d_{1i}	$e_i = E_{H_0}(d_{1i})$	$v_i = Var_{H_0}(d_{1i})$
$t_1 = 8$		D	S		2	$\frac{2(12)}{25} = 0.96$	$\frac{2(23)(12)(13)}{(25)^2(25-1)} = 0.48$
	1	2	10	12			
	2	0	13	13			
		2	33	25			
$t_2 = 13$		D	S		0	$\frac{1(10)}{23} = 0.43$	$\frac{(22)(10)(13)}{(23)^2(23-1)} = 0.25$
	1	0	10	10			
	2	1	12	13			
		1	22	23			
$t_3 = 18$		D	S		0	0.45	0.25
	1	0	10	10			
	2	1	11	12			
		1	21	22			
$t_4 = 23$		D	S		0	0.48	0.25
	1	0	10	10			
	2	1	10	11			
		1	20	21			
$t_5 = 52$		D	S		1	0.50	0.25
	1	1	9	10			
	2	0	10	10			
		1	19	20			
$t_6 = 63$		D	S		2	0.95	0.47
	1	2	7	9			
	2	0	10	10			
		2	17	19			
$t_7 = 70$		D	S		0	0.41	0.24
	1	0	7	7			
	2	1	9	10			
		1	16	17			
$t_8 = 76$		D	S		0	0.44	0.25
	1	0	7	7			
	2	1	8	9			
		1	15	16			
$t_9 = 180$		D	S		0	0.47	0.25
	1	0	7	7			
	2	1	7	8			
		1	14	15			
$t_{10} = 195$		D	S		0	0.50	0.25
	1	0	7	7			
	2	1	6	7			
		1	13	14			
$t_{11} = 210$		D	S		0	0.54	0.25
	1	0	7	7			

Note that:

$$D_1 = \sum d_{1i} = 6,$$

$$E_1 = \sum e_{1i} = 8.34,$$

$$V = \sum v_i = 4.17,$$

$$\chi^2_{obs} = \frac{(|D_1 - E_1| - 0.50)^2}{V} = \frac{(|16 - 8.34| - 0.50)^2}{4.17} = 0.81$$

Peto, et al (1977) give a different value of χ^2. This is primarily because they did not use the continuity correction factor.

9.21.6 Adjusted Log Rank Statistic

The following section covers a test of no significance in survival distributions from the previous section. Use the following legend. 'N = Normal', 'I = Impaired.' The procedure is as follow:

1. Group the trial times by Renal Function.
2. Calculate D_{1i}, E_{1i}, and V_{1i}, for stratum $i = 1, 2$.
3. Determine $D = D_1 + D_2$, $E = E_1 + E_2$, $V = V_1 + V_2$, and $\chi^2 = \frac{(|D-E|-0.50)^2}{V}$ with one degree of freedom.

9.21.7 Adjusted Log Rank Statistic

The survival distribution of a population of patients may depend on some explanatory variable. For example, blood pressure, gender, heart rate, etc may affect a patient's survival time. If patients are grouped by different levels of some explanatory variable, then patients with-in groups may be more homogeneous with respect to the survival time than the group as a whole. Adjusting the log rank statistic by grouping on the basis of an explanatory variable often gives a more sensitive test or a treatment effect. We use the example of Peto, et al to show the effect of grouping.

The $n = 25$ patients were classified at baseline as having normal (N) or impaired (I) renal kidney function. Our previous analysis of this data indicated no difference in survival distributions for the two treatment groups

when testing at $\alpha = 0.05$. This analysis however ignored the possible effect on survival of a good or poor prognosis. What kinds of measurements can be used as a basis for grouping the data and using the adjusted log rank statistic? There is nearly universal agreement that any baseline measure (i.e. pre-randomization) can be used as an explanatory variable. This would include age, gender, good/poor prognosis, high/low blood pressure, etc. Measurements taken after randomization are generally viewed as outcomes and not appropriate as explanatory variables.

Consider the regression model for two treatment groups:

$$y = \alpha + \beta x + \epsilon$$

for the pairs $(x_i, y_i), i = 1, 2, ..., n_1$, and the pairs $(x_i, y_i), i = 1, 2, ..., n_2$. The model can be re-written as

$$y = \beta x + \delta z + \epsilon,$$

for $z = 1$ for group 1, and for $z = 2$ for group 2. The hypotheses of interest are $H_0 : \delta = 0$, versus $H_1 : \delta \neq 0$. The procedure is as follow:

1. Group the trial times by renal function.

2. Calculate D_{1i}, E_{1i}, V_{1i} for strata $i = 1, 2$.

3. Determine $D = D_1 + D_2, E = E_1 + E_2$, and $V = V_1 + V_2$ and $\chi^2 = \frac{(|D-E|-0.50)^2}{V}$ with 1 degree of freedom.

Ordered Trial Times	Treatment Group	Renal Kidney Function
8	1	I
8	1	N
13	2	I
18	2	I
23	2	I
52	1	I
63	1	I
63	1	I
70	2	N
76	2	N
180	2	N
195	2	N
210	2	N
220	1	N
365^+	1	N
632	2	N
700	2	N
852^+	1	N
1296	2	N
1296^+	1	N
1328^+	1	N
1460^+	1	N
1976^+	1	N
1990^+	2	N
2240^+	2	N

Trial Times of the Impaired Group

Ordered Trial Times	Treatment Group	Treatment Group 1		Treatment Group 2	
		No. at Risk	No. Deaths	No. at Risk	No. Deaths
8	1	4	1	3	0
13	2	3	0	3	1
18	2	3	0	2	1
23	2	3	0	1	1
52	1	3	1	0	0
63	1	2	2	0	0
63	1				

t_i		D	S		d_{1i}	$e_i = E_{H_0}(d_{1i})$	$v_i = Var_{H_0}(d_{1i})$
$t_1 = 8$		D	S		1	0.57	$\frac{1(6)(4)(3)}{(17)^2(6)} = 0.24$
	1	1	3	4			
	2	0	3	3			
		1	6	7			
$t_2 = 13$		D	S		0	0.50	$\frac{1(5)(3)(3)}{(6)^2(5)} = 0.25$
	1	0	3	3			
	2	1	2	3			
		1	5	6			
$t_3 = 18$		D	S		0	0.60	$\frac{1(4)(3)(2)}{(5)^2(4)} = 0.24$
	1	0	3	3			
	2	1	1	2			
		1	4	5			
$t_4 = 23$		D	S		0	0.75	$\frac{1(3)(3)(1)}{(4)^2(3)} = 0.19$
	1	0	3	3			
	2	1	0	1			
		1	3	4			
$t_5 = 52$		D	S		1	1	$\frac{1(2)(3)(0)}{(3)^2(2)} = 0$
	1	1	2	3			
	2	0	0	0			
		1	2	3			
$t_6 = 63$		D	S		2	2	$\frac{1(1)(2)(0)}{(2)^2(1)} = 0$
	1	2	0	2			
	2	0	0	0			
		2	0	2			

Then, $D_1 = \sum d_{1i} = 4$, $E_1 = \sum e_i = 5.42$, and $V = \sum v_i = 0.92$.

Trial Times of the Normal Renal Group

Ordered Trial Times	Treatment Group	Treatment Group 1		Treatment Group 2	
		No. at Risk	No. Deaths	No. at Risk	No. Deaths
8	1	8	1	10	0
70	2	7	0	10	1
76	2	7	0	9	1
180	2	7	0	8	1
195	2	7	0	7	1
210	2	7	0	6	1
220	1	7	1	5	0
365^+	1				
632	2	5	0	5	1
700	2	5	0	4	1
852^+	1				
1296^+	2	4	0	3	1
1328^+	1				
1460^+	1				
1976^+	1				
1990^+	2				
2240^+	2				

t_i					d_{1i}	$e_i = E_{H_0}(d_{1i})$	$v_i = Var_{H_0}(d_{1i})$
$t_1 = 8$		D	S		1	0.44	$\frac{1(17)(10)(8))}{(18)^2(17)} = 0.25$
	1	1	7	8			
	2	0	10	10			
		1	17	18			
$t_2 = 70$		D	S		0	0.41	$\frac{1(16)(7)(10)}{(17)^2(16)} = 0.24$
	1	0	7	7			
	2	1	9	10			
		1	16	17			
$t_3 = 76$		D	S		0	0.44	$\frac{1(15)(7)(9)}{(16)^2(15)} = 0.25$
	1	0	7	7			
	2	1	8	9			
		1	15	16			
$t_4 = 180$		D	S		0	0.47	$\frac{1(14)(7)(8)}{(15)^2(14)} = 0.25$
	1	0	7	7			
	2	1	7	8			
		1	14	15			
$t_5 = 195$		D	S		0	0.50	$\frac{1(13)(7)(7)}{(14)^2(13)} = 0.25$
	1	0	7	7			
	2	1	6	7			
		1	13	14			
$t_6 = 210$		D	S		0	0.54	$\frac{1(12)(7)(6)}{(13)^2(12)} = 0.25$
	1	0	7	7			
	2	1	5	6			
		1	12	13			
$t_7 = 220$		D	S		1	0.58	$\frac{1(11)(5)(7)}{(12)^2(11)} = 0.24$
	1	1	6	7			
	2	0	5	5			
		1	11	12			
$t_8 = 632$		D	S		0	0.50	$\frac{1(9)(5)(5)}{(10)^2(9)} = 0.25$
	1	0	5	5			
	2	1	4	5			
		1	9	10			
$t_9 = 700$		D	S		0	0.56	$\frac{1(8)(5)(4)}{(9)^2(8)} = 0.25$
	1	0	5	5			
	2	1	3	4			
		1	8	9			
$t_{10} = 1296$		D	S		0	0.57	$\frac{1(6)(4)(3)}{(7)^2(6)} = 0.24$
	1	0	4	4			
	2	1	2	3			
		1	6	7			

Then, $D_1 = \sum d_{1i} = 2$, $E_1 = \sum e_i = 5.01$, and $V = \sum v_i = 2.47$.

Summary for the Adjusted Log Rank Statistic Example

$D = D_1 + D_2 = 4 + 2 = 6$, $E = E_1 + E_2 = 5.42 + 5.01 = 10.43$, $V = V_1 + V_2 = 0.92 + 2.47 = 3.39$, $\chi^2_{obs} = \frac{(|D-E|-0.50)^2}{V} = 4.55$. For $\alpha = 0.05$, reject H_0 if $\chi^2_{obs} \geq 3.84$. Thus, reject H_0.

9.22 Asymptotic Distribution of the Log Rank Statistic

Let $Z = \frac{D_1 - E_1}{\sqrt{V}}$, where D_1, E_1, and V were defined earlier under the proportional hazard rate model and the assumption that the survival times have a continuous distribution. It has been shown that (Biometrika, Schoenfeld, 1981, p316) that Z has an approximate limiting Normal distribution with a variance of 1. Let the mean,

$$\mu = (\log \phi)\sqrt{Q_1 Q_2 np}$$

where

$$Q_i = \frac{n_i}{n}, i = 1, 2,$$

$$\phi = \frac{\lambda_1(x)}{\lambda_2(x)}, x \geq 0,$$

p is the population proportion of subjects in the combined treatment groups that die before the trial ends which is $Q_1 p_1 + Q_2 p_2$ where p_i is the proportion in group i that die before the trial ends and $n = n_1 + n_2$. The log rank test for one sided alternatives is as follow. Consider the problem of testing H_0 : no difference in survival distributions for groups 1 and 2, versus H_1 : treatment 1 increases survival time relative to treatment 2. We express these hypotheses in terms of $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$ which is the death rate in group 1 divided by the death rate in group 2. So the hypotheses can be restated as $H_0 : \phi = 1$, versus $H_1 : \phi < 1$ for the test statistic $Z = \frac{D_1 - E_1}{\sqrt{V}}$. Since $\log \phi$ gives a negative number, and $E(z) = \log \phi \sqrt{Q_1 Q_2 np}$, then the implied the rejection region is $z_{obs} < z_\alpha$. If H_0 is true then $\phi = 1$ which implies that $\mu = E(z) = 0$, and z has a standard normal distribution.

9.22.1 Confidence Limits for the Relative Hazard Rate Model

The following is implied by the asymptotic result given by Schoenfeld (1981). $z' = z - (\log \phi)\sqrt{V}$ has an approximate standard normal distribution where

$$Z = \frac{D_1 - E_1}{\sqrt{V}}.$$

On page 241 of the text book, the confidence limits for $\log \phi$ can be found. Choose the limit $z_{\alpha/2}$ from the table for the normal distribution so that

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1 - \alpha.$$

$$P\left(-z_{\alpha/2} < z - \log\phi\sqrt{V} < z_{\alpha/2}\right) =$$

$$P\left(-z_{\alpha/2} - z < -\log\phi\sqrt{V} < z_{\alpha/2} - z\right) =$$

$$P\left(\frac{-z_{\alpha/2} - z}{\sqrt{V}} < -\log\phi < \frac{z_{\alpha/2} - z}{\sqrt{V}}\right) =$$

$$P\left(\frac{z - z_{\alpha/2}}{\sqrt{V}} < \log\phi < \frac{z + z_{\alpha/2}}{\sqrt{V}}\right) =$$

On pages 7-9 of Peto, et al, for the unadjusted log rank statistic, $n = 25$, $n_1 = 12, n_2 = 13, D_1 = 6, E_1 = 8.34, V = 4.17$. The 95% confidence interval implies:

$$\frac{z}{\sqrt{V}} \pm \frac{z_{\alpha/2}}{\sqrt{V}} \Rightarrow \frac{D_1 - E_1}{V} \pm \frac{z_{\alpha/2}}{\sqrt{V}}$$

$$\frac{6 - 8.34}{4.17} \pm \frac{1.96}{\sqrt{4.17}} \Rightarrow (-0.56 \pm 0.96)$$

or $-1.52 < \log\phi < 0.40$ for the hazard rate $e^{-1.52} < \phi < e^{0.40}$. For $H_0 : \phi = 1$, versus $H_1 : \phi \neq 1$, $\frac{D_1 - E_1}{V}$ is the point estimate of $\log\phi \Rightarrow \log\widehat{\phi} = \frac{D_1 - E_1}{V}$ but is not usually estimated this way.

9.23 Sample Size and the Power of the Log Rank Test

Consider the problem of testing for no differences in survival distributions, $H_0 : \phi = 1$ and $H_1 : \phi < 1$ where the death rate in group 1 is less than that in group 2. Recall that $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$ and the test statistic is $z = \frac{D_1 - E_1}{\sqrt{V}}$. Reject H_0 if $z_{obs} < -z_\alpha$. Recall that the general sample size-power equation is

$$|\mu_1 - \mu_0| = z_\alpha \sigma_{0n} + z_\beta \sigma_{1n}$$

where z_α is replaced by $z_{\alpha/2}$ for a 2 sided alternative. Recall from Schoenfeld (1981) that z has an approximate normal distribution with a variance equal to 1 as $n \to \infty$, and $\mu = \log \phi \sqrt{Q_1 Q_2 np}$. In Part I of the Peto paper, the sensitivity of the log rank test is similar, but does not give the calculation.

$$\mu_1 = E_{H_1}(z) = \log \phi_1 \sqrt{Q_1 Q_2 np}, \text{ where } Q_i = \frac{n_i}{n},$$

$$\mu_0 = E_{H_0}(z) = 0,$$

$$\sigma_{0n}^2 = Var_{H_0}(z) = 1,$$

$$\sigma_{1n}^2 = Var_{H_1}(z) = 1.$$

Then by substitution, $|\log \phi_1 \sqrt{Q_1 Q_2 np}| = z_\alpha + z_\beta$ where ϕ_1 is a particular alternative hypothesis for which a high statistical power is desired. Let p_i be the proportion of patients dying in group i before their trial time ends. And let $\theta = np$ be the combined expected number of deaths occuring before4 the trial ends. Then, the above equation can be re-written as $|\log \phi_1| \sqrt{Q_1 Q_2 \theta} = z_\alpha + z_\beta$ where

$$\theta = \frac{(z_\alpha + z_\beta)}{(\log \phi_1)^0 Q_1 Q_2}.$$

Assuming equal allocation of subjects to the two treatment groups, we have $Q_1 = Q_2 = \frac{1}{2}$, and $d = \frac{4(z_\alpha + z_\beta)^2}{(\log \phi_1)^2}$. To summarize the literature so far, we need to determine the sample size $n = n_1 + n_2$, needed, so that an α level, for a one sided test has the statistical power of $1 - \beta$ at a particular alternative ϕ_1. We first solve for d, then solve for n by using $n = \frac{d}{p}$. The power of the log rank test depends critically upon the expected number of

deaths ($d = np$). If the probability p that a patient dies before the trial ends is quite small, then n must be quite large so that the log rank test has adequate power. According to Peto (1977), clinical trials where the time of death is of prime interest should rarely be undertaken unless either:

1. There is some hope that the death rate can be halved by the new treatment (i.e. $\phi \leq 0.50$).

2. The trial will be able to continue until at least, 100 patients have died (i.e. $d \geq 100$) which usually requires enrolling well over 100 patients.

9.24 Estimating the Proportion of Deaths Occuring in a Maximum Duration Trial

Case 1: All the subjects enter the trial at the same point in time. Assume that there are no losses due to follow-up other than those subjects who survive past the termination date. The notation is as follow. T is the trial duration. $\bar{F}_i(x)$ is the survival function of subjects in group $p_i, i = 1, 2$. $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$. p is the combined proportion of subjects who die in the interval $(0, T) = Q_1 p_1 + Q_2 p_2$. Then $p_1 = P(x \leq t) = F_1(T) = 1 - \bar{F}_1(T)$. $p_2 = P(x \leq t) = F_2(T) = 1 - \bar{F}_2(T)$. The proportional hazard rate model states that

$$\bar{F}_1(t) = [\bar{F}_2(t)]^{\phi}, \forall t \geq 0$$

$$\Rightarrow p_1 = 1 - \bar{F}_1(T) = 1 - [\bar{F}_2(T)]^{\phi} = 1 - (1 - p_2)^{\phi}.$$

If treatment 2 is a standard treatment (e.g. a placebo), then p_2 can probably be estimated from a previous study or by looking in the literature.

Case 2: All the subjects are staggered entry over a time period of length R. The assumptions are as follow.

1. The entry time u of a subject is uniformly distributed over the interval $(0, R)$. That is, the pdf of U is

$$g(U) = \begin{cases} \frac{1}{R}, & \text{if } 0 \leq u \leq R. \\ 0, & \text{otherwise.} \end{cases}$$

2. The survival time X of subjects in group i has an exponential distribution with cdf

$$F_i(x) = \begin{cases} 1 - e^{-\lambda_i x}, & \text{if } x \geq 0. \\ 0, & \text{if } x < 0. \end{cases}$$

3. The relative hazard rate $F_i(x) = P(X \leq x) \Rightarrow \bar{F}_i(x) = 1 - F_i(x)$ which has the distribution

$$\bar{F}_i(x) = \begin{cases} e^{-\lambda_i x}, & \text{if } x \geq 0. \\ 1, & \text{if } x < 0. \end{cases}$$

$\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$. $\bar{F}_1(x) = [\bar{F}_2(x)]^\phi$. $e^{-\lambda_1 x} = [e^{-\lambda_2 x}]^\phi$. $e^{-\lambda_1 x} = [e^{-\lambda_2 \phi x}], \forall x. \Rightarrow$ $\lambda_1 = \lambda_2 \phi \Rightarrow \phi = \frac{\lambda_1}{\lambda_2}$, a constant in the exponential model. Let X be the survival time and $T-U$ be the patient's exposure time. p_i is the probability that a subject in group i dies during the exposure period.

$$p_i = P(X < T - U) = \int_0^R P(X \leq T - U) g(u) du =$$

$$\int_0^R F_i(T - U) g(u) du = \int_0^R \frac{[1 - e^{-\lambda_i (T-U)}]}{R} du =$$

$$\frac{1}{R} U \Big|_0^R - \frac{e^{-\lambda_i T}}{R} \frac{e^{\lambda_i U}}{\lambda_i} \Big|_0^R =$$

$$1 - \frac{1}{R} \frac{e^{-\lambda_i T}}{\lambda_i} [e^{\lambda_i R} - 1] =$$

$$1 - \frac{[e^{-\lambda_i (T-R)} - e^{-\lambda_i T}]}{R \lambda_i} = p_i$$

9.25 Sample Size and Power of the Log Rank Test

Given that $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$, the hypotheses tests are $H_0 : \phi = 1$, versus $H_0 : \phi < 1$ (the death rate in group 1 is less than the death rate in group 2). The test statistic is $Z = \frac{D_1 - E_1}{\sqrt{V}}$. We reject H_0 if $z_{obs} \leq -z_\alpha$. The sample size power equation is $|\mu_1 - \mu_0| = z_\alpha \sigma_{0n} + z_\beta \sigma_{1n}$. Assuming equal allocation $Q_i = \frac{1}{2}, i = 1, 2$, the expected number of deaths is $\theta = np \Rightarrow n = \frac{\theta}{p}$.

9.25.1 Calculating p

1. Consider the case where all of the subjects enter the trial at the same point in time. Assume no losses due to follow-up other than those that survive past time T. Let p be the combined proportion of patients dying in the time interval $(0, T)$. Then, $p = Q_1 p_1 + Q_2 p_2 = \frac{p_1 + p_2}{2}$ where $p_1 = F_1(t)$, and $p_2 = F_2(t)$. Then under the PHR model $\bar{F}_1(t) = [\bar{F}_2(t)]^\phi \Rightarrow p_1 = (1 - p_2)^\phi$. Using p_2 as the standard treatment, p_2 can be guessed or estimated from a previous study.
   Example: How many deaths must be observed when a trial ends so that a 5% level test of $H_0 : \phi = 1$ versus $H_0 : \phi < 1$ has a power of 0.90 at the particular alternative $\phi_1 = \frac{1}{3}$? Solution: $\theta = \frac{4(z_\alpha + z_\beta)^2}{(\ln \phi_1)^2}, z_\alpha = 1.65, z_\beta = 1.28, \theta = \frac{4(1.65+1.28)^2}{(\ln \frac{1}{3})^2} = 208.87$.

2. Consider the case where the entry into the study over a time period of length R : 1) The entry time U has a uniform distribution over $(0, R)$, 2) the survival times have exponential distributions $\bar{F}_i(x) = e^{-\lambda_i x} \Rightarrow \phi = \frac{\lambda_1}{\lambda_2}$, and $p_i = 1 - \frac{[e^{-\lambda_i (T-R)} - e^{-\lambda_i T}]}{\lambda_i R}, i = 1, 2$.
   Example: In a trial of duration $T = 5$ years, it is estimated that 42% of the subjects receiving the standard treatment will die during the 5 year period. Assuming all subjects enter the trial at the same point in time, then determine the total number of patients n that must be randomized so that a 5% confidence level of a one-sided test has a power of 0.90. Solution: $p_2 = 0.42, \phi_1 = \frac{1}{3}, p_1 = 1 - (1 - p_2)^{\frac{1}{3}} = 1 - (1 - 0.42)^{\frac{1}{3}} = 0.17 \Rightarrow p = \frac{p_1 + p_2}{2} = \frac{0.17 + 0.42}{2} = 0.295 \Rightarrow n = \frac{\theta}{p} = \frac{208.87}{0.295} = 708.03$ or $n = 709$.

3. Consider a trial of duration of 5 years and the following two groups of subjects: Group A contains those subjects entering after the trial starting date, and Group B contains those subjects entering the trial before the starting date. Let the variables n_A be the number of

group A subjects and n_B be the number of group B subjects, and $n = n_A + n_B$. The assumptions are as follow:

(a) None of the subjects are lost to follow-up except those that live past the stopping date.

(b) Group B participants are all randomized on the trial starting date and have a maximum exposure time of 5 years.

(c) Group A participants enter the trial over a uniform rate of 2 years ($R = 2$).

(d) The survival times of subjects in the two groups have exponential distributions with hazard rates λ_1, and λ_2.

Example: If $n_B = 500$ subjects have been recruited before the trial starts, how many eligible subjects must be recruited during the 2 year period so the test in (1) has a power of 0.90? Solution: p_A is the proportion of group A subjects that die before the trial ends. p_B is the proportion of group B subjects that die before the trial ends. $p_B = 0.295$ from Question (2). We know that $n_A p_A + n_B p_B$ is the combined expected number of deaths occuring before the trial ends. Find n_A. $p_A = Q_1 p_1 + Q_2 p_2 = \frac{p_1+p_2}{2}$, $p_i = 1 - \frac{e^{-\lambda_i(T-R)} - e^{-\lambda_i T}}{\lambda_i R}, i = 1, 2$. We know that $R = 2, T = 5, \phi = \frac{1}{3}$. Since, $\phi = \frac{\lambda_1}{\lambda_2} \Rightarrow \lambda_1 = \phi\lambda_2$. $F_2(5) = 0.42 \Rightarrow 1 - F_2(5) = 0.58 \Rightarrow \bar{F}_2(5) = 0.58 \Rightarrow e^{-\lambda_2(5)} = 0.58, -\lambda_2(5) = \ln 0.58, \lambda_2 = 0.1089 \Rightarrow \lambda_1 = \frac{1}{3}(0.1089) = 0.0363$. $p_1 = 1 - \frac{[e^{-(0.0363)(3)} - e^{-(0.0363)(5)}]}{(0.0363)(2)} = 0.1350$. In a similar way, $p_2 = 0.3518$. Then, $p_A = \frac{p_1+p_2}{2} = \frac{0.1350+0.3518}{2} = 0.2434 \Rightarrow n_A p_A + n_B p_B = \theta \Rightarrow n_A(0.2434) + 500(0.295) = 208.87 \Rightarrow n_A = 252.11 \Rightarrow n = n_A + n_B = 253 + 500 = 753$.

9.26 Sequential Monitoring

Reference the technical report and the JAMA article for this section. A beta blocker is one of a number of drugs that can block increased sympathetic stimulation of the heart that occurs during a heart attack. These drugs decrease the oxygen demand of the heart and the susceptibility to cardiac arrythmias. The design and analysis features of the BHAT study are as follow:

- Multi-center (31 centers).

- Randomized.
- Double blinded.
- Placebo controlled trial.
- Adherence to the treatment was monitored.
- Analysis by intent to treat (page 1709 of the article).
- Sequential monitoring at calendar times 11, 16, 21, 28, 34, 40, 48 months. The trial was stopped at 40 months.
- The maximum duration of the trial was $T = 48$ months.
- The recruitment period $R = 27$ months.
- The starting date was June 19, 1978.
- 16,400 patients were recruited and checked for eligibility. Only 23% or 3,837 patients were randomized due to the following reasons: 18% could not take the study treatment, 18% already or were likely to receive the study treatment by prescription, 26% because of study design considerations, competing risks, etc, and 15% did not consent to participation.

9.26.1 Introduction

In many clinical trials, patients enter serially in time and responses to the treatment from the patients also become available serially in time. For scientific and ethical reasons, the results of the trial are reviewed periodically as they become available. Interim monitoring is now required by all trials sponsored by the NIH. Based on such monitoring, early termination of the trial may be recommended if important differences become apparent. An early decision enables switching subjects to the most beneficial treatment as was don in the BHAT study. The *multiple test of significance problem:* suppose we perform p different tests of $H_0 : \delta = 0$, versus $H_1 : \delta \neq 0$ each at the significance level α. If the tests are independent, then, overall the significance level is equal to p the probability of making a Type I error. $P(\text{test 1 rejects } H_0 \text{ or test two rejects } H_0 \text{ or ...}) = 1 - P_{H_0}(\text{all tests accept } H_0) = 1 - (1 - \alpha)^p$. If $\alpha = 0.05$, then overall the significance level is $1 - (0.95)^p$ and note that $(0.95)^p \to 0$ as p increases. Thus, as p increases we become almost certain to discover a treatment effect when none exists. The following table assumes independence.

p	Overall α
2	0.0975
3	0.1427
4	0.1855
5	0.2263

9.26.2 Repeated Tests

The tests in this section are repeated based on accumulated data that are not independent. The following calculations are based on the Lan-DeMets program disk which assumes a particular form of independence. Suppose we test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ and each test has $\alpha = 0.05$. Reject H_0 if $|z_{obs}| \geq 1.96$. Then, this leads to the following table with an increasing p and α.

p	Overall α
2	0.07
3	0.09
4	0.11
5	0.14

9.26.3 Some Limitations

The classical methods were introduced by Wald (1947). They have not been widely used in clinical trials for the following reasons.

1. Open Design — No upper limit on the number of subjects that must be enrolled. Classical methods require that a trial continue until a decision is reached (accepted/rejected H_0).

2. The subjects must be paired; one member of each pair randomized to a new treatment group.

3. A new pair of subjects can be enrolled only after the response variable outcome is known for all previously enrolled pairs. This implies the response to the treatment must be observed over a short time period.

4. The data must be monitored continuously so that a decision can be made before enrolling the next pair.

9.26.4 Layout of the Data

Calendar Time	Treatment 1	Treatment 2	Test Statistic	Accumulated Information
t_1	$n_1(t_1)$	$n_2(t_1)$	$z(t_1)$	$I(t_1)$
t_2	$n_1(t_2)$	$n_2(t_2)$	$z(t_2)$	$I(t_2)$
t_3	$n_1(t_3)$	$n_2(t_3)$	$z(t_3)$	$I(t_3)$
$\vdots$				
t_p	$n_1(t_p)$	$n_2(t_p)$	$z(t_p)$	$I(t_p)$

The information fractions are defined as $\tau(t_1), \tau(t_2), \tau(t_3), ..., \tau(t_p)$, where $\tau(t_p) = 1$. t_p is the time the trial terminates. $n_i(t)$ is the number of patients randomized to treatment i by time $t, i = 1, 2$. $n_i = n_i(t_p)$ and $n = n_1 + n_2$ is the total sample size. $z(t)$ is the standardized form of some test statistic based on the data accumulated by time t. $I(t)$ is the information accumulated by time t. $I = I(t_p)$ is the total information available when the trial is planned to terminate. $\tau(t) = \frac{I(t)}{I}, 0 \leq \tau(t) \leq 1$.

9.26.5 Information Fractions

As the calendar time passes, units of information (i.e. subjects) are collected but not necessarily at the same rate as the passing of the calendar time. For example, suppose that a trial is designed to last 4 years and recruit $n = 1,000$ subjects. If an interim analysis is conducted 2 years after the start of the trial and if a) each subject contributes one unit of information, and b) only 400 subjects have been randomized during the 2 year period, then 50% of the trial duration has passed. 40% of the total information has been collected. As implied by this example,

1. $I(t)$ is a function of the number of subjects that have entered the trial.
2. $I(t)$ is a non-decreasing function in t.
3. $\tau(t) = \frac{I(t)}{I}$ is also a non-decreasing function in t.
4. $I(t)$ is the sum of the information units contributed by individual subjects.

The term *information* also refers to information about some parameter δ contained in an estimator of $\widehat{\delta}$. The information about δ contained in $\widehat{\delta}$ is $I = \frac{1}{Var(\widehat{\delta})}$. Let $z = \frac{\widehat{\delta} - \delta}{\sqrt{Var(\widehat{\delta})}} = \sqrt{I}(\widehat{\delta} - \delta)$.

9.26.6 Sequential Monitoring of Clinical Trials

This section and sub-section considers the cases of comparing means, proportions, and survival function slopes. The notation from the previous section will be used.

Comparing Means

$\tau(t) = \frac{I(t)}{I}, \delta = \mu_1 - \mu_2, \widehat{\delta}(t) = \bar{x}_1(t) - \bar{x}_2(t).$

$$Var(\widehat{\delta}(t)) = \sigma^2 \left[\frac{1}{n_1(t)} + \frac{1}{n_2(t)} \right],$$

$$I(t) = \frac{1}{Var(\widehat{\delta}(t))} = \frac{1}{\sigma^2 \left[\frac{1}{n_1(t)} + \frac{1}{n_2(t)} \right]},$$

$$z(t) = \frac{\bar{x}_1(t) - \bar{x}_2(t) - \delta}{\sigma \sqrt{\left[\frac{1}{n_1(t)} + \frac{1}{n_2(t)} \right]}} = \sqrt{I(t)} \left[\widehat{\delta}(t) - \delta \right].$$

$$I = I(t_p) = \frac{1}{\sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}.$$

$$\tau(t) = \frac{I(t)}{I} = \frac{\left[\frac{1}{n_1} + \frac{1}{n_2} \right]}{\left[\frac{1}{n_1(t)} + \frac{1}{n_2(t)} \right]}.$$

Note that equal allocation to the two groups implies that $n_1 = n_2$ and $n_1(t) = n_2(t), \forall t$. Then,

$$\tau(t) = \frac{\frac{2}{n_1}}{\frac{2}{n_1(t)}} = \frac{2n_1(t)}{2n_1}$$

Note that $\tau(t) = \frac{n_1(t)+n_2(t)}{n_1+n_2}$ is a common formula.

Comparing Proportions

The following notation is used for comparing proportions.

$$\delta = p_1 - p_2.$$

$$\widehat{\delta}(t) = \widehat{p}_1(t) - \widehat{p}_2(t).$$

$$Var(\widehat{\delta}(t)) = \frac{p_1(1-p_1)}{n_1(t)} + \frac{p_2(1-p_2)}{n_2(t)}$$

The hypothesis test of interest is $H_0 : p_1 = p_2 \Rightarrow p_1 = p_2 = \bar{p}$. The test statistic for testing the null hypothesis is

$$z = \frac{\widehat{p}_1(t) - \widehat{p}_2(t) - \delta}{\sqrt{Var_{H_0}(\widehat{\delta}(t))}},$$

where

$$Var_{H_0}(\widehat{\delta}(t)) = \frac{\bar{p}(1-\bar{p})}{n_1(t)} + \frac{\bar{p}(1-\bar{p})}{n_2(t)}.$$

The information function under H_0 accumulated by time t is given by

$$I_0(t) = \frac{1}{Var_{H_0}(\widehat{\delta}(t))} = \frac{1}{\bar{p}(1-\bar{p})\left[\frac{1}{n_1(t)} + \frac{1}{n_2(t)}\right]}.$$

$$\tau(t) = \frac{I_0(t)}{I}.$$

$$I = I(t_p) = \bar{p}(1-\bar{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right].$$

$$\tau(t) = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{\frac{1}{n_1}(t) + \frac{1}{n_2}(t)}.$$

Note that under equal allocation,

$$\tau(t) = \frac{n_1(t) + n_2(t)}{n_1 + n_2}.$$

Comparing Survival Distributions

$$\phi = \frac{\lambda_1(x)}{\lambda_2(x)}, \forall x \geq 0.$$

The test statistic is

$$z(t) = \frac{D_1(t) - E_1(t)}{\sqrt{V(t)}},$$

where D_1, E_1, and V are calculated as if the trial were terminating at time t. Recall that in the limit ($n \to \infty$),

$$Var[z(t)] = 1,$$

and

$$E[z(t)] = \ln(\phi)\sqrt{Q_1(t)Q_2(t)d(t)}$$

where

$$Q_i(t) = \frac{n_i(t)}{n_1(t) + n_2(t)}, i = 1, 2.$$

$$I(t) = Q_1(t)Q_2(t)d(t).$$

Equal allocation implies that

$$Q_1(t) = Q_2(t) = \frac{1}{2} \Rightarrow I(t) = \frac{d(t)}{4}$$

where $d(t)$ is the expected number of deaths in the combined groups by time t. $d(t) = n \times p(t)$ where $p(t)$ is the probability of dying by time t in the combined groups. During the course of a trial, $I(t)$ is estimated by $\widehat{I}(t) = \frac{\text{observed no. deaths by time } t}{4}$.

$$\tau(t) = \frac{I(t)}{I},$$

$$I = \frac{d(t_p)}{4} \Rightarrow \tau(t) = \frac{d(t)}{d(t_p)}.$$

Note that we observe the number of deaths at time t but we may not have an estimate of $d(t_p)$.

Comparing Slopes

$$\delta = \theta_1 - \theta_2.$$

$$\widehat{\delta}(t)) = \widehat{\theta}_1(t) - \widehat{\theta}_2(t)$$

$$z = \frac{\widehat{\delta}(t)) - \delta}{\sqrt{Var[\widehat{\delta}(t)]}},$$

$$Var[\widehat{\delta}(t))] = Var(\widehat{\theta}_1(t)) + Var(\widehat{\theta}_2(t)) =$$

$$\left[\sum_{i=1}^{n_1(t)} V_{1i}^{-1}\right]^{-1} + \left[\sum_{i=1}^{n_2(t)} V_{2i}^{-1}\right]^{-1}$$

where

$$V_{1i} = \sigma_\theta^2 \left[1 + \frac{R}{S_{1i}}\right],$$

and

$$V_{2i} = \sigma_\theta^2 \left[1 + \frac{R}{S_{2i}}\right].$$

Let

$$I_1(t) = \sum_{i=1}^{n_1(t)} V_{1i}^{-1}$$

and

$$I_2(t) = \sum_{i=1}^{n_2(t)} V_{2i}^{-1}$$

Then,

$$Var[\widehat{\delta}(t)] = \frac{1}{I_1(t)} + \frac{1}{I_2(t)}.$$

We know by definition that

$$I(t) = \frac{1}{Var[\widehat{\delta}(t)]}.$$

Let

$$I(t) = \left[\frac{1}{I_1(t)} + \frac{1}{I_2(t)}\right]^{-1}.$$

Then,

$$\frac{1}{I(t)} = \frac{1}{I_1(t)} + \frac{1}{I_2(t)} = Var[\widehat{\delta}(t)]$$

and

$$\tau(t) = \frac{I(t)}{I}$$

where

$$I = \left[\frac{1}{I_1(t_p)} + \frac{1}{I_2(t_p)}\right]^{-1}.$$

9.26.7 Formulation as a Sequential Testing Problem

Let δ be a parameter representing a treatment difference. Let $z(t)$ be the test statistic. At certain selected calendar times, we wish to decide between $H_0 : \delta = 0$ versus $H_1 : \delta > 0$. A decision at time t_i requires a boundary b_i such that $z(t_i) \geq b_i \Rightarrow$ reject H_0. To determine b_i we must specify how much of the Type I error rate we wish to spend at each interim analysis. That is, we must specify $\alpha(t_1), \alpha(t_2), ..., \alpha(t_p)$ where these quantities are defined as follow:

$$\begin{aligned}
\alpha(t_1) = &\quad P_{H_0}(z(t_1) \geq b_1) \\
\alpha(t_2) = &\quad P_{H_0}(z(t_2) \geq b_2 \text{ or } z(t_1) \geq b_1) \\
&\vdots \\
\alpha(t_p) = &\quad P_{H_0}(z(t_p) \geq b_p \text{ or } ... \text{ or } z(t_1) \geq b_1)
\end{aligned}$$

Notes:

1. The '*' in the articles defines a system of p equations with p unknown variables.
2. The overall significance level is $\alpha = \alpha(t_p)$.
3. $\alpha(t_i), i = 1, 2, ..., p$ is a non-decreasing sequence with an upper limit of α.
4. Specifying α alone will not uniquely determine the boundaries.

Two-Tailed Tests

When testing $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, H_0 is rejected at time t_i if $|z(t_i)| \geq b_i$. In this case, then $\alpha(t_1) = P_{H_0}(|z(t_1)| \geq b_1), \alpha(t_2) = P_{H_0}(|z(t_1)| \geq b_1 \text{ or } |z(t_2)| \geq b_2)$, and so on. Also, $z(t_i) \leq -b_i$ or $z(t_i) \geq b_i)$. Thus the lower boundaries can be determined by symmetry from the upper boundaries and by taking α to be $\frac{\alpha}{2}$ for a two-tailed test.

9.27 Comparing Slopes in a Linear Random Effects Model with Repeated Measures

Recall that subject i visits a clinic at times $x_j, j = 1, 2, ..., L_i$. Let $L_{1i}(t)$ be the number of visits by subject i in group 1 by time t. Let $L_{2i}(t)$ be the number of visits by subject i in group 2 by time t.

$$\bar{x}_{1i}(t) = \frac{\sum_{j=1}^{L_{1i}(t)} x_j}{L_{1i}(t)}$$

$$\bar{x}_{2i}(t) = \frac{\sum_{j=1}^{L_{2i}(t)} x_j}{L_{2i}(t)}$$

$$S_{1i}(t) = \sum_{j=1}^{L_{1i}(t)} [x_j - \bar{x}_{1i}(t)]^2$$

$$S_{2i}(t) = \sum_{j=1}^{L_{2i}(t)} [x_j - \bar{x}_{2i}(t)]^2$$

If the follow-up times are equally spaced and there are no missed visits, then these quantities vary from subject to another only because of differences in $L_{1i}(t)$ and $L_{2i}(t)$. Let the slope difference $\delta = \theta_1 - \theta_2$ be estimated by $\widehat{\delta}(t) = \widehat{\theta}_1(t) - \widehat{\theta}_2(t)$ where $\widehat{\theta}_i(t), i = 1, 2$ are estimates based on data observed by time t.

$$z(t) = \frac{\widehat{\delta}(t) - \delta}{\sqrt{Var[\widehat{\delta}(t)]}} = \sqrt{I(t)}[\widehat{\delta}(t) - \delta]$$

where

$$I(t) = \frac{1}{Var[\widehat{\delta}(t)]},$$

$$Var[\widehat{\delta}(t)] = Var[\widehat{\theta}_1(t)] + Var[\widehat{\theta}_2(t)] =$$

$$\frac{1}{\sum_{i=1}^{n_1(t)} V_{1i}^{-1} + \sum_{i=1}^{n_2(t)} V_{2i}^{-1}} = \frac{1}{I_1(t) + I_2(t)}.$$

$n_1(t)$ is the number of group 1 subjects with at least one visit. $n_2(t)$ is the number of group 2 subjects with at least one visit.

$$V_{1i} = \sigma_\theta^2[1 + R/S_{1i}].$$

$$V_{2i} = \sigma_\theta^2[1 + R/S_{2i}].$$

$$R = \frac{\sigma_\epsilon^2}{\sigma_\theta^2}.$$

We have

$$I(t) = \frac{1}{I_1(t) + I_2(t)}$$

where

$$I_1(t) = \sum_{i=1}^{n_1(t)} \frac{1}{\sigma_\theta^2[1 + R/S_{1i}]},$$

$$I_2(t) = \sum_{i=1}^{n_2(t)} \frac{1}{\sigma_\theta^2[1 + R/S_{2i}]}.$$

Let $I = I(t_p)$. In general, $\tau(t) = \frac{I(t)}{I}$ does not have a simple form but it can be estimated from the data accumulated by time t. If all subjects complete all planned visits by time t_p and have the same spacings between visits, then $S_{1i} = S_{2i} = S$ and

$$I_1(t_p) = \frac{n_1}{\sigma_\theta^2[1 + R/S]},$$

$$I_2(t_p) = \frac{n_2}{\sigma_\theta^2[1 + R/S]},$$

and

$$I = \frac{1}{\frac{1}{I_1(t_p)} + \frac{1}{I_2(t_p)}} = \frac{1}{\sigma_\theta^2[1 + R/S]\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

9.28 α Spending Functions

Lan and DeMets (1983) proposed viewing $\alpha(t)$ as a function of the cumulative information fraction $\tau(t) = \frac{I(t)}{I}, 0 \leq \tau(t) \leq 1$. In particular, they proposed the following α spending functions:

1. $\alpha_1(\tau) = 1 - \phi(z_\alpha/\tau), 0 \leq \tau \leq 1$, (Obrien-Flemming)
2. $\alpha_2(\tau) = \alpha \ln[1 + (e-1)/\tau], 0 \leq \tau \leq 1$, (Pocock).
3. $\alpha_3(\tau) = \alpha\tau^p, 0 \leq \tau \leq 1, p > 0$.

Each of these functions increases to α as $\tau \to 1$. The spending functions do not require specifying in advance the number or timing or the interim analyzes. The O'Brien-Flemming bounds are popular because of a low spending rate for small values of α which implies that extreme bounds initially when very little information is available. Chapter 15 of the text book show a graph of τ_i versus α. If we choose α and any one of the spending functions, then $\alpha(t_1), \alpha(t_2), ..., \alpha(t_p)$ are completely determined and it follows that the boundaries $b_1, b_2, ..., b_p$ are going to be uniquely determined. So also are the following quantities:

$$\Pi_1 = \alpha(t_1),$$

$$\Pi_2 = \alpha(t_2) - \alpha(t_1),$$

$$\vdots$$

$$\Pi_p = \alpha(t_p) - \alpha(t_{p-1}).$$

$$\alpha(t_k) = \Pi_1 + \Pi_2 + \cdots + \Pi_k.$$

We have a choice of specifying either the vector $(\Pi_1, \Pi_2, ..., \Pi_p)$ or the vector $(\alpha(t_1), \alpha(t_2), ..., \alpha(t_p)) \Rightarrow \alpha(t) = \alpha[\tau(t)], \alpha(t_p) = \alpha(1) = \alpha$. We claim that

$$\Pi_1 = P_{H_0}(z(t_1) \geq b_1),$$

$$\Pi_2 = P_{H_0}(z(t_1) < b_1 \text{ and } z(t_2) \geq b_2),$$

$$\Pi_3 = P_{H_0}(z(t_1) < b_1 \text{ and } z(t_2) < b_2 \text{ and } z(t_3) \geq b_3),$$

and so on. Proof: Let us consider only the case Π_2. Let A_1 be the event that $z(t_1) \geq b_1$, and A_2 be the event that $z(t_2) \geq b_2$.

$$\alpha(t_2) = P_{H_0}(A_1 \cup A_2),$$

$$\alpha(t_1) = P_{H_0}(A_1),$$

$$\Pi_2 = \alpha(t_2) - \alpha(t_1) = P(A_1 \cup A_2) - P(A_1) =$$

$$P(A_1) + P(A_1' \cap A_2) - P(A_1) = P(A_1' \cap A_2) =$$

$$P(z(t_1) < b_1 \text{ and } z(t_2) \geq b_2),$$

$$\Pi_1 + \Pi_2 + \cdots + \Pi_p = \alpha(t_p) = \alpha.$$

9.28.1 An Approximate Solution to the Sequential Testing Problem

To determine the bounds required, we know the null distribution of $z(t_1), z(t_2), ..., z(t_p)$. The bounds can be determined approximately from the fact that as $n \to \infty$, $z(t_1), z(t_2), ..., z(t_p)$ has an approximate limiting multivariate normal pdf or equivalently that

$$S_i = \sqrt{\tau(t_i)} z(t_i), i = 1, 2, ..., p$$

has a multivariate normal distribution. The limiting normal distribution is related to Brownian motion in the following way.

$$B[\tau(t)] = \sqrt{\tau(t)} z(t).$$

In all of the four examples, we discussed earlier, Brownian motion gives the approximate limiting distribution. The Brownian motion (also called the Weiner process) with a drift parameter θ and a unit variance is a family of random variables $\{B(t), 0 \leq t \leq 1\}$ with the following properties:

1. $B(0) = 0$.
2. For $0 \leq S \leq t \leq 1$, $B(t) - B(S)$ has a normal distribution with a mean $E[B(t) - B(S)] = \theta(t - S)$ and a variance $Var[B(t) - B(S)] = t - S$.

3. $B(t)$ has independent increments that is for $0 = t_0 < t_1 < t_2 < \cdots < t_n = 1$. $B(t_i) - B(t_{i-1}), i = 1, 2, ..., n$ are independent.

4. $B(t)$ must be a continuous function of t.

So, some obvious questions arise. How are sequential boundaries computed? And, do the boundaries require knowing the total information function $I = I(t_p)$? Brownian motion implies $x_1 = S_1, x_2 = S_2 - S_1, ..., x_p = S_p - S_{p-1}$. Then, $B[\tau(t_i)] = \sqrt{\tau(t_i)}z(t_i)$. To calculate the mean, $x_1 = B[\tau(t_1)] = \sqrt{\tau(t_1)}z(t_1)$, $E(x_1) = \theta\tau(t_1)$, $Var(x_1) = \tau(t_1); E(x_2) = \theta[\tau(t_2) - \tau(t_1)]$, $Var(x_2) = \tau(t_2) - \tau(t_1), ..., E(x_i) = \theta[\tau(t_i) - \tau(t_{i-1})], Var(x_i) = \tau(t_i) - \tau(t_{i-1})$. Under $H_0, \delta = 0$.

$$E_{H_0}[z(t)] = 0,$$

$$z(t) = \frac{\bar{x}_1(t) - \bar{x}_2(t) - \delta}{\sigma\sqrt{\frac{1}{n_1(t)} + \frac{1}{n_2(t)}}} \Rightarrow$$

$$E_{H_0}[B[\tau(t)]] = 0$$

which corresponds to $\theta = 0$ for the drift parameter. In general, the bounds must be determined by numerical integration. The following quantities are assumed to be known.

1. $\Pi_1 = \alpha(t_1), \Pi_2 = \alpha(t_2), ...$

2. $\tau_i = \tau(t_i)$. We discussed how the bounds are computed in terms of $S_i = \sqrt{\tau(t_i)}z(t_i)$.

Suppose the exit probabilities at the first two interim analyzes are specified as $\Pi_1 = \Pi_2 = 0.01$. How do we calculate b_1 and b_2? $H_1 : \delta > 0$. Reject H_0 if $z(t_1) \geq b_1$ or $z(t_2) \geq b_2$. $\alpha(t_1) = \Pi_1, \alpha(t_2) - \alpha(t_1) = \Pi_2$. $\alpha(t_1) = 0.01$. $\alpha(t_2) = 0.02$. $\Pi_1 = 0.01 = P_{H_0}(z(t_1) \geq b_1) \Rightarrow b_1 = 2.33$. To find b_2, $\Pi_2 = 0.01 = P_{H_0}(z(t_1) < b_1 \text{ and } z(t_2) \geq b_2)$ where b_1 has already been determined.

$$0.01 = P_{H_0}(\sqrt{\tau(t_1)}z(t_1) < \sqrt{\tau(t_1)}b_1 \text{ and } \sqrt{\tau(t_2)}z(t_2) \geq \sqrt{\tau(t_2)}b_2) =$$

$$P(S_1 < a_1 \text{ and } S_2 \geq a_2)$$

where

$$a_1 = \sqrt{\tau(t_1)}b_1,$$

and

$$a_2 = \sqrt{\tau(t_2)}b_2.$$

Since a_1 is a known number, all we need to do is solve for a_2.

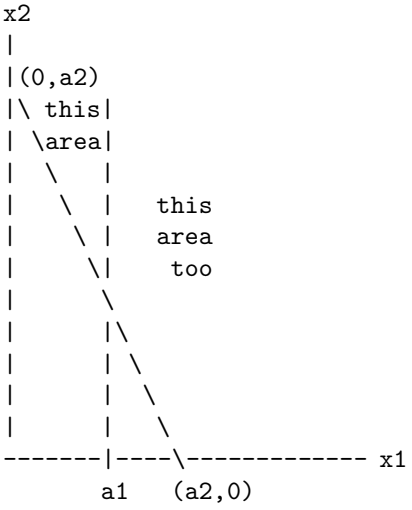
$$0.01 = P_{H_0}(x_1 < a_1 \text{ and } x_1 + x_2 > a_2),$$

$H_0 \Rightarrow E_{H_0}(x_1) = E_{H_0}(x_2) = 0$ and $Var(x_1) = \tau(t_1), Var(x_2) = \tau(t_2) - \tau(t_1)$. The pdf of x_1 is

$$f_1(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}} = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{x_1^2}{\sigma_1^2}}, -\infty < x_1 < \infty.$$

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\frac{x_2^2}{\sigma_2^2}}, -\infty < x_2 < \infty.$$

We need to find the limits of integration from the following graph.



We know that $x_1 + x_2 = a_2$.

$$\int_{-\infty}^{a_1} \int_{a_2 - x_1}^{\infty} f_1(x_1) f_2(x_2) dx_2 dx_1 =$$

$$\int_{-\infty}^{a_1}\int_{a_2}^{\infty} f_1(x_1)f_2(x_2-x_1)dx_2dx_1.$$

Using integration by parts $u = x_2 - x_1$ and $du = dx_2$. Then we have

$$\int_{-\infty}^{a_1}\int_{a_2-x_1}^{\infty} f_1(x_1)f_2(u)dudx_1 =$$

which is identical. Changing the order of integration,

$$\int_{a_2}^{\infty}\int_{-\infty}^{a_1} f_1(x_1)f_2(x_2-x_1)dx_1dx_2 =$$
$$\int_{a_2}^{\infty} h(x_2)dx_2$$

where

$$h(x_2) = \int_{-\infty}^{a_1} f_1(x_1)f_2(x_2-x_1)dx_1.$$

Do the bounds require knowing the total information $I = I(t_p)$? During the course of a trial, the total information $I = I(t_p)$ may not be known. Recall that the information fractions $\tau(t_i) = \frac{I(t_i)}{I}$ depend on I. Thus, if I is not known and interim analyzes are performed at times t_1 and t_2, say, then we cannot compute $\tau(t_1)$ and $\tau(t_2)$. Do the bounds b_1 and b_2 require that we know $\tau(t_1)$ and $\tau(t_2)$? Clearly b_1 does not depend on $\tau(t_1)$ since $\Pi_1 = P_{H_0}(z(t_1) \geq b_1)$ and $\Pi_1 = 0.01 \Rightarrow b_1 = 2.33$. Thus, the value of $\tau(t_1)$ was not used to determine b_1.

Now let us consider whether b_2 depends on knowing $\tau(t_1)$ and $\tau(t_2)$.

$$\Pi_2 = P(z(t_1) < b_1 \text{ and } z(t_2) \geq b_2)$$

where $z(t_1)$ and $z(t_2)$ have standard normal distributions but are not independent. The only way that b_2 can depend on $\tau(t_1)$ and $\tau(t_2)$ is through the covariance between $z(t_1)$ and $z(t_2)$.
THEOREM:

$$Cov(z(t_1), z(t_2)) = \sqrt{\frac{\tau(t_1)}{\tau(t_2)}}$$

which does not depend on I because $\tau(t_1) = \frac{I(t_1)}{I}$ and $\tau(t_2) = \frac{I(t_2)}{I}$. In other words, the I term is divided-out.

9.29 Computing Boundaries and Sample Size Using LAND and GLAN

According to Roboussin, et. al. (1995), the program disk can be used to do the following:

1. Compute boundaries for a given α spending function.
2. Compute probabilities or power for given bounds.
3. Compute confidence limits.

These are options on the program disk. Reference page 16 of the technical report. However, the program disk included with the technical report seems to be an earlier version of the program. It can be used to solve (1) and (2) above, but not (3). The program disk contains two executable files LAND and GLAN. LAND can be used to solve (1) above while GLAN can be used to solve (2).

Example 1: Use the program disk to determine sequential boundaries. The boundaries are determined by the choice of

1. Spending function $\alpha^*(t)$.
2. $\alpha = \alpha^*(1)$.
3. Whether the test is one or two sided.
4. The number p of interim analyzes and the information fractions $\tau_1 < \tau_2 < \cdots < \tau_p$.

The sequential boundaries determined by using the program disk do not depend on the choice of the standardized test statistic $Z(t)$. That is, the same boundaries can be used to compare means, proportions, survival functions, and slopes. The boundaries are computed by using the program LAND and specifying items (1) through (4) listed above. The following print-out shows a LAND session.

```
Is this an interactive session? (1=yes,0=no)
1
interactive =  1
Overall significance level? (>0 and <=1)
```

```
.05
alpha = .050
One(1)- or two(2)-sided symmetric?
1
1-sided test
Use function? (1-5)
(1) OBrien-Fleming type
(2) Pocock type
(3) alpha * t
(4) alpha * t^1.5
(5) alpha * t^2
1
Use function alpha-star 1
Number of interim analyzes in the past (0 if this is the first):
2
2 pervious analyzes.
Enter times (or information fractions > 0 and < 1) in the past:
.6 .8
Previous times (or info fractions) are .6000 .8000
Time or info fraction (<= 1) of the current interim analysis?
1.0
Current time (or information fraction) is 1.000
Do you wish to input the exact or estimated information? (e.g.
number of patients or number of events, as in Lan & DeMets 89?)
(1=yes,0=no)
0
Delta will be taken to be zero.


This program generates one-sided boundaries.
n = 3
alpha =   .050
use function = 1
        Time        Bounds     alpha(i)-alpha(i-1)    cum alpha
        .60     -8.0000 2.2769        .01140            .01140
        .80     -8.0000 1.9591        .01703            .02843
       1.00     -8.0000 1.7387        .02157            .05000
```

Thus our bounds are $b_1 = 2.2769, b_2 = 1.9591, b_3 = 1.7387$.

Example 2: Use the program disk to sequentially analyze a trial. By using α spending functions, an analysis can be conducted at any time dur-

ing the course of a trial while preserving the overall Type I error rate. However, the use of α spending functions requires that we estimate the cumulative information fractions $\tau(t) = \frac{I(t)}{I}$. The total information I may not be known at the time an interim analysis is desired. If bounds are based on an under-estimate of I, the result may be that we overspend α before the trial actually terminates. An example where the value of I was not known is the BHAT trial discussed in the UW Technical Report. To circumvent this problem, it has been proposed (see the Technical Report, page 12) that calendar time be used to specify the α spending rate and that cumulative information (e.g. cumulative numbers of deaths) be used to specify correlation between test statistics. The following table comes from the Beta Blocker Heart Attack trial. See page 23 of the Technical Report.

Calendar Time (Months)	Cumulative Time Fractions	Cumulative No. of Deaths $d(t)$
11	$\frac{11}{48} = 0.229$	56
16	$\frac{16}{48} = 0.250$	77
21	$\frac{21}{48} = 0.438$	126
28	$\frac{28}{48} = 0.583$	177
34	$\frac{34}{48} = 0.708$	247
40	$\frac{40}{48} = 0.833$	318
48	?	?

Example 3: Use the program disk (LAND) to enter two time scales (calendar and information) to determine sequential boundaries for a two-sided test, $\alpha = 0.05$, spending rate $\alpha^* = \alpha\tau$ (i.e. spend α by calendar time, $t_p = 48$ months).

```
Is this an interactive session? (1=yes,0=no)
1
interactive =  1
Overall significance level? (>0 and <=1)
.05
alpha = .050
One(1)- or two(2)-sided symmetric?
2
2-sided test
Use function? (1-5)
(1) OBrien-Fleming type
(2) Pocock type
(3) alpha * t
```

```
(4) alpha * t^1.5
(5) alpha * t^2
3
Use function alpha-star 3
Number of interim analyzes in the past (0 if this is the first):
5
5 pervious analyzes.
Enter times (or information fractions > 0 and < 1) in the past:
.229 .250 .438 .583 .708
Previous times (or info fractions) are .229 .250 .438 .583 .708
Time or info fraction (<= 1) of the current interim analysis?
.833
Current time (or information fraction) is .833
Do you wish to input the exact or estimated information? (e.g.
number of patients or number of events, as in Lan & DeMets 89?)
(1=yes,0=no)
1
Entering information.
Information for past analyzes:
56 77 126 177 247
Previous information 56.000 77.000 126.000 177.000 247.000
Information for current analysis:
318
Current information 318.000

Delta will be taken to be zero.


This program generates two-sided symmetric boundaries.
n = 6
alpha =   .050
use function for the lower boundary = 3
use function for the upper boundary = 3
  Time    Information       Bounds       alpha(i)-alpha(i-1)    cum alpha
  .23         56.00     -2.5287 2.5287        .01145             .01145
  .25         77.00     -2.9598 2.9598        .00105             .01250
  .44        126.00     -2.5011 2.5011        .00940             .02190
  .58        177.00     -2.4826 2.4826        .00725             .02915
  .71        247.00     -2.4988 2.4988        .00625             .03540
  .83        318.00     -2.4607 2.4607        .00625             .04165
```

Example 4: Use the program disk for study design. The power of a sequential test depends upon the timing and frequency of the interim analyzes as well as on the total number of trial participants. Examples of sample size calculations are given in the UW Technical Report (pages 5-9). To determine the sample size, we must specify the following:

1. Whether the test is one-sided or two-sided.
2. The α level and the α spending function.
3. The number of interim analyzes.
4. The information fractions.
5. The desired power.

After specifying (1) through (5), the program disk can be used to calculate the value of the drift parameter θ that gives the desired power. There are 3 ways for calculating the sample size: a) using LAND, enter the information in (1) through (4) to compute the bounds as before, b) use GLAN and the bounds obtained in step (a) to determine the value of the drift parameters θ that gives the power specified in (5); the program actually computes power for specified values of θ; thus, we must repeatedly enter different values of θ until the desired power is attained, and c) use a formula that relates sample size to the drift parameter θ to solve for n; the formula that relates sample size to θ is given in the next table.

Comparison	δ	I	n
Means	$\mu_1 - \mu_2$	$\frac{n}{4\sigma^2}$	$\frac{4\theta^2\sigma^2}{\delta^2}$
Proportions	$p_1 - p_2$	$\frac{n}{4\bar{p}(1-\bar{p})}$	$\frac{4\theta^2\bar{p}(1-\bar{p})}{\delta^2}$
Survival Functions	$\log\left[\frac{\lambda_1(x)}{\lambda_2(x)}\right]$	$\frac{d}{4}$	$n = \frac{d}{p}, d = \frac{4\theta^2}{\delta^2}$
Slopes	$\theta_1 - \theta_2$	$\frac{n}{\sigma_\theta^2[1+R/S]}$	$\frac{4\theta^2\sigma_\theta^2[1+R/S]}{\delta^2}$

Assumption: Subjects are allocated equally to the two groups. p is the probability that a subject in the combined groups dies before the trial ends. Repeated measures: $R = \sigma_\epsilon^2/\sigma_\theta^2$. Assume equal allocation, equally spaced visits, and no missed visits. Then,

$$S = \frac{Dk(k+1)}{12(k-1)},$$

where k is the number of planned visits, and D is the time between the first and last visit.

Example 5: Let the group 1 subjects receive a new drug and the group 2 subjects receive a placebo. A baseline response is observed when each subject is randomized and a final response is observed six months later. Subject accrual ends after 1.5 years. The trial duration is 2.0 years. Interim monitoring is to be done at calendar times 1.0, 1.5, and 2.0 years. The parameters are μ_i is the mean drop in cholesterol level over a 6 month period for subjects in group i, $\delta = \mu_1 - \mu_2$, and σ^2 is the common variance of 6 month changes in cholesterol levels.

a. Determine boundaries for a two-sided test with $\alpha = 0.05$, the O'Brien-Flemming spending function, and three interim analyzes, with information fractions 0.50, 0.75, and 1.00.

b. What sample size $n = n_1 + n_2$ is needed so the 5% level test in (a) has the power equal to 0.90 when $\sigma = 50$ and $\delta = 10$? Determine the value of θ accurate to the nearest hundredth of a unit.

The following interactive session with LAND solves Exercise (5a).

```
Is this an interactive session? (1=yes,0=no)
1
interactive =  1
Overall significance level? (>0 and <=1)
.05
alpha = .050
One(1)- or two(2)-sided symmetric?
2
2-sided test
Use function? (1-5)
(1) OBrien-Fleming type
(2) Pocock type
(3) alpha * t
(4) alpha * t^1.5
(5) alpha * t^2
1
Use function alpha-star 1
Number of interim analyzes in the past (0 if this is the first):
```

```
2
2 pervious analyzes.
Enter times (or information fractions > 0 and < 1) in the past:
.5 .75
Previous times (or info fractions) are .500 .750
Time or info fraction (<= 1) of the current interim analysis?
1.0
Current time (or information fraction) is 1.000
Do you wish to input the exact or estimated information? (e.g.
number of patients or number of events, as in Lan & DeMets 89?)
(1=yes,0=no)
0
Delta will be taken to be zero.


This program generates one-sided boundaries.
n = 3
alpha =   .050
use function for lower boundary = 1
use function for upper boundary = 1
        Time        Bounds     alpha(i)-alpha(i-1)    cum alpha
         .50     -2.9626 2.9626        .00305            .00305
         .75     -2.3590 2.3590        .01625            .01930
        1.00     -2.0140 2.0140        .03070            .0500
```

The upper boundaries are needed in GLAN. The GLAN session for solving Exercise (5b) is given next.

```
GLAN
Is this an interactive session? (1=yes,0=no)
1
interactive = 1
Number of interim analyzes?
3
Times of interim analyzes:
.5 .75 1.0
Analysis times: .5000 .750 1.000
Do you wish to use drift parameter (Delta) other than zero? (1=yes,0=no)
1
Enter non centrality parameter:
3.0
```

```
Delta = 3.
One(1)- or two(2) sided?
2
2-sided test
Symmetric bounds? (1=yes,0=no)
1
Two sided symmetric bounds.
Enter upper bounds in standardized form:
2.9626 2.3590 2.0140
n = 3, delta = 3.000

look    time     lower  upper    alpha(i)-alpha(i-1)    cum alpha
1        .50    -2.9626 2.9626          .20010             .20010
2        .75    -2.3590 2.3590          .39790             .59799
3       1.00    -2.0140 2.0140          .24617             .84417

Do you wish to recompute using a new drift parameter (delta) (1=yes,0=no)?
1

------------------------------
Enter new drift parameter:
3.3
Recomputing with delta = 3.3
n = 3, delta = 3.3000

look    time     lower  upper    alpha(i)-alpha(i-1)    cum alpha
1        .50    -2.9626 2.9626          .26463             .26463
2        .75    -2.3590 2.3590          .42946             .69409
3       1.00    -2.0140 2.0140          .21087             .90496

Do you wish to recompute using a new drift parameter (delta) (1=yes,0=no)?
1

------------------------------
Enter new drift parameter:
3.3
Recomputing with delta = 3.27
n = 3, delta = 3.2700

look    time     lower  upper    alpha(i)-alpha(i-1)    cum alpha
1        .50    -2.9626 2.9626          .25773             .25773
2        .75    -2.3590 2.3590          .42720             .68493
```

```
3       1.00    -2.0140 2.0140          .21488              .89981
```

Note that the last entry in the "cum alpha" column contains the desired power. The drift parameter must be entered using trial-and-error until the desired power is achieved. So, the final run gives a power of 0.89981.

9.30 Designing a Trial with Sequential Monitoring

The power of a sequential test depends on the timing and frequency of the interim analyzes. We now describe how these quantities are related given the following:

1. The α spending rate $\alpha = \alpha(1)$.
2. The number of interim analyzes.
3. The timing (i.e. $\tau(t_1), \tau(t_2), ..., \tau(t_p)$).
4. The desired power.

Then the drift parameter θ is completely determined. We later show that the power is an increasing function of θ.

Relation Between Sample Size and the Drift Parameter

Recall from Brownian motion that $B[\tau(t)] = \sqrt{\tau(t)}z(t)$. In general, $E[z(t)] = \sqrt{I(t)}\delta$.

$$z = \frac{\widehat{\delta}(t) - \delta}{\sqrt{Var(\widehat{\delta}(t))}},$$

$$\widehat{\delta}(t) = \bar{x}_1(t) - \bar{x}_2(t).$$

$$\delta = \mu_1 - \mu_2,$$

$$Var[\widehat{\delta}(t)] = \frac{\sigma_1^2}{n_1(t)} + \frac{\sigma_2^2}{n_2(t)} = \sigma^2 \left[\frac{1}{n_1(t)} + \frac{1}{n_2(t)}\right]$$

with a common variance.

$$I(t) = \frac{1}{Var[\widehat{\delta}(t)]} \Rightarrow$$

$$z(t) = \sqrt{I(t)}\widehat{\delta}(t) \Rightarrow$$

$$E[z(t)] = \sqrt{I(t)}E[\widehat{\delta}(t)] = \sqrt{I(t)}\delta.$$

$$E[B[\tau(t)]] = \theta\tau(t).$$

So,

$$\theta\tau(t) = \sqrt{\tau(t)}\sqrt{I(t)}\delta$$

$$\theta = \frac{\sqrt{I(t)}\delta}{\sqrt{\tau(t)}} = \frac{\sqrt{I(t)}\delta}{\sqrt{\frac{I(t)}{I}}} = \theta = \sqrt{I}\delta.$$

Assuming equal allocation of subjects to the two groups (see page 10 of the handout), then,

$$\theta = \sqrt{I}\delta = \sqrt{\frac{n}{4\sigma^2}}\delta,$$

$$\frac{\theta}{\delta} = \sqrt{\frac{n}{4\sigma^2}} \Rightarrow n = \frac{4\theta^2\sigma^2}{\delta^2}.$$

This is done for each I.

9.31 Homework and Answers

Complete by next Monday.

1. Let x and y be any random variables and a and b be any constants. Show each of the following is true.

   (a) $Cov(ax, by) = abCov(x, y)$. Solution: $Cov(ax, by) = E(abxy) - E(ax)E(by) = ab[E(xy) - E(x)E(y)] = abCov(x, y)$.

   (b) $Cov(x, x + y) = Var(x) + Cov(x, y)$. Solution: $Cov(x, x + y) = E[x(x+y)] - E(x)E(x+y) = E(x^2+xy) - E(x)[E(x)+E(y)] = E(x^2) + E(xy) - [E(x)]^2 - E(x)E(y) = Var(x) + Cov(x, y)$.

2. Let $S_i = \sqrt{\tau(t_i)}$. $Z(t_i), i = 1, 2, ..., p$ where $\tau(t_i)$ are constants with $0 \leq \tau(t_1) < \tau(t_2) < \cdots < \tau(t_p) \leq 1$. Assume that $X_i = S_i - S_{i-1}, i = 1, 2, ..., p$ are independent random variables with means and variances $E(X_i) = 0, i = 1, 2, ..., p$ and $Var(X_i) = \tau(t_i) - \tau(t_{i-1}), i = 1, 2, ..., p$. In particular, $E(X_1) = 0, Var(X_1) = \tau(t_1)$. Show that $Cov[z(t_1), z(t_2)] = \sqrt{\frac{\tau(t_1)}{\tau(t_2)}}$. Hint, use the results in Exercise 1. Solution:

$$Cov(z(t_1), z(t_2)) = Cov\left[\frac{1}{\sqrt{\tau(t_1)}}S_1, \frac{1}{\sqrt{\tau(t_2)}}S_2\right] =$$

$$\frac{1}{\sqrt{\tau(t_1)\tau(t_2)}}Cov(x_1, x_1 + x_2) =$$

$$Var(x_1) + Cov(x_1, x_2) = Var(x_1)$$

because x_1 and x_2 are independent. Thus,

$$Cov(z(t_1), z(t_2)) = \frac{1}{\sqrt{\tau(t_1)\tau(t_2)}}Var(x_1) =$$

$$\frac{\tau(t_1)}{\sqrt{\tau(t_1)\tau(t_2)}} = \sqrt{\frac{\tau(t_1)}{\tau(t_2)}}$$

3. Consider a one sample clinical study to evaluate a new treatment. The standard treatment has success probability 0.50. The study is conducted to test $H_0 : \Pi = 0.50$, versus $H_1 : \Pi > 0.50$. Let n denote the total number of subjects that are planned to enter the study. For the i-th subject, let $x_i = 1$ if the treatment is a success and $x_i = -1$ otherwise. Then, $x_1, x_2, ...$ are iid random variables with the following discrete distribution.

x	-1	1
$f(x)$	$1 - \Pi$	Π

Let the total planned sample size be $n = 1,000$. The entry of subjects is staggered over a 2 year period with the trial terminating after 3

years. Let $n(t)$ denote the number of responses observed by calendar time t. Let $\widehat{\delta}(t) = \sum_{i=1}^{n(t)} x_i/n(t)$. Determine

(a) $E(\widehat{\delta}(t))$. Solution:

$$E(x) = 1(\pi) + (-1)(1-\pi) = 2\pi - 1.$$

$$Var(x) = E(x^2) - [E(x)]^2 = 1 - (2\pi - 1)^2 = 4\pi(1-\pi)$$

since

$$E(x^2) = \pi + 1 - \pi = 1.$$

So,

$$E[\widehat{\delta}(t)] = E(x) = 2\pi - 1.$$

(b) $Var_{H_0}(\widehat{\delta}(t))$. Solution:

$$Var_{H_0}(\widehat{\delta}(t)) = \frac{Var_{H_0}(\widehat{\delta}(x))}{n(t)} = \frac{4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}{n(t)} = \frac{1}{n(t)}.$$

(c) $I_0(t) = [Var_{H_0}(\widehat{\delta}(t))]^{-1}$. Solution:

$$I_0(t) = n(t).$$

(d) $\tau(t) = \frac{I_0(t)}{I}$ where I is the information obtained when the responses of all $n = 1,000$ subjects are observed. Solution:

$$I = I_0(t_p) = n,$$

$$\tau(t) = \frac{I_0(t)}{I} = \frac{n(t)}{n}.$$

4. Let group 1 be the subjects that are given a new treatment, and group 2 be the subjects that are given a standard treatment. Each subjects' response (success or failure) can only be observed 6 months after the initiation of treatment, which begins the moment a subject is randomized. The parameters: p_i is the success probability for subjects in group i. $n_i(t)$ is the number of observations accumulated on subjects in group i by time t. $n = n_1 + n_2$ where n_i is the total number of observations accumulated on subjects in group i when the trial terminates. Assumptions:

(a) Approximately equal allocation is maintained during the course of the trial (i.e. $n_1(t) = n_2(t)$ for all t).

(b) Subjects are randomized every 6 months in groups of equal size m until the end of a 2 year period. Thus, $n = 5m$, and the trial duration is 2.5 years.

(c) The standard treatment is known to have success probability $p_2 = 0.20$. It is anticipated that the new treatment will have a success rate $p_1 = 0.30$.

(a) Determine the boundaries needed to monitor the trial results at 6, 12, 18, 24 and 30 months after the trial starting date. The boundaries are to be determined for a one-sided test of $H_0 : \delta = 0$, versus $H_1 : \delta > 0$ with $\alpha = 0.05$, $\delta = p_1 - p_2$, and when using the O'Brien Flemming spending rate function. The last group of subjects is randomized at 24 months but their response to treatment can not be observed until 30 months after the trial begins. You must first determine $\tau(t)$ at $t = 6, 12, ...$ under assumptions (1) and (2) from the formula

$$\tau(t) = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{\frac{1}{n_1(t)} + \frac{1}{n_2(t)}}.$$

Solution:

$$\tau(t) = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{\frac{1}{n_1(t)} + \frac{1}{n_2(t)}}.$$

So, if equal allocation is maintained approximately at all times, then,

$$n_1 = n_2, n_1(t) = n_2(t), \tau(t) = \frac{\frac{2}{n_1}}{\frac{2}{n_1(t)}} \Rightarrow$$

$$\tau(t) = \frac{n_1(t)}{n_1} = \frac{n_1(t) + n_2(t)}{n_1 + n_2},$$

$$n = n_1 + n_2 = 5M, \tau(t_i) = \frac{iM}{5M}, i = 1, 2, 3, 4, 5.$$

Calendar Time	$\tau(t_i)$	Boundaries
6	$\frac{M}{5M} = 0.20$	4.23
12	$\frac{2M}{5M} = 0.40$	2.89
18	$\frac{3M}{5M} = 0.60$	2.30
24	$\frac{4M}{5M} = 0.80$	1.96
30	$\frac{5M}{5M} = 1.00$	1.74

(b) Use the boundaries obtained in Part (a) and the program disk to determine the value of the drift parameter so the test in Part (a) has a power of 0.90. Determine the value of θ accurate to the nearest hundredth of a unit. Solution:

θ	Power
3.00	0.90448
2.97	0.89933
2.98	0.90107

So, choose $\theta = 2.97$ since it is closest to 0.90.

(c) What sample size n is needed so the one-tailed test in Part (a) has a power of 0.90? When $p_1 = 0.30$ and $p_2 = 0.20$? Solution:

$$n = \frac{4\theta^2 \bar{p}(1-\bar{p})}{\delta^2},$$

where

$$\bar{p} = \frac{p_1 + p_2}{2} = \frac{0.30 + 0.20}{2} = 0.25.$$

$$\delta = p_1 - p_2 = 0.30 - 0.20 = 0.10.$$

Then,

$$n = \frac{4(2.97)^2(0.25)(0.75)}{(0.10)^2} = 661.57$$

9.32 Final Exam Review

Our final exam is scheduled for Monday, May 5, 7-10pm, BAL 408. The exam consists of about 7 short answer questions and 3 problems on the following topics:

1. Definitions of some terms common to the area of clinical trials: double blinded trial, trial time, analysis by intent to treat, baseline measurements, withdrawals.

2. Blocked randomization: disadvantages of a block size too small and too large. Be able to explain how to use blocked randomization to allocate subjects to two treatment groups.

3. Be able to clearly distinguish between phase II and phase III trials.

4. Know the general criteria used to define the study population.

5. Which of the following variables should not be used as an explanatory variable to adjust the log rank statistic? Measurements after randomization, analysis by compliance, etc.

6. Be able to apply the analysis by intent to treat principle to a specific example to decide on an appropriate course of action.

7. Be able to apply the δ method to some statistic.

8. Be able to interpret the hazard ration $\phi = \frac{\lambda_1(x)}{\lambda_2(x)}$ for the purpose of determining the correct rejection region for a one-tailed log rank test (see Peto, et. al, 1977).

9. Sample size calculation.

10. Sequential methods: information fractions, fractions, exit probabilities, $\alpha(t_i)$.

9.33 References

1. $\beta-$Blocker Heart Attack Trial Research Group, "A Randomized Trial of Propranolol in Patients with Acute Myocardial Infarction," *Journal of the American Medical Association,* March 26, 1982, Vol 247, No. 12.

2. Gordon Lan, K. K., and Zucker, David M., "Sequential Monitoring of Clinical Trials: The Role of Information and Brownian Motion," *Statistics in Medicine,* Vol. 12, pages 753-765, 1993.

3. Peto, R., Pike, M. C., Armitage, P., et al, "Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient," *Br. J. Cancer,* 1977, Vol 35, No. 1.

4. Reboussin, David M., DeMets, David L., Kim, KyungMann, Gordon Lan, K. K., "Programs for Computing Group Sequential Bounds Using the Lan-Demets Method, Version 2," University of Wisconsin, Department of Biostatistics, *Technical Report # 95,* October 1995.

Chapter 10

Mathematical Statistics I

Dr. Morgan, Old Dominion University

STAT 625, Fall 1996

Text used: Hogg, Robert V. and Allen T. Craig, *Introduction to Mathematical Statistics, 5-th edition,* Prentice Hall, Upper Saddle River, NJ, 1995

10.1 Notions from Set Theory

$$A \cap B = \{x : x \in A, \text{ and } x \in B\}$$

$$\bigcap_{j=1}^{\infty} A_j = \{x : x \in A_j \forall j\}$$

$$A \cup B = \{x : x \in A \text{ and } x \in B\}$$

$$\bigcup_{j=1}^{\infty} A_j = \{x : x \in A_j \text{ for some } j\}$$

$$A - B = A \cap B^*,$$

where B^* is the complement of B.

- Distributive Property:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

- DeMorgan's Laws:

$$(A \cup B)^* = A^* \cap B^*.$$

$$(A \cap B)^* = A^* \cup B^*.$$

- Cumulative Property:

$$A \cup B = B \cup A.$$

$$A \cap B = B \cap A.$$

- Associative Property:

$$A \cup (B \cup C) = (A \cup B) \cup C.$$

$$A \cap (B \cap C) = (A \cap B) \cap C.$$

10.2 Introduction to Probability

The following is an observed phenomena: there are sets of circumstances which can be repeated, but whose outcome cannot be predicted. But after long sequences of repetitions, the relative frequencies of various events seem to approach limiting values. These limits, called *expected long-run relative frequencies,* are how we interpret probability.

Example: A coin toss.

Example: Mass measurements.

theory of statistical inference is built on a foundation of probability. Inferences are, in fact, probability statements. We begin this course with the goal

of studying those fundamental statistical entities called *random variables.* So, how do we construct a theory of probability? We have experiments of the type discussed above, called *random* or *chance experiments.* These experiments produce results called *outcomes.* The collection of all possible outcomes is called the *sample space,* denoted by Ω. Subsets of Ω are called *events.*

Example: Roll a die once. $\Omega = \{1, 2, 3, 4, 5, 6, \}$. An event might be $A = \{1, 2\}$ which is a roll less than 3.

We wish to be able to assign probabilities to the various events in a consistent and coherent manner. Certainly, the following requirements must be met: If $P(C = c)$ is the probability of the set C, then

1. $P(C = c) \geq 0$.
2. If C_1 and C_2 are two events which cannot occur simultaneously, then $P(C_1 \cup C_2) = P(C_1) + P(C_2)$ which is the probability that at least one of the events occur. More generally, if $C_1, C_2, ...$ are disjoint events, the
$$P\left(\bigcup_{j=1}^{\infty}\right) = \sum_{j=1}^{\infty} P(C_j).$$
3. $P(\Omega) = 1$.

The relationship P that assigns a probability to each event which satisfies the three conditions is called *a probability set function.* It is a naive definition because we have still not said which subsets or events the probability is defined for. Let $\Im$ be a collection of subsets of Ω. $\Im$ is said to be a $\sigma-$field of subsets of Ω if

1. $\Omega \in \Im$.
2. $C \in \Im \Rightarrow C^* \in \Im$.
3. $C_1, C_2, C_3, ... \in \Im \Rightarrow \bigcup_{n=1}^{\infty} C_n \in \Im$.

Properties of $\Im$: If $C_1, C_2, ...$ are members of $\Im$, then so are

1. $\emptyset$.

2. $C_1 \cup C_2$.
3. $C_1 \cap C_2$.
4. $C_1 - C_2$.
5. $\bigcup_{j=1}^{\infty} C_j$.
6. $\bigcap_{j=1}^{\infty} C_j$.
7. $\bigcup_{j=1}^{n} C_j$.
8. $\bigcap_{j=1}^{n} C_j$.

Our more rigorous definition of probability is as follow: a *probability set function* or *probability measure* P is a function from a $\sigma-$field $\Im$ of subsets Ω to $\Re$ satisfying:

1. $P(C) \geq 0, \forall C \in \Im$.
2. $P(\Omega) = 1$.
3. If $C_1, C_2, ...$ are disjoint members of $\Im$, then

$$P\left(\bigcup_{i=1}^{\infty} C_i\right) = \sum_{i=1}^{\infty} P(C_i).$$

A *probability space* is a triple $(\Omega, \Im, P)$ where Ω is the sample space of possible outcomes of an experiment. $\Im$ is a $\sigma-$field of subsets of Ω. P is a probability set function in $\Im$. Such a triple is inherent in every discussion of probability.

Example: Suppose $\Im$ is a $\sigma-$field containing three sets A, B, and C which intersect such a way that none of the 8 possible intersections are empty. The, $\Im$ also contains all the 8 regions and all unions of the 8 regions. $|\Im| = 2^8 = 256$.

Example: Let Ω be any set. The family of all subsets of Ω is a $\sigma-$field denoted 2^{Ω}. The family consisting of just the 2 sets $\emptyset$ and Ω is a $\sigma-$field.

Example: Let $C \subset \Omega$ and suppose $C \neq \emptyset, C \neq \Omega$. Then, $\Im = \{\emptyset, C, C^*, \Omega\}$ is a $\sigma-$field.

Example: Ω is any infinite set. $\Im$ is the family of finite complements. Then, $\Im$ is not a $\sigma-$field. e.g. $\Omega = Z$. $A_i = \{5i \ni i = 1, 2, 3, ...\}$. $B = \bigcup_{i=1}^{\infty} A_i$ is infinite. B^* is also infinite. Note that $B \notin \Im$ because $\Im$ is not a $\sigma-$field.

Why consider any $\sigma-$field other than 2^{Ω} which is the collection of all subsets of Ω? The problem is that if Ω is too large, it is sometimes impossible to construct a probability set function on 2^{Ω}. In this case there will be some events(subsets) that we cannot assign a probability to. Such problems are left to higher level courses. We briefly consider $\Omega = \Re$. The two $\sigma-$fields we will be most concerned with are:

1. The family 2^{Ω} of all subsets of Ω, whenever Ω is finite and countable.

2. The smallest $\sigma-$field containing all subsets of $\Re$ of the form $(-\infty, a]$. This is called the *Borel $\sigma-$field.* Elements are called *Borel sets.*

What kind of subsets are Borel sets? Clearly $\Re = (-\infty, \infty)$ and all sets of the form $(-\infty, a)$ by definition. We also have:

1. (a, ∞) by complementation is $(-\infty, a]$.
2. $(-\infty, a) = \bigcup_{n=1}^{\infty}(-\infty, a - \frac{1}{n}]$.
3. $[a, \infty) = (-\infty, a)^*$.
4. $[a, b] = [a, \infty) \cap (-\infty, b)$.
5. $(a, b) = (a, \infty) \cap (-\infty, b)$.
6. $(a, b] = (a, \infty) \cap (-\infty, b]$.
7. $[a, b) = [a, \infty) \cap (-\infty, b)$.

Thus, all intervals are in this $\sigma-$field. So are all singletons: $\{a\} = (-\infty, a] \cap [a, \infty)$.

Example: Three tosses of a coin. $\Omega = \{HHH, TTT, HHT, HTT, HTH, THH, TTH, THT\}$. $\Im = 2^{\Omega}$ is the collection of all subsets of Ω. Some example events are $A =$ two tails $= \{HTT, TTH, THT\}$, and $B =$ 1-st and 3-rd toss are

the same $= \{HHH, TTT, HTH, THT\}$. Here is a probability set function for this experiment:

$$P(\{c\}) = \frac{1}{8} \text{for each } c \in \Omega.$$

$$P(c) = \sum_{c \in \Omega} P(\{c\}) = \frac{|c|}{8} \Rightarrow P(A) = \frac{3}{8}, P(B) = \frac{1}{2}.$$

Suppose we have the following probabilities:

$$P(\{HHH\}) = \frac{27}{64}, P(\{TTT\}) = \frac{1}{64},$$

$$P(\{HHT\}) = P(\{HTH\}) = P(\{THH\}) = \frac{9}{64},$$

$$P(\{TTH\}) = P(\{THT\}) = P(\{HTT\}) = \frac{3}{64},$$

$$P(\Omega) = \sum_{c \in \Omega} P(\{C\}).$$

Then,

$$P(A) = \frac{3}{64} + \frac{3}{64} + \frac{3}{64} = \frac{9}{64}.$$

$$P(B) = \frac{27}{64} + \frac{1}{64} + \frac{9}{64} + \frac{3}{64} = \frac{5}{8}.$$

There are many other possibilities above.

Example: Choose a number at random from the interval $[0, 1]$. $\Omega = [0, 1]$. $\Im = ß(\text{Borel set}) \cap [0, 1] = \{B \cap [0, 1] \exists : B \in ß\}$. For any $A \in \Im$, define $P(A) = \int_A dx$. Then, P as defined is a probability set function. PROOF: For any $A \in \Im$,

$$P(A) = \int_A dx \geq \int_\emptyset dx \geq 0,$$

$$P(\Omega) = \int_{[0,1]} dx = 1,$$

Let $A_1, A_2, ... \in \Im$ and each A_i are disjoint. Then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \int_{\cup_{i=1}^{\infty} A_i} dx =$$

$$\sum_{i=1}^{\infty}\int_A dx = \sum_{i=1}^{\infty} P(A_i).$$

Another choice for P could be:

$$P(A) = \int_A 2x dx.$$

The properties of a probability set function are as follow:

1. For each $C \in \Im$, $P(C) = 1 - P(C^*)$. PROOF: $\Omega = C \cup C^*$, and $C \cap C^* = \emptyset$. $1 = P(\Omega) = P(C \cup C^*) = P(C) + P(C^*) \Rightarrow 1 - P(C^*) = P(C)$.

2. $P(\emptyset) = 0$. PROOF: In (1), put $C = \Omega$ and $C^* = \emptyset$.

3. If $C_1, C_2 \in \Im \ni: C_1 \subseteq C_2$, then $P(C_1) \leq P(C_2)$. PROOF: $C_2 = C_1 \cup (C_1^* \cap C_2)$ and $C_1 \cap (C_1^* \cap C_2) = \emptyset$. $\Rightarrow P(C_2) = P(C_1) + P(C_1^* \cap C_2) \Rightarrow P(C_2) \geq P(C_1)$.

4. For each $C \in \Im, 0 \leq P(C) \leq 1$. PROOF: Note that $\emptyset \subseteq C \subset \Omega$. By, Theorems (2), and (3), $0 = P(\emptyset) \leq P(C)$. Also Theorem (3iii) implies $P(C) \leq P(\Omega) \leq 1$.

5. If $C_1, C_2 \in \Im$, then $P(C_1) \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2)$.

Theorem 1A: Let $A_1, A_2, ...$ be a finite or countable family of events in the probability space. Then,

$$P\left(\bigcup_{j=1} A_j\right) \subseteq \sum_j P(A_j).$$

PROOF: Define $B_1 = A_1, B_2 = A_2 - A_1, B_3 = A_3 - (A_1 \cup A_2), ..., B_j = A_j - (\bigcup_{k=1}^{j-1} A_k)$. Then, (a) $B_j \subseteq A_j \Rightarrow P(B_j) \leq P(A_j) \Rightarrow \sum_j P(B_j) \leq \sum_j P(A_j)$, (b) $B_j's$ are disjoint implies $\sum_j P(B_j) = P(\bigcup_j B_j)$, (c) $\bigcup_j B_j = \bigcup_j A_j$: proof: clearly $\bigcup_j B_j \subseteq A_j$. Given $w \in \bigcup_j A_j$ then $w \in A_k$ for some k. Let j be the smallest such k. Then, $w \in A_j$ and $w \notin \bigcup_{k=1}^{j-1} A_k \Rightarrow w \in B_j \Rightarrow w \in \bigcup_j B_j \Rightarrow \bigcup_j A_j \subseteq \bigcup_j B_j$. Combining (a), (b), (c) gives

$$P\left(\bigcup_j A_j\right) = P\left(\bigcup_j B_j\right) = \sum_j P(B_j) \leq \sum_j P(A_j).$$

Definition: A sequence of sets $A_1, A_2, ...$ is said to be *non-decreasing* if $A_j \subseteq A_{j+1}, \forall j$ and *non-increasing* if $A_{j+1} \subseteq A_j, \forall j$.

Definition:

$$\lim_{j\to\infty} A_j = \bigcup_{j=1}^{\infty} A_j$$

if the $A_j's$ are non-decreasing.

$$\lim_{j\to\infty} A_j = \bigcap_{j=1}^{\infty} A_j$$

if the $A_j's$ are non-increasing.

Theorem 2A: If $\{A_j\}$ is a monotone sequence of events in a probability space, then

$$P\left(\lim_{j\to\infty} A_j\right) = \lim_{j\to\infty} P(A_j).$$

PROOF: Suppose the $A_j's$ are non-decreasing. Define $B_1 = A_1, B_2 = A_2 - A_1, B_j = A_j - A_{j-1}$. Then the $B_j's$ are pairwise disjoint(i.e. $B_l \cap B_m = \emptyset$ for $l \neq m$). $A_j = \bigcup_{k=1}^{j} B_k$.

$$\lim_{j\to\infty} A_j = \bigcup_{j=1}^{\infty} A_j = \bigcup_{j=1}^{\infty} B_j.$$

So,

$$P\left(\lim_{j\to\infty} A_j\right) = P\left(\lim_{j\to\infty} B_j\right) =$$

$$\sum_{j=1}^{\infty} P(B_j) = \lim_{j\to\infty} \sum_{k=1}^{j} P(B_k) =$$

$$\lim_{j\to\infty} P\left(\bigcup_{k=1}^{j} B_k\right) = \lim_{j\to\infty} P(A_j).$$

The proof for non-increasing can be quickly done given the above proof.

10.2.1 Finite Sample Spaces

Let $\Omega = \{c_1, c_2, ..., c_n\}$ be finite. We take $\Im = 2^{\Omega}$ to be all subsets of Ω. Let p be a real value function defined on Ω satisfying

1. $p(c_i) = 0, i = 1, 2, 3, ..., n.$
2. $\sum_{i=1}^{n} p(c_i) = 1.$

Then p produces a probability set function on $\Im$ by

$$P(c) = \sum_{c_i \in c} p(c_i), \forall c \in \Im.$$

Note:

1. $P(c) \geq 0$ is clear.
2. $$P\left(\bigcup_{j=1}^{k} A_j\right) = \sum_{c_i \in \bigcup_{j=1}^{k} A_j} p(c_i) = \sum_{j=1}^{k} \sum c_i \in A_j p(c_i) = \sum_{j=1}^{k} P(A_j).$$
3. $$P(c) = \sum_{c \in \Omega} p(c_i) = \sum_{i=1}^{n} p(c_i) = 1.$$

So, given a finite sample space, it is sufficient to specify the probability for the individual outcomes $c_1, ..., c_n$. From this a probability set function P is easily specified. A special case of the special case is when all members of Ω are equalikely.

$$p(c_i) = \frac{1}{n}, i = 1, 2, ..., n.$$

Then from any event C:

$$P(C) = \sum_{c_i \in C} p(c_i) = \sum_{c_i \in C} \frac{1}{n} = \frac{|C|}{|\Omega|}.$$

With equalilikely outcomes , calculating probabilities is as simple as counting.

Example: Five cards are dealt from an ordinary deck of 52 cards. Count the number of aces the deal produces. What is the probability that this

number is 2? $\Omega =$ { all possible subsets of 5 cards } = { all possible 5-card hands }.

$$|\Omega| = \left(\begin{array}{c} 52 \\ 5 \end{array} \right) = 2598960.$$

Let $C =$ { all 5 card hands containing exactly 2 aces. }

$$|C| = \left(\begin{array}{c} 4 \\ 2 \end{array} \right) \left(\begin{array}{c} 48 \\ 3 \end{array} \right) = 103776.$$

$$P(C) = \frac{|C|}{|\Omega|} = \frac{103776}{2598960} = 0.0399.$$

10.2.2 Interpretations of Probabilities

In the above examples, nothing dictated the choices of values of P that we made. A probability space is a probability space, regardless of its validity as a model of reality. But, to make the subject worth our effort, we need to find interpretations of the abstract notions which are such that the theorems of the abstract subject become verifiable statements about the real world. A comparison with Euclidean geometry is instructive: points and lines are undefined notions satisfying certain axioms, but they do not exist in the real world. Nevertheless, there are interpretations of the notions of "point" and "line" such that the theorems in geometry become useful real-world facts. Their status as facts, of course, is less secure in the world than it is in geometry; a theorem may be indisputably true, but ultimately it is true only about the abstractions it speaks of.

A convenient real-world interpretation for the probability of an event is provided by the following observed fact: In many situations the outcome of an experiment seems to vary even though the experiment is performed repeatedly under identical conditions. but it is found that if, after each trial, the ratio of the number of occurrences of a certain outcome to the number of performances of the experiment to date is recorded, then these "relative frequencies" of occurrence of the outcome appear to tend to a limit.

Note: The above fact is not a mathematical fact; its truth cannot be proved. For example, it may be impossible to repeat an experiment under identical conditions. Indeed, it may be this impossibility that causes the observed variability. And it is impossible to know whether a finite sequence of relative frequencies is tending to a limit. But these objections do not matter; it is the apparent truth of the observed fact that makes the interpretation

possible.

With the above in mind, we agree to interpret P(A) as the expected relative frequency of occurrence of A over a long series of trials of the experiment. Thus, for example, in the coin toss example, we let A be the event "1-st toss shows H, 2-nd and 3-rd tosses agree with each other," ie $A = \{HHH, HTT\}$, then we say $P(A) = \frac{30}{64}$. We mean that if n is very large, then we expect n repetitions of the experiment to result in event A occurring on about $\frac{30}{64}n$ times.

We do not expect exactly $\frac{30}{64}n$ occurrences of A; indeed, we may occasionally observe large deviations from this expected number. But we expect large deviations to be rare. It is because of this that one way to test the validity of the function P is to perform the experiment a large number of times and see whether the observed relative frequencies of occurrence of the outcomes are close to the probabilities we assigned. In this connection, we might say that probability theory is the subject that constructs probability spaces and deduces theorems, while statistics is the subject that tests the validity of given models by comparing observed results(data) with assertions made in accordance with the above interpretation.

Example: For a group of r randomly selected people, what is the probability that their birthdays are different?

$\Omega = \{$all r-tuples of birthdays chosen from 365 days$\}$.

$|\Omega| = 365^r$.

$C = \{$all r-tuples where all members are different$\}$.

$$P(C) = \frac{|C|}{|\Omega|}.$$

r	probability
2	0.9973
5	0.9729
10	0.8831
20	0.5886
22	0.5243
23	0.4927
30	0.2937
35	0.1856

Example: The digits $1, 2, ..., n$ are arranged in random order. Find the probability that $1, 2, ..., k$ occur consecutively in standard order.

$$\frac{(n-k+1)(n-k)!}{n!} = \frac{(n-k+1)!}{n!}.$$

Note that for the j-th observation, $j + k - 1 \leq n \Rightarrow j \leq n - k + 1$.

10.2.3 Conditional Probability and Independence

Let $C_1, C_2 \in \Im$ and suppose $P(C_1) > 0$.

Definition: The *conditional probability* of C_2 given C_1, denoted by $P(C_2|C_1)$, is the number

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)}.$$

Motivation: The number is intended to represent our revised view of the probability of $C_2's$ occurrence given that we know that C_1 has occurred. In terms of long run relative frequency:

- In n trials, C_1 occurs about $nP(C_1)$ times.
- Among those trials where C_1 occurs, C_2 occurs about $nP(C_1 \cap C_2)$ times.
- So, the relative frequency of C_2 among those trials where C_1 occurs is about

$$\frac{nP(C_1 \cap C_2)}{nP(C_1)} = \frac{P(C_1 \cap C_2)}{P(C_1)}.$$

Example: Given that 3 tosses of a fair coin produce at least 1 head, what is the probability that the first toss is a tail.

$$\Omega = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}.$$

$$C_1 = \{HHH, HHT, HTH, THH, TTH, THT, HTT\}.$$

$$C_2 = \{THH, TTH, TTT, THT\}.$$

$$C_1 \cap C_2 = \{THH, THT, TTH\}.$$

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)} = \frac{\frac{3}{8}}{\frac{7}{8}} = \frac{3}{7}.$$

Find the probability that the first toss is a tail, given exactly one head.

$C_1 = \{HTT, THT, TTH\}$.

$C_1 \cap C_2 = \{THT, TTH\}$.

$$P(C_2|C_1) = \frac{\frac{2}{8}}{\frac{3}{8}} = \frac{2}{3}.$$

Theorem 3A: Let P be a probability set function for $(\Omega, \Im)$ and let $C_1 \in \Im$ such that $P(C_1) > 0$. Then, $P(\cdot|C_1)$ is also a probability set function for $(\Omega, \Im)$. Proof:

1. Let $C_2 \in \Im$.

$$P(C_2|C_1) = \frac{P(C_1 \cap C_2)}{P(C_1)} \geq 0.$$

2. Let $C_2, C_3, ... \in \Im$ be pairwise disjoint. Then,

$$P\left(\bigcup_{i=2}^{\infty} C_i|C_1\right) =$$

$$\frac{P\left(\bigcup_{i=2}^{\infty} C_i \bigcap C_1\right)}{P(C_1)} =$$

$$\frac{P\left(\bigcup_{i=2}^{\infty} C_i \bigcap C_1\right)}{P(C_1)} =$$

$$\sum_{i=2}^{\infty} \frac{P(C_i \cap C_1)}{P(C_1)} =$$

$$\sum_{i=2}^{\infty} P(C_i|C_1).$$

3. $P(\Omega|C_1) =$

$$\frac{P(\Omega \cap C_1)}{P(C_1)} = \frac{P(C_1)}{P(C_1)} = 1.$$

Theorem 4a: For any events $C_1, C_2, ..., C_n$ such that $P(C_1 \cap C_2 \cap \cdots \cap C_{n-1}) > 0$, we have

$$P(C_1 \cap C_2 \cap \cdots \cap C_n) =$$

$$P(C_1)P(C_2|C_1)P(C_3|C_1 \cap C_2)P(C_4|C_1 \cap C_2 \cap C_3) \cdots P(C_n|C_1 \cap C_2 \cap \cdots \cap C_{n-1}).$$

The proof is simple. This theorem is called the *multiplication rule.*

Example: The older child paradox. Pick a family at random from all families with 2 children. Assume that all 4 gender distributions FF, FM, MF, MM are equalilikely. Define the following events:

$$B = \text{ both children are girls } = \{FF\}.$$

$$A = \text{ at least 1 girl } = \{FF, FM, MF\}.$$

$$C = \text{ older child is a girl } = \{FM, FF\}.$$

Then,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)}{P(A)} =$$

$$\frac{P(\{FF\})}{P(\{FF, FM, MF\})} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

And,

$$P(B|C) = \frac{P(B)}{P(\{FM, FF\})} = \frac{\frac{1}{4}}{\frac{2}{4}} = \frac{1}{2}.$$

Example: Polya urn scheme. An urn has r red balls and b blue balls. A ball is drawn and replaced along with c balls of the same color. This is repeated as often as desired. NOTE: $c = 0 \Rightarrow$ sampling with replacement and $c = -1 \Rightarrow$ sampling without replacement. Let R_j be such that j-th ball drawn is red, and B_j be such that j-th ball drawn is blue which is the same as R_j^*. Then,

$$P(R_1) = \frac{r}{r+b}.$$

$$P(B_1) = \frac{b}{r+b}.$$

$$P(R_2|R_1) = \frac{r+c}{r+b+c}.$$

$$P(B_2|R_1) = \frac{b}{r+b+c}.$$

$$P(R_3|R_1 \cap B_2) = \frac{r+c}{r+b+2c}.$$

Then,

$$P(R_1 \cap B_2 \cap R_3) = P(R_1)P(B_2|R_1)P(R_3|R_1 \cap B_2) =$$

$$\left(\frac{r}{r+b}\right)\left(\frac{b}{r+b+c}\right)\left(\frac{r+c}{r+b+2c}\right).$$

Definition: A collection of sets $C_1, C_2, ...$ is a partition of Ω if $\bigcup_{i=1}^{\infty} C_i = \Omega$ and $C_i \bigcap C_j = \emptyset$.

Theorem 5A: (Law of Total Probability) Let $C_1, C_2, ...$ be a finite or countable partition of Ω into events of positive probability. Then for any event A,

$$P(A) = \sum_j P(A|C_j)P(C_j).$$

proof:

$$\sum_j P(A|C_j)P(C_j) = \sum_j P(A \cap C_j) =$$

$$P\left(\bigcup_j A \cap C_j\right) = P(A \cap \Omega) = P(A).$$

Example: In the Polya's urn scheme, what is P(2nd ball drawn is blue)?

$$P(B_2) = P(B_2|R_1)P(R_1) + P(B_2|B_1)P(B_1) =$$

$$\left(\frac{b}{r+b+c}\right)\left(\frac{r}{r+b}\right) + \left(\frac{b+2}{r+b+c}\right)\left(\frac{b}{r+b}\right) =$$

$$\frac{b}{r+b} = P(B_1).$$

The next famous result, which first appeared in the 1760's, is just a combination of Theorem 4A and Theorem 5A.

Theorem 6A: (Baye's Theorem) Let $C_1, C_2, ...$ be a finite or countable partition of Ω into events of positive probability and let A be an event of positive probability. Then for any n,

$$P(C_n|A) = \frac{P(A|C_n)P(C_n)}{\sum_j P(A|C_j)P(C_j)}.$$

Definition: Events $A, B \in \Im$ are *independent* if $P(A \cap B) = P(A)P(B)$.

Motivation: If either $P(A) = 0$ or $P(B) = 0$, then $P(A \cap B) = 0 = P(A)P(B)$. Otherwise A and B are independent implies

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

So, $P(A|B) = P(B)$. Now let $C_1, C_2, ..., C_n \in \Im$. We say that $C_1, C_2, ..., C_n$ are *mutually independent* if for every subset k of $\{1, 2, ..., n\}$

$$P\left(\bigcap_{j \in k} C_j\right) = \prod_{j \in k} P(C_j).$$

Pairwise independence does not imply mutual independence. But, mutual independence does imply pairwise independence.

Example: Two rolls of a fair die. Define the following events:

$C_1 =$ 1st roll is odd,

$C_2 =$ 2nd roll is odd,

$C_3 =$ sum is odd.

Then, $P(C_1) = \frac{18}{36}$, $P(C_2) = \frac{18}{36}$, and $P(C_3) = \frac{18}{36}$.

$P(C_1 \cap C_2) = P(\{11, 13, 15, 31, 33, 35, 51, 53, 55\}) =$

$$\frac{9}{36} = \frac{1}{4} = P(C_2)P(C_1).$$

$P(C_1 \cap C_3) = P(\{12, 14, 16, 32, 34, 36, 52, 54, 56\}) =$

$$\frac{9}{36} = \frac{1}{4} = P(C_1)P(C_3).$$

$$P(C_2 \cap C_3) = P(\{21, 41, 61, 23, 43, 63, 25, 45, 65\}) =$$

$$\frac{9}{36} = \frac{1}{4} = P(C_2)P(C_3).$$

But,

$$P(C_1 \cap C_2 \cap C_3) = P(\emptyset) = 0 \neq \frac{1}{8} = P(C_1)P(C_2)P(C_3).$$

Thus, no mutual independence.

Furthermore, on mutual independence,

$$P\left(\bigcap_{i=1}^{n} C_i\right) = \prod_{i=1}^{n} P(C_i).$$

Example: Define the set $C = \{1, 2, 3, 4\}$ such that

$$P(\{1\}) = \frac{2\sqrt{2} - 1}{4},$$

$$P(\{2\}) = \frac{1}{4},$$

$$P(\{3\}) = \frac{3 - 2\sqrt{2}}{4},$$

$$P(\{4\}) = \frac{1}{4}.$$

Define $C_1 = \{1, 3\}$, $C_2 = \{2, 3\}$, $C_3 = \{3, 4\}$. Then,

$$C_1 \cap C_2 \cap C_3 = \{3\}.$$

$$P(C_1 \cap C_2 \cap C_3) = P(\{3\}) = \frac{3 - 2\sqrt{2}}{4},$$

$$P(C_1)P(C_2)P(C_3) = \left(\frac{1}{2}\right)\left(\frac{2 - \sqrt{2}}{2}\right)\left(\frac{2 - \sqrt{2}}{2}\right) =$$

$$\frac{4 - 4\sqrt{2} + 2}{8} = \frac{3 - 2\sqrt{2}}{4} = P(C_1 \cap C_2 \cap C_3).$$

But,

$$P(C_1 \cap C_2) = \frac{3 - 2\sqrt{2}}{4} \neq \frac{2 - \sqrt{2}}{4} = P(C_1)P(C_2).$$

Theorem 7A: If $C_1, C_2, ..., C_n$ are mutually independent, the so are $C_1^*, C_2, ..., C_n$. proof: Let $\{j_2, ..., j_k\} \subset \{2, 3, ..., n\}$. Then, $C_{j_2} \cap C_{j_3} \cap \cdots \cap C_{j_k} = (C_1 \cap (C_{3_2} \cap C_{j_3} \cap \cdots \cap C_{j_k})) \cup (C_1^* \cap (C_{j_2} \cap \cdots \cap C_{j_k}))$. The two sets on the right side are disjoint, so

$$P(C_{j_2} \cap C_{j_3} \cap \cdots \cap C_{j_k}) =$$

$$P(C_1 \cap C_{j_2} \cap \cdots \cap C_{j_k}) + P(C_1^* \cap C_{j_2} \cap \cdots \cap C_{j_k}) \Rightarrow$$

$$P(C_{j_2})P(C_{j_3}) \cdots P(C_{j_k}) =$$

$$P(C_1)P(C_{j_2}) \cdots P(C_{j_k}) + P(C_1^* \cap C_{j_2} \cap \cdots \cap C_{j_k}) =$$

$$(1 - P(C_1))P(C_{j_2})P(C_{j_3}) \cdots P(C_{j_k}) =$$

$$P(C_1^* \cap C_{j_2} \cap \cdots C_{J_k}) =$$

$$P(C_1^*)P(C_{j_2})P(C_{j_3}) \cdots P(C_{j_k}).$$

Clearly, the above extends to any number of complements.

10.2.4 Discrete Random Variables

It is often the case that it is difficult or bothersome to write down probability statements in terms of subsets of Ω. For instance, with the example of tossing a coin 3 times, the probability of 2 heads is written:

$$P(\{HHT, HTH, THH\}).$$

Suppose we define a function X on Ω where $X(c) = \#$ heads in c for each $c \in \Omega$. Then, instead of the above expression, we could write the probability of 2 heads as $P(X = 2) = P(\{c : X(c) = 2\})$. X as just defined is a random variable, which is a function from the sample space to the real line $\Re$ satisfying certain properties.

Example: Let Ω be all 5 card hands from a standard deck of 52 cards. If we want to know about the number of aces found in a hand, define $X(c) = \#$ aces in $c \in \Omega$. X has possible values $\{0, 1, 2, 3, 4\}$. $P(X = 3) = P(\{c : X(c) = 3\})$.

Two primary purposes for using random variables are:

1. Convince — It is much easier to write down probability statements in terms of numbers, than in terms of the sample space.
2. Restriction of attention to events of interest — Not every event can necessarily be expressed in terms of values of a particular random variable. In the coin toss experiment, "heads on first toss" can not be expressed in terms of $X =$ "# heads in 3 tosses." But, if we are only concerned with the total number of heads, why be bothered with the other event?

Definition: A random variable X on a sample space Ω with σ-field $\Im$ is a function assigning one real number $X(c)$ to each $c \in \Omega$ so that $\{X \leq x\} = \{c : X(c) \leq x\} \in \Im$ for every $x \in \Re$(ie the inverse under X of each Borel set of $\Re$ is in the σ-field $\Im$).

Notation: $\{X \leq x\}$ is notation for $\{c : X(c) \leq x\}$. We can also write this as $X^{-1}(-\infty, x]$. In general, $\{X \in B\}$ means $\{c : X(C) \in B\}$ which can be written as $X^{-1}(B)$. Probability statements in terms of X correspond to subsets of Ω that our probability set function is defined for.

Theorem 8A: Let X be a random variable on sample space Ω with σ-field $\Im$ and probability set function P. Then, the set function P_x on the Borel sets of $\Re$, defined by $P_x(B) = P(X^{-1}(B))$ is a probability set function. proof:

1. For any Borel set B, $P_x(B) = P(X^{-1}(B)) \geq 0$, where P is a probability set function.
2. Let $B_1, B_2, ...$ be disjoint Borel sets. Then,

$$P_x(\cup_i B_i) = P(\{c : X(c) \in \cup_i B_i\}) =$$

$$P(\cup_i \{c : X(c) \in B_i\}) = \sum_i P(\{c : X(c) \in B_i\}) =$$

$$\sum_i P(X \in B_i) = \sum P_x(B_i).$$

3. $P_x(\Re) = P(X^{-1}(\Re)) = P(\Omega) = 1$ since P is a probability set function.

We have started with the triple $(\Omega, \Im, P)$. For reasons of convince, given the above, we introduce the random variable X. this gives us a new triplet $(\Re, ß, P_x)$ via the correspondence of Theorem 8A. We can now forget about $(\Omega, \Im, P)$, and think wholly in terms of $(\Re, ß, P_x)$ hence dealing only with numbers instead of the subset of some arbitrary space. It is common to drop the x subscript and write $Pr(X = 2)$ instead of $P_x(2)$. NOTE: X does not necessarily take on all real values.

Example: $(\Omega, \Im, P)$ for any fixed $A \in \Im$, define

$$I_A(c) = \begin{cases} 1 & \text{if } c \in A. \\ 0 & \text{if } c \notin A. \end{cases}$$

This is called the *indicator* of the set A. To show that I_A is a random variable, we must prove that

$$I_A^{-1}(-\infty, x] \in \Im \forall x \in \Re,$$

$$I_A^{-1}(-\infty, x] = \{c : I_A(c) \leq x\} =$$

$$\begin{cases} \emptyset, & x \leq 0. \\ A^*, & 0 \leq x < 1. \\ \Omega, & x \geq 1. \end{cases}$$

The space a of I_A is $a = \{0, 1\}$.

Example: $\Omega = \{HH, TH, HT, TT\}$, $\Im = 2^\Omega$, $X(c) = \#$ heads in c, $X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0$. Check that X is a random variable:

$$X^{-1}(-\infty, x] = \begin{cases} x < 0 \Rightarrow \emptyset \\ 0 \leq x \leq 1 \Rightarrow \{HH\} \\ 1 \leq x < 2 \Rightarrow \{HT, TH, TT\} \\ x \geq 2 \Rightarrow \Omega \end{cases}$$

$a = \{0, 1, 2\}$.

Theorem 9a: Let X be a random variable on $(\Omega, \Im, P)$. Let $g : \Re \to \Re$ such that $g^{-1}(B) \in ß$ for $B \in ß$. Then, $y = g(x)$ is also a random variable. proof: given as a homework problem.

One special class of random variables is composed of what are called *discrete random variables.* A random variable with probability set function P_x on $\Re$ is called discrete if there is a function f on $\Re$ which is non-zero only on some finite or countable set a(the space of X) and

$$P_x(A) = Pr(x \in A) = \sum_{x \in A \cap a} f(x)$$

for every Borel set A.

Note:

1. $f(x) \geq 0, \forall x$ and $f(x) = 0$ for $x \notin a$.
2. $\sum_{x \in a} f(x) = 1$.

Example: $a = \{1, 3, 5, 6\}$. $f(1) = \frac{1}{4}$, $f(3) = \frac{1}{8}$, $f(5) = \frac{1}{8}$, $f(6) = \frac{1}{2}$. $f(x) = 0$ otherwise.

$$P_x((-\infty, 3]) = Pr(x \leq 3) = \sum_{x \in (-\infty, 3] \cap a} f(x) = f(1) + f(3) = \frac{3}{8}.$$

$$P_x\left([2\frac{1}{2}, 3\frac{1}{2}]\right) = Pr\left(2\frac{1}{2} \leq x \leq 3\frac{1}{2}\right) = \sum_{x \in [2\frac{1}{2}, 3\frac{1}{2}] \cap a} f(x) = f(3) = \frac{1}{8}.$$

Example:

$$f(x) = \begin{cases} \left(\frac{1}{2}\right)^x, & x = 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

$$P_x((-\infty, 5]) = Pr(x \leq 5) =$$

$$\sum_{x \in (-\infty, 5] \cap a} = f(1) + f(2) + f(3) + f(4) + f(5) = \frac{31}{32}.$$

$$Pr(\ x \text{ is odd}) = \sum_{x=1,3,5} \left(\frac{1}{2}\right)^x =$$

$$\sum_{y=0}^{\infty} \left(\frac{1}{2}\right)^{2y+1} = \frac{1}{2} \sum_{y=0}^{\infty} \left(\frac{1}{4}\right)^y = \frac{2}{3}.$$

Terminology: The function $f(x)$ is called the *probability density function* of the random variable x, sometimes abbreviated pdf or called simply the

density. Some authors call this a *probability mass function.* Note: It contains all of the probability information for x.

Fact: A necessary and sufficient condition for a real valued function $f(x)$ to be the density function for a discrete random variable is

1. $f(x) \geq 0, \forall x \in \Re$ with $f(x)$ 0 only on some countable set a.
2. $\sum_{x \in a} f(x) = 1$.

Proof: Simple(restate the definition).

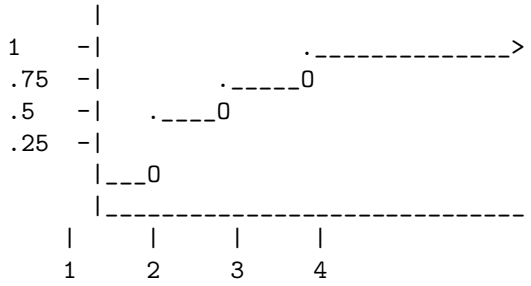
The distribution function of a random variable is defined as $F(x) = P(X \leq x), -\infty < x < \infty$. More details on the general properties of this function are in the next section. For a discrete random variable, it can be written

$$F(x) = \sum_{w \in a \text{ and } w \leq x} f(w).$$

Example: 3 tosses of a fair coin. X is the number of heads.

$$f(x) = \begin{cases} \frac{1}{8}, & x = 0. \\ \frac{3}{8}, & x = 1. \\ \frac{3}{8}, & x = 2. \\ \frac{1}{8}, & x = 3. \\ 0, & \text{otherwise.} \end{cases}$$

$$F(x) = \begin{cases} 0, & x < 0. \\ \frac{1}{8}, & x \leq 0. \\ \frac{1}{2}, & x \leq 1 < 2. \\ \frac{7}{8}, & x \leq 2 < 3. \\ 1, & x \geq 3. \end{cases}$$



The graph is

1. A step function with jumps at the points of positive probability.
2. F is non-decreasing.
3. F is right-continuous.

$$\lim_{x \to 1^+} x = \frac{1}{2}, \text{from the right.}$$

$$\lim_{x \to 1^-} x = \frac{1}{8}, \text{from the left.}$$

10.2.5 Continuous Random Variables

A random variable with probability set function P on $\Re$ is called *continuous* if there is a non-negative function $f(x)$ on $\Re$ such that

$$P(X \in A) = P(A) = \int_A f(x)dx, \text{ for every Borel set } A.$$

Note:

1. $f(x) \geq 0, \forall x \in \Re$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Note: High level courses call this type of random variable *absolutely continuous* since the probability function P is absolutely continuous with respect to Lebseque measure.

Example:

$$f(x) = \begin{cases} e^{-x}, & x > 0. \\ 0, & x \leq 0. \end{cases}$$

$$P([3,4]) = Pr(3 \leq x < 4) =$$

$$\int_3^4 e^{-x}dx = -e^{-x}\Big|_3^4 = -e^{-4} + e^{-3} = 0.0315.$$

$$P((-\infty, 1.5)) = \int_{-\infty}^{1.5} f(x)dx =$$

$$\int_{-\infty}^{0} e^{-x}dx + \int_0^{1.5} e^{-x}dx = 1 - e^{-1.5} = 0.7767.$$

Terminology: As with discrete random variables, all of the probability information is contained in the function $f(x)$, which is called the *probability density function* or simply the *density.* Some authors distinguish the discrete case by calling the density the *probability mass function.*

Fact: A set of necessary and sufficient conditions for a real valued function f to be a density for a continuous random variable is

1. $f(x) \geq 0, \forall x \in \Re$.
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

The proof requires Aratheodory's extension theorem. As with discrete random variables, we can find the distribution function $F(x) = Pr(X \leq x)$.

Example: X is a continuous random variable with density

$$f(x) = \begin{cases} \frac{2}{x^3}, & 1 < x < \infty. \\ 0, & x \leq 1. \end{cases}$$

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(w)dw = \begin{cases} 0, & x \leq 1. \\ 1 - \frac{1}{x^2}, & x > 1. \end{cases}$$

1. This function is continuous — not a step function.
2. This function is non-decreasing.
3. This function is continuous.

10.2.6 Properties of the Distribution Function

X is any random variable. $P(A), A \in ß$ gives the probability for X defined on the Borel sets of $\Re$. The *distribution function* of X is defined as

$$P(X) = Pr(X \leq x) = P((-\infty, x])$$

abbreviated as d.f. For discrete X,

$$F(X) = \sum_{w \leq x} f(w).$$

For continuous X,

$$F(X) = \int_{-\infty}^{x} f(w)dw.$$

These are only 2 types of all possible random variables and hence of all the possible distribution functions. In fact, distribution functions *characterize* probability set functions on the real line. The properties of distribution functions are:

1. $0 \leq F(x) \leq 1$. Proof: $F(x) = Pr(X \leq x)$ is a probability.

2. $F(x)$ is a non-decreasing function of X. Proof: Let $x_1 \leq x_2$. Then,

   $$F(x_1) - F(x_2) = Pr(X \leq x_1) - Pr(X \leq x_2) =$$

   $$P((-\infty, x_1]) - P((-\infty, x_2) \leq 0$$

   since $(-\infty, x_1] \subseteq (-\infty, x_2]$ by Theorem 3.

3. $F(\infty) = \lim_{x\to\infty} F(x) = 1$ and $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$. Proof:

   $$\lim_{x\to\infty} F(x) = \lim_{x\to\infty} P((-\infty, x]).$$

   Let A_x be equal to $(-\infty, x]$. The $A'_x s$ are a monotone sequence of events. So by Theorem 2a,

   $$P\left(\lim_{x\to\infty}(-\infty, x]\right) = P((-\infty, \infty)) = 1.$$

   The proof of $F(-\infty)$ is similar.

4. $F(x)$ is a right continuous function of x. Proof:

   $$\lim_{h\downarrow 0} F(x+h) = \lim_{h\downarrow 0} P((-\infty, x+h]).$$

   Let $h = \frac{1}{j}$. Then, as $j \to \infty$, $h \downarrow 0$. Let $A_j = (-\infty, x + \frac{1}{j}]$. Then,

   $$\lim_{h\downarrow 0} F(x+h) = \lim_{j\to\infty} P(A_j) =$$

   $$P\left(\lim_{j\to\infty} A_j\right) = P\left(\bigcap_{j=1}^{\infty} A_j\right) = P((-\infty, x]) = F(x).$$

Definition: A function $g(x)$ is *right continuous* at x if $\lim_{h\downarrow 0} g(x+h) = g(x)$.

We introduce two convent notations: $F(a^+) = \lim_{h\downarrow 0} F(a+h)$, and $F(a^-) = \lim_{h\uparrow 0} F(a+h)$. Right continuity says that $F(x^+) = F(x), \forall x$. However, it is not necessarily true that $F(x^-) = F(x)$. Following are probabilities for all intervals. Let $a \leq b$ with $\pm\infty$ permitted.

$$P((a,b]) = F(b) - F(a).$$

$$P((a,b)) = f(b-) - F(a).$$

$$P([a,b)) = f(b-) - F(a-).$$

$$P([a,b]) = F(b) - F(a-).$$

$$P(\{a\}) = F(a) - F(a-).$$

Important Fact: It may be shown that distribution functions are characterized by these 3 properties:

1. Non-decreasing.
2. Right continuous.
3. $F(\infty) = 1$, $F(-\infty) = 0$.

That is, a function is a cdf iff it has these 3 properties. There is a 1-to-1 correspondence between functions with these 3 properties and probability set functions on $(\Re, ß)$. In general, the cdf's of discrete random variables will be step functions while those of continuous random variables will be continuous functions. In fact, the Fundamental Theorem of Calculus says if $F(x) = \int_{-\infty}^{x} f(w)dw$, then

1. $F(x)$ is continuous at all x.
2. $F(x)$ is differentiable at least at those x for which $f(x)$ is continuous.
3. If f is continuous at x, then $\frac{\partial F(x)}{\partial x} = f(x)$.

Different densities can produce the same cdf, but those densities can differ only at a finite or countable set of points.

Example:

$$f(x) = \begin{cases} 0, & x \leq 0. \\ 4x, & 0 < x < \frac{1}{2}. \\ 0, & \frac{1}{2} \leq x < 1. \\ \frac{3x^2}{14}, & 1 \leq x \leq 2. \\ 0, & x > 2. \end{cases}$$

$$F(x) = \int_{-\infty}^{x} f(w)dw =$$

$$\begin{cases} 0, & x \leq 0. \\ 2x^2, & 0 < x < \frac{1}{2}. \\ \frac{1}{2}, & \frac{1}{2} < x < 1. \\ \frac{1}{2} + \frac{1}{14}(x^3 - 1), & 1 \leq x \leq 2. \\ 1, & x > 2. \end{cases}$$

The function is non-differentiable at $\frac{1}{2}$. This F has properties of a discrete and continuous random variable. It is a cdf, however.

10.2.7 Mathematical Expectation

Let x be a random variable of discrete or continuous type with pdf $f(x)$. Let $u(x)$ be a function of x. Define

$$u^+(x) = \begin{cases} u(x), & \text{if } u(x) \geq 0. \\ 0, & \text{if } u(x) < 0. \end{cases}$$

$$u^-(x) = \begin{cases} -u(x), & \text{if } u(x) < 0. \\ 0, & \text{if } u(x) \geq 0. \end{cases}$$

Clearly, $u(x) = u^+(x) - u^-(x)$. If at least one of

$$\int_{-\infty}^{\infty} u^+(x)f(x)dx$$

or

$$\int_{-\infty}^{\infty} u^-(x)f(x)dx$$

is finite, then we define the *mathematical expectation* or *expected value* of $u(x)$ as

$$E(u(x)) = \int_{-\infty}^{\infty} u(x)f(x)dx = E(u^+(x)) - E(u^-(x)).$$

And,

$$\sum_x u^+(x)f(x),$$

$$\sum_x u^-(x)f(x),$$

$$\sum_x u(x)f(x) = E(u(x))$$

for discrete cases. Otherwise, expectation is undefined. This differs from the text, which allows the expectation to be defined only if both $E(u^+(x))$ and $E(u^-(x))$ are finite. In our definition,

$$\int_{-\infty}^{\infty} |u(x)|f(x)dx$$

is sufficient but not necessary for existence of the expectation. For the test, this is necessary and sufficient. Properties of mathematical expectation:

1. $u(x) = k$, then $E(u(x)) = E(k) = k$.
2. $u(x) = kv(x)$, then $E(u(x)) = E(kv(x)) = kE(v(x))$.
3. $u(x) = k_1v_1(x) + k_2v_2(x)$ where $v_1(x)$ and $v_2(x)$ do not have opposite infinite expectations. Then, $E(u(x)) = k_1E(v_1(x)) + k_2E(v_2(x))$ or if they do, k_1 and k_2 have opposite signs, then this can be generalized to a finite linear combination of functions.

Review:

$$E(u(x)) = \int u^+(x)f(x)dx - \int u^-(x)f(x)dx$$

provided at least one of the two integrals is finite. Otherwise, the expectation does not exist.

Definition: Let m be a positive integer. The *m-th moment* of the random variable x is $E(x^m)$ provided this expectation exists.

Definition: The *n-th central limit* of the random variable x is $E[(x - E(x))^n]$, provided the expectation exists and $E(x)$ is finite.

The first moment is called the *mean* of x and is denoted by μ. So, the m-th central moment is $E((x - \mu)^m)$.

Theorem 10a: Let x be a random variable with finite mean μ.

1. If $E(x^m)$ is finite, so are

   $$E(x^j), j = 1, 2, ..., m - 1,$$

   and

   $$E((x - \mu)^j), j = 1, 2, ..., m.$$

2. If $E((x-\mu)^m)$ is finite, so are

$$E(x^j), j = 1, 2, ..., m,$$

and

$$E((x-\mu)^j), j = 1, 2, ..., m-1.$$

The second central moment of x is called the *variance* of x and is denoted by σ^2.

$$\sigma^2 = E[(x-\mu)^2].$$

Note that

$$E[(x-\mu)^2] = E(x^2) - 2\mu E(x) + \mu^2 = E(x^2) - \mu^2.$$

The square root of the variance is called the *standard deviation* and is denoted by σ.

$$\sigma = \sqrt{E(x^2) - \mu^2}.$$

It is a widely used measure of dispersion.

$$\int (x-\mu)^2 f(x) dx.$$

The moment generating function: Let x be a random variable and suppose there exists an interval $(-h, h)$ such that the expectation $E(e^{tx})$ is finite for all $t \in (-h, h)$. Clearly this expectation is a function of t, and we write $M(t) = E(e^{xt})$. This is called the *moment generating function* of x. If x is a discrete random variable,

$$M(t) = \sum_x e^{tx} f(x).$$

If x is a continuous random variable,

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

Example: Let

$$f(x) = \begin{cases} e^{-x}, & x > 0. \\ 0, & x \leq 0. \end{cases}$$

$$M(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx =$$

$$\int_0^\infty e^{tx}e^{-x}dx = \int_0^\infty e^{(t-1)x}dx =$$

$$\frac{1}{t-1}e^{(t-1)x}\Big|_0^\infty = -\frac{1}{t-1}, \text{for } t < 1,$$

ie

$$M(t) = \frac{1}{t-1}, t \in (-1, 1).$$

Fact: Moment generating functions(mgf), when they exist, are unique and completely determine the distribution. So, two random variables with the same mgf have the same distribution.

Now, for the continuous case(similar for discrete cases):

$$\frac{\partial M(t)}{\partial t} = M'(t) = \frac{\partial}{\partial t}\int e^{tx}f(x)dx =$$

$$\int_{-\infty}^\infty \frac{\partial}{\partial t}e^{tx}f(x)dx = \int_{-\infty}^\infty e^{tx}xf(x)dx$$

$$\Rightarrow M'(0) = \int_{-\infty}^\infty xf(x)dx = \mu.$$

$$M''(t) = \frac{\partial^2}{\partial t^2}M(t) = \int_{-\infty}^\infty e^{tx}x^2f(x)dx =$$

$$\Rightarrow M''(0) = \int_{-\infty}^\infty x^2f(x)dx = E(x^2).$$

In general,

$$M^{(m)}(t) = \frac{\partial^m}{\partial t^m}M(t) = \int_{-\infty}^\infty e^{tx}x^mf(x)dx$$

$$\Rightarrow M^{(m)}(0) = E(x^m),$$

the m-th moment. Hence, the name "moment generating" function. $M(t)$ is intimately related to the Laplace transform of the density $f(x)$. We have the following result form analysis:
Theorem 11a: If $\exists h \ni M(t)$ is finite for $t \in (-h, h)$, then

1. $M(t)$ is infinitely differentiable at zero.

2. $E(x^m)$ is finite for $m = 0, 1, 2, ...$
3. $E(x^m) = M^{(m)}(0), m = 0, 1, 2, ...$

Corollary: If x has an infinite moment, then its mgf does not exist.

Example:

$$f(x) = \begin{cases} e^{-x}, & x > 0. \\ 0, & x \leq 0. \end{cases}$$

$$M(t) = \frac{1}{1-t}, |t| < 1.$$

$$M'(t) = \frac{1}{(1-t)^2}.$$

$$M'(0) = 1\mu.$$

$$M''(t) = \frac{2}{(1-t)^3}.$$

$$M''(0) = 2 = E(x^2) \Rightarrow \sigma^2 = 2 - (1)^2 = 1.$$

Example:

$$f(x) = \begin{cases} \frac{6}{\pi^2 x^2}, & x = 1, 2, 3, ... \\ 0, & \text{elsewhere.} \end{cases}$$

$$E(x) = \sum_{x=1}^{\infty} \frac{6x}{\pi^2 x^2} = \frac{6}{\pi^2} \sum_{x=1}^{\infty} \frac{1}{x} = \infty$$

$\Rightarrow$ the mgf does not exist.

Using a McClaurin series expansion,

$$M(t) = \sum_{m=0}^{\infty} \frac{E(x^m) t^m}{m!}.$$

Can two different distributions have all the same moments and they are all finite? If so, mgf does not exist. If not, then the moments determine the distribution.

Housdolff Moment Theorem: (see Fellier, vol 2). If a distribution is concentrated on a bounded interval, it is determined by its moments. Now consider two densities:

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi x}} e^{-\frac{1}{2}(\log x)^2}, & x > 0. \\ 0, & x \leq 0. \end{cases}$$

$$g(x) = \begin{cases} f(x)[1 + \sin(2\pi \log x)], & x > 0. \\ 0, & x \leq 0. \end{cases}$$

You can check that

$$\int_0^\infty x^m f(x) \sin(2\pi \log x) dx = 0$$

when $m = 0, 1, 2.... \Rightarrow$ The two distributions which are clearly different, have the same moments.

Now that we have done the mathematics of expectation, let's look deeper at the motivation. Let x be a discrete random variable with density $f(x)$ on $a = \{x_1, x_2, ..., x_n\}$. According to the long run "relative frequency" interpretation of probability, $Pr(X = x_j) = f(x_j)$ leads us to "expect" that in N trials, approximately $Nf(x_j)$ of the trials will result in x_j as the value of x. So in N trials, the average value of X should approach

$$\frac{x_1 N f(x_1) + x_2 N f(x_2) + \cdots + x_n N f(x_n)}{N} = \sum_{j=1}^{n} x_j f(x_j).$$

Similarly, if $u(x)$ is a function of x, we expect about $Nf(x_j)$ occurences of $u(x_j)$ in N trials. So the average observed value of $u(x)$ will be approximately

$$\frac{u(x_1) N f(x_1) + u(x_2) N f(x_2) + \cdots + u(x_n) N f(x_n)}{N} = \sum_{j=1}^{n} u(x_j) f(x_j).$$

Example: Toss a fair coin. You win x dollars if first head appears on toss x. What is your average winnings?

$$f(x) = \begin{cases} \left(\frac{1}{2}\right)^x, & x = 1, 2, ... \\ 0, & \text{elsewhere.} \end{cases}$$

The average winnings are

$$\sum_{x=1}^{\infty} x \left(\frac{1}{2}\right)^x = 2 \text{ dollars.}$$

Example: x has the pdf

$$f(x) = \begin{cases} \frac{2}{3^j}, & x = 3^j, j = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

$$\sum_x f(x) = \sum_{j=1}^{\infty} \frac{2}{3^j} = 2 \left(\frac{\frac{1}{3}}{1 - \frac{1}{3}} \right) = 1.$$

Then,

$$\sum_x xf(x) = \sum_{j=1}^{\infty} 3^j \left(\frac{1}{3^j}\right) = \sum_{j=1}^{\infty} 1 = \infty.$$

Example:

$$f(x) = \begin{cases} \frac{1}{3^j}, & x = -3^j, j = 1, 2, 3, ... \\ \frac{1}{3^j}, & x = 3^j, j = 1, 2, 3, ... \end{cases}$$

Obviously, $\sum_x f(x) = 1$. But,

$$\sum_x xf(x) \sum_{j=1}^{\infty} -3^j \left(\frac{1}{3^j}\right) + \sum_{j=1}^{\infty} 3^j \left(\frac{1}{3^j}\right) =$$

$$\sum_{j=1}^{\infty} -1 + \sum_{j=1}^{\infty} 1 = -\infty + \infty \Rightarrow \text{undefined.}$$

10.2.8 Chebyshev's Inequality

Theorem 6: Let $u(x)$ be a non-negative function of x. Let c be a positive constant. Then

$$Pr(u(x) \geq c) \leq \frac{E(u(x))}{c}.$$

Proof: We prove this for the continuous case(discrete case is similar).

$$E(u(x)) = \int_{-\infty}^{\infty} u(x)f(x)dx =$$

$$\int_A u(x)f(x)dx + \int_{A^c} u(x)f(x)dx$$

where $A = \{x : u(x) \geq c\}$.

$$\Rightarrow E(u(x)) \geq \int_A u(x)f(x)dx \geq$$

$$\int_A cf(x)dx = c\int_A f(x)dx = cPr(u(x) \geq c).$$

Theorem 7: (Chebychev's Theorem) Let x be a random variable with finite variance σ^2. Then for every $k > 0$,

$$Pr(|x - \mu| \geq \ k\sigma) \leq \frac{1}{k^2}.$$

Proof: In Theorem 6, let $u(x) = (x-\mu)^2$ and let $c = k^2\sigma^2$. Then, Theorem 6 says

$$Pr((x-\mu)^2 \geq k^2\sigma^2) \leq \frac{E[(x-\mu)^2]}{k^2\sigma^2},$$

$$Pr(|x-\mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Note: The inequality is typically not sharp. See Example 2 on page 70 of the text book where the bound is sharp.

10.3 Multivariate Distributions

Example: Consider the distribution of two random variables. Toss a coin 3 times. Define the following events:

x_1 = number of heads on the first two tosses.

x_2 = number of heads on all three tosses.

The vector $\underline{x} = (x_1, x_2)$.

Outcome	x_1, x_2
HHH	(2,3)
HHT	(2,2)
HTH	(1,2)
THH	(1,2)
TTH	(0,1)
THT	(1,1)
HTT	(1,1)
TTT	(0,0)

The probability distribution of (x_1, x_2) is

(x_1, x_2)	$P(x_1, x_2)$
(0,0)	$\frac{1}{8}$
(0,1)	$\frac{1}{8}$
(1,1)	$\frac{2}{8}$
(1,2)	$\frac{2}{8}$
(2,2)	$\frac{1}{8}$
(2,3)	$\frac{1}{8}$

A pair of random variables (x, y) is said to be discrete, or to have a discrete distribution if there exists a function $f(x, y)$, which is non-zero only on some countable or finite 2-dimensional set a such that,

$P(A) = Pr((x, y) \in A) = \sum_{(x,y)\in A\cap a} \sum f(x, y)$

for every 2 dimensional Borel set A.

Example:

x	y	$f(x, y)$
0	2	0.1
1	4	0.2
2	6	0.3
3	8	0.4

where $f(x, y) = 0$ elsewhere.

$a = \{(0, 2), (1, 4), (2, 6), (3, 8)\}$

$Pr(x < 2, y \geq 3) = 0.2 = \{(1, 4)\}$

$Pr(0 \leq x \leq 1, y \leq 6) = 0.3 = \{(0, 2), (1, 4)\}$

Note: $f(x, y)$ must have

$$f(x, y) \geq 0, \forall (x, y), \sum_{(x,y)\in a} f(x, y) = 1.$$

Take the continuous case on the previous page, if $\exists$ a function $f(x, y)$ on

$$\Re^2 \ni: Pr((x, y) \in A) = \int_A \int f(x, y)dxdy$$

for every 2-dimensional Borel set A, $f(x, y)$ is called the pdf of (x, y).

Example:

$$f(x, y) = \begin{cases} 6x^2y, & 0 < x < 1, \text{ or } 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

$$P\left(0 < x < \frac{3}{4}, \frac{1}{3} < y < 2\right) =$$

$$\int_{\frac{1}{3}}^{2}\int_{0}^{\frac{3}{4}} f(x,y)dxdy = \int_{\frac{1}{3}}^{2}\int_{0}^{\frac{3}{4}} 6x^2 y dxdy = \int_{\frac{1}{3}}^{2} y2x^3\Big|_{0}^{\frac{3}{4}} dy =$$

$$2x^3\Big|_{0}^{\frac{3}{4}} \times \frac{y}{2}\Big|_{\frac{1}{3}}^{1} = \frac{3}{8}.$$

$$Pr(x<y) = \int_{x<y}\int f(x,y)dxdy = \int_{0}^{1}\int_{0}^{y} 6x^2 y dxdy =$$

$$\int_{0}^{1} y2x^3\Big|_{0}^{y} dy = \int_{0}^{1} 2y^4 dy = \frac{2}{5}y^5\Big|_{0}^{1} = \frac{2}{5}.$$

Note: Borel sets of $\Re^2$ consist of those sets in the smallest σ-field containing all sets of the form $(-\infty, x_1] \times (-\infty, x_2]$, (all x_1, x_2). Equivalently, this is all products $B_1 \times B_2$ where B_1 and B_2 are Borel sets of $\Re^1$.

Example:

$$f(x,y) = \begin{cases} 2, & 0 < x < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Is this a density?

$$\int_{0}^{1}\int_{0}^{y} 2dxdy = \int_{0}^{1} 2x\Big|_{0}^{y} dy = \int_{0}^{1} 2ydy = y^2\Big|_{0}^{1} = 1.$$

Find

$$Pr\left(x > \frac{1}{2}\right) = Pr\left(x > \frac{1}{2}, x < y < 1\right) = \int_{\frac{1}{2}}^{1}\int_{\frac{1}{2}}^{y} 2dxdy = \int_{\frac{1}{2}}^{1} 2x\Big|_{\frac{1}{2}}^{y} dy =$$

$$\int_{\frac{1}{2}}^{1} (2y-1)dy = y^2 - y\Big|_{\frac{1}{2}}^{1} = 1 - 1 - \frac{1}{4} + \frac{1}{2} = \frac{1}{4}.$$

10.4 Homework

10.5 More on Two Random Variables

Let (x, y) be a pair of random variables. We define their distribution function as

$$F(x,y) = Pr(X \le x, Y \le y).$$

If x and y are both continuous, then

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(v,w)dwdv.$$

If they are both discrete, then

$$F(x,y) = \sum_{V\le x}\sum_{W\le y} f(v,w).$$

The following can be checked:

1. $F(x,y)$ is non-decreasing in each argument.
2. $\lim_{x\to\infty,y\to\infty}F(x,y) = 1.$

   $$\lim_{x\to-\infty} F(x,y) = 0 = \lim_{y\to-\infty} F(x,y).$$

3. $F(x,y)$ is right continuous in each argument.
4. Prove for Homework: For any $A < b$ and $C < d$,

   $$Pr(A < x \le b, C < y \le d) = F(b,d) - F(b,C) - F(A,d) + F(A,d).$$

Example: Let

$$f(x,y) = \begin{cases} 2e^{-x-2y}, & x>0, y>0 \\ 0, & \text{otherwise} \end{cases}$$

$$\int_0^X\int_0^y 2e^{-v-2w}dwdv = \int_0^x e^{-v}\left[\int_0^y 2e^{-2w}dw\right]dv =$$

$$\int_0^x e^{-v}\left[-e^{-2w}\right]_0^y dv = (1-e^{-2y})\int_0^x e^{-v}dv = (1-e^{-2y})(1-e^{-x}).$$

i.e.

$$F(x,y) = \begin{cases} (1-e^{-2y})(1-e^{-x}), & x>0, y>0 \\ 0, & \text{elsewhere} \end{cases}$$

When discussing the pair of random variables (x,y), we will speak of the *joint density function* $f(x,y)$ and the *joint distribution function* $F(x,y)$. This is to distinguish them from the corresponding functions for x and y

alone which will be called *marginal densities* and *distributions.* For example, the joint pdf of (x, y) is $f(x, y)$. Let's find the density for y alone:

$$Pr(Y \leq y) = Pr(-\infty < x < \infty, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f(x, y) dx dy,$$

where

$$\int_{-\infty}^{\infty} f(x, y) dx$$

is the density for y. We make two observations:

1. The marginal density for y is

   $$f(y) = \int_{-\infty}^{\infty} f(x, y) dx,$$

   and for x

   $$f(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

2. The marginal distribution function for y is

   $$F(y) = \lim_{x \to \infty} F(x, y),$$

   and for x

   $$F(x) = \lim_{y \to \infty} F(x, y).$$

Example:

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y}, & x > 0, y > 0 \\ 0, & \text{elsewhere} \end{cases}$$

The marginal density of x is

$$f(x) = \begin{cases} \int_0^{\infty} 2e^{-x}e^{-2y} dy = e^{-x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

The marginal cdf of x is

$$F(x) = \lim_{y\to\infty} F(x,y) = \lim_{y\to\infty}(1-e^{-x})(1-e^{-2y}) =$$

$$\begin{cases} 1-e^{-x}, & x>0 \\ 0, & \text{elsewhere} \end{cases}$$

Example: Toss a coin 3 times. Let x_1 be the number of heads on the first 2 tosses. Let x_2 be the total number of heads. The joint distribution is as follow:

$f(x,y)$	0	1	2
0	$\frac{1}{8}$	0	0
1	$\frac{1}{8}$	$\frac{1}{4}$	0
2	0	$\frac{1}{4}$	$\frac{1}{8}$
3	0	0	$\frac{1}{8}$

$$f_1(x_1) = \sum_{x_2} f(x_1,x_2) = \begin{cases} \frac{1}{4}, & x_1=0 \\ \frac{1}{2}, & x_1=1 \\ \frac{1}{4}, & x_1=2 \end{cases}$$

$$f_2(x_2) = \sum_{x_1} f(x_1,x_2) = \begin{cases} \frac{1}{8}, & x_2=0 \\ \frac{3}{8}, & x_2=1 \\ \frac{3}{8}, & x_2=2 \\ \frac{1}{8}, & x_2=3. \end{cases}$$

10.6 Conditional Distributions and Expectations

Now we can define conditional probabilities for values of random variables in terms of densities. First consider x_1 and x_2 are discrete. Let A_1 and A_2 be events $A_1 = \{X_1 = x_1\} = \{X_1 = x_1, -\infty < x_2 < \infty\}$ and $A_2 = \{X_2 = x_2\} = \{X_2 = x_2, -\infty < x_1 < \infty\}$. By definition,

$$P(A_2|A_1) = \frac{P(A_1 \bigcap A_2)}{P(A_1)} = \frac{P(X_1 = x_1 \text{ and } X_2 = x_2)}{P(X_1 = x_1)} = \frac{f(x_1,x_2)}{f_1(x_1)}.$$

Hence, define the conditional density x_2 given x_1 as

$$\frac{f(x_1,x_2)}{f(x_1)}, (\text{whenever } f_1(x_1) > 0) = f(x_2|x_1).$$

Likewise, the conditional density of x_1 given x_2 is

$$\frac{f(x_1, x_2)}{f(x_2)} = f(x_1|x_2), \text{ when } f_(x_2) > 0.$$

Suppose x_1 and x_2 are discrete. $f(x_1, x_2)$ is the joint density. The conditional density of x_2 given x_1 is

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \text{defined when } f_1(x_1) > 0.$$

The c.d. of x_1 given x_2 is

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \text{defined when } f_2(x_2) > 0.$$

Conditional densities for discrete random variables given the probability for one random variable given that the other takes a specific value. A conditional density for discrete random variables has the properties on any discrete random variable density:

1. $f(x_2, |x_1) \geq 0, \forall x_2$ and $f(x_2|x_1) > 0$ only for x_2 in a countable set.
2. $\sum_{x_2} f(x_2|x_1) = \sum_{x_2} \frac{f(x_1,x_2)}{f_1(x_1)} = \frac{1}{f_1(x_1)} \sum_{x_2} f(x_1, x_2) = 1.$

In particular, $P(X_2 = x_2|X_1 = x_1) = P(x_2|x_1)$. We now extend the above definitions to the continuous case. Let x_1 and x_2 be continuous random variables. Define:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}, \text{when } f(x_2) > 0.$$

as the conditional density of x_1 given $X_2 = x_2$. Also,

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f(x_1)}, \ f(x_1) > 0 \text{ is}$$

defined as the conditional density of x_2 given $X_1 = x_1$. NOTE WELL: $f(x_1, x_2)$ and $F(x_2|x_1)$ are not probabilities. They are densities for continuous random variables and probabilities for the continuous case are found by integration.

$$Pr(a < x_1 < b|X_2 = x_2) = \int_a^b f(x_1|x_2)dx_1.$$

We CAN NOT write

$$\frac{Pr(a < x_1 < b)}{Pr(X_2 = x_2)},$$

because the denominator is zero. Nevertheless, we are interested in the distribution of x_1 given x_2 when x_2 takes a particular value. This information is in the conditional density $f(x_1|x_2)$. We note that the conditional densities for conditional random variables satisfy the properties of any continuous random variable density:

1. $f(x_1|x_2) \geq 0, \forall x.$
2. $\int_{-\infty}^{\infty} f(x_1|x_2)dx_1 = \int_{-\infty}^{\infty} \frac{f(x_1,x_2)}{f(x_2)}dx_1 = 1 = \frac{f(x_2)}{f(x_2)}.$

Once we have the density functions in discrete or continuous cases, we can also talk about expectation. The mean for the distribution with density $f(x_2|x_1)$ is called the conditional expectation of X_2 given x_1. It's value is

$$E(x_2|x_1) = \sum_{x_2} x_2 f(x_2|x_1),$$

or

$$E(x_2|x_1) = \int_{-\infty}^{\infty} x_2 f(x_2|x_1)dx_2$$

which is a function of x_1. In general, if $u(x_2)$ is any function of x_2, then the conditional of $u(x_2)$ given $X_1 = x_1$ is

$$E(u(x_2)|x_1) = \sum_{x_2} u(x_2) f(x_2|x_1)$$

or

$$\int_{-\infty}^{\infty} u(x_2) f(x_2|x_1)dx_2.$$

In particular, the conditional variance of x_2 given $X_1 = x_1$ is

$$Var(x_2|x_1) = E\{[x_2 - E(x_2|x_1)]^2|x_1)\} = \sum_{x_2} [x_2 - E(x_2|x_1)]^2 f(x_2|x_1)$$

or

$$\int_{-\infty}^{\infty} [x_2 - E(x_2|x_1)]^2 f(x_2|x_1)dx_2.$$

Example: $f(x_1,x_2) = \begin{cases} x_1+x_2, & 0<x_1<1, 0<x_2<1. \\ 0, & \text{otherwise.} \end{cases}$

$$f_1(x_1) = \int_0^1 x_1 + x_2 dx_2 = x_1x_2 + \frac{x_2^2}{2}\Big|_0^1 = \begin{cases} x_1+\frac{1}{2}, & 0<x_1<1. \\ 0, & \text{otherwise.} \end{cases}$$

$$f_2(x_2) = \int_0^1 f(x_1,x_2) = \begin{cases} x_2+\frac{1}{2}, & 0<x_2<1. \\ 0, & \text{otherwise.} \end{cases}$$

$$E(x_2|x_1) = \int_0^1 \frac{x_2(x_1+x_2)}{x_1+\frac{1}{2}} dx_2 = \frac{3x_1+2}{6x_1+3}.$$

$$E(x_2^2|x_1) = \int_0^1 \frac{x_2^2(x_1+x_2)}{x_1+\frac{1}{2}} dx_2 = \frac{4x_1+3}{12x_1+6}.$$

$$\Rightarrow Var(x_2|x_1) = E(x_2^2|x_1) - [E(x_2|x_1)]^2 =$$

$$\frac{4x_1+3}{12x_1+6} - \left(\frac{3x_1+2}{6x_1+3}\right)^2 = \frac{2(6x_1^2+6x_1+1)}{(12x_1+6)^2}.$$

The mean and variance of the conditional distribution are functions of the variable conditioned on. Hence, they may well be treated as random variables.

Theorem 1A:

a) $E(x_2) = E[E(x_2|x_1)]$. Proof:

$$E[E(x_2|x_1)] = E(u(x_1)) = \int u(x_1) f_1(x_1) dx_1 =$$

$$\int f_1(x_1) \int x_2 x_2 f(x_2|x_1) dx_2 dx_1 = \int f_1(x_1) \int \frac{x_2 f(x_1,x_2)}{f_1(x_1)} dx_2 dx_1 =$$

$$\int\int x_2 f(x_1,x_2) dx_2 dx_1 = \int x_2 \int f(x_1,x_2) dx_1 dx_2 = \int x_2 f(x_2) dx_2 = E(x_2).$$

b) $Var(x_2) = E[Var(x_2|x_1)] + Var[(E(x_2|x_1)]$. Proof:

$$Var(x_2|x_1) = E(x_2^2|x_1) - [E(x_2|x_1)]^2. \Rightarrow$$

$$E\{E(x_2^2|x_1) - E[E(x_2|x_1)]^2\} = E(x_2^2) - E[E(x_2|x_1)]^2.$$

Also,

$$Var(E(x_2|x_1)) = E\{[E(x_2|x_1)]^2\} - \{E[E(x_2|x_1)]\}^2 = E[E(x_2|x_1)]^2 - [E(x_2)]^2.$$

Combining the two expressions,

$$\Rightarrow Var(x_2|x_1) = E(x_2^2) - [E(x_2)]^2.$$

A consequence of this is the following: Let $y = E(x_2|x_1)$. Then, y has the same mean as x_2 while the variance of y satisfies $Var(y) \leq Var(x_2)$.

Example:

$x_1 =$ the number of heads on the first of two tosses.
$x_2 =$ the number of heads on three tosses.

The joint distribution is

		x_1		
		0	1	2
x_2	0	$\frac{1}{8}$	0	0
	1	$\frac{1}{8}$	$\frac{1}{4}$	0
	2	0	$\frac{1}{4}$	$\frac{1}{8}$
	3	0	0	$\frac{1}{8}$

The marginal distributions are:

x_1	0	1	2
$f_1(x_1)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

and

x_2	0	1	2	3
$f_2(x_2)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Then,

	x_1			
$f(x_2\|x_1)$	0	1	2	
0	1	0	0	$\frac{3}{12}$
1	$\frac{1}{3}$	$\frac{2}{3}$	0	$\frac{3}{12}$
2	0	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{3}{12}$
3	0	0	1	$\frac{3}{12}$
	$\frac{4}{12}$	$\frac{4}{12}$	$\frac{4}{12}$	

$$E(x_1|x_2) = \left\{ \begin{array}{ll} 0, & x_2 = 0. \\ \frac{2}{3}, & x_2 = 1. \\ \frac{4}{3}, & x_2 = 2. \\ 2, & x_2 = 3. \end{array} \right\} = \frac{2x_2}{3}$$

and,

$$Var(x_1|x_2) = \left\{ \begin{array}{ll} 0, & x_2 = 0. \\ \frac{1}{3}(\frac{2}{3})^2 + (\frac{1}{3})^2\frac{2}{3}, & x_2 = 1. \\ \frac{2}{9}, & x_2 = 2. \\ 0, & x_2 = 3. \end{array} \right.$$

Then, $E(x_1) = 1$ and $E(x_2) = \frac{3}{2}$.

$$E[E(x_1|x_2)] = E\left(\frac{2x_2}{3}\right) = \frac{2}{3}E(x_2) =$$

$$\frac{2}{3} \times \frac{3}{2} = 1 = E(x_1).$$

10.7 Correlation Coefficient

Let x and y be random variables with finite variances σ_1^2 and σ_2^2. Hence, they also have finite means μ_1 and μ_2. Let $u_{xy} = (x - \mu_1)(y - \mu_2)$ and consider

$$E[u(x,y)] = E[(x - \mu_1)(y - \mu_2)] = E[xy - \mu_1 y - \mu_2 x + \mu_1\mu_2] =$$

$$E(x,y) - \underbrace{\mu_1 E(y)}_{\mu_2} - \underbrace{\mu_2 E(x)}_{\mu_1} + \mu_1\mu_2 = E(x,y) - \mu_1\mu_2.$$

This is called *the covariance of* x, y, abbreviated $Cov(x,y)$. If $\sigma_1 > 0$ and $\sigma_2 > 0$, we also define the *correlation coefficient* of x, y as

$$Corr(x,y) = \frac{Cov(x,y)}{\sigma_1\sigma_2} = \rho.$$

If one of σ_1 or σ_2 is equal to zero, then we do not define the correlation.

Example:

$$f(x,y) = \left\{ \begin{array}{ll} \frac{1}{50}(x^2 + y^2), & 0 < x < 2. \\ 0, & \text{elsewhere.} \end{array} \right.$$

Let's find $Corr(x, y)$.

$$E(x,y) = \int_1^4 \int_0^2 \frac{xy}{50}(x^2+y^2)dxdy = \frac{315}{100}.$$

$$f_1(x) = \int_1^4 \frac{x^2+y^2}{50}dy = \begin{cases} 3x^2+21, & 0<x<2. \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_2(y) = \int_0^2 \frac{x^2+y^2}{50}dx = \begin{cases} \frac{\frac{8}{3}+2y^2}{50}, & 1<y<4. \\ 0, & \text{elsewhere.} \end{cases}$$

$$\mu_1 = E(x) = \int_0^2 \frac{x(3x^2+21)}{50}dx = \frac{54}{50}.$$

$$\mu_2 = E(y) = \int_1^4 \frac{y(\frac{8}{3}+2y^2)}{50}dy = \frac{295}{100}.$$

$$\Rightarrow Cov(x,y) = E(x,y) - \mu_1\mu_2 = \frac{315}{100} - \frac{54}{50} \times \frac{295}{100} = -\frac{18}{500} = -0.036.$$

$$E(x^2) = \int_0^2 \frac{x^2}{50}(3x^2+21)dx = \frac{188}{125}.$$

$$\Rightarrow Var(x) = \frac{188}{125} - \left(\frac{54}{50}\right)^2 = \frac{211}{625} = 0.3376.$$

Similarly for y :

$$Var(y) = 0.6015.$$

$$\Rightarrow Corr(x,y) = \frac{-0.036}{\sqrt{0.3376}\sqrt{0.6015}} = -0.0799.$$

Usage: ρ is commonly regarded as a measure of the intensity of the linear relationship between x and y. In the homework you will show that $-1 \leq \rho \leq 1$, if $Pr(X = a, Y = b) = 1$ for some constants a, b. Then either $\rho = 1$, or $\rho = -1$ as a is positive or negative. Hence, the limiting values of ρ are found for the pair x, y having all values on a straight line; the weaker the linear relation, the smaller $|\rho|$ will be.

Notes:

1. $Cov(x,y) = E[(x-\mu)(x-\mu)] = E[(x-\mu)^2] = Var(x). \Rightarrow Corr(x,y) = 1$

2. $Cov(ax,by) = E\{[ax - a\mu_1][by - b\mu_2]\} = abE[ax - \mu_1)(y - \mu_2)] = abCov(x,y). \Rightarrow Corr(x,y) = \frac{Cov(ax,by)}{\sqrt{Var(ax)}\sqrt{Var(by)}} =$
$\frac{abCov(x,y)}{ab\sigma_x\sigma_y} = \frac{Cov(x,y)}{\sigma_x\sigma_y} = Corr(x,y).$

Let $x_1, ..., x_n$ be a random sample with finite variances. Then,

$$Var\Big(\sum_{i=1}^{n} x_i\Big) = E\Big[\Big(\sum_{i+1}^{n} x_i\Big)^2\Big] - \Big[E\Big(\sum x_i\Big)\Big]^2 =$$

$$E\Big[\sum_{i=1}^{n} x_i^2 + \sum^{n}\sum_{i\neq j}^{n} x_i x_j\Big] - \Big[\sum_{i=1}^{n} E(x_i) + 2\sum^{n}\sum_{i<j}^{n} E(x_i)E(x_j)\Big] =$$

$$\sum_{i=1}^{n} E(x_i^2) - \sum_{i=1}^{n}\Big[E(x_i)\Big]^2 + 2\Big[\sum\sum_{i<j}^{n} E(x_i x_j) - \sum\sum_{i<j}^{n} E(x_i)E(x_j)\Big] =$$

$$\sum_{i=1}^{n}\Big[E(x_i^2) - E(x_i)^2\Big] + 2\sum^{n}\sum_{i<j}^{n}\Big[E(x_i,x_j) - E(x_i)E(x_j)\Big] =$$

$$\sum_{i=1}^{n} Var(x_i) + 2\sum^{n}\sum_{i<j}^{n} Cov(x_i,x_j).$$

In particular, $Var(x+y) = Var(x) + Var(y) + 2Cov(x,y)$. Now, let $y_i = a_i x_i$. Then,

$$Var\Big(\sum_{i=1}^{n} a_i x_i\Big) = Var\Big(\sum_{i=1}^{n} y_i\Big) =$$

$$\sum_{i=1}^{n} Var(y_i) + 2\sum^{n}\sum_{i<j}^{n} Cov(y_i,y_j) =$$

$$\sum_{i=1}^{n} a_i^2 Var(x_i) + 2\sum^{n}\sum_{i<j}^{n} a_i a_j Cov(x_i,x_j).$$

10.8 Joint Moment Generating Functions

If $\exists h_1, h_2$ such that $E(e^{t_1x+t_2y})$ is finite for all $t_1 \in (-h_1, h_1)$ and for $t_2 \in (-h_2, h_2)$, then we define the joint moment generating function of x and y as

$$E(e^{t_1x+t_2y}) = M(t_1, t_2).$$

Note that $M(t_1, 0) = E(e^{t_1x}) = M(t_1)$, the mgf of x. $M(0, t_2) = E(e^{t_2y}) = M(t_2)$ the mgf of y. Also,

$$\frac{\partial^{k+m} M(t_1, t_2)}{\partial t_1^k \partial t_2^m} = \frac{\partial^{k+m} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x+t_2y}}{\partial t_1^k \partial t_2^m} f(x, y)dxdy =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m e^{t_1x+t_2y} f(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k y^m f(x, y)dxdy = E(x^k y^m)$$

which is also called *a moment of order* $k + m$ for (x, y). It follows that,

$$Cov(x, y) = \frac{\partial^2}{\partial_{t_1}\partial_{t_2}} M(t_1, t_2)\Big|_{t_1=t_2=0} - \frac{\partial M(t_1, o)}{\partial_{t_1}}\Big|_{t_1=0} \times \frac{\partial}{\partial t_2} M(0, t_2)\Big|_{t_2=0}.$$

10.9 Independent Random Variables

We now extend our notation of independent events (section 1.4 of the text book), to independent random variables. We will do this in terms of *distribution functions* since they exist for all random variables, not just discrete or continuous.

Definition: Let the random variables x_1 and x_2 have distribution functions $F_1(x_1)$ and $F_2(x_2)$, and a joint cdf $F(x_1, x_2)$. We say that x_1 and x_2 are *independent* iff $F(x_1, x_2) = F_1(x_1)F_2(x_2), \forall x_1, x_2$. This says that $Pr(X_1 \leq x_1, X_2 \leq x_2) = Pr(X_1 \leq x_1)Pr(X_2 \leq x_2)$. Recalling that Borel sets are just countable unions, intersections, and compliments of intervals of these forms $(-\infty, x]$, it follows that x_1, x_2 are independent iff $Pr(X_1 \in A_1, X_2 \in A_2) = Pr(X_1 \in A_1)Pr(X_2 \in A_2)$ for any Borel sets A_1 and A_2.

This implies that $Pr(X_1 \in A_1 | X_2 \in A_2) = Pr(X_1 \in A_1)$, and $Pr(X_2 \in A_2 | X_1 \in A_1) = Pr(X_2 \in A_2)$. So, the probability for events concerning one variable are not affected by knowledge of the other variable.

Now, suppose x_1, x_2 are continuous (replace $\int by \sum$) or discrete with densities $f_1(x_1)$ and $f_2(x_2)$. Now if continuous x_1, x_2 are independent iff

$$\int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(x_1,x_2)dx_1dx_2 = F(x_1,x_2) = F_1(x_1)F_2(x_2) =$$

$$\int_{-\infty}^{x_1} f_1(x_1)dx_1 \int_{-\infty}^{x_2} f_2(x_2)dx_2 =$$

$$\int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_1(u_1)f_2(x_2)du_1du_2, \forall x_1, x_2$$

$\Rightarrow f(x_1,x_2) = f_1(x_1)f_2(x_2)$ is the joint density of (x_1,x_2). Moreover since

$$f_1(x_1|x_2) = \frac{f(x_1,x_2)}{f_2(x_2)}$$

it follows that independence implies (and is implied by) $f(x_1|x_2) = f_1(x_1)$, and $f_2(x_2|x_1) = f_2(x_2)$ i.e the conditional and marginal densities are the same.

10.10 Homework Answers

8.6) (Note: There's something wrong with the first part of the proof). Prove that:

$$E(|x-b|) = E(|x-m|) + 2\int_m^b (b-m)f(x)dx.$$

proof:

$$E(|x-b|) = \int_{-\infty}^{\infty} |x-b|f(x)dx =$$

$$\int_{-\infty}^{m} |x-b|f(x)dx + \int_m^b |x-b|f(x)dx + \int_b^{\infty} |x-b|f(x)dx.$$

$$\int_{-\infty}^{m} |x-b|f(x)dx + \int_b^{\infty} |x-b|f(x)dx =$$

$$\int_{-\infty}^{m} (b-x)f(x)dx + \int_b^{\infty} (x-b)f(x)dx + \int_m^b (b-x)f(x)dx =$$

$$\int_{-\infty}^{m}(b-x)f(x)dx+\int_{b}^{\infty}(x-b)f(x)dx-\int_{m}^{b}(b-x)f(x)dx+2\int_{m}^{b}(b-x)f(x)dx.$$

$$\int_{-\infty}^{m}(b-m)+(m-x)f(x)dx+\int_{m}^{\infty}(x-m)(m-b)f(x)dx=$$

$$\int_{-\infty}^{m}(m-x)f(x)dx+\int_{m}^{\infty}(x-m)f(x)dx+\int_{-\infty}^{m}(b-m)f(x)dx+$$

$$\int_{m}^{\infty}(m-b)f(x)dx=$$

$$\overbrace{\int_{-\infty}^{m}|x-m|f(x)dx+\int_{m}^{\infty}|x-m|f(x)dx}^{\int_{-\infty}^{\infty}|x-m|f(x)dx}+\overbrace{(b-m)F(m)+(m-b)(1-F(m))}^{0\text{ since }F(m)=\frac{1}{2}}.$$

$$\Rightarrow E(|x-m|)+2\int_{m}^{b}(b-x)f(x)dx.$$

1a) Prove that:

$$E(|x|^s)=E(|x|^r)+1, 0\leq s\leq r.$$

Proof:

$$E(|x|^s)=\int|x|^sf(x)dx.$$

$$\int_{-1}^{1}|x|^sf(x)dx+\int_{x\in[-1,1]}|x|^sf(x)dx\leq$$

$$\int_{-1}^{1}f(x)dx+\int_{x\notin[-1,1]}|x|^rf(x)dx\leq$$

$$\int_{-\infty}^{\infty}f(x)dx+\int_{-\infty}^{\infty}|x|^rf(x)dx=$$

$$1+E(|x|^r).$$

1b) Prove that:

$$E(|x|^m)<\infty\Leftrightarrow E(x^m)<\infty.$$

Obvious if m is even. If m is odd,

$$E(x^m) = -\int_{-\infty}^{0} -x^m f(x)dx + \int_0^{\infty} x^m f(x)dx =$$

$$E(u^+(x)) - E(u(x)),$$

$$E(|x|^m) = \int_{-\infty}^{0} -x^m f(x)dx + \int_0^{\infty} x^m f(x)dx =$$

$$E(u^+(x)) + E(u^-(x))$$

The result is that $E(x^m)$ is finite.

1c) Given $E(x^m)$ is finite, then:

i $E(x^j) < \infty$ for $j = 1, 2, ..., m-1$. $E(x^m)$ being finite implies $E|x|^m$ is finite. Then, $E|x|^j, j < m$. Then, $E|x|^j$ is finite for $j = 1, 2, ...m-1$.

ii $E(x-\mu)^j < \infty, j = 1, 2, ..., m$.

$$E(x-\mu)^j = E\left[\sum_{k=0}^{j} \begin{pmatrix} j \\ k \end{pmatrix} x^{k-\mu+j-k}\right] =$$

$$\sum_{k=0}^{j} \overbrace{\begin{pmatrix} j \\ k \end{pmatrix}}^{<\infty} \overbrace{(-\mu)^{j-k}}^{<\infty} \overbrace{E(x^k)}^{\text{finite by part (i)}} .$$

which is finite.

$E(x-b)^2$ is minimum at $b = \mu$ given $E(x)$ is finite. $E(x-b)^2 = E(x^2 - 2bx + b^2) = E(x^2) - 2bE(x) + b^2$. The $E(x^2)$ could be infinite. If $E(x^2) = \infty$, then the function is constant in b.

2.3) The Law of Total Probability: $Pr(a < x \leq b, c < y \leq d) = Pr(a < x \leq b, y \leq d) - Pr(a < x \leq b, y \leq c)$ where there is no overlap. Then, $Pr(a \leq x \leq b, y \leq d) - Pr(a < x \leq b, y \leq y \leq c)$. $Pr(x \leq b, y \leq d) - Pr(x \leq a, y \leq d) - [Pr(x \leq b, y \leq c) - Pr(x \leq a, y \leq c)] = F(b,d) - F(a,d) - F(b,c) + F(a,c)$.

10.11 Test and Answers

Work any five problems. Only the first five turned in will be graded. Turn in your copy of the test with your answers. It will be returned with your graded paper.

I. Answer the following:

(a) Define what it means for two random variables X and Y to be independent. The joint distribution function equals the product of the marginal distribution functions; i.e. $F(x,y) = F_1(x)F_2(y), \forall x, y$.

(b) List three equivalent sets of conditions for independence: two for arbitrary X and Y, and one for the special case of continuous X and Y. See notes — we made a list.

(c) Use one of your conditions in (c) to prove the following : If X and Y are continuous random variables, and u and v are functions for which the expectations exist, then,

$$E[u(X)v(Y)] = E[u(X)]E[v(Y)].$$

For continuous, independent $\Leftrightarrow f(x,y) = f_1(x)f_2(y)$. So

$$E[u(x)v(y)] = \int\int u(x)v(y)f(x,y)dxdy = E[u(x)]E(v(y)].$$

(d) Define the joint mgf $M(t_1,t_2)$ of the continuous random variables X and Y. If X and Y are independent, use (c) to prove that $M(t_1,t_2) = M(t_1,0)M(0,t_2)$.

$$M(t_1,t_2) = E(e^{t_1x+t_2y}) \Rightarrow \begin{cases} M(t_1,0) = E(e^{t_1x}). \\ M(0,t_2) = E(e^{t_2y}). \end{cases}$$

Then, if x and y are independent,

$$M(t_1,t_2) = E[u(x)v(y)] = E[u(x)]E(v(y)] =$$

$$M(t_1,0)M(0,t_2).$$

II. There are 15 cards on a table, arranged in three stacks of five cards each. One stack contains three spades, one stack contains four spades, and one stack contains all spades. Each stack has been thoroughly shuffled. A stack is randomly selected and its first two cards are revealed: they are both spades.

"S" represents spades. Let $c_i = i$-th card turned over, and $s_j = j$-th stack selected.

(a) If the third card in the stack is revealed, what is the probability that it is a spade?

$$P(C_3 = S|C_1 = C_2 = S) = \frac{P(C_1 = C_2 = C_3 = S)}{P(C_1 = C_2 = S)}.$$

Find the numerator and denominator using the total law of probability.

$$P(C_1 = C_2 = C_3 = S) =$$

$$\sum_{j=1}^{3} P(C_1 = C_2 = C_3 = S|S_j)P(S_j) =$$

$$\frac{\binom{3}{3}\binom{2}{0}}{\binom{5}{3}}\frac{1}{3} + \frac{\binom{4}{3}\binom{1}{0}}{\binom{5}{3}}\frac{1}{3} + \frac{\binom{5}{3}\binom{0}{0}}{\binom{5}{3}}\frac{1}{3} =$$

$$\left(\frac{1}{10} + \frac{4}{10} + \frac{10}{10}\right)\frac{1}{3} = \left(\frac{15}{10}\right)\left(\frac{1}{3}\right).$$

$$P(C_1 = C_2 = S) = \sum_{j=1}^{3} P(C_1 = C_2 = S|S_j)P(S_j) =$$

$$\frac{\binom{3}{2}\binom{2}{0}}{\binom{5}{2}}\frac{1}{3} + \frac{\binom{4}{2}\binom{1}{0}}{\binom{5}{2}}\frac{1}{3} + \frac{\binom{5}{2}\binom{0}{0}}{\binom{5}{2}}\frac{1}{3} =$$

$$\left(\frac{3}{10} + \frac{6}{10} + \frac{10}{10}\right)\frac{1}{3} = \left(\frac{19}{10}\right)\left(\frac{1}{3}\right).$$

Therefore,

$$P(C_3 = S|C_1 = C_2 = S) = \frac{\left(\frac{15}{10}\right)\frac{1}{3}}{\left(\frac{19}{10}\right)\frac{1}{3}} = \frac{15}{19}.$$

(b) What is the probability that the chosen stack is the one with the four spades? Using Baye's theorem,

$$P(S_2|C_1 = C_2 = S) = \frac{P(C_1 = C_2 = S|S_2)P(S_2)}{P(C_1 = C_2 = S)} =$$

$$\frac{\left(\frac{6}{10}\right)\frac{1}{3}}{\left(\frac{19}{10}\right)\frac{1}{3}} = \frac{6}{19}.$$

III. X and Y are discrete random variables with joint density

		y		
$f(x,y)$	0	1	2	$f_1(x)$
0	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{3}$
1	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{3}$
2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{3}$
$f_2(y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	

(a) Give two proofs for the statement "X and Y are dependent."
Proof 1:

$$Pr(x=0,y=0)=f(0,0)=0.$$

$$Pr(x=0)Pr(y=0)=f_1(0)f_2(0)=\frac{1}{3}\frac{1}{3}=\frac{1}{9}.$$

$$0\neq\frac{1}{9}.$$

Proof 2:

$$Pr(x=1|y=2)=\frac{Pr(x=1,y=2)}{Pr(y=2)}=\frac{f(1,2)}{f_2(2)}=\frac{0}{\frac{1}{3}}=0.$$

$$Pr(x=1)=f_1(1)=\frac{1}{3}.$$

$$0\neq\frac{1}{3}.$$

(b) Find $E(X|Y=y)$ and $Var(X|Y=y)$.

		y	
$f(x\|y)$	0	1	2
0	0	$\frac{1}{4}$	$\frac{3}{4}$
1	$\frac{1}{2}$	$\frac{1}{2}$	0
2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

$$F(x|Y=y)=\begin{cases}E(x|y=0)=0(0)+1(\frac{1}{2})+2(\frac{1}{2})=1.5\\E(x|y=1)=0(\frac{1}{4})+1(\frac{1}{2})+2(\frac{1}{4})=1.0\\E(x|y=2)=0(\frac{3}{4})+1(0)+2(\frac{1}{4})=0.5\end{cases}$$

$$\begin{cases}E(x^2|y=0)=0^2(0)+1^2(\frac{1}{2})+2^2(\frac{1}{2})=2.5\\E(x^2|y=1)=0^2(\frac{1}{4})+1^2(\frac{1}{2})+2^2(\frac{1}{4})=1.5\\E(x^2|y=2)=0^2(\frac{3}{4})+1^2(0)+2^2(\frac{1}{4})=1.0\end{cases}$$

Therefore,

$$\begin{cases} Var(x|y=0) = 2.5 - (1.5)^2 = 0.25 \\ Var(x|y=1) = 1.5 - (1.0)^2 = 0.50 \\ Var(x|y=2) = 1.0 - (0.5)^2 = 0.75 \end{cases}$$

(c) Show by calculating both sides directly that $E[E(X|Y)] = E(X)$.

$$E\{E(x|y)\} = \sum_y E(x|Y=y) f_2(y) =$$

$$1.5(\frac{1}{3}) + 1.0(\frac{1}{3}) + 0.5(\frac{1}{3}) = 1.0.$$

$$E(x) = \sum_x x f_1(x) = 0(\frac{1}{3}) + 1(\frac{1}{3}) + 2(\frac{1}{3}) = 1.0.$$

IV. Define the *distribution function* for a random variable X. Now find the distribution function for each of the following random variables:

(a) W = the result of one toss of a fair die.

$$Pr(W=w) = \begin{cases} 0 & w < 1. \\ \frac{1}{6} & 1 \leq w < 2. \\ \frac{2}{6} & 2 \leq w < 3. \\ \frac{3}{6} & 3 \leq w < 4. \\ \frac{4}{6} & 4 \leq w < 5. \\ \frac{5}{6} & 5 \leq w6. \\ 1 & w \geq 6. \end{cases}$$

(b) Y = a number randomly selected from the interval $[0, 1]$.

$$Pr(Y \leq y) = \int_0^y du = \begin{cases} 0, & y \leq 0. \\ y, & 0 < y < 1. \\ 1, & y \geq 1. \end{cases}$$

(c) The function X of Y and W defined by

$$X = \begin{cases} Y & \text{if } W = 1. \\ X & \text{if } W > 1. \end{cases}$$

$$Pr(X \leq x) = Pr(X = x|w=1)Pr(w=1) + Pr(X \leq x|w>1)Pr(w>1) =$$

$$Pr(Y \leq x)\frac{1}{6} + Pr(1 < W \leq x)\frac{5}{6} =$$

$$\begin{cases} 0, & x < 0. \\ \frac{x}{6}, & 0 \le x < 1. \\ \frac{1}{6}, & 1 \le x < 2. \\ \frac{2}{6}, & 2 \le x < 3. \\ \frac{3}{6}, & 3 \le x < 4. \\ \frac{4}{6}, & 4 \le x < 5. \\ \frac{5}{6}, & 5 \le x < 6. \\ 1, & x \ge 6. \end{cases}$$

V. P is a probability set function defined on the σ-field $\Im$ of subsets of the sample space Ω.

(a) What are the properties that define a σ-field? See notes.

(b) Use the properties listed in (a) to prove that

$$C_1, C_2 \in \Im \Rightarrow C_1 - C_2 \in \Im.$$

Proof:

$$C_1 - C_2 = C_1 \cap C_2^* = (C_1^* \cup C_2)^*,$$

$$C_1 \in \Im \Rightarrow \left.\begin{array}{l} C_1^* \in \Im. \\ C_2 \in \Im. \end{array}\right\} \Rightarrow$$

$$C_1^* \cup C_2 \in \Im \Rightarrow (C_1^* \cup C_2)^* \in \Im$$

(c) What are the properties that *define* a probability set function? See notes.

(d) Use the properties listed in (c) to prove that

$$C_1, C_2 \in \Im \; C_1 \subseteq C_2 \Rightarrow P(C_1 - C_2) = P(C_1)P(C_2).$$

Proof:

$$C_2 \subset C_1 \Rightarrow C_1 = C_2 \cup (C_1 1 \cap C_2^*)$$

and

$$C_2 \cap (C_1 \cap C_2^*) = \emptyset$$

$$\Rightarrow P(C_1) = P(C_2) + P(C_1 \cap C_2^*).$$

$$\Rightarrow P(C_1) - P(C_2) = P(\overbrace{C_1 \cap C_2^*}^{C_1 - C_2})$$

(e) Give a counterexample to prove that the property proved in (d) does not hold if the condition $C_1 \subseteq C_2$ is omitted. **Proof:** Take any experiment. Take $C_1 = \emptyset, C_2 = \Omega$.

$$P(C_1 - C_2) = P(\emptyset) = 0.$$

$$P(C_1) - P(C_2) = 0 - 1 = -1.$$

VI. Consider the function

$$f(x) = \begin{cases} \frac{a^{x-1}e^{-a}}{(x-1)!}, & \text{if x = 1, 2, 3, ...} \\ 0, & \text{otherwise.} \end{cases}$$

Here $a \in (0, \infty)$ is a constant.

(a) Prove that f is a density for a discrete random variable X. **Proof:** Clearly $f(x) \geq 0, \forall x$. Also,

$$\sum_{x=1}^{\infty} \frac{a^{x-1}e^{-a}}{(x-1)!} = e^{-a} \sum_{x=1}^{\infty} \frac{a^{x-1}}{(x-1)!} =$$

$$e^{-a} \sum_{y=0}^{\infty} \frac{a^y e^{-a}}{y!} = e^{-a}e^{a} = 1.$$

(b) Find the moment generating function of X.

$$M(t) = E(e^{tx}) = \sum_{x=1}^{\infty} \frac{e^{tx}a^{x-1}e^{-a}}{(x-1)!} =$$

$$e^{t-a} \sum_{x=1}^{\infty} \frac{(ae^t)^{x-1}}{(x-1)!} = e^{t-a}e^{ae^t}$$

(c) Use the moment generating function to find $E(X)$ and $var(X)$.

$$\frac{\partial}{\partial t} M(t) = M(t)[1 + ae^t]$$

Set $t = 0$:

$$M(0)(1 + a) = 1 + a = E(x).$$

$$\frac{\partial^2}{\partial t^2} M(t) = M'(t)[1 + ae^t] + M(t)ae^t.$$

Set $t = 0$:

$$(1 + a)^2 + a = E(x^2) \Rightarrow$$

$$E(x^2) - [E(x)]^2 = (1 + a)^2 + a - (1 + a)^2 = a = Var(x).$$

(d) Find $E(X)$ directly.

$$F(x) = \sum_{x=1}^{\infty} \frac{xa^{x-1}e^{-a}}{(x-1)!} =$$

$$a\sum_{x=2}^{\infty} \frac{a^{x-2}e^{-a}}{(x-2)!} + \sum_{x=1}^{\infty} \frac{a^{x-1}e^{-a}}{(x-1)!} = a+1.$$

10.12 Independent RVs, Expectations, and MGFs

If x_1 and x_2 are random variables, then the following statements are equivalent notions of independence:

1. $F(x_1, x_2) = F_1(x_1)F_2(x_2), \forall x_1, x_2$.
2. $Pr(X_1 \in A_1, X_2 \in A_2) = Pr(X_1 \in A_1)Pr(X_2 \in A_2), \forall A_1, A_2$ Borel sets.
3. $Pr(X_1 \in A_1 | X_2 \in A_2) = Pr(X_1 \in A_1)$ and $Pr(X_2 \in A_2 | X_1 \in A_1) = Pr(X_2 \in A_2), \forall A_1, A_2$ Borel sets.
4. If x_1 and x_2 are either discrete or continuous, then $f(x_1, x_2) = f_1(x_1)f_2(x_2)$.
5. If x_1 and x_2 are either discrete or continuous, then $f(x_1|x_2) = f_1(x_1)$, and $f(x_2|x_1) = f_2(x_2)$.

Example:

$$f(x,y) = \begin{cases} \frac{1}{50}(x^2+y^2), & 0 < x < 2. \\ 0, & \text{elsewhere.} \end{cases}$$

We already found that

$$f_1(x) = \begin{cases} \frac{1}{50}(3x^2+21), & 0 < x < 2. \\ 0, & \text{elsewhere.} \end{cases}$$

$$f_2(y) = \begin{cases} \frac{1}{50}(\frac{8}{3}+2y^2), & 1 < y < 4. \\ 0, & \text{elsewhere.} \end{cases}$$

Then,

$$f_1(x)f_2(y) = \frac{1}{2500}(8x^2+42y^2+56+6x^2y^2) \neq \frac{1}{50}(x^2+y^2)$$

$\Rightarrow x, y$ are dependent.

Also, $f(x|y) = \frac{x^2+y^2}{\frac{8}{3}+2y^2} \neq f_1(x)$. We can make the task for determining independence of x, y even simpler in the continuous and discrete cases.

Theorem 1: Let x_1 and x_2 have a joint pdf $f(x_1, x_2)$. Then, x_1, x_2 are independent iff $f(x_1, x_2) = g(x_1)h(x_2)$ where $g(x_1) > 0$ and $h(x_2) \geq 0$. Proof: If x_1, x_2 are independent then $f(x_1, x_2) = f_1(x_1)f_2(x_2) = g(x_1)h(x_2)$. Suppose next that

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_2 = \int_{-\infty}^{\infty} g(x_1)h(x_2)dx_2 =$$

$$g(x_1)\overbrace{\int_{-\infty}^{\infty} f(x_1, x_2)dx_2}^{c_1} = g(x_1)c_1, \text{ where } c_1 \text{ is positive and } c_1 > 0.$$

Similarly,

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)dx_1 = \int g(x_1)h(x_2)dx_1 =$$

$$h(x_2)\int_{-\infty}^{\infty} g(x_1)dx_1 = h(x_2)c_2, c_2 > 0.$$

Also,

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)dx_1dx_2 =$$

$$\int\int g(x_1)h(x_2)dx_1dx_2 = \overbrace{\int g(x_1)dx_1}^{\frac{1}{c_1}}\overbrace{\int h(x_2)dx_2}^{\frac{1}{c_2}}$$

$c_1c_2 = 1$. Hence $f_1(x_1)f_2(x_2) = g(x_1)c_1h(x_2)c_2 = g(x_1)h(x_2) =$

$f(x_1, x_2)$. $\Rightarrow x_1, x_2$ are independent.

Example:

$$f(x, y) = \begin{cases} \frac{1}{5}(x^2 + y^2), & 0 < x < 2, 1 < y < 4. \\ 0, & \text{elsewhere.} \end{cases}$$

The function can not be factored out as $g(x)h(y)$. Thus, x, y are not independent.

Example:

$$f(x_1, x_2) = \begin{cases} 6e^{-2x_1-3x_2}, & x_1 > 0, x_2 > 0. \\ 0, & \text{elsewhere.} \end{cases}$$

$$g(x_1) = \begin{cases} 6e^{-2x_1}, & x_1 > 0. \\ 0, & \text{elsewhere.} \end{cases}$$

$$h(x_2) = \begin{cases} 6e^{-3x_2}, & x_2 > 0. \\ 0, & \text{elsewhere.} \end{cases}$$

Then,

$$g(x_1) = e^{-2x_1} I_{x_1>0},$$

$$h(x_2) = 6e^{-3x_2} I_{x_2>0},$$

Then, $f(x_1, x_2) = g(x_1)h(x_2) \Rightarrow x_1, x_2$ are independent.

Example:

$$f(x_1, x_2) = \begin{cases} 8x_1x_2, & 0 < x_1 < x_2 < 1. \\ 0, & \text{elsewhere.} \end{cases}$$

Then,

$$8x_1x_2 I_{0<x_1<x_2<1}$$

does not factor into functions of x_1, x_2 alone. Look at the range for independence.

Theorem 2: (slightly modified) If x_1, x_2 are independent, then $Pr(a < x_1 \leq b, c < x_2 \leq d) = Pr(a < x_1 \leq b)Pr(c < x_2 \leq d)$. Proof: (from the homework)

$$F(bd) - F(ad) - F(bc) + F(ac) \overbrace{=}^{indep.}$$

$$F_1(b)F_2(d) - F_1(a)F_2(d) - F_1(b)F_2(c) + F_1(a)F_2(c) =$$

$$[F(_1(b) - F_1(a)][F_2(d) - F_2(c)] = Pr(a < x_1 \leq b)Pr(c < x_2 \leq d).$$

Independence can simplify expectations also.

Theorem 3: Let x_1, x_2 be independent random variables with marginals $f_1(x_1)$ and $f_2(x_2)$. Then,

$$E[u(x_1)v(x_2)] = E[u(x_1)]E[(v(x_2)].$$

Proof: (for the continuous case. the discrete case is similar)

$$\int\int u(x_1)v(x_2)f_1(x_1)f_2(x_2)dx_1dx_2 =$$

$$\left[\int u(x_1)f_1(x_1)dx_1\right]\left[\int v(x_2)f_2(x_2)dx_2\right].$$

Suppose x, y have finite variances. If they are independent, then $Cov(x, y) = E(x, y) - E(x)E(y) = E(x)E(y) - E(x)E(y) = 0$. And so,

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} = 0.$$

Therefore, x, y are uncorrelated random variables. The converse is not true. Un-correlation does not imply independence.

Example: Let x have the pdf:

$$f(x) = \begin{cases} \frac{1}{2}, & -1 < x < 1. \\ 0, & \text{elsewhere.} \end{cases}$$

You can check that $E(x) = 0$, and $E(x^3) = 0$. Let $y = x^2$. Then, $Cov(x, y) = E(xy) - E(x)E(y) = E(x^3) - E(x)E(x^2) = 0 - 0 = 0$. Therefore, the variables are uncorrelated.

Theorem 4: Let x_1, x_2 have the joint density $f(x_1, x_2)$ with marginals $f_1(x_1)$ and $f_2(x_2)$. Also, let $M(t_1, t_2)$ be the mgf for the joint distribution. Then, x_1 and x_2 are independent iff $M(t_1, t_2) = M(t_1, 0) \times M(0, t_2)$. **Proof:** If x_1, x_2 are independent, the

$$M(t_1, t_2) = E[e^{t_1x_1+t_2x_2}] = E[e^{t_1x_1}e^{t_2x_2}] =$$

$$E(e^{t_1x_1 1})E(x^{t_2x_2}) = M(t_1, 0)M(0, t_2).$$

Part II: If $M(t_1, t_2) = M(t_1, 0) \times M(0, t_2)$, then

$$M(t_1, t_2) = M(t_1, 0)m(0, t_2) =$$

$$\left[\int_{-\infty}^{\infty} e^{t_1 x_1} f_1(x_1) dx_1\right]\left[\int_{-\infty}^{\infty} e^{t_2 x_2} f_2(x_2) dx_2\right] =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x_1} e^{t_2 x_2} f_1(x_1) f_2(x_2) dx_1 dx_2 =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1 x_1 + t_2 x_2} f_1(x_1) f_2(x_2) dx_1 dx_2.$$

Therefore, by the uniqueness of mgf's, $f_1(x_1)f_2(x_2)$ is the joint pdf of x_1, x_2. Otherwise, $M(t_1, t_2)$ would be the mgf of some other distribution. $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Therefore, x_1, x_2 are independent.

10.12.1 Extension to Several Random Variables

Sometimes we want to consider more than two random variables for the same expression and may want to ask probability questions about the simultaneous values of several variables. Write $\underline{x}(c) = (x_1(c), x_2(c), ..., x_n(c))$ where $x_1, x_2, ..., x_n$ are each real valued functions on Ω with $\sigma-$field $\Im$.

Definition: The Borel $\sigma-$field of subsets of $\Re^n$ is the smallest $\sigma-$field containing all sets of the form

$$(-\infty, a_1] \times (-\infty, a_2] \times .... \times (-\infty, a_n] =$$

$$\{(x_1, x_2, ..., x_n) \ni x_1 \leq a_1, x_2 \leq a_2, ..., x_n \leq a_n\}$$

It will be denoted by $ß_n$.

Definition: $\underline{x} = (x_1, ..., x_n)$ mapping Ω to $\Re^n$ is a *n-dimensional random vector* if $\underline{x}^{-1}(B) \in \Im, \forall B \in ß_n$. Here, $\underline{x}^{-1}(b) = \{c : (x_1(c), x_2(c), ..., x_n(c)) \in B\}$. It may be shown that:

Result: $\underline{x} = (x_1, x_2, ..., x_n)$ is a random vector iff each x_i is a random variable.

We define a probability set function $\underline{x}$ just as we did for a single random variable: iff B is any Borel set of $\Re^n$, i.e. $B = \{(x_1, x_2, ..., x_n) \to x_1 \in B_1, x_2 \in B_2, ..., x_n \in B_n\} = B_1 \times B_2 \times ... \times B_n$ where $B_i \in ß$, then

$$P_{\underline{x}}(B) = P(\underline{x}^{-1}(B)) = P(\{c : \underline{x}(c) \in B\}) =$$

$$P(\{c : x_1(c) \in B_1, x_2(c) \in B_2, ..., x_n(c) \in B_n\}).$$

10.12.2 Discrete and Continuous Cases

The n random variables $(x_1, ..., x_n)$ are jointly discrete iff $\exists$ a function $f(x_1, x_2, ..., x_n)$ which is non-zero only on some finite or countable subset $a \in \Re^n$, such that

$$Pr((x_1, ..., x_n) \in A) = \sum ... \sum_{(x_1, x_2, ...x_n) \in A \cap a} f(x_1, ..., x_n), \forall A \in ß_n.$$

Note:

1. $f(x_1, ..., x_n) \geq 0, \forall (x_1, ..., x_n)$ only if $(x_1, ..., x_n) \in a$.

2. $\sum ... \sum_{(x_1, ...x_n) \in a} f(x_1, ..., x_n) = 1$.

(1) and (2) are necessary and sufficient for $f(x_1, ..., x_n)$ to be a pdf for a set of discrete variables.

The n random variables $x_1, x_2, ...x_n$ are said to be jointly continuous of $(x_1, ..., x_n)$ is said to be a *continuous random vector* if there exists a function $f(x_1, ..., x_n)$ on $\Re^n : Pr(x_1, ..., x_n \in A) =$

$$\int ... \int f(x_1, ..., x_n) dx_1 dx_2 ... dx_n$$

for every n dimensional Borel set $A \in ß^n$. Note that if f must satisfy

$$f(x_1, x_2, ..., x_n) \geq 0, \forall x_1, ..., x_n,$$

$$\int \int_{\Re^n} f(x_1, ..., x_n) dx_1 ... dx_n = 1.$$

f is called the *joint density* of $x_1, x_2, ..., x_n$.

10.12.3 Distribution Functions for n Random Variables

Given n random variables, $x_1, ..., x_n$ or the random vector $(x_1, ..., x_n)$ (think of them all at the same time not individually) be the discrete or continuous, or otherwise their *joint density function* is defined to be

$$F(x_1, ..., x_n) = Pr(X_1 \leq x_1, X_2 \leq x_2, ...X_n \leq x_n).$$

The following can be checked:

1. $F(x_1, x_2, ..., x_n)$ is non decreasing in each x_i.
2. $\lim_{x_1 \to \infty x_2 \to \infty, ... x_n \to \infty} F(x_1, x_2, ..., x_n) = 1$.

   $$\lim_{x_i \to -\infty} F(x_1, x_2, ..., x_n) = 0 \text{ for each } x_i.$$

3. $F(x_1, x_2, ..., x_n)$ is right continuous in each x_i.
4. For any set $(a_1, b_1] \times (a_2, b_2] \times ... \times (a_n, b_n]$,

   $$\sum \begin{pmatrix} + \\ - \end{pmatrix} F(c_1, c_2, ..., c_n)$$

   is non-negative where $\sum$ is over all 2^n combinations of $c_i = a_i$ and the sign is + iff the numerator of $c_i s = a_i$ is even.

Example: $n = 1$. Then, $F(b_1) - F(a_1) = Pr(a_1 < x_1 \leq b_1)$.

Example: $n = 2$. Then,

$$F(b_1, b_2) - \overbrace{F(a_1, b_2)}^{\text{number of } a_i\text{'s is odd}} - F(b_1, a_2) + \overbrace{F(a_1, a_2)}^{\text{number of } a_i\text{'s is even}} =$$

$$Pr(a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2).$$

It can be shown that (1) thru (4) are *necessary and sufficient* for a function $F(x_1, x_2, ..., x_n)$ to be a distribution function. Another important property is that letting any subset of the $x_i s$ go to ∞ produces the distribution function of the remaining xs. So, e.g., $F(\infty, \infty, x_3, ..., \infty)$ is the distribution function of x_3. $F(\infty, x_2, x_3, ..., x_n)$ is the distribution function of $(x_2, x_3, ..., x_n)$.

Example:

$$f(x, y, z) = \begin{cases} 4e^{-x-2y-2z}, & x > 0, y > 0, z > 0. \\ 0, & \text{otherwise.} \end{cases}$$

is the pdf for (x, y, z).

$$\Rightarrow F(x, y, z) = \int_0^z \int_0^y \int_0^x f(u, v, w) du dv dw =$$

$$\begin{cases} (1 - e^{-x})(1 - e^{-2y})(1 - e^{-2z}), & x > 0, y > 0, z > 0. \\ 0, & \text{otherwise.} \end{cases}$$

$F(\infty, y, z) = (1 - e^{-2y})(1 - e^{-2z})$ is the Distribution function of (y, z). $F(\infty, \infty, z) = (1 - e^{-2z})$ is the distribution function of z.

10.12.4 Marginal Densities

$f(x_1, x_2, ..., x_n)$ is the pdf for $(x_1, x_2, ..., x_n)$. Assume continuous (discrete is similar):

$$Pr(a < x_1 < b) = Pr(a < x_1 < b, -\infty < x_2 < \infty, ..., -\infty < x_n < \infty) =$$

$$\underbrace{\int_a^b \int_{-\infty}^{\infty} .... \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n) dx_n dx_{n-1} ... dx_1}_{\text{integrating this function of } x_1 \text{ gives the probability for } x_1.}$$

Therefore, the marginal density of x_1 is

$$f(x_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} .... \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n) dx_2 dx_3 ... dx_n.$$

Likewise, choose any $k < n$ of the x_i's then the joint pdf of these k variables is found by integrating over the other $n - k$ variables.

Example: $n = 6$. The joint pdf of x_2, x_4, x_5 is

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3, x_4, x_5, x_6) dx_1 dx_3 dx_6 =$$

$$f_{2,4,5}(x_2, x_4, x_5).$$

The conditional density of x_1, x_3, x_6 given x_2, x_4, x_5 is defined as x_1, x_3, x_6 given x_2, x_3, x_5 is defined as

$$\frac{f(x_1, x_2, x_3, x_4, x_5, x_6)}{f_{2,4,5}(x_2, x_4, x_5)} = f(x_1, x_3, x_6 | x_2, x_4, x_5).$$

Likewise, we can define the conditional density for $n-k$ of the x's given the other k. Conditional expectations follow naturally. e.g.

$$E(u(x_1, x_3, x_6)(x_2, x_4, x_5) =$$

$$\int \int \int u(x_1, x_3, x_6) f(x_1, x_3, x_6 | x_2, x_4, x_5) dx_1 dx_3 dx_6.$$

10.12.5 Joint Independence

This is the generalization of independent n variables $x_1, x_2, ..., x_n$.

Definition: Let $x_1, x_2, ..., x_n$ have marginal distributions $F_1(x_1), ..., F_n(x_n)$ and a joint distribution function $F(x_1, ..., x_n)$. These are *mutually independent* iff $F_1(x_1)F_2(x_2)...F_n(x_n) = F(x_1, x_2, ..., x_n)$. It can be shown that the following are equivalent notions of *mutual independence* of $x_1, x_2, ..., x_n$.

1. $F(x_1, x_2, ..., x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n)$.
2. $Pr(x_1 \in A_1 \text{ and } x_2 \in A_2 \text{ and } \cdots \text{ and } x_n \in A_n) = Pr(x_1 \in A_1)Pr(x_2 \in A_2) \cdots Pr(x_n \in A_n)$.
3. If $x_{i1}, x_{i2}, ..., x_{ik}$ and $x_{j1}, x_{j2}, ..., x_{jL}$ are any partition of $x_1, x_2, ..., x_n$ then,

$$Pr(x_{i1} \in A_{i1}, x_{i2} \in A_{i2}, ..., x_{ik} \in A_{ik} | x_{j1} \; in A_{j1}, ..., x_{jL} \in A_{jL}) =$$

$$Pr(x_{i1} \in A_{i1}, x_{i2} \in A_{i2}, ..., x_{ik} \in A_{ik})$$

for all Borel sets $A_{i1}, A_{i2}, ..., A_{iL}$.

Furthermore, if $x_1, ..., x_n$ are discrete or continuous, then

4 $(x_1, ..., x_n) = f(x_1)f(x_2) \cdots f(x_n)$.

5 $(x_{i1}, x_{i2}, ..., x_{ik} | x_{j1}, x_{j2}, ..., x_{jL}) = f(x_{i1}, x_{i2}, ..., x_{ik})$ for any partition as described in (3).

Theorems (1) - (4) can all be generalized to n mutually independent random variables. Briefly stated (see page 111 of the text book):

1. The joint density $f(x_1, x_2, ..., x_n)$ factors into the product of n non-negative functions. $f(x_1, x_2, ..., x_n) = g(x_1)h(x_2) \cdots k(x_n)$ iff mutual independence holds.
2. The probabilities for intervals multiply or are found by multiplying the marginal probabilities.
3. $E(u_1(x_1), u_2(x_2), ..., u_n(x_n)) = E(u_1(x_1))E(u_2(x_2)) \cdots E(u_n(x_n))$.

4. $$M(t_1, t_2, ..., t_n) = \prod_{i=1}^{n} M(0, 0, 0, t_i, 0, 0, ...)$$

i.e. the joint mgf factors iff $x_1, x_2, ..., x_n$ are mutually independent.

Example: Bad luck in life: Let x_0 be your loss (pain, etc) in a particular situation. Let $x_1, ..., x_n$ be the losses of others in the same situation. Assume $x_0, x_1, ..., x_n$ are independent with the same continuous distribution. Define N to be the smallest n for which $x_{n+1} > x_0$. Now, the event $(N \geq n)$ is x_0 is the largest of $x_0, x_1, ..., x_n$. By the above assumptions,

$$Pr(N \geq n) = \frac{1}{n+1} \Rightarrow Pr(N = n) =$$

$$Pr(N \geq n) - Pr(N \geq n-1) = \frac{1}{n+1} - \frac{1}{n+2} = \frac{1}{(n+1)(n+2)} \Rightarrow$$

$$E(N) = \sum_{n=0}^{\infty} n \left(\frac{1}{(n+1)(n+2)} \right) = \infty.$$

10.13 Some Homework Answers

$$P(A) = P(A_1 \cup A_3 \cup A_5) = P(A_1) + P(A_3) + P(A_5),$$

$$P(A_1) = \frac{1}{6},$$

$$P(A_3) = P(\text{roll 1 = not 6, roll 2} < 5, \text{ roll 3} \geq 4) =$$

$$P(\text{roll 1 = not 6})P(\text{roll 2} < 5)P(\text{roll 3} \geq 4) =$$

$$\frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} =$$

$$P(A_5) = P(\text{ roll 1 = not 6, roll 2} < 5, \text{ roll 3} \geq 4, \text{ roll 4} < 3, \text{ roll 5} \geq 2) =$$

$$\frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} \times \frac{2}{6} \times \frac{5}{6}.$$

10.14 Indicators

For any $A \subseteq \Omega$, we define the *indicator function* of A :

$$I_A(c) = \begin{cases} 1, & \text{if } c \in A. \\ 0, & \text{if } c \notin A. \end{cases}$$

We have already seen that I_A is a random variable iff $A \in \Im$. We say that A has the *Bernoulli distribution* with parameter $p = P(A)$. I_A is a discrete random variable with pdf

$$f(x) = \begin{cases} p, & x = 1. \\ 1-p, & x = 0. \\ 0, & \text{otherwise.} \end{cases}$$

$$E(I_A) = \sum_x x f(x) = 1(p) + 0(1-p) = p = P(A).$$

$$E(I_A^2) = 0^2 f(0) + 1^2 f(1) = p. \Rightarrow$$

$$Var(I_A) = p - p^2 = (1-p)p.$$

Next, let events $A_1, A_2, ..., A_n$ be any n events ($A_i \in \Im$). Then,

1. $I_{A_1 \cap A_2 \cap \cdots A_n} = I_{A_1} I_{A_2} \cdots I_{A_n}$.
2. $I^*_{A_i} = 1 - I_{A_i}$.
3. $A_1, A_2, ..., A_n$ are mutually independent events iff $I_{A_1}, I_{A_2}, ..., I_{A_n}$ are mutually independent random variables.
4. $$Cov(I_{A_i}, I_{A_j}) = E(I_{A_i}, I_{A_j}) - E(I_{A_i})E(I_{A_j}) =$$
   $$E(I_{A_i \cap A_j}) - E(I_{A_i})E(I_{A_j}) =$$
   $$P(A_i \cap A_j) - P(A_i)P(A_j).$$

Example: An urn contains N balls of which m are red and $N - m$ are white. n balls are drawn from the urn at random without replacement. Let x be the number of red balls in the same sample. The distribution of X is called the *hyper geometric distribution.*

$$Pr(X = x) =$$

$$\begin{cases} \dfrac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, & x = \max(0, m(N-m))... \min(n, m). \\ 0, & \text{otherwise.} \end{cases}$$

Use indicators to find the mean and variance. Let

$$I_j = \begin{cases} 1, & \text{if red ball } j \text{ is drawn.} \\ 0, & \text{otherwise.} \end{cases}$$

where the red balls in the urn are numbered $1, 2, ..., M$. Then,

$x = I_1 + I_2 + \cdots + I_m$.

$E(x) = E\left(\sum_{j=11}^{m} I_j\right) = \sum_{j=1}^{m} E(I_j) =$

$\sum_{j=1}^{m} Pr(\text{red ball } j \text{ drawn}) = \sum_{j=1}^{m} Pr(\text{red ball 1 drawn}) =$

$mPr(I_1 = 1) = \dfrac{mn}{N}$. see next page.

$Var(x) = Var\left(\sum_{j=1}^{m} I_j\right) =$

$\sum_{j=1}^{m} Var(I_j) + 2\sum\sum_{j<j'} Cov(I_j, I_{j'}) =$

$\sum_{j=1}^{m} P(I_j = 1)[1 - P(I_j = 1)] +$

$2\sum\sum_{j<j'} [P(I_j I_{j'} = 1) - P(I_j = 1)P(I_{j'} = 1)] =$

$MPr(I_1 = 1)[(1 - Pr(I_1 = 1)] +$

$2\binom{m}{2}\Big[Pr(I_1 I_2 = 1) - Pr(I_1 = 1)Pr(I_2 = 1)\Big],$

$$Pr(I_1 = 1) = \frac{\left(\begin{array}{c} 1 \\ 1 \end{array} \right)\left(\begin{array}{c} N-1 \\ n-1 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} = \frac{n}{N},$$

$$Pr(I_1 = 1)Pr(I_2 = 1) = \frac{\left(\begin{array}{c} 2 \\ 2 \end{array} \right)\left(\begin{array}{c} N-2 \\ n-2 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} = \frac{n(n-1)}{N(N+1)}.$$

Then,

$$\frac{mn}{N}\frac{(N-n)}{N} + 2\left(\begin{array}{c} M \\ 2 \end{array} \right)\left[\frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} \right] =$$

$$n\left(\begin{array}{c} M \\ N \end{array} \right)\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right) = Var(x).$$

10.15 Binomial, Negative Binomial, and Multivariate Distributions

Consider an experiment with two possible outcomes (mutually exclusive and exhaustive) which for purposes of identification will be called *success* and *failure*. Let $p = P(\text{success})$ and $1 - p = P(\text{failure}), 0 \leq p \leq 1$. Define a random variable x to be the number of successes in n independent trials of the experiment. x is called the *binomial* random variable with parameters n, p. Clearly, x is discrete with space $0, 1, ..., n$. The density of x is

$$f(x) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, & x = 0, 1, .., n. \\ 0, & \text{otherwise.} \end{cases}$$

To verify this is a density, recall the binomial formula.

$$(a+b)^n = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) a^x b^{n-x}$$

where a, b are any two real numbers and n is an integer. So,

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} =$$

$$[p + (1-p)]^n = 1^n = 1.$$

The mgf of x is

$$M(t) = E(e^{tx}) = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} e^{tx} =$$

$$\sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) (pe^t)^x (1-p)^{n-x} =$$

$$[pe^t + 1 - p]^n = [1 - p + pe^t]^n.$$

The text book uses this to get μ and σ^2. We calculate directly,

$$E(x) = \sum_{x=0}^{n} x f(x) = \sum_{x=0}^{n} x \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} =$$

$$n \sum_{x=1}^{n} \left(\begin{array}{c} n-1 \\ x-1 \end{array} \right) p^x (1-p)^{n-x} =$$

$$np \sum_{x=1}^{n} \left(\begin{array}{c} n-1 \\ x-1 \end{array} \right) p^{x-1} (1-p)^{n-x} =$$

Let $y = x - 1$, then,

$$np \sum_{y=0}^{n-1} \left(\begin{array}{c} n-1 \\ y \end{array} \right) p^y (1-p)^{n-1-y} = np.$$

$$E(x^2) = \sum_{x=0}^{n} x^2 \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^{n} x^2 \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} =$$

$$\sum_{x=1}^{n} \frac{xn!}{(x-1)!(n--x)!} p^x (1-p)^{n-x} =$$

$$np \sum_{y=0}^{n-1} (y+1) \left(\begin{array}{c} n-1 \\ y \end{array} \right) p^y (1-p)^{n-1-y} = np[(n-1)p+1]. \Rightarrow$$

$$Var(x) = np[(n-1)p+11] - n^2 p^2 = np\{np - p + 1 - np\} = np(1-p),$$

where $y = x - 1 = E(y+1)$ where y is $bin(n-1, p)$.

Let,

$$f(x) = \left\{ \begin{array}{ll} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, & x = 0, 1, 2, ..., n. \\ 0, & \text{otherwise.} \end{array} \right.$$

The special case of $n = 1$ is called the *Bernoulli Distribution* (which we encountered previously as the distribution of an indicator variable). For $n = 1$,

$$f(x) = \left(\begin{array}{c} 1 \\ x \end{array} \right) p^x (1-p)^{1-x}, x = 0, 1 \Big\} =$$

$$\left\{ \begin{array}{ll} 1-p, & x = 0. \\ p, & x = 1. \end{array} \right.$$

Clearly, the Binomial is the number of successes in n independent Bernoulli trials:

$$x = \sum_{j=1}^{n} I_j$$

where I_j is an indicator that trial j is a success, $x \sim b(n, p)$.

Example: Draw n cards from a standard deck of cards with replacement. Let x be the number that are Jack, Queen, King, and Ace.

$$p = \frac{4}{13} = \frac{16}{52}$$

With $n = 3$,

$$P(\text{Jack or better}) = \frac{4}{3} \left(\frac{4}{13} \right)^2 \left(\frac{9}{13} \right)^1$$

A different random variable is defined as follows. We fix the number of successes, r, and ask how many independent trials before the $r-th$ success occurs. Let y be the total number of failures before success r. $y+r$ is the number of trials to produce the $r-th$ success. Let's derive the distribution of y.

$$P(Y=y) =$$

$$P(r-1 \text{ successes on 1st } y+r-1 \text{ trials and } (y+r)\text{th try success}) =$$

$$P(r-1 \text{ success on 1st } y+r-1 \text{ trys}) \times P(\text{success on try } y+r) =$$

$$\left(\begin{array}{c} y+r-1 \\ r-1 \end{array} \right) p^{r-1}(1-p)^{y+r-1}p =$$

$$\left\{ \begin{array}{ll} \left(\begin{array}{c} y+r-1 \\ r-1 \end{array} \right) p^r(1-p)^{y+r-1} & y=0,1,2,... \\ 0, & \text{otherwise.} \end{array} \right.$$

y is called the *negative binomial* random variable. The special case of $r=1$ is called the *geometric distribution.*

$$g(y) = \left\{ \begin{array}{ll} p(1-p)^y, & y=0,1,2,... \\ 0, & \text{otherwise.} \end{array} \right.$$

Now we generalize the binomial. Suppose an experiment can result in 1 of k mutually exclusive and exhaustive outcomes: $c_1, c_2, ..., c_k$. $c_i \cap c_j = \emptyset, \cup c_i = \Omega$. Let

$$p_i = P(c_i) \Rightarrow \sum_{i=1}^{k} p_i = 1.$$

We will perform n independent trials of this experiment and count $x_1, x_2, ..., x_{k-1}$ where x_i is the number of the n trials that result in c_i. For any $x_1, x_2, ..., x_{k-1} \ni$: $\sum_{i=1}^{k-1} x_i \leq n$ and $x_i \geq 0, c = 1, 2, ..., k-1$.

$$Pr(X_1 = x_1, X_2 = x_2, ..., X_{k-1} = x_{k-1}) =$$

$$\frac{n!}{x_1!x_2!x_{k-1}!(n-\sum_{i=1}^{k-1})!} c_1 c_2 \cdots c_{k-1} c_k$$

where

$$c_1 = p_1^{x_1}, c_2 = p_2^{x_2}, ..., c_{k-1} = p_{k-1}^{x_{k-1}}, c_k = \left(1 - \sum_{i=1}^{k-1} p_i^{(n-\sum_{i=1}^{k-1} x_i)} \right)$$

$$p_1^{x_1}p_2^{x_2}\cdots p_{k-1}^{x_{k-1}}\left(1-\sum_{i=1}^{k-1}p_i^{(n-\sum_{i=1}^{k-1}x_i)}\right)=$$

$$\frac{n!}{x_1!x_2!\cdots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k},$$

$$x_k=n-\sum_{i=1}^{k-1}x_i.$$

This is called the *multivariate density.* For the special case $k=2$

$$Pr(X_1=x_1)=\frac{n!}{x_1!(n-x_1)!}p_1^{x_1}(1-p_1^{n-x_1})$$

is the *binomial distribution.* It is the pdf for the *multinomial distribution* of $(x_1,x_2,...,x_{k-1})$. When $k=3$, it is called the *trinomial distribution.* The density is

$$Pr(X_1=x_1,X_2=x_2,X_3=x_3)=$$

$$\begin{cases}\frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_3^{x_3}, & x_1\geq 0,x_2\geq 0,x_3\geq 0\\ 0, & \text{otherwise.}\end{cases}$$

$x_1+x_2+x_3=n, x_i\geq 0.$

Example: Take n draws with replacement from a standard deck of cards. Define the following random variables:
$x_1=$ number of aces.
$x_2=$ number of kings.
$x_3=$ number of queens and jacks.
$x_4=$ number of 10's, 9's, and 8's.
For $k=5$:

$$p_1=\frac{1}{13},p_2=\frac{1}{13},p_3=\frac{2}{13},p_4=\frac{3}{13},p_5=1-\sum_{i=11}^{4}p_i=\frac{6}{13}.$$

For $n=10$:

$$Pr(x_1=2,x_2=1,x_3=2,x_4=4)=$$

$$\frac{10!}{2!1!2!4!1!}\left(\frac{1}{13}\right)^2\left(\frac{1}{13}\right)^1\left(\frac{2}{13}\right)^2\left(\frac{3}{13}\right)^4\left(\frac{6}{13}\right)^1.$$

The following results can be established for the multinomial: (see homework later on):

1. The distribution of any subset of $x_1, x_2, ..., x_{k--1}$ is also multinomial.

2. Conditionals are multinomial.

10.16 The Poisson Distribution

Recall that:

$$\sum_{x=0}^{\infty} \frac{m^x}{x!} = e^m \text{ (for any } m).$$

Hence,

$$f(x) = \begin{cases} \frac{m^x e^{-m}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

is the density for a discrete random variable. Such a random variable is said to have a *Poisson Distribution.* This distribution is empirically known to accurately model many discrete time processes, i.e. the number of occurences of some phenomenon within a given time period interval. In fact, with appropriate assumptions, this density can be derived for such a situation:

1. The chance of one occurrence in a time interval is proportional to the length of the interval.

2. The chance of more than one occurences is proportional to the number of small intervals.

3. Disjoint intervals are independent.

See the Remark on pages 126-128 of the text book. We derive the mgf:

$$M(t) = E(e^{tx}) = \sum_{x=0}^{\infty} \frac{e^{tx} m^x e^{-m}}{x!} =$$

$$e^{-m} \sum_{x=0}^{\infty} \frac{(e^t m)^x}{x!} = e^{-m} e^{e^t m} = exp\{m(e^t - 1)\}.$$

$$\frac{\partial}{\partial t} M(t) = m e^t M(t) \Big|_{t=0} = M(0) m e^0 = m.$$

$$\frac{\partial^2}{\partial t^2}M(t) = m\{M(t)me^t + M(t)e^t\}_{t=0}.$$

$$m(m+1) = m^2 + m = E(x^2).$$

$$Var(x) = m^2 + m - m^2 = m.$$

Since $m = \mu$, it is common to write

$$f(x) = \begin{cases} \frac{\mu^x e^{-\mu}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{elsewhere.} \end{cases}$$

10.17 The Gamma Distribution

For $\alpha > 0$, define the *gamma distribution* as:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy.$$

Note that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$.

$$\int_0^\infty y^{\alpha-1}e^{-y}dy.$$

Let $u = y^{\alpha-1}$. Then using calculus,

$$\int v = e^{-y}.$$

$$du = (\alpha-1)y^{\alpha-2}dy.$$

$$v = e^{-y}.$$

$$\overbrace{-y^{\alpha-1}e^{-y}\Big|_0^\infty}^{=0} + \int_0^\infty (\alpha-1)y^{\alpha-2}e^{-y}dy =$$

$$(\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}dy = (\alpha-1)\Gamma(\alpha-1).$$

For $\Gamma(1) = 1$,

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) = (\alpha-1)(\alpha-2)\Gamma(\alpha-1) =$$

$$(\alpha-1)(\alpha-2)\cdots 2\Gamma(1) \text{ if } \alpha \text{ is an integer} = (\alpha-1)!.$$

In the integral, let $x = \beta y$ for some $\beta > 0$. Then,

$$y = \frac{x}{\beta},$$

$$dy = \frac{1}{\beta}dx,$$

$$\Gamma(\alpha) = \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\frac{x}{\beta}} \frac{1}{\beta} dx$$

or

$$1 = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x\beta} dx.$$

Definition: A random variable has a *gamma distribution* with parameters $\alpha > 0$ and $\beta > 0$ if it's pdf is

$$f(x) = \begin{cases} \frac{1}{\Gamma(x)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0. \\ 0, & \text{otherwise.z} \end{cases}$$

Definition: The *Gamma function is*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

If α is an integer, then $\Gamma(\alpha) = (\alpha - 1)!$.

Definition: The Gamma Distribution has the density

$$f(x) = \begin{cases} \frac{1}{\Gamma(x)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0. \\ 0, & x \leq 0. \end{cases}$$

where $\alpha > 0$ and $\beta > 0$. The mgf for the Gamma distribution is

$$M(t) = E(e^{tx}) = \int_0^\infty \frac{e^{x(t-\frac{1}{\beta})}}{\Gamma(\alpha)\beta^\alpha} = x^{-\alpha-1} dx =$$

$$\frac{1}{\beta^\alpha} \int_0^\infty \frac{1}{\Gamma(\alpha)} e^{-\frac{1}{\beta}(1-\beta t)x} x^{\alpha-1} dx =$$

$$\frac{1}{\beta^\alpha} \left[\frac{\beta}{1-\beta t}\right]^\alpha \int_0^\infty \frac{1}{\Gamma(\alpha)(\frac{\beta}{1-\beta t})^\alpha} e^{-\frac{x}{\beta/(1-betat)}} x^{\alpha-1} dx$$

The integral part is the gamma density with parameters α and $\frac{\beta}{1-\beta t}$ where

$$\frac{\beta}{1-\beta t} > 0.$$

$$\Rightarrow \beta t < 1 \Rightarrow t < \frac{1}{\beta} = (1-\beta t)^{-\alpha}, t < \frac{1}{\beta},$$

$$M(t) = (1-\beta t)^{-\alpha}, |t| < \frac{1}{\beta},$$

$$E(x) = \frac{\partial}{\partial t} M(t)\Big|_{t=0} = -\alpha(1-\beta t)^{-\alpha-1}(-\beta) =$$

$$\alpha\beta(1-\beta t)^{-\alpha-1}.$$

Given,

$$M'(0) = \alpha\beta,$$

$$E(x^2) = \frac{\partial^2}{\partial t^2} M(t)\Big|_{t=0} =$$

$$-\alpha\beta(\alpha+1)(1-\beta t)^{-\alpha-2}(-\beta)\Big|_{t=0}.$$

$$\Rightarrow M''(0) = \alpha\beta^2(\alpha+1),$$

$$\Rightarrow Var(x) = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 =$$

$$\alpha^2\beta^2 + \alpha\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

When $\alpha = 1$, the gamma distribution is called the *exponential distribution.*

$$f(x) = \begin{cases} \frac{1}{\beta}e^{-\frac{x}{\beta}}, & x > 0. \\ 0, & \text{otherwise.} \end{cases}$$

The mgf is

$$M(t) = \frac{1}{1-\beta t}.$$

It is sometimes written with ($\lambda = \frac{1}{\beta}$) as

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0. \\ 0, & \text{otherwise.} \end{cases}$$

The new mgf is

$$M(t) = \frac{\lambda}{\lambda - t}.$$

Another special case arises when $\alpha = \frac{r}{2}$, where r is an integer and $\beta = 2$. This is called the *Chi-square distribution* denoted by

$$f(x) = \begin{cases} \frac{1}{\Gamma(\frac{1}{2})2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}}, & x > 0. \\ 0, & \text{otherwise.} \end{cases}$$

r is called the *degrees of freedom.* The moment generating function is

$$M(t) = (1-2t)^{-\frac{r}{2}}, |t| < \frac{1}{2}.$$

Example: Bad luck at banks. Person A and Person B enter the bank and join separate lines of equal length simultaneously. Let x be the time til service for Person A, and y be the time til service for Person B. x and y are independent with the same exponential distribution. Consider $\frac{x}{y}$.

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)}, & x > 0, y > 0. \\ 0, & \text{otherwise.} \end{cases}$$

$Pr(\text{Person A waits at least 3 times as long as Person B}) =$

$$Pr\left(\frac{x}{y} > 3\right) = \int_0^\infty \int_{3y}^\infty \lambda^2 e^{-\lambda(x+y)} dx dy =$$

$$\int_0^\infty \lambda e^{-\lambda y} \left(-e^{-\lambda x}\right)_{3y}^\infty dy =$$

$$\int_0^\infty \lambda e^{-\lambda y} e^{-3\lambda y} dy =$$

$$\int_0^\infty \lambda e^{-4\lambda y} dy = \frac{1}{4} \overbrace{\int_0^\infty 4\lambda e^{-4\lambda y} dy}^{=1} = \frac{1}{4}.$$

Given the same problem, what is the expected ratio?

$$E\left(\frac{x}{y}\right) = \int_0^\infty \int_0^\infty \frac{x}{y} e^{-\lambda(x+y)} dx dy =$$

$$\int_0^\infty \lambda \frac{1}{y} e^{-\lambda y} \overbrace{\int_0^\infty \lambda x e^{-\lambda x} dx}^{=\frac{1}{\lambda}} dy,$$

with $\alpha = 1$, and $\beta = \frac{1}{\lambda}$,

$$\int_0^\infty \frac{1}{y} e^{-\lambda y} dy = \int_0^1 \frac{1}{y} e^{-\lambda y} dy + \int_1^\infty \frac{1}{y} e^{-\lambda y} dy \geq$$

$$\int_1^\infty \frac{1}{y} e^{-\lambda y} dy + \int_0^1 \frac{1}{y} e^{-\lambda} dy =$$

$$e^{-\lambda} \log y \Big|_0^1 \geq 0 + \cdots = \infty + \geq 0.$$

In other words $E(\frac{x}{y}) = \infty$.

The *Beta Distribution:* For $\alpha > 0$, and $\beta > 0$ define the *Beta function* by

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy.$$

Then,

$$f(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1. \\ 0, & \text{otherwise.} \end{cases}$$

is a density for a continuous random variable. Such a random variable is said to have a *Beta Distribution.*

Fact:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

(this will be proven in chapter 4 of the text book). This allows for easy calculation of the moments of the Beta Distribution.

$$E(x^k) = \frac{1}{B(\alpha, \beta)} \int_0^1 x^k x^{\alpha-1}(1-x)^{\beta-1} dx =$$

$$\frac{B(\alpha + k, \beta)}{B(\alpha, \beta)} \overbrace{\frac{1}{B(\alpha + k, \beta)} \int_0^1 x^{(\alpha+k)-1}(1-x)^{\beta-1} dx}^{=1} =$$

$$\frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+k\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} =$$

$$\frac{\Gamma(\alpha+k)}{\Gamma(\alpha+k\beta)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}.$$

Note that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$.

$$\Rightarrow \frac{(\alpha+k-1)(\alpha+k-2)(\alpha+k-3)\cdots\alpha}{(\alpha+\beta+k-1)(\alpha+\beta+k-2)\cdots(\alpha+\beta)}.$$

So, for $k=1, \Rightarrow E(x) = \frac{\alpha}{\alpha+\beta}$. For $k=2, \Rightarrow E(x^2) = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)((\alpha+\beta)} \Rightarrow$

$$Var(x) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} =$$

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Special Case: When $\alpha = \beta = 1$, the Beta Distribution is called the *Uniform Distribution.* The density is

$$f(x) = \begin{cases} 1, & 0 < x < 1. \\ 0, & \text{otherwise.} \end{cases}$$

$$E(x) = \frac{1}{2},$$

$$Var(x) = \frac{1}{12}.$$

10.18 The Normal Distribution

We have from calculus (or page 138 of the text book):

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} dy = \sqrt{2\pi}.$$

Let $x = a + by$ where $b > 0$. Then, $y = \frac{x-a}{b}$, and $dy = \frac{1}{b}dx$. Then,

$$\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}\frac{(x-a)^2}{b^2}}}{b\sqrt{2\pi}} dx = 1.$$

Definition: A random variable has a *Normal Distribution* with parameters a, b if its density is

$$f(x) = \left\{ \begin{array}{l} \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-a)^2}{b^2}}, \quad -\infty < x < \infty. \end{array} \right.$$

Let's find the mgf:

$$E(e^{tx}) = \frac{1}{\sqrt{2\pi}} \frac{1}{b} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}\frac{(a-b)^2}{b^2}} dx =$$

$$\frac{1}{\sqrt{2\pi}} \frac{1}{b} \int_{-\infty}^{\infty} e^{-\frac{1}{2b^2}[x^2-2ax+a^2-2b^2tx]} dx =$$

$$\frac{1}{b} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2b^2}[x^2-2ax+a^2-2b^2tx]} dx =$$

$$\frac{1}{b} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2b^2}[(x-(a+b^2t))^2+a^2-(a+b^2t)^2]} dx =$$

$$e^{-\frac{1}{2b^2}[-2ab^2t-b^4t^2]} \overbrace{\frac{1}{b\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2b^2}[x-(a+b^2t)]^2} dx}^{=1}$$

We have the *Normal Density* with parameters $(a + b^2t)$ and $b = 1$.

$$M(t) = e^{at+\frac{b^2t^2}{2}}.$$

$$E(x) = \frac{\partial}{\partial t} M(t) \Big|_{t=0} = M(t)[a + b^2t] = a + 0 = a.$$

$$E(x^2) = \frac{\partial^2}{\partial^2 t} M(t) \Big|_{t=0} = M(t)[a + b^2t]^2 + M(t)b^2 \Big|_{t=0} = a^2 + b^2.$$

$$\Rightarrow Var(x) = a^2 + b^2 - a^2 = b^2.$$

$$\Rightarrow \mu = a, \sigma^2 = b^2.$$

Definition(another): A random variable is said to have a *Normal Distribution* with mean μ and variance σ^2 if the density is

$$f(x) = \left\{ \begin{array}{l} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty. \end{array} \right.$$

Notation: If $x \sim N(\mu, \sigma^2)$, then the mgf is

$$e^{(\mu t + \frac{1}{2}\sigma^2 t^2)}$$

Theorem 1: If x is $N(\mu, \sigma^2)$, then

$$w = \frac{x - \mu}{\sigma}$$

is $N(0, 1)$. Proof: The mgf of w is

$$E(e^{tw}) = E(e^{t\frac{(x-\mu)}{\sigma}}) =$$

$$E(e^{\frac{tx}{\sigma}} e^{-\frac{\mu t}{\sigma}}) = e^{-\frac{\mu t}{\sigma}} E(e^{\frac{tx}{\sigma}}) = e^{-\frac{\mu t}{\sigma}} E(e^{t^* x}),$$

where $t^* = \frac{t}{\sigma}$. Then,

$$= e^{(-\frac{\mu t}{\sigma} + \frac{\mu t}{\sigma} + \frac{1}{2}\sigma^2 \frac{t^2}{\sigma^2})} = e^{\frac{t^2}{2}}$$

which is the mgf of $N(0, 1)$.

Consequences: For any set A,

$$Pr\left(w \in \frac{A - \mu}{\sigma}\right) = Pr(w \in A^*)$$

where A^* is the set A shifted by μ and then all elements divided by σ. $N(0, 1)$ is called the *standard normal distribution.*

Theorem 2: If $x \sim N(0, 1)$, then

$$v = \frac{(x - \mu)^2}{\sigma^2} \sim \chi^2(1).$$

Proof: Note $v = w^2$ where $w \sim N(0, 1)$. The distribution function of $v, v > 0$ is

$$Pr(V \leq v) = Pr(w^2 \leq V) =$$

$$Pr(-\sqrt{v} \leq w \leq \sqrt{v}) =$$

by symmetry of $N(0, 1)$,

$$2 \times Pr(0 \leq w \leq \sqrt{v}) =$$

$$2\frac{1}{\sqrt{2\pi}}\int_0^{\sqrt{v}} e^{-\frac{1}{2}w^2}dw$$

Let $y = w^2, \Rightarrow w = \sqrt{y}, dw = \frac{1}{2\sqrt{y}}dy,$

$$= \frac{1}{\sqrt{2\pi}}\int_0^{v} e^{-\frac{1}{2}y}y^{\frac{1}{2}-1}dy$$

Then, the density for v is

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y}y^{\frac{1}{2}-1}, & y > 0. \\ 0, & y \leq 0. \end{cases}$$

which is the *Gamma Density* with $\alpha = \frac{1}{2}, \beta = 2$, i.e. $\chi^2(1)$.

10.19 Bivariate Normal Distribution

Definition: x, y are bivariate normal random variables if their joint density is

$$f(x,y) = \begin{cases} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-[-\frac{1}{2(1-\rho^2)}(\frac{x-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})\frac{(y-\mu_2)}{\sigma_2} + (\frac{y-\mu_2}{\sigma^2})^2]}, \\ -\infty < x < \infty, -\infty < y < \infty. \end{cases}$$

We will show several properties of $f(x,y)$ and of the distribution by a certain factorization. Consider inside the []. It can be written

$$\left\{\frac{y-\mu_2}{\sigma_2} - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right\}^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2 =$$

$$\left(\frac{y-bx}{\sigma_2}\right)^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2$$

where $bx = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$. So,

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e - \frac{1}{2(1-\rho^2)}[(\frac{y-bx}{\sigma_2})^2 + (1-\rho^2)(\frac{x-\mu_1}{\sigma_1})^2]$$

$$= \overbrace{\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}}}^{\text{Density for } N(\mu_1,\sigma_1^2)} \times \overbrace{\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-bx}{\sqrt{1-\rho^2}\sigma_2})^2}\frac{1}{\sqrt{1-\rho^2}}}^{\text{for fixed } x,\ N(bx,(1-\rho^2)\sigma_2^2)}$$

Therefore,

1. $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx =$$

$$\int_{-\infty}^{\infty}\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}}dx\overbrace{\int_{-\infty}^{\infty}\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-bx}{\sqrt{1-\rho^2}\sigma_2})^2}\frac{1}{\sqrt{1-\rho^2}}dy}^{=1}$$

$$\int_{-\infty}^{\infty}\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}}dx = 1$$

So, we have verified $f(x,y)$ is a density.

2. From (1),

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy =$$

$$\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}}\overbrace{\int_{-\infty}^{\infty}\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-bx}{\sqrt{1-\rho^2}\sigma_2})^2}\frac{1}{\sqrt{1-\rho^2}}dy}^{=1} =$$

$$\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma_1^2}}, -\infty < x < \infty.$$

$x \sim N(\mu_1, \sigma_1^2)$. Likewise, $y \sim N(\mu_2, \sigma_2^2)$.

3. Based on (2), we now see that factorization has this form $f(x,y) = f_1(x)h(x,y)$

$$\Rightarrow h(x,y) = \frac{f(x,y)}{f_1(x)} \Rightarrow h(x,y) = f(y|x)$$

$$\Rightarrow y|X = x \sim N\left(\frac{\overbrace{\mu_2 + \rho\sigma_2(x-\mu_1)}^{bx}}{\sigma_1}, \sigma_2^2(1-\rho^2)\right).$$

Likewise,

$$X|Y = y \sim N\left(\mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y-\mu_2), (1-\rho^2)\sigma_1^2\right).$$

$$M(t_1,t_2)=E(e^{t_1x+t_2y})=$$

$$\int\int e^{t_1x+t_2y}f(x,y)dxdy=$$

$$\int e^{t_1x}f_1(x)\int e^{t_2y}f_2(y|x)dydx$$

where

$$\int e^{t_2y}f_2(y|x)dy$$

is the mgf of $N(\mu_2+\frac{\rho\sigma_2}{\sigma_1}(x-\mu_1),(1-\rho^2)\sigma_2^2)$.

$$\exp\left\{\mu_2t_2+\frac{\rho\sigma_2}{\sigma_1}(x-\mu_1)t_2+\frac{1}{2}(1-\rho^2)\sigma_2^2t_2^2\right\}=$$

$$\exp\left\{\mu_2t_2+\frac{1}{2}(1-\rho^2)\sigma_2^2t_2^2-\frac{\rho\sigma_2}{\sigma_1}\mu_1t_2\right\}\times$$

$$\int\exp\left\{t_1x+\frac{\rho\sigma_2}{\sigma_1}xt_2\right\}f_1(x)dx=$$

$$\exp\left\{\mu_2t_2+\frac{1}{2}(1-\rho^2)\sigma_2^2t_2^2-\frac{\rho\sigma_2}{\sigma_1}\mu_1t_2\right\}\times$$

$$\overbrace{\int\exp\left\{(t_1+\frac{\rho\sigma_2}{\sigma_1}t_2)x\right\}f_1(x)dx}^{M(t_1+\frac{\rho\sigma_2t_2}{\sigma_1})}$$

where $M(t)$ is the mgf of $N(\mu_1,\sigma_1^2)$.

$$=\exp\Big\{-\frac{\rho\sigma_2}{\sigma_1}\mu_1t_1+\mu_2t_2+\frac{1}{2}(1-\rho^2)\sigma_2^2t_2^2$$

$$+\mu_1[t_1+\frac{\rho\sigma_2}{\sigma_1}t_2]+\frac{1}{2}\sigma_1^2[t_1+\frac{\rho\sigma_2t_2}{\sigma_1}]^2\Big\}=$$

$$\exp\Big\{\mu_1t_1+\mu_2t_2+\frac{1}{2}(1-\rho^2)\sigma_2^2t_2^2+\frac{1}{2}\sigma_1^2t_1^2+$$

$$\frac{\sigma_1^2\rho\sigma_2}{\sigma_1}t_1t_2+\frac{1}{2}\frac{\sigma_1^2\rho^2t_2^2}{\sigma_1^2}\Big\}=$$

$$\exp\left\{(\mu_1 t_1 + \frac{\sigma_1^2 t_1^2}{2}) + (\mu_2 t_2 + \frac{1}{2}\sigma_2^2 t_2^2) + \rho\sigma_1\sigma_2 t_1 t_2\right\} =$$

$$M(t_1, 0)M(0, t_2)\exp\left\{\rho\sigma_1\sigma_2 t_1 t_2\right\}$$

Find

$$E(x, y) = \frac{\partial^2}{\partial t_1 \partial t_2} M(t_1, t_2)\Big|_{t_1=t_2=0} =$$

$$\frac{\partial}{\partial t_1}\left\{M(t_1, t_2)[\mu_2 + \sigma_2^2 t_2 + \rho\sigma_1\sigma_2 t_1]\Big|_{t_1=t_2=0}\right\} =$$

$$M(t_1, t_2)[\mu_1 + \sigma_1^2 t_1 + \rho\sigma_1\sigma_2 t_2][\mu_2 + \sigma_2^2 t_2 + \rho\sigma_1\sigma_2 t_1] + M(t_1, t_2)\rho\sigma_1\sigma_2 =$$

$$1[\mu_1][\mu_2] + 1\rho\sigma_1\sigma_2.$$

$$\Rightarrow Cov(x, y) = E(xy) - E(x)E(y) =$$

$$\mu_1\mu_2 + \rho\sigma_1\sigma_2 - \mu_1\mu_2 = \rho\sigma_1\sigma_2.$$

$$\Rightarrow Corr(x, y) = \frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_2} = \rho.$$

Theorem 3: Let x, y be bivariate normal. Then they are independent iff they are uncorrelated (i.e. iff $\rho = 0$). Proof: Independence implies that the correlation is zero. Suppose x, y have $\rho = 0$. $M(t_1, t_2) = M(t_1, 0)M(0, t_2)$. That implies independence.

10.20 Distribution of Functions of RVs

Let $x_1, x_2, ..., x_n$ be random variables and let $y = u(x_1, x_2, ..., x_n)$ be a function of $x_1, x_2, ..., x_n$.

Definition: A function of one or more random variables that does not depend on any unknown parameter is called a *statistic.*

Example:

1. $y = \sum_{i=1}^{n} \frac{x_i}{n}$ is a statistic called the *sample mean* denoted by $\bar{x}$.

2. $y = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$

is a statistic called the *variance* of the sample and denoted as s^2.

3. $y = \sum_{i=1}^{n}(x_i - E(x_i))^2$

is not a statistic unless each $E(x_i)$ is known.

We generally think of statistics in the following setting: a sample $x_1, x_2, ..., x_n$ of size n from some population has been obtained. For purposes of description, estimating, testing, etc, statistics such as $\bar{x}$ and s^2 are calculated. Often we have a *random sample* defined as follow:

Definition: $x_1, x_2, ..., x_n$ is said to be a *random sample* if they are mutually independent and each x_i has the same distribution.

In practice, selection of $x_i's$ randomly from a given population ensures that the requirements for a random sample hold. Such $x_1, x_2, ..., x_n$ are also said to be iid (identically and independently distributed). In this chapter of the text book, we study the techniques for finding the Distribution of $y = u(x_1, x_2, ..., x_n)$ assuming we know the distribution of $x_1, x_2, ..., x_n$. The simplest (and perhaps least powerful) technique is the *distribution function technique* (used in proof of theorem 2, chapter 3 of the text book). Find the distribution function of

$$y = G(y) = Pr(Y \leq y) = Pr(u(x_1, x_2, ..., x_n) \leq y)$$

and simplify.

Example: (a new proof of Theorem 1, in Chapter 3 of the text book). $x \sim N(0, 1)$, and $y = ax + b$ where $a, b > 0$ are constants. Find the distribution of y.

$$G(y) = Pr(Y \leq y) = Pr(ax + b \leq y) =$$

$$Pr\left(x \leq \frac{y-b}{a}\right) = \int_{-\infty}^{\frac{y-b}{a}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx.$$

Let $y^* = ax + b$. Then,

$$x = \frac{y^* - b}{a},$$

$$dx = \frac{1}{a} dy^*,$$

Then,

$$\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}a} e^{-\frac{1}{2}(\frac{y^*-b}{a})^2} dy^*$$

which is the density for $N(b, a^2)$. Then, $y \sim N(b, a^2)$.

10.21 1:1 Transformations

Suppose we have two sets A and B, and the function $u : A \to B$.

Definition: u is 1:1 if $a_1, a_2 \in A, a_1 \neq a_2 \Rightarrow u(a_1) \neq u(a_2)$. i.e. each $a \in A$ has a different image under u.

Definition: u is 'onto' if $b \in B \Rightarrow \exists a \in A \ni: u(a) = b$. i.e. each $b \in B$ is the image of some $a \in A$.

Definition: u is a *1:1 transformation* if it is both 1:1 and onto. In this case, each $a \in A$ is uniquely associated with exactly one $b \in B$ and vice-versa. Likewise, each subset of a is uniquely associated with a subset of B, and vice-versa.

Example: $A = [-1, 1], B = [-3, 3]$, and $u(a) = 3a$ is a 1:1 transformation.

Example: $A = [-1, 1], B = [0, 1]$, and $u(a) = a^2$ is not a 1:1 transformation.

Example: $A = [-1, 1], B = [0, 1]$, and $u(a) = a + 1$ is not unto but a 1:1 transformation.

A lecture is missing here.

Fact: If u is a 1:1 transformation where $\exists w : B \to A \ni w(u(a)) = a, \forall a \in A$ and $u(w(b)) = b, \forall b \in B$, then w is called the *inverse* of u.

10.22 Change of Variables Technique

Example:

$$h(x,y) = \begin{cases} 2(1-x-y+2xy), & 0<x<1. \\ 0, & \text{otherwise.} \end{cases}$$

Find the distribution of $u = x - y$. Define $w = x$ for the second variable. $\Rightarrow x = w \Rightarrow y = w - u$.

$$J = \begin{vmatrix} 1 & -1 \\ 1 & 0 \end{vmatrix} = 1.$$

$$g(u,w) = h(w, w-u) \times 1 = 2(1 - w - (w-u) + 2w(w-u)) =$$

$$\begin{cases} 2(1+u-2wu-2w+2w^2), & 0<w<1, -1<u<1. \\ 0, & \text{otherwise.} \end{cases}$$

$$\begin{array}{lll} x = w & 0 < w < 1 & u < w \\ y = w - u & \Rightarrow \quad 0 < w - u < 1 & \Rightarrow \quad w < u + 1 \end{array} \Rightarrow$$

Finally, find the marginal for u.

$$g_1(u) = \int g(u,w)dw,$$

if $u < 0$,

$$g_1(u) = \int_0^{1+u} 2(1+u-2wu-2w+2w^2)dw =$$

$$2\left(w + uw - w^2u - w^2 + \frac{2w^3}{3}\right)\Big|_0^{1+u} =$$

$$2(1+u+u(1+u) - (1+u)^2u - (1+u)^2 + \frac{2(1+u)^3}{3}) =$$

$$\frac{4}{3} + 2u - \frac{2u^3}{3}$$

If $u \geq 0$, then

$$g_1(u) = \int_u^1 g(u,w)dw =$$

$$\int_u^1 2(1+u-2wu-2w+2w^2)dw = \frac{4}{3}-2u+\frac{2u^3}{3}.$$

Therefore,

$$\begin{cases} \frac{4}{3}+2u-\frac{2u^3}{3}, & -1<u<0. \\ \frac{4}{3}-2u+\frac{2u^3}{3}, & 0\le u<1. \\ 0, & \text{otherwise.} \end{cases}$$

10.23 The Beta, T, and F Distributions

Example: Let x_1, x_2 be independent gamma random variables with the joint density

$$h(x_1,x_2)=\begin{cases} \frac{1}{\Gamma(\alpha)\Gamma(\beta)}x_1^{\alpha-1}x_2^{\beta-1}e^{-x_1-x_2}, & x_1>0, x_2>0. \\ 0, & \text{otherwise.} \end{cases}$$

Find the marginal distribution of $y_1 = x_1+x_2$ and $y_2=\frac{x_1}{x_1+x_2}$. Given that, then solve for x_1 and x_2 :

$$x_1 = y_1y_2,$$

$$x_2 = y_1 - y_1y_2.$$

Note that the $x's$ are put in terms of the $y's$.

$$J=\begin{vmatrix} y_2 & y_1 \\ 1-y_2 & -y_1 \end{vmatrix} = -y_1y_2 - y_1(1-y_2) = -y_1$$

$$g(y_1,y_2) = h(y_1y_2, y_1-y_1y_2)|-y_1|,$$

$$\begin{matrix} x_1\ge 0 \\ x_2\ge 0 \end{matrix} \Rightarrow \begin{matrix} y_1y_2>0 \\ y_1-y_1y_2>0 \end{matrix} \Leftrightarrow \begin{matrix} y_1>0 \\ y_1>y_1y_2 \end{matrix} \Rightarrow \ 1>y_2>0$$

$$g(y_1,y_2)=\frac{1}{\Gamma(\alpha)\Gamma(\beta)}(y_1y_2)^{\alpha-1}(y_1-y_1y_2)^{\beta-1}e^{-y_1}y_1 =$$

$$\begin{cases} \frac{1}{\Gamma(\alpha)\Gamma(\beta)}y_1^{\alpha+\beta-1}e^{-y_1}y_2^{\alpha-1}(1-y_2)^{\beta-1}, & y_1>0, 0<y_2<1. \\ 0, & \text{otherwise.} \end{cases}$$

Find the marginal densities.

$$g_2(x_2)=\frac{1}{\Gamma(\alpha)\Gamma(\beta)}y_2^{\alpha-1}(1-y_2)^{\beta-1}\overbrace{\int_0^\infty y_1^{\alpha+\beta-1}e^{-y_1}dy_1}^{\Gamma(\alpha+\beta)} =$$

$$\begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha-1}(1-y_2)^{\beta-1}, & 0 < y_2 < 1. \\ 0, & \text{otherwise.} \end{cases}$$

which is the Beta distribution.

$$\Rightarrow B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Since y_1 and y_2 are independent,

$$g_1(y_1) = \frac{g(y_1,y_2)}{g_2(y_2)} =$$

$$\begin{cases} \frac{1}{\Gamma(\alpha+\beta)} y_1^{\alpha+\beta-1} e^{-y_1}, & y_1 > 0. \\ 0, & \text{otherwise.} \end{cases}$$

which is the Gamma distribution with parameters $\alpha+\beta, 1$. Hence,

1. y_1 and y_2 are independent.
2. y_1 is $Gamma(\alpha+\beta, 1)$, the sum of 2 gammas with the second parameter as 1 is another gamma.
3. y_2 is $Beta(\alpha,\beta)$.
4. It has been proven that $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Example: Let $w \sim N(0,1)$ and $v \sim \chi^2(r)$. Assume that w and v are independent. Find the distribution of $T = \frac{w}{\sqrt{\frac{v}{r}}}$. Let's use $u = v$ for the second variable. Then,

$$v = u,$$

$$w = T\sqrt{\frac{u}{r}},$$

$$\begin{vmatrix} 1 & 0 \\ \frac{t}{2\sqrt{ru}} & \sqrt{\frac{u}{r}} \end{vmatrix} = \sqrt{\frac{u}{r}}.$$

$$h(v,w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2} \frac{1}{\Gamma(\frac{r}{2}) 2^{\frac{r}{2}}} v^{\frac{r}{2}-1} e^{-\frac{v}{2}}, v > 0, -\infty < w < \infty.$$

$$g(t,u) = h\left(u, t\sqrt{\frac{u}{r}}\right)\sqrt{\frac{u}{r}}, u > 0, -\infty < t < \infty.$$

$$g(t,u) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}t^2\frac{u}{r}\right\}\frac{1}{\Gamma(\frac{r}{2})2^{\frac{r}{2}}}u^{\frac{r}{2}-1}e^{-\frac{u}{2}}\sqrt{\frac{u}{r}} =$$

$$\frac{1}{\sqrt{2\pi}\Gamma(\frac{r}{2})2^{\frac{r}{2}}\sqrt{r}}\exp\left\{-\frac{u}{2}\left[\frac{t^2}{r}+1\right]\right\}u^{\frac{r-1}{2}}$$

$$g_1(t) = \frac{1}{\sqrt{2\pi}\Gamma(\frac{r}{2})2^{\frac{r}{2}}}\int_0^u u^{(\frac{r+1}{2})-1}\exp\left\{\frac{-u}{2/(\frac{t^2}{r}+1)}du\right\} =$$

$$\left[\frac{2}{\frac{t^2}{r}+1}\right]^{r+1}\Gamma\left(\frac{r+1}{2}\right)\overbrace{\int_0^\infty \frac{1}{\Gamma(\frac{r+1}{2})[2/(\frac{t^2}{r}+1)]^{\frac{r+1}{2}}}\exp\left\{\frac{2}{\frac{t^2}{r}+1}\right\}du}^{=1} =$$

$$\frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma(\frac{r}{2})}\frac{1}{\sqrt{\pi r}}\left(1+\frac{t^2}{r}\right)^{-\frac{r+1}{2}}, -\infty < t < \infty$$

which is the density for the t distribution with r degrees of freedom which arise as a distribution. $N(0,1)$ divided by the square root of an independent χ^2 over it's degrees of freedom.

Example: Let u and v be independent χ^2 random variables with r_1 and r_2 degrees of freedom. The joint density of u, v is

$$h(u,v) = \frac{1}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})2^{\frac{r_1+r_2}{2}}}u^{\frac{r_1}{2}-1}v^{\frac{r_2}{2}-1}e^{-\frac{1}{2}(u+v)}.$$

Find the distribution of

$$F = \frac{\frac{u}{r_1}}{\frac{v}{r_2}}, u > 0, v > 0.$$

Define,

$$z = v,$$

Then,

$$v = z,$$

$$u = \frac{r_1}{r_2}Fz,$$

$$J = \left| \begin{array}{cc} 0 & 1 \\ z\frac{r_1}{\sqrt{r_2}} & z\frac{r_1}{r_2} \end{array} \right| = \left| -z\frac{r_1}{r_2} \right|.$$

$$g(F,z) = h\left(\frac{r_1}{r_2}Fz, z\right)\frac{r_1}{r_2}z =$$

$$\frac{1}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})2^{\frac{r_1+r_2}{2}}}\left(\frac{r_1}{r_2}\right)\left(\frac{r_1}{r_2}Fz\right)^{\frac{r_1}{2}-1} z^{\frac{r_2}{2}-1}e^{-\frac{1}{2}z(\frac{r_1}{r_2}F+1)}z,$$

$$\left.\begin{array}{l} \frac{r_1}{r_2}Fz > 0 \\ z > 0 \end{array}\right\} \Rightarrow \begin{array}{l} F > 0 \\ z > 0 \end{array}$$

$$= \frac{1}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})2^{\frac{r_1+r_2}{2}}}\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{r_2}} F^{\frac{r_1}{2}-1}z^{\frac{r_1+r_2}{2}-1}\exp\left\{\frac{-z}{[2(\frac{1}{\frac{r_1}{r_2}F+1})]}\right\} \Rightarrow$$

$$g_1(f) =$$

$$\Gamma(\frac{r_1+r_2}{2})\frac{1}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})2^{\frac{r_1+r_2}{2}}}\left(\frac{r_2}{r_1}\right)^{\frac{r_1}{2}} F^{\frac{r_1}{2}-1}\times$$

$$\overbrace{\int_0^\infty \frac{z^{\frac{r_1+r_2}{2}-1}}{\Gamma(\frac{r_1+r_2}{2})[\frac{2}{\frac{r_1}{r_2}F+1}]^{\frac{r_1+1}{2}}}e^{-z/2(\frac{1}{\frac{r_1}{r_2}F+1})}dz}^{=1} =$$

$$\left\{\begin{array}{ll} \frac{\Gamma(\frac{r_1+r_2}{2})}{\Gamma(\frac{r_1}{2})(\Gamma(\frac{r_2}{2})}\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} F^{\frac{r_1}{2}-1}\left(\frac{r_1}{r_2}F+1\right)^{-\frac{r_1+r_2}{2}}, & F > 0. \\ 0, & \text{otherwise.} \end{array}\right.$$

which is the density of the F distribution with r_1 and r_2 degrees of freedom.

Fact: If T has a t distribution with r degrees of freedom, then $F = t^2$ has an F distribution with 1 and r degrees of freedom. The proof is left for homework.

10.23.1 Extensions

Let $x_1, x_2, ..., x_n$ be continuous random variables with joint density $h(x_1, x_2, ..., x_n)$ and n dimensional space $A = \{(x_1, ..., x_n)\} \rightarrow: h(x_1, x_2, ..., x_n) > 0$. Let $y_i = u_i(x_1, x_2, ..., x_n), i = 1, 2, ..., n$ with n dimensional space. ß be such that the u_i's are a 1:1 transformation from A to ß. Hence, $\exists w_i, i = 1, 2, .., n$

such that $x_i = w_i(y_1, y_2, ..., y_n)$ and the w_i's are a 1:1 transformation from ß to A. The first partials of the inverse functions are continuous i.e. $\frac{dx_i}{dy_i}$ are continuous for all i, j, and the Jacobian,

$$J = \begin{vmatrix} \frac{dx_1}{dy_1} & \frac{dx_1}{dy_2} & \frac{dx_1}{dy_3} & \cdots & \frac{dx_1}{dy_n} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \frac{dx_n}{dy_1} & \frac{dx_n}{dy_2} & \frac{dx_n}{dy_3} & \cdots & \frac{dx_n}{dy_n} \end{vmatrix}$$

$\forall (y_1, y_2, ..., y_n) \in$ ß. Then, the joint density $g(y_1, y_2, ..., y_2)$ of $y_1, y_2, ..., y_n$ is

$$g(y_1, y_2, ..., y_n) = h(w_1(y_1, ..., y_n), w_2(y_1, ..., y_n), ..., w_n(y_1, ..., y_n))|J|$$

Example: Let the random variables x_1, x_2, x_3 be iid $N(0, 1)$. Find the joint density of $y_1 = x_1 + x_2$, $y_2 = x_1 + x_2 + x_3$ and $y_3 = x_1 - x_2$. Then,

$$x_1 = \frac{y_1 + y_3}{2},$$

$$x_2 = \frac{y_1 - y_3}{2},$$

$$x_3 = y_2 - y_1.$$

The Jacobian is

$$J = \begin{vmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ -1 & 1 & 0 \end{vmatrix} \begin{matrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 0 \\ -1 & 1 \end{matrix} =$$

$$0 + 0 + \frac{1}{4} - 0 - -\frac{1}{4} - 0 = \frac{1}{2}.$$

$$h(x_1, x_2, x_3) = \frac{1}{(2\pi)^{\frac{3}{2}}} \exp\left\{-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right\},$$

$$\left.\begin{array}{l} -\infty < \frac{y_1+y_3}{2} < \infty \\ -\infty < \frac{y_1-y_3}{2} < \infty \\ -\infty < y_2 - y_1 < \infty \end{array}\right\} \Leftrightarrow -\infty < y_i < \infty, i = 1, 2, 3$$

$$\Rightarrow g(y_1, y_2, y_3) = \frac{1}{2} h\left(\frac{y_1 + y_3}{2}, \frac{y_1 - y_3}{2}, y_2 - y_1\right) =$$

$$\frac{1}{2(2\pi)^{\frac{3}{2}}}\exp\left\{-\frac{1}{2}\left(\left(\frac{y_1+y_3}{2}\right)^2+\left(\frac{y_1-y_3}{2}\right)^2+(y_2-y_1)^2\right)\right\}=$$

$$\frac{1}{2(2\pi)^{\frac{3}{2}}}\exp\left\{-\frac{1}{2}\left(\frac{y_3^2}{2}+\frac{3y_1^2}{2}-2y_1y_2+y_2^2\right)\right\}=$$

$$\frac{1}{\sqrt{2\pi}\sqrt{2}}\exp\left\{-\frac{1}{2}\left(\frac{y_3}{\sqrt{2}}\right)^2\right\}\frac{1}{\sqrt{2}(2\pi)}\exp\left\{-\frac{1}{2}\left(\frac{3y_1^2}{2}-2y_1y_2+y_2^2\right)\right\}$$

where

$$\overbrace{\frac{1}{\sqrt{2}(2\pi)}\exp\left\{-\frac{1}{2}\left(\frac{3y_1^2}{2}-2y_1y_2+y_2^2\right)\right\}}^{\text{Bivariate Normal}}$$

and

$$\overbrace{\frac{1}{\sqrt{2\pi}\sqrt{2}}\exp\left\{-\frac{1}{2}\left(\frac{y_3}{\sqrt{2}}\right)^2\right\}}^{N(0,1)}.$$

The bivariate normal exponent can be written as

$$-\frac{3}{2}\left[\left(\frac{y_1-0}{\sqrt{2}}\right)^2-\frac{4}{\sqrt{6}}\left(\frac{y_1-0}{\sqrt{2}}\right)\left(\frac{y_2-0}{\sqrt{3}}\right)+\left(\frac{y_2-0}{\sqrt{3}}\right)^2\right].$$

So, put $\sigma_1=\sqrt{2},\sigma_2=\sqrt{3},\rho=\frac{2}{\sqrt{6}}=\sqrt{\frac{2}{3}},\mu_1,\mu_2=0$. Then it is a bivariate normal.

Example: Let $x_1,x_2,...,x_n$ be independent and continuous random variables with densities $h_1(x_1),x_1\in A_1;h_2(x_2),x_2\in A_2;...;h_n(x_n),x_n\in A_n$. Define $y_i=u_i(x_i),i=1,2,...,n$ where $u_i,i=1,2,...,n$ is invertible. Then, $x_i=w_i(y_i),i=1,2,...,n$

$$J=\begin{vmatrix}\frac{dw_1}{dy_1} & 0 & 0 & \cdots & 0\\ 0 & \frac{dw_2}{dy_2} & 0 & \cdots & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \frac{dw_n}{dy_n}\end{vmatrix}=\left|\frac{dw_1}{dy_1}\right|\left|\frac{dw_2}{dy_2}\right|\cdots\left|\frac{dw_n}{dy_n}\right|.$$

$$h(x_1, x_2, ..., x_n) = h_1(x_1)h_2(x_2) \cdots h_n(x_n),$$

$$g(y_1, y_2, ..., y_n) = h(w_1(y_1), w_2(y_2), ..., w_n(y_n))|J| =$$

$$h(w_1(y_1))h(w_2(y_2)) \cdots h(w_n(y_n)) \left|\frac{dw_1}{dy_1}\right| \left|\frac{dw_2}{dy_2}\right| \cdots \left|\frac{dw_n}{dy_n}\right|,$$

where $y_1 \in ß_1, y_2 \in ß_2, ..., y_n \in ß_n$, and $ß_i = u_i(a_i)$.

$$= h_1(w_1(y_1)) \left|\frac{dw_1}{dy_1}\right| h_2(w_2(y_2)) \left|\frac{dw_2}{dy_2}\right| \cdots h_n(w_n(y_n)) \left|\frac{dw_n}{dy_n}\right|$$

$\Rightarrow y_1, y_2, ..., y_n$ are independent. Functions of independent random variables are also independent. Suppose we want to consider functions which are not 1:1 transformations from A to ß i.e. more than one point of A is mapped to the same point in ß. We can still use the change of variables technique if we can partition A into disjoint subsets

$$A = \bigcup_{i=1}^{k} A_i,$$

where the function from A_i to ß is a 1:1 transformation. The new density is

$$g(y_1, y_2, ..., y_n) = \sum_{i=1}^{k} |J_i| h(w_{1i}(y_1, ..., y_n), w_{2i}(y_1, ..., y_n), ..., w_{ni}(y_1, ..., y_n))$$

where $w_{1i}, ..., w_{ni}$ are the inverse of the transformation from A_i to $ß_i$ and J_i is the corresponding Jacobian.

Example:

$$h(x) = \begin{cases} \frac{1}{2}\cos(x), & -\frac{\pi}{2} < x < \frac{\pi}{2}. \\ 0, & \text{otherwise.} \end{cases}$$

$$y = |x|,$$

$$A = \left(-\frac{\pi}{2}, \frac{\pi}{2}\right),$$

$$ß = \left[0, \frac{\pi}{2}\right].$$

Let

$$A_1 = \left(0, \frac{\pi}{2}\right)$$

$$A_2 = \left(-\frac{\pi}{2}, 0\right)$$

Under $A_1 : y = x$, and $x = y, J = 1$. Under $A_2 : y = -x$ and $x = -y, J = -1$.

$$g(y) = \sum_{i=1}^{2} |J_i| h(w_i(y_i)) =$$

$$\frac{1}{2}\cos(y) + \frac{1}{2}\cos(-y) =$$

$$\begin{cases} \cos(y), & 0 < y < \frac{\pi}{2}. \\ 0, & \text{otherwise.} \end{cases}$$

10.24 Extension of Substitution of Variables

Suppose we do not have a 1:1 transformation.

Example: Let x_1, x_2 be iid random variables that are $N(0, 1)$.

$$h(x_1, x_2) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x_1^2 + x_2^2)\right\}$$

Let $y_1 = x_1^2, y_2 = |x_1 + x_2|$

$$A = \{(x_1, x_2) \ni: -\infty < x_1 < \infty, -\infty < x_2 < \infty.\}$$

$$ß = \{(y_1, y_2) \ni: 0 < y_1 < \infty, 0 < y_2 < \infty\}$$

Example: $(1, 0)$ and $(-1, 0)$ map to $(1, 1)$. Let

$$A_1 = (x_1 > 0, x_1 + x_2 > 0),$$

$$A_2 = (x_1 > 0, x_1 + x_2 < 0),$$

$$A_3 = (x_1 < 0, x_1 + x_2 > 0),$$

$$A_4 = (x_1 < 0, x_1 + x_2 < 0).$$

Then, under A_1 :

$$x_1 = \sqrt{y_1}, x_2 = y_2 - \sqrt{y_1},$$

under A_2 :

$$x_1 = \sqrt{y_1}, x_2 = -y_2 - \sqrt{y_1},$$

under A_3 :

$$x_1 = -\sqrt{y_1}, x_2 = y_2 + \sqrt{y_1},$$

under A_4 :

$$x_1 = -\sqrt{y_1}, x_2 = -y_2 + \sqrt{y_1}.$$

The Jacobian under A_1 is

$$J_1 = \begin{vmatrix} \frac{1}{2\sqrt{y_1}} & 0 \\ -\frac{1}{2\sqrt{y_1}} & 1 \end{vmatrix} = \frac{1}{2\sqrt{y_1}},$$

$$J_2 = J_3 = J_4 = \left| \frac{1}{2\sqrt{y_1}} \right|.$$

$$g(y_1, y_2) = \frac{1}{2\pi} \frac{1}{2\sqrt{y_1}} \exp\left\{ -\frac{1}{2} \left(y_1 + (y_2 - \sqrt{y_1})^2 \right) \right\} +$$

$$\exp\left\{ -\frac{1}{2} (y_1 + (-y_2 - \sqrt{y_1})^2) \right\} +$$

$$\exp\left\{ -\frac{1}{2} (y_1 + (y_2 + \sqrt{y_1})^2) \right\} +$$

$$\exp\left\{ -\frac{1}{2} (y_1 + (-y_2 + \sqrt{y_1})^2) \right\} =$$

$$\frac{1}{2\pi} \frac{1}{2\sqrt{y_1}} \left\{ 2 \exp\left\{ -\frac{1}{2} [y_1 + (y_2 - \sqrt{y_1})^2] \right\} + \right.$$

$$\left. 2 \exp\left\{ -\frac{1}{2} [y_1 + (y_2 + \sqrt{y_1})^2] \right\} \right\},$$

$$g(y_1, y_2) = \frac{e^{-\frac{1}{2} y_1}}{2\pi y_1} \left\{ \exp\left[-\frac{1}{2} (y_2 - \sqrt{y_1})^2 \right] + \right.$$

$$\left. \exp\left[-\frac{1}{2} (y_2 + \sqrt{y_1})^2 \right] \right\}, y_1 > 0, y_2 > 0.$$

10.25 Ordered Statistics

Let $x_1, x_2, ..., x_n$ be iid random variables i.e. they are the random sample. Define

$$y_1 = \min\{x_1, x_2, ..., x_n\}$$

as the *first order statistic.* Define y_2 as the second smallest of $x_1, x_2, ..., x_n$ as the *second order statistic.* Define

$$y_n = \max(x_1, x_2, ..., x_n)$$

as the *n-th order statistic.* Then, $y_1 \leq y_2 \leq \cdots \leq y_n$. We will find the distribution for order statistics. The text book only does the continuous case and for x_i's having finite spaces (A, b). We will derive more generally by the distribution function technique. Consider the distribution function of y_k being the k-th smallest of $x_1, x_2, ..., x_n$. Then,

$$G_k(y_k) = Pr(Y_k \leq y_k) = Pr(\text{at least } k \text{ of the } x\text{'s are} \leq y_k\ ) =$$

$$\sum_{j=k}^{n} \overbrace{Pr(\text{exactly } j \text{ of the } x\text{'s are} \leq y_k)}^{\text{Binomial probability}}$$

$$\sum_{j=k}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) [F(y_k)]^j \, [1 - F(y_k)]^{n-j}$$

which is the distribution function of the k-th order statistic. "Success $= x_i \leq y_k$," n trials $= nx_i$'s

$$Pr(\text{success}) = Pr(x \leq y_k) = F(y_k) = \text{distribution function of } x.$$

Now suppose we are in the continuous case. Then, we can differentiate $F_k(y_k)$ to get a pdf for y_k.

$$\sum_{j=k}^{n} \left(\begin{array}{c} n \\ j \end{array} \right)^j [F(y_k)]^{j-1} \, f(y_k) \, [1 - F(y_k)]^{n-j} +$$

$$\sum_{j=k}^{n-1} \left(\begin{array}{c} n \\ j \end{array} \right) [F(y_k)]^j (n-j) [1 - F(y_k)]^{n-j-1} [-f(y_k)].$$

Notice that

$$\left(\begin{array}{c} n \\ j \end{array} \right)^j = \frac{n! j}{j!(n-j)!} = \frac{n!}{(j-1)!(n-j)!} =$$

$$n\left(\begin{array}{c} n-1 \\ y-1 \end{array}\right).$$

Notice that

$$\left(\begin{array}{c} n \\ j \end{array}\right)(n-j) = \frac{n!(n-j)}{j!(n-j)!} = \frac{n!}{j!(n-j-1)!} = n\left(\begin{array}{c} n-1 \\ j \end{array}\right).$$

So,

$$g_k(y_k) = nf(y_k)\left\{\sum_{j=k}^{n}\left(\begin{array}{c} n-1 \\ j-1 \end{array}\right)[F(y_k)]^{j-1}[1-F(y_k)]^{n-j} -\right.$$

$$\left.\sum_{j=k}^{n-1}\left(\begin{array}{c} n-1 \\ j \end{array}\right)[F(y_k)]^{j}[1-F(y_k)]^{n-j-1}\right\} =$$

$$nf(y_k)\left(\begin{array}{c} n-1 \\ k-1 \end{array}\right)[F(y_k)]^{k-1}[1-F(y_k)]^{n-k} =$$

$$\frac{n!}{(k-1)!(n-k)!}f(y_k)[F(y_k)]^{k-1}[1-F(y_k)]^{n-k}, -\infty < y_k < \infty,$$

which is the density for the k-th order statistic (must know this for the exam).

Example: Let n be odd and consider a random sample of size n from a continuous distribution with a distribution function F and the median $m = F^{-1}(\frac{1}{2})$. Find

$$Pr(y_{\frac{n+1}{2}} < m).$$

$$g_{\frac{n+1}{2}}\left(y_{\frac{n+1}{2}}\right) = \frac{n!f(y_{\frac{n+1}{2}})[F(y_{\frac{n+1}{2}})]^{\frac{n-1}{2}}}{\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)!\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)!}[1-F(y_{\frac{n+1}{2}})]^{\frac{n-1}{2}} \Rightarrow$$

$$Pr(y_{\frac{n+1}{2}} \leq m) =$$

$$\int_{-\infty}^{m}\frac{n!f(y_{\frac{n+1}{2}})[F(y_1\frac{n+1}{2})]^{\frac{n-1}{2}}}{\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)!\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)!}[1-F(y_{\frac{n+1}{2}})]^{\frac{n-1}{2}}dy_{\frac{n+1}{2}}.$$

Let

$$u = F(y_{\frac{n+1}{2}})$$

$$du = f(y_{\frac{y+1}{2}})dy_{\frac{n+1}{2}},$$

Then,

$$\int_0^{\frac{1}{2}} \frac{n!}{\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)!\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)!} u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}}du =$$

(looks like a Beta density)

$$\frac{1}{2}\overbrace{\int_0^1 \frac{\Gamma(n+1)}{\Gamma(\frac{n+1}{2})\Gamma(\frac{n+1}{2})} u^{\frac{n+1}{2}-1}(1-u)^{\frac{n+1}{2}-1}du}^{=1} = \frac{1}{2}.$$

In other words,

$$Pr(y_{\frac{n+1}{2}} \leq m) = \frac{1}{2}.$$

Hence, $y_{\frac{n+1}{2}}$ is called the *sample median.* It has the same median as the original population. Using similar calculations above, for deriving the distribution function $G_k(y_k)$ and the density $g(y_k)$, we can show that the joint density for y_i and y_j in the continuous case is

$$g_{ij}(y_i, y_j) =$$

$$\frac{n!f(y_i)f(y_j)[F(y_i)]^{i-1}}{(i-1)!(j-i-1)!(n-j)!}[F(y_i)-F(y_j)]^{j-i-1}[1-F(y_j)]^{n-j},$$

$$-\infty < y_i < y_j < \infty$$

(know this on the exam).

Example: Let x_1, x_2, x_3, x_4 be an iid sequence.

$$f(x) = \begin{cases} 2x, & 0 < x < 1. \\ 0, & \text{otherwise.} \end{cases}$$

$$F(x) = \begin{cases} 0, & x \leq 0. \\ x^2, & 0 < x < 0. \\ 1, & x \geq 1. \end{cases}$$

Find the probability that the range of the sample is less than $\frac{1}{2}$. i.e.

$$Pr\left(y_4 - y_1 < \frac{1}{2}\right).$$

Let,

$$w = y_4 - y_1,$$

$$z = y_4.$$

Let $w = y_4 - y_1, z = y_4$. Then, $y_1 = z - y, y_4 = z$.

$$J = \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.$$

$$g_{14}(y_1, y_4) = \frac{4!2y_1 2y_4}{0!2!0!}[y_4^2 - y_1^2]^2, 0 < y_1 < y_4 < 1.$$

$$f(w, z) = g_{14}(z - w, z)1 = 48(z - w)z[z^2 - (z - w)^2]^2 =$$

$$48w^2[4z^4 - 8z^3w + 5z^2w^2 - zw^3], 0 < w < z < 1.$$

$$f_1(w) = \int_w^1 f(w, z)dz =$$

$$48w^2\left\{\frac{w^5}{30} - \frac{w^3}{2} + \frac{5w^2}{3} - 2w + \frac{4}{5}\right\}, 0 < w < 1.$$

$$Pr(\text{range} < \tfrac{1}{2}) = \int_0^{\frac{1}{2}} f_1(w)dw =$$

$$\frac{w^8}{5} - 4w^6 + 16w^5 - 24w^4 + \frac{64w^3}{5}\Bigg|_0^{\frac{1}{2}} = 0.538.$$

10.26 Homework Answers

Answers from Chapter 4 of the text book.

5) $G(y) = Pr(Y \le y) = Pr(-2\ln x^4 \le y) =$

$$Pr\left(\ln x^4 \ge -\frac{y}{2}\right) =$$

$$Pr(x^4 \geq e^{-\frac{y}{2}}) = Pr(x \geq e^{-\frac{y}{8}}) =$$

$$1 - \int_0^{e^{-\frac{y}{8}}} x dx.$$

$$\cdots \Rightarrow y \sim \chi^2(2).$$

7) $$Pr\left(\frac{x_1}{x_2} \leq \frac{1}{2}\right) = Pr\left(x_1 \leq \frac{x_2}{2}\right) =$$

$$\int_0^1 \int_0^{\frac{x_2}{2}} 4x_1 x_2 dx_1 dx_2 = \int_0^1 2x_1^2 x_2 \Big|_0^{\frac{x_2}{2}} dx_2 =$$

$$\int_0^1 \frac{x_2^2}{2} x_2 dx_2 = \frac{x_2^4}{8}\Big|_0^1 = \frac{1}{8}.$$

$$Pr\left(x_1 x_2 \geq \frac{1}{4}\right) = Pr\left(x_1 \geq \frac{1}{4x_2}\right) =$$

$$\int_{\frac{1}{4}}^1 \int_{\frac{1}{4x_2}}^1 4x_1 x_2 dx_1 dx_2 = \cdots$$

13) For $0 < y < 2$:

$$Pr(Y \leq y) = Pr(x_1 + x_2 \leq y) = Pr(x_1 \leq y - x_2) =$$

$$\int_0^y \int_0^{y-x_2} \frac{1}{4} dx_1 dx_2 = \cdots.$$

Must be done for $y > 2$ also.

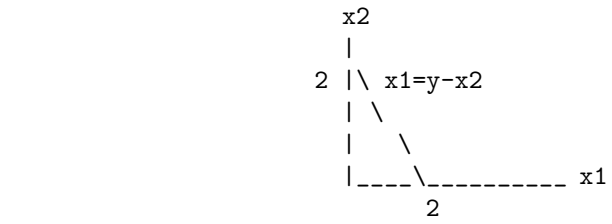
13) $$f(x) = \begin{cases} \frac{1}{2}, & 0 < x < 2. \\ 0, & \text{otherwise.} \end{cases}$$

$$f(x_1, x_2) = \begin{cases} \frac{1}{4}, & 0 < x_1 < 2, 0 < x_2 < 2. \\ 0, & \text{otherwise.} \end{cases}$$

$$y = x_1 + x_2,$$

$$G(y) = Pr(Y \leq y) = Pr(x_1 + x_2 \leq y) =$$

$$\int_{\{x_1 + x_2 \leq y\}} \int \frac{1}{4} dx_1 dx_2.$$



Case(1): $y < 2$.

$$G(y) = \int_0^y \int_0^{y-x_2} \frac{1}{4} dx_1 dx_2 = \frac{y^2}{8}.$$

Case(2): $2 < y < 4$.

$$1 - \int_{y-2}^{2} \int_{y-x_2}^{2} \frac{1}{4} dx_1 dx_2.$$

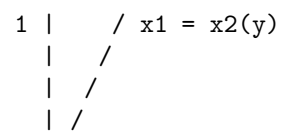
$$G(y) = \begin{cases} 0, & y \leq 0. \\ \frac{y^2}{8}, & 0 < y \leq 2. \\ y - 1 - \frac{y^2}{8}, & 2 < y < 4. \\ 1, & y \geq 4. \end{cases}$$

14) $$f(x) = \begin{cases} 1, & 0 < x < 1. \\ 0, & \text{otherwise.} \end{cases}$$

$$y = \frac{x_1}{x_2}.$$

$$Pr(Y \leq y) = G(y) = Pr\left(\frac{x_1}{x_2} \leq y\right) =$$

$$\int_{\frac{x_1}{x_2} \leq y} \int f(x_1, x_2) dx_1 dx_2$$



Case (1): $0 < y < 1 \Rightarrow$

$$G(y) = \int_0^1 \int_0^{yx_2} dx_1 dx_2 = \frac{y}{2}.$$

Case (2): $y > 1 \Rightarrow$

$$G(y) = \int_0^1 \int_{\frac{x_1}{y}}^1 dx_2 dx_1 = 1 - \frac{1}{2y} \Rightarrow$$

$$G(y) = \begin{cases} 0, & y \leq 0. \\ \frac{y}{2}, & 0 < y \leq 1. \\ 1 - \frac{1}{2y}, & y > 1. \end{cases}$$

$$g(y) = \begin{cases} 0, & y \leq 0. \\ \frac{1}{2}, & 0 < y \leq 1. \\ \frac{1}{2y^2}, & y > 1. \end{cases}$$

17) $Pr(y = 2x + 1) = Pr\left(\frac{y}{2} = x + 1\right) =$

$$Pr\left(\frac{y}{2} - 1 = x\right) =$$

$$\begin{cases} \frac{1}{3}, & y = 3, 5, 7. \\ 0, & \text{otherwise.} \end{cases}$$

19) $Pr(y = x^3) = Pr(\sqrt[3]{y} = x) =$

$$\begin{cases} \left(\frac{1}{2}\right)^{\sqrt[3]{y}}, & y = 1, 8, 27, ... \\ 0, & \text{otherwise.} \end{cases}$$

25) $y = x^3 \Rightarrow x = \sqrt[3]{y}$.

$$J = \frac{dx}{dy} = \frac{d(y^{\frac{1}{3}})}{dy} = \frac{1}{3}y^{-\frac{2}{3}} = \frac{1}{3y^{\frac{2}{3}}}.$$

$$g(y) = f(y^{\frac{1}{3}})|J| = \frac{y^{\frac{2}{3}}}{9}\frac{1}{3y^{\frac{2}{3}}} =$$

$$\begin{cases} \frac{1}{27}, & 0 < y < 27. \\ 0, & \text{otherwise.} \end{cases}$$

26) $y = x^2 \Rightarrow 0 < y < \infty, x = y^{\frac{11}{2}}$.

$$J = \frac{dx}{dy} = \frac{1}{2y^{\frac{1}{2}}}.$$

$$g(y) = f(y^{\frac{1}{2}})|J| = \cdots$$

28) $x \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

$$f(x) = \frac{1}{\frac{\pi}{2} + \frac{\pi}{2}} = \begin{cases} \frac{1}{\pi}, & -\frac{\pi}{2} < x < \frac{\pi}{2}. \\ 0, & \text{otherwise.} \end{cases}$$

$$y = \tan x,$$

$$x = \tan^{-1} y,$$

$$\frac{dx}{dy} = \frac{1}{1+y^2},$$

$$g(y) = f(\tan^{-1} y)\left|\frac{1}{1+y^2}\right| =$$

$$\left\{\frac{1}{\pi}\left(\frac{1}{1+y^2}\right), -\infty < y < \infty.\right.$$

30) $|J| = -\frac{1}{2}$.

$$h(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{x_1^2}{2} - \frac{x_2^2}{2}}, -\infty < x_1 < \infty, -\infty < x_2 < \infty.$$

$$g(y_1, y_2) = h\left((\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2) \right) |J| =$$

$$\frac{1}{4\pi} e^{-\frac{(y_1+y_2)}{2} - \frac{(y_1-y_2)}{2}} \quad -\infty < y_1 < \infty, -\infty < y_2 < \infty.$$

Simplify the exponent:

$$-\frac{\frac{1}{4}(y_1^2 + 2y_1y_2 + y_2^2)}{2} - \frac{\frac{1}{4}(y_1^2 - 2y_1y_2 + y_2^2)}{2} =$$

$$-\frac{\frac{1}{4}(2y_1^2 + 2y_2^2)}{2} = -\frac{1}{2}\left[\frac{y_1^2}{2} + \frac{y_2^2}{2} \right].$$

Then,

$$g(y_1, y_2) = \frac{1}{4\pi} e^{-\frac{1}{2}[\frac{y_1^2}{2} + \frac{y_2^2}{2}]}$$

$$\frac{1}{\sqrt{4\pi}} e^{-\frac{1}{2}\frac{y_1^2}{2}} \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{2}\frac{y_2^2}{2}}$$

$$\frac{1}{\sqrt{4\pi}} e^{-\frac{1}{2}\left(\frac{y_1}{\sqrt{2}}\right)^2} \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{2}\left(\frac{y_2}{\sqrt{2}}\right)^2}$$

Both are $N(0, 2)$ and separable into functions of y_1 and y_2. So, $g(y_1, y_2) = g_1(y_1)g_2(y_2) \Rightarrow$ independence.

44) $T = \dfrac{w}{\sqrt{\frac{v}{r}}}; y = T^2$ is not a 1:1 transformation.

$$T^2 = \frac{w^2}{\frac{v}{r}} = \frac{\frac{u}{1}}{\frac{v}{r}} \sim F(1, r),$$

where $u \sim \chi^2(1)$.

46) $$y = \frac{1}{1 + \frac{r_1}{r_2} w},$$

$$1 + \frac{r_1}{r_2} w = \frac{1}{y},$$

$$w = \left(\frac{1}{y} - 1 \right) \frac{r_2}{r_1},$$

$$\frac{dw}{dy} = -\frac{r_2}{r_1 y^2}.$$

$$g(y) = f\left(\left[\frac{1}{y} - 1 \right] \frac{r_2}{r_1} \right) \left| -\frac{r_2}{r_1 y^2} \right| =$$

$$\frac{\Gamma(\frac{r_1+r_2}{2})(\frac{r_1}{r_2})^{\frac{r_1}{2}}}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})} \frac{[(\frac{1}{y} - 1)\frac{r_2}{r_1}]^{\frac{r_1}{2}-1}}{[1 + \frac{r_1}{r_2}(\frac{1}{y} - 1)\frac{r_2}{r_1}]^{\frac{r_1+r_2}{2}}} \frac{r_2}{r_1 y^2} =$$

$$\frac{\Gamma(\frac{r_1+r_2}{2})(\frac{1}{y} - 1)^{\frac{r_1}{2}-1}}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})(\frac{1}{y})^{\frac{r_1+r_2}{2}}} \frac{1}{y^2} = \cdots$$

116) $$E(F) = E\left(\frac{r_2 u}{r_1 v} \right) = \frac{r_2}{r_1} E(u) E\left(\frac{1}{v} \right) =$$

$E(u) = r_1$. So, $E(\frac{u}{r_1}) = 1$.

$$E\left(\frac{r_2}{v} \right) = \frac{r_2}{r_2 - 2}$$

1) A stationary train with N cars will be hit by n bombs. Each bomb will hit a car or the place where a previously destroyed car was. Each car (or place) is equally likely to be hit by each bomb, and the bombs are independent of one another. Each bomb destroys all cars that still exist within k cars on either side of the car or place hit ($2k+1 < N$). Find the expected number of cars destroyed. HINT: Use indicators. $n = 1$.

$$I_j = \begin{cases} 1, & \text{if car } j \text{ is destroyed.} \\ 0, & \text{otherwise.} \end{cases}$$

Let x be the number of cars destroyed. Then,

$$E(x) = E\left(\sum_{j=1}^{N} I_j\right) = \sum_{j=1}^{N} Pr(I_j = 1),$$

$$Pr(I_1 = 1) = Pr(\text{bomb hits cars } 1, 2, ..., k+1) =$$

$$\frac{k+1}{N} = Pr(I_N = 1).$$

$$Pr(I_2 = 1) = Pr(I_{N-1} = 1) = \frac{k+2}{N}$$

$$\sum_{j=1}^{N} Pr(I_j = 1) = 2k + 1 - \left(\frac{k^2+k}{N}\right)$$

for 1 bomb ($n = 1$).

$$f(x_1, x_2) = \begin{cases} \frac{x_1 x_2}{36}, & x_1 = 1, 2, 3; x_2 = 1, 2, 3. \\ 0, & \text{otherwise.} \end{cases}$$

$$y_1 = x_1 x_2,$$

$$y_2 = x_2.$$

$$ß = \{(y_1, y_2) \in$$

$$(1,1), (2,2), (3,3), (2,1), (4,2), (6,3), (3,1), (6,2), (9,3)\}$$

10.27 Moment Generating Function Technique

Let $y = u(x_1, x_2, ..., x_n)$. Suppose we know the joint density for $x_1, x_2, ..., x_n$ and we want to find the distribution of y. Perhaps we can do this by finding the mgf of y. Calculate

$$E(e^{ty}) = E(e^{tu(x_1, x_2, ..., x_n)}) =$$

$$\int\int \cdots \int e^{tu(x_1, x_2, ..., x_n)} f(x_1, x_2, ..., x_n) dx_1 dx_2 \cdots dx_n$$

or in the discrete case:

$$\sum \sum \cdots \sum e^{tu(x_1,x_2,...,x_n)} f(x_1, x_2, ..., x_n)$$

Example:

$$f_1(x_1) = \begin{cases} \frac{1}{2}, & x_1 = \pm 1. \\ 0, & \text{otherwise.} \end{cases}$$

$$f_2(x_2) = \begin{cases} \frac{1}{n}, & x_2 = 1, 2, ..., n. \\ 0, & \text{otherwise.} \end{cases}$$

Let x_1, x_2 be independent. $y = x_1 + x_2$,

$$E(e^{ty}) = E(e^{t(x_1+x_2)}) =$$

$$E(e^{tx_1} e^{tx_2}) = E(e^{tx_1}) E(e^{tx_2}) =$$

$$\left(\frac{e^t + e^{-t}}{2}\right) \left(\frac{1}{n} \sum_{k=1}^{n} e^{tk}\right) =$$

$$\frac{1}{2n}(1 + e^t + e^{tn} + e^{t(n+1)}) + \sum_{k=2}^{n-1} e^{tk} \Rightarrow$$

$$g(y) = \begin{cases} \frac{1}{2n}, & y = 0, 1, ..., n, n+1. \\ \frac{1}{n}, & y = 2, 3, ..., n-1. \end{cases}$$

Theorem 1: Let $x_1, x_2, ..., x_n$ be mutually independent random variables with $x_i \sim N(\mu_i, \sigma_i^2)$. Then,

$$y = \sum_{i=1}^{n} k_i x_i$$

is

$$N\left(\sum_{i=1}^{n} k_i \mu_i, \sum_{i=1}^{n} k_i^2 \sigma_i^2\right).$$

Proof:

$$E(e^{ty}) = E(e^{t \sum k_i x_i}) = E\left(\prod_{i=1}^{n} e^{t k_i x_i}\right)$$

$$\prod_{i=1}^{n} E(e^{tk_i x_i}) = \prod_{i=1}^{n} M_i(tk_i)$$

where $M_i(t)$ is the mgf of x_i.

$$\prod_{i=1}^{n} \exp\left\{\mu_i t k_i + \frac{1}{2}\sigma_i^2 t^2 k_i^2\right\} =$$

$$\exp\left\{\left(\sum_{i=1}^{n} k_i \mu_i\right) t + \frac{1}{2}\left(\sum_{i=1}^{n} k_i^2 \sigma_i^2\right) t^2\right\}$$

which is the mgf of

$$N\left(\sum k_i \mu_i, \sum k_i^2 \sigma_i^2\right).$$

Notice that this technique works nicely for linear combinations of independent random variables. If $x_1, x_2, ..., x_n$ are independent and

$$y = \sum_{i=1}^{n} k_i x_i,$$

then,

$$E(e^{ty}) = E(e^{t\sum k_i x_i}) = \prod_{i=1}^{n} E(e^{tk_i x_i}) = M_i(tk_i).$$

Theorem 2: If $x_1, x_2, ..., x_n$ are independent and x_i has an mgf $M_i(t)$, then

$$y = \sum_{i=1}^{n} a_i x_i$$

has an mgf

$$\prod_{i=1}^{n} M_i(a_i t).$$

Theorem 1 is a special case of Theorem 2.

Theorem 3: Let $x_1, x_2, ..., x_n$ be mutually independent and $x_i \sim \chi^2(r_i)$. Then,

$$y = \sum_{i=1}^{n} x_i$$

is

$$\chi^2\left(\sum_{i=1}^{n} r_i\right).$$

Proof: By Theorem 2, the mgf of y is

$$M(t) = \prod_{i=1}^{n}(1-2t)^{-\frac{r_i}{2}} = (1-2t)^{-\frac{\sum r_i}{2}}$$

which is the mgf of

$$\chi^2\left(\sum r_i\right).$$

Theorem 4: Let $x_1, x_2, ..., x_n$ be a random sample from $N(\mu, \sigma^2)$. Then,

$$y = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2$$

is $\chi^2(n)$. **Proof:** By Theorem 2 of chapter 3 in the text book,

$$\left(\frac{x_i - \mu}{\sigma}\right)^2 \sim \chi^2(1).$$

By the results in section 4.3 in the text book,

$$\left(\frac{x_i - \mu}{\sigma}\right)^2$$

are independent for $i = 1, 2, ..., n$. So, Theorem 3 says the sum is $\chi^2(n)$.

10.28 Distribution of $\bar{x}$ and s^2

Notes:

- $s^2 = \frac{\sum(x_i - \bar{x})}{n}$.
- The square of $N(0,1)$ is $\chi^2(1)$.
- The T statistic is a normal divided by the square root of a chi-square.

The purpose of this section is to prove the following: if $x_1, x_2, ..., x_n$ is a random sample from a normal population $N(\mu, \sigma^2)$, then

1. $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

2. $\frac{ns^2}{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n-1)$.

3. $\bar{x}$ and s^2 are independent.

We will prove this by use of moment generating functions. The joint mgf of $\bar{x}, x_1 - \bar{x}, x_2 - \bar{x}, ..., x_n - \bar{x}$ is

$$M(t_0, t_1, ..., t_n) = E\left\{\exp[t_0\bar{x} + t_1(x_1 - \bar{x}) + \cdots + t_n(x_n - \bar{x})]\right\}.$$

The quantity in the exponent is

$$t_0\bar{x} + \sum_{j=1}^{n} t_j x_j - \bar{x}\sum_{j=1}^{n} t_j =$$

$$\sum_{i=1}^{n}\left(\frac{t_0}{n} + t_i - \bar{t}\right)x_i,$$

where,

$$\bar{t} = \frac{\sum t_i}{n}.$$

Write

$$t_i^* = \frac{t_0}{n} + t_i - \bar{t}.$$

Then,

$$M(t_0, t_1, ..., t_n) = E\left\{\exp\left[\sum t_i^* x_i\right]\right\} =$$

By Theorem 2,

$$\prod_{i=1}^{n} M_i(t_i^*)$$

where M_i is the mgf of x_i

$$\prod_{i=1}^{n} \exp\left\{\mu t_i^* + \frac{1}{2}\sigma^2 t_i^{*2}\right\} =$$

$$\exp\left\{\mu\sum t_i^* + \frac{1}{2}\sigma^2\sum t_i^{*2}\right\}.$$

Now,

$$\sum_{i=1}^{n} t_i^* = 0$$

since

$$n\bar{t} - n\bar{t} = 0.$$

$$\sum_{i=1}^{n} t_i^{*2} = \frac{t_0^2}{n} + \sum_{i=1}^{n}(t_i - \bar{t})^2 \Rightarrow$$

$$M(t_0, t_1, ..., t_n) =$$

$$\exp\left\{\mu t_0 + \frac{1}{2n}\sigma^2 t_0^2 + \frac{1}{2}\sigma^2\sum(t_i - \bar{t})^2\right\} =$$

$$\overbrace{\exp\left\{\mu t_0 + \frac{1}{2n}\sigma^2 t_0^2\right\}}^{\text{mgf of } N(\mu, \sigma^2/n).} \exp\left\{\frac{1}{2}\sigma^2\sum(t_i - \bar{t})^2\right\}.$$

And we know from Theorem 1 that $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. Thus, we have

$$M(t_0, t_1, ..., t_n) = \overbrace{M(t_0, 0, 0, ..., 0)}^{\text{mfg of } \bar{x}} \times$$

$$\overbrace{M(0, t_1, t_2, ..., t_n)}^{\text{joint mgf of } x_1 - \bar{x}, ..., x_n - \bar{x}} \qquad \Rightarrow$$

$\bar{x}$ is independent of $x_1 - \bar{x}, ..., x_n - \bar{x}$. $\Rightarrow \bar{x}$ is independent of

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

It remains to show that $\frac{ns^2}{\sigma^2} \sim \chi^2(n-1)$. Easily checked is that

$$\overbrace{\sum_{i=1}^{n}\left(\frac{(x_i - \mu)}{\sigma}\right)}^{\chi^2(n) \text{ by Theorem 4}} \quad =$$

$$\overbrace{\frac{ns^2}{\sigma^2}}^{\text{independent}} + \overbrace{\frac{(\bar{x}-\mu)^2}{\sigma^2/n}}^{\text{independent;}\chi^2(1)\text{. Theorem 2, Chapter 3}} .$$

Hence, the mgf of the left-hand-side is the product of 2 mgf's for the right-hand-side.

$$(1-2t)^{-\frac{n}{2}} = M(t)(1-2t)^{-\frac{1}{2}}$$

where $M(t)$ is the mgf of $\frac{ns^2}{\sigma^2}$. $\Rightarrow$

$$M(t) = (1-2t)^{-\frac{(n-1)}{2}}$$

which is the mgf of $\chi^2(n-1)$. $\Rightarrow$

$$\frac{ns^2}{\sigma^2} \sim \chi^2(n-1).$$

10.29 Convergence in Distribution

In this chapter we study distributions that depend on an integer n. In particular, we study the distribution as n grows large.

Example: Let $\bar{x}$ be from a normal random sample

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Example: Let x be binomial. The mgf is $(1-p+pe^t)^n$.

Usually distributions that are a function of n arise from random variables calculated from samples of size n. Sometimes it is difficult to find the distribution for any fixed n, but simplifications arise as $n \to \infty$, for large samples. That is our motivation. If the distribution depends on n, we will indicate this by subscripting the distribution function as $F_n(x)$. For example for $N\left(\mu, \frac{\sigma^2}{n}\right)$,

$$F_n(x) = \int_{-\infty}^{x} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} e^{-\frac{n}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy.$$

Definition: Suppose that

$$\lim_{n\to\infty} F(y) = F(Y),$$

at every point y for which F is continuous and suppose further that F is a distribution function. Then, the random variable y_n with the Distribution function $F_n(y)$ is said to have a *limiting distribution* with the distribution function $F(y)$. y_n is said to *converge in distribution* with the distribution function $F(y)$.

Example: If $x_1, x_2, ..., x_n$ are a random sample from a distribution with a density

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta, \theta > 0 \text{ is fixed.} \\ 0, & \text{otherwise.} \end{cases}$$

$y_n = \max(x_1, x_2, ..., x_n)$. From Chapter 4, the distribution for y_n is

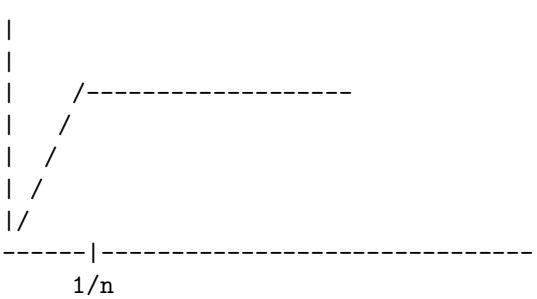
$$F_n(y) = \begin{cases} 0, & y \leq 0. \\ \frac{y^n}{\theta^n}, & 0 < y < \theta. \\ 1, & \theta \leq y. \end{cases}$$

$$\lim_{n \to n} F_n(y) = \begin{cases} 0, & y \leq 0. \\ 0, & 0 < y < \theta. \\ 1, & y \geq \theta. \end{cases}$$

Call this function $F(y)$. It is a distribution function with all its probability on the point θ. y_n converges in distribution to θ.

Example:

$$F_n(y) = \begin{cases} 0, & y \leq 0. \\ ny, & 0 < y < \frac{1}{n}. \\ 1, & y \geq \frac{1}{n}. \end{cases}$$



$$\lim_{n \to n} F_n(y) = \begin{cases} 0, & y \leq 0. \\ 1, & y > 0. \end{cases}$$

$F_n(y)$ is not a distribution function since it is not right continuous. Manually fit the following distribution:

$$F(y) = \begin{cases} 0, & y < 0. \\ 1, & y \geq 0. \end{cases}$$

to the definition of convergence. Convergence in distribution to the trivial distribution with all probability at zero.

Example: Let $y_n = \max(x_1, x_2, ..., x_n)$, and let $x_1, x_2, ..., x_n$ be iid uniform on $(0, \theta)$. This is the same as the first example. Let $z_n = n(\theta - y_n)$. Then it is easily shown that

$$G_n(z) = \begin{cases} 0, & z < 0. \\ 1 - \left(1 - \frac{z}{n\theta}\right)^n, & 0 \leq z \leq n\theta. \\ 1, & z \geq n\theta. \end{cases}$$

Remember that

$$\lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^n = e^{-a}.$$

Then,

$$\lim_{n\to\infty} G_n(z) = \begin{cases} 0, & z < 0. \\ 1 - e^{-\frac{z}{\theta}}, & z \geq 0. \end{cases}$$

which is the distribution function of an exponential random variable. Hence z converges in distribution to an exponential distribution.

10.30 Homework Answers

5.2 $f(x) = e^{-(x-\theta)}, \theta < x < \infty.$

$z_n = n(y_1 - \theta).$

$$g(y_1) = \frac{n! f(y_1)[F(y_1)]^0[1 - F(y_1)]^{n-1}}{(1-1)!(n-1)!},$$

$$f(y_1) = e^{-(y_1-\theta)}$$

$$F(y_1) = \int_\theta^{y_1} e^{-(x-\theta)} dx = -e^{-(x-\theta)}\Big|_\theta^{y_1} =$$

$1 - e^{-(y_1-\theta)}$.

$g(y_1) = ne^{-(y_1-\theta)}[1 - 1 + e^{-(y_1-\theta)}]^{n-1} = ne^{-n(y_1-\theta)}$.

Let

$y_1 = \frac{z_n}{n} + \theta$.

$\Rightarrow ne^{-n(\frac{z_n}{n}+\theta-\theta)} = ne^{-\frac{nz_n}{n}}$,

$J = \frac{1}{n}$,

$|J|g(y_1) = e^{-z_n}$,

$\int_0^{z_n} e^{-w}dw = 1 - e^{-z_n}$,

$\lim_{n\to\infty} 1 - e^{z_n} = 1$.

5.3 $z_n = n[1 - F(y_n)]$

$\frac{z_n}{n} = 1 - F(y_n)$

$F(y_n) = 1 - \frac{z_n}{n}$

$f(y_n) = F' = -\frac{1}{n}$

$Pr(Y_n \leq y_n) = \frac{n!f(y_n)[F(y_n)]^{n-1}}{(n-1)!(n-n)!} = n[F(y_n)]^{n-1}$

$g(y_n) = g\left(1 - \frac{z_n}{n}\right) = |J|f(y_n) = [F(y_n)]^{n-1}$.

10.31 Homework Answers

49) $y_1 = \dfrac{x_1}{x_1 + x_2},$

$y_1(x_1 + x_2) = x_1$

$x_1 + x_2 = \dfrac{x_1}{y_1}$

$x_1 - \dfrac{x_1}{y_1} = -x_2$

$x_1 \left(1 - \dfrac{1}{y_1}\right) = -x_2$

$x_1 = \dfrac{-x_2}{1 - \frac{1}{y_1}}$

$x_1 = \dfrac{x_2}{\frac{1}{y_1} - 1} = \dfrac{y_1 x_2}{1 - y_1}, 0 < y_1 < 1.$

$y_2 = \dfrac{\frac{y_1 x_2}{1-y_1} + x_2}{\frac{y_1 x_2}{1-y_1} + x_2 + x_3}$

$y_2 \left(\dfrac{y_1 x_2}{1 - y_1} + x_2 + x_3\right) = \dfrac{y_1 x_2}{1 - y_1} + x_2$

$\dfrac{y_1 x_2}{1 - y_1} + x_2 + x_3 = \dfrac{y_1 x_2}{y_2(1 - y_1)} + \dfrac{x_2}{y_2}$

$x_3 = \dfrac{y_1 x_2}{y_2(1 - y_1)} + \dfrac{x_2}{y_2} - x_2 - \dfrac{y_1 x_2}{1 - y_1}, 0 < y_2 < 1.$

$y_3 = \dfrac{y_1 x_2}{y_2(1 - y_1)} + \dfrac{x_2}{y_2}$

$y_3 = \left[\dfrac{y_1}{y_2(1 - y_1)} + \dfrac{1}{y_2}\right] x_2$

$x_2 = \dfrac{y_3}{\frac{y_1}{y_2(1-y_1)} + \frac{1}{y_2}} = \dfrac{y_3 y_2(1 - y_1)}{y_1 + 1 - y_1}, y_3 > 0.$

$x_2 = y_3y_2(1-y_1),$

$x_1 = y_1y_2y_3.$

$\frac{1}{1-y_1}y_1y_2y_3(1-y_1) + y_2y_3(1-y_1) + x_3 =$

$\frac{y_1y_2y_3(1-y_1)}{y_2(1-y_1)} + \frac{y_2y_3(1-y_1)}{y_2},$

$y_1y_2y_3 + y_2y_3 - y_1y_2y_3 + x_3 =$

$y_1y_3 + y_3 - y_1y_3,$

$x_3 = -y_2y_3 + y_3.$

$g(y_1, y_2, y_3) = e^{-y_1y_2y_3}e^{-y_2y_3(1-y_1)}e^{y_2y_3-y_3} =$

$e^{-y_1y_2y_3-y_2y_3+y_1y_2y_3+y_2y_3-y_3} = e^{-y_3}.$

$$J = \begin{vmatrix} y_2y_3 & -y_2y_3 & 0 \\ y_1y_3 & y_3(1-y_1) & -y_3 \\ y_1y_2 & y_2(1-y_1) & 1-y_2 \end{vmatrix} =$$

$y_2y_3^2(1-y_1)(1-y_2) + y_1y_2^2y_3^2 + 0 - [0 - y_2^2y_3^2(1-y_1) - (1-y_2)y_1y_2y_3^2] =$

$[y_2y_3^2 - y_1y_2y_3^2](1-y_2) + y_1y_2^2y_3^2 + y_2^2y_3^2 - y_1y_2^2y_3^2 + y_1y_2y_3^2 - y_1y_2^2y_3^2 =$

$y_2y_3^2$

$$\Rightarrow |J|g(y_1, y_2, y_3) = \begin{cases} y_2y_3^2e^{-y_3}, & 0 < y_1 < 1, 0 < y_2 < 1, y_3 > 0. \\ 0, & \text{otherwise.} \end{cases}$$

53) $f(x) = \frac{1}{2}, -1 < x < 1.$

$y = x^2 \Rightarrow x = \pm\sqrt{y}.$

Let $A_1 = (-1 < x < 0), A_2 = (0 < x < 1).$

$A_1 : x = -\sqrt{y}, 0 < y < 1.$

$A_2 : x = \sqrt{y}, 0 < y < 1.$

$J_1 = -\dfrac{1}{2\sqrt{y}},$

$J_2 = \dfrac{1}{2\sqrt{y}}.$

$g(y) = \dfrac{1}{2}\dfrac{1}{2\sqrt{y}} + \dfrac{1}{2}\dfrac{1}{2\sqrt{y}} = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 < y < 1. \\ 0, & \text{otherwise.} \end{cases}$

54) $y_1 = x_1^2 + x_2^2,$

$y_2 = x_2.$

$y_1 = x_1^2 + y_2^2,$

$y_1 - y_2^2 = x_1^2,$

$x_1 = \pm\sqrt{y_1 - y_2^2},$

$x_2 = y_2.$

$A_1 : (x_1^2 + x_2^2 > 0, x_1 > 0),$

$A_2 : (x_1^2 + x_2^2 > 0, x_1 < 0).$

$A_1 : x_1 = \sqrt{y_1 - y_2^2}, x_2 = y_2,$

$A_2 : x_1 = -\sqrt{y_1 - y_2^2}, x_2 = y_2.$

$J_1 = \begin{vmatrix} \frac{1}{2}(y_1 - y_2^2)^{-\frac{1}{2}} & 0 \\ \frac{1}{2}(-2y_2)(y_1 - y_2^2)^{-\frac{1}{2}} & 1 \end{vmatrix} = \dfrac{1}{2\sqrt{y_1 - y_2^2}},$

$J_2 = \begin{vmatrix} -\frac{1}{2}(y_1 - y_2^2)^{-\frac{1}{2}} & 0 \\ \frac{1}{2}(2y_2)(y_1 - y_2^2)^{-\frac{1}{2}} & 1 \end{vmatrix} = -\dfrac{1}{2\sqrt{y_1 - y_2^2}}.$

$$h(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)},$$

$$g(y_1, y_2) = \frac{1}{2\pi} \frac{e^{-\frac{1}{2}[(y_1 - y_2^2) + y_2^2]}}{2\sqrt{y_1 - y_2^2}} + \frac{1}{2\pi} \frac{e^{-\frac{1}{2}(y_1 - y_2^2 + y_2^2)}}{2\sqrt{y_1 - y_2^2}} =$$

$$\frac{1}{4\pi\sqrt{y_1 - y_2^2}} [e^{-\frac{1}{2}y_1} + e^{-\frac{1}{2}y_1}] =$$

$$\left\{ \frac{1}{2\pi} \frac{e^{-\frac{1}{2}y_1}}{\sqrt{y_1 - y_2^2}}, -\sqrt{y_1} < y_2 < \sqrt{y_1}, 0 < y_1 < \infty. \right.$$

$$g(y_1) = \int_{-\sqrt{y_1}}^{\sqrt{y_1}} \frac{1}{2\pi} \frac{e^{-\frac{1}{2}y_1}}{\sqrt{y_1 - y_2^2}} dy_2$$

is the integral of

$$\int \sin^{-1}\left(\frac{y_2}{\sqrt{y_1}}\right)$$

56) $f(x) = e^{-x}$,

$f(y_4) = e^{-y_4}$,

$$\int_0^{y_4} e^{-z} dx = -e^{-z}\Big|_0^{y_4} = 1 - e^{-y_4} = F(y_4).$$

$$Pr(3 \leq y_4) = \frac{4!(e^{-y_4})[1 - e^{-y_4}]^3[1 - 1 + e^{-y_4}]^0}{(3)!(0)!} =$$

$4e^{-y_4}[1 - e^{-y_4}]^3$.

Integrate from 0 to 3.

62) Find

$$Pr\left(y_4 - y_1 < \frac{1}{2}\right)$$

Let $w = y_4 - y_1, x = y_4$. Then,

$$z = y_4,$$

$$y_1 = z - w,$$

$$y_4 = z.$$

$$J = \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.$$

$$f(x) = 1.$$

$$F(x) = x,$$

$i = 1, j = 4, n = 4.$

$$g_{1,4}(y_1, y_4) = \frac{4!1 \times 1[y_1]^0[y_4 - y_1]^{4-1-1}[1 - y_4]^0}{0!2!0!} =$$

$$\begin{cases} 12(y_4 - y_1)^2, & 0 < y_1 < y_4 < 1. \\ 0, & \text{otherwise.} \end{cases}$$

$$f(w, z) = g_{1,4}(z - w, z)1 =$$

$$12(z - z + w)^2 = 12w^2, 0 < z - w < z < 1, \text{ or } 0 < w < z < 1.$$

$$f_1(w) = \int_0^1 f(w, z)dz = \int_w^1 12w^2 dz =$$

$$12w^2 z\Big|_w^1 = 12w^2 - 12w^3, 0 < w < 1.$$

$$Pr\left(y_4 - y_1 < \frac{1}{2}\right) = \int_0^{\frac{1}{2}} f_1(w)dw =$$

$$\int_0^{\frac{1}{2}} 12w^2 - 12w^3 dw = \frac{12w^3}{3} - \frac{12w^4}{4}\Big|_0^{\frac{1}{2}} =$$

$$4\frac{1}{8} - 3\frac{1}{16} = \frac{1}{2} - \frac{3}{16} = \frac{5}{16}.$$

78) Given,

$$x_1 \sim B\left(n_1, \frac{1}{2}\right),$$

$$x_2 \sim B\left(n_2, \frac{1}{2}\right).$$

$$y = x_1 - x_2 + n_2,$$

$$f_1(x_1) = \begin{cases} \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \left(\frac{1}{2}\right)^{x_1} (1-p)^{n_1 - x_1}, & x_1 = 0, 1, , ..., n_1. \\ 0, & \text{otherwise.} \end{cases}$$

$$f_2(x_2) = \begin{cases} \begin{pmatrix} n_2 \\ x_2 \end{pmatrix} \left(\frac{1}{2}\right)^{x_2} \left(\frac{1}{2}\right)^{n_2 - x_2}, & x_2 = 0, 1, , ..., n_2. \\ 0, & \text{otherwise.} \end{cases}$$

$$E(e^{ty}) = E(e^{t(x_1 - x_2 + n_2)}) = E(e^{tx_1}) E(e^{t(n_2 - x_2)}),$$

$$E(e^{tx_1}) = \sum_{x_1=0}^{n_1} e^{tx_1} \begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \left(\frac{1}{2}\right)^{x_1} \left(\frac{1}{2}\right)^{n_1 - x_1} = \left[e^t + \frac{1}{2}\right]^{n_1},$$

$$E(e^{t(n_2 - x_2)}) = \sum_{x_2=0}^{n_2} e^{t(n_2 - x_2)} \begin{pmatrix} n_2 \\ x_2 \end{pmatrix} \left(\frac{1}{2}\right)^{x_2} \left(\frac{1}{2}\right)^{n_2 - x_2} =$$

$$\sum_{x_2=0}^{n_2} \begin{pmatrix} n_2 \\ x_2 \end{pmatrix} \left(\frac{1}{2}\right)^{x_2} \left(e^t \frac{1}{2}\right)^{n_2 - x_2} = \left[\frac{1}{2} + \frac{1}{2} e^t\right]^{n_2},$$

$$E(e^{ty}) = \left[\frac{1}{2} + \frac{1}{2} e^t\right]^{n_1 + n_2} \sim Bin\left(n_1 + n_2, \frac{1}{2}\right)$$

81) $$f_1(x_1) = \begin{cases} \frac{x_1^2 e^{-\frac{x_1}{3}}}{\Gamma(3) 3^3}, & x_1 > 0. \\ 0, & \text{otherwise.} \end{cases},$$

$$f_2(x_2) = \begin{cases} \frac{x_2^4 e^{-\frac{x_2}{1}}}{\Gamma(5) 1^5}, & x_2 > 0. \\ 0, & \text{otherwise.} \end{cases}.$$

$a) y = 2x_1 + 6x_2,$

$E(e^{ty}) = E(e^{(2x_1+6x_2)t}) = E(e^{2x_1 t})E(e^{6x_x t}),$

$$E(e^{2x_1 t}) = \int_0^\infty \frac{e^{2x_1 t} x_1^2 e^{-\frac{x_1}{3}}}{\Gamma(3)27} dx_1 =$$

$$\frac{1}{\Gamma(3)27} \int_0^\infty x_1^2 e^{-x_1(\frac{1}{3}-2t)} dx_1 =$$

$$\frac{1}{27} \frac{1}{\left(\frac{1}{3}-2t\right)^3} \int_0^\infty \frac{x_1^2 e^{\frac{-x_1}{\frac{1}{(\frac{1}{3}-2t)}}}}{\Gamma(3)(\frac{1}{\frac{1}{3}-2t})^3} dx_1 =$$

$$\frac{1}{27} \left(\frac{1}{\frac{1}{3}-2t}\right)^3 =$$

$$\frac{1}{27} \left(\frac{3}{1-6t}\right)^3 = \left(\frac{1}{1-6t}\right)^3, |t| < \frac{1}{6}.$$

$$E(e^{6x_2 t}) = \int_0^\infty \frac{e^{6x_2 t} x_2^4 e^{-x_2}}{\Gamma(5)} dx_2 =$$

$$\int_0^1 \frac{x_2^4 e^{-x_2(1-6t)} dx_2}{\Gamma(5)} dx_2 =$$

$$\int_0^1 \frac{x_2^4 e^{\frac{-x_2}{\frac{1}{(1-6t)}}}}{\Gamma(5)} dx_2,$$

$$\left(\frac{1}{1-6t}\right)^5 \int_0^1 \frac{x_2^4 e^{\frac{-x_2}{\frac{1}{(1-6t)}}}}{\Gamma(5)(\frac{1}{1-6t})^5} dx_2,$$

$$\left(\frac{1}{1-6t}\right)^5, |t| < \frac{1}{6}.$$

$$E(e^{yt}) = \left(\frac{1}{1-6t}\right)^8, |t| < \frac{1}{6}.$$

b) Gamma $\alpha = 8, \beta = 6$.

$$g(y) = \begin{cases} \frac{y^7 e^{-\frac{y}{6}}}{\Gamma(8)6^8}, & y > 0. \\ 0, & \text{otherwise.} \end{cases}$$

84b) $x_i \sim Posiison(\mu_i), i = 1, 2, ..., n.$ $y = x_1 + x_2 + \cdots + x_n$.

$$E(e^{ty}) = E(e^{t\sum_{i=1}^n x_i}) = E(\prod_{i=1}^n e^{tx_i}) =$$

$$\prod_{i=1}^n E(e^{tx_i}).$$

$$f(x_i) = \begin{cases} \frac{\mu_i^{x_i} e^{-\mu_i}}{x_i!}, & x_i = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

$$\prod_{i=1}^n E(e^{tx_i}) = \prod_{i=1}^n e^{\mu_i(e^t-1)} = e^{\sum_{i=1}^n \mu_i(e^t-1)}$$

which is Poisson $(\mu_1 + \mu_2 + \cdots + \mu_n)$.

Chapter 11

Modeling Project

Immudyne Project[1] Data Analysis

Roger Goodwin
3909 Forest Glen Road
Virginia Beach, VA 23452
Fax: (757) 481-6177
Voice: (757) 431-2654
Email: rogergoodwin@CompuServe.com

Introduction

A known weight of powered yeast is treated with two factors: a disruption method(Factor A) and a digestion method(Factor B). The disruption method has three levels: A1) liquid nitrogen, A2) ground mechanically, and A3) unground (control). The digestion method has two levels: B1) water(H_2O), and B2) sodium hydroxide (NaOH). The purpose of treating the yeast is to break-up the cell walls of the yeast and to extract a compound called 1,3-beta-glucan to be used in manufacturing women's makeup. Given this, the ideal method or combination of methods should produce the highest yield of 1,3-beta-glucan.

Yeast was exposed to each level of the two factors in the following manner:

[1] Center for Biotechnology, Old Dominion University

A1B1, A1B2, A2B1, A2B2, A3B1, A3B2. Thus, there are six treatments in this study, the yeast being the experimental unit. Yeast exposure to the treatments was replicated 3 times each. Measurements of the solution(in terms of area under the absorbance curve cm^2) occured at 5 minute intervals starting with 5 minutes and ending at 90 minutes. Thus, there are 18 measurements taken of each treatment per replication. This gives a total of $18(3) = 54$ measurements for each treatment and a total of $54(6) = 324$ measurements in the experiment. The questions to be answered are as follow:

1. Which disruption method(Factor A) is the most effective?
2. Which digestion method(Factor B) is the most effective?
3. Which of the factors or combination of factors produced the fastest rate of extraction?
4. What is the optimal extraction time?
5. Are either of the two extraction methods more effective than the control?

Questions 1, 2, and 5 can definitely be answered with a statistical model. Consider time as another factor in the experiment, and Question 4 can be answered. However, Question 3 suggests finding the reaction rate of each treatment which would fall in the realm of Chemistry not Statistics. Moreover, the treatment with the fastest reaction rate does not necessarily imply the highest yield will be extracted. Question 3 will be omitted from the analysis.

11.1 Model Selection

Measurements were consistently taken from the same solution in time increments. This may lead to a repeated measures model depending on the correlation coefficient. Modeling the 18 time increments as another factor(Factor C) gives the following choice of models with possible interaction:

1. In the case that the correlation coefficient is *insignificant,* the data can be represented in a 3-way crossed design(ANOVA):

$$\left.\begin{array}{rl} y_{ijkm} = & \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \\ & (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkm}, \\ & i = 1, 2, 3; j = 1, 2; k = 1, 2, ..., 18; m = 1, 2, 3. \end{array}\right\} (11.1)$$

μ is the overall mean. α_i is the effect of the i-th level of Factor A. β_j is the effect of the j-th level of Factor B. γ_k is the effect of the k-th level of Factor C. $(\alpha\beta)_{ij}$ is the interaction of the i-th level of Factor A with the j-th level of Factor B. $(\alpha\gamma)_{ik}$ is the interaction of the i-th level of Factor A with the k-th level of Factor C. $(\beta\gamma)_{jk}$ is the interaction of the j-th level of Factor B with the k-th level of Factor C. $(\alpha\beta\gamma)_{ijk}$ is the interaction of the i-th level of Factor A, the j-th level of Factor B and the k-th level of Factor C. ϵ_{ijkm} is random error and $\epsilon \sim N(0, \sigma^2)$.

2. In the case that the correlation coefficient is *significant,* the data can be represented in a multivariate design(MANOVA) as:

$$y_{ijk} = \left.\begin{array}{l} \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \\ i = 1, 2, 3; j = 1, 2; k = 1, ..., 18; \end{array}\right\} \tag{11.2}$$

μ is the overall mean. α_i is mean effect of the i-th level of Factor A. β_j is the mean effect of the j-th level of Factor B. $(\alpha\beta)_{ij}$ is the mean effect of interaction between the i-th level of Factor A and the j-th level of Factor B. ϵ_{ijk} is random error and $\epsilon \sim N(0, \sigma^2 I)$.

Once the correlation coefficient has been quantified, the proper model can be selected and an analysis can be performed on the data.

11.1.1 Correlation Analysis

Eighteen dependent SAS variables, **Y1, Y2, ...,Y18,** were created to represent each measurement at time $i, i = 1, 2, ..., 18$. Two independent SAS variables, **DISRUPT** and **DIGEST,** were created to represent Factor A and Factor B. A repeated measures analysis was performed to determine if correlation existed among the 18 dependent variables and to determine if the correlation was constant. Looking at a subset of the partial correlation matrix in Appendix A.2.1 on page 1243, it is obvious that the first 6 random variables are correlated(partial correlation was calculated for all 18 variables and all the partial correlations were high). Why only 6 variables appear on the printout will become apparent in the next paragraph. For now, it can be concluded that a model similar to equation 11.2 should be used.

The next problem is to determine if the correlation structure is constant. The sphericity test can be used to test for equal correlation among the 18 random variables. In a multivariate model, the hypotheses are H_0 : $\sigma^2 I$, versus H_1 : correlation is not equal(Λ.) Specifically with the given

data, there were not enough degrees of freedom to run the sphericity test. However, if H_0 is rejected on some subset of the 18 random variables, then it can be concluded that H_1 is true. This approach gives enough degrees of freedom for the sphericity test. Since the p-value of the sphericity test on page 1243 in Appendix A.2.1 is 0.0000, reject H_0 using the subset **Y1, Y2, Y3, Y4, Y5, Y6.** ¿From this, it can be concluded that a different approach must be taken to analyze the data. PROC MIXED in SAS will be used.

11.2 Data Analysis

PROC MIXED gives many ways of choosing the structure for ϵ when $\epsilon \sim N(0, \Lambda)$. To analyze the data using PROC MIXED, a new data set was created. The dependent variable $\mathbf{Y}$ was created and the dependent variables **Y1**–**Y18** were dropped from the data set. $\mathbf{Y}$ is a 324×1 column vector containing all the information in **Y1**–**Y18.** AR(1) was used to model the correlation structure. The parameter estimate of the correlation is $\widehat{\rho} = 0.8922$ and the parameter estimate of σ^2 is $\widehat{\sigma}^2 = 1362.33$. PROC MIXED did give an estimate of ρ which PROC GLM did not do. The hypotheses tests in Appendix A.2.2 on pages 1244 and 1245 using PROC GLM match those in Appendix A.3.2 on page 1247 using PROC MIXED: the digestion methods and time levels are both significant, and the disruption methods are insignificant. The next problem is to determine which digestion level and which time level are most significant using PROC MIXED.

The disruption methods(Factor A) had an overall p-value of 0.0677. See Appendix A.3.2 on page 1247. None of the different levels of disruption are statistically significant at the 95% confidence level. So, the recommended disruption method is unground yeast.

The digestion methods(Factor B) had an overall p-value of 0.0001. See Appendix A.3.2 on page 1247. Digestion is a significant factor in the experiment. Using the Tukey pairwise comparison test, there is a statistically significant difference between H_2O and NaOH. See Appendix A.3.4 on page 1249. Since NaOH has the higher mean(255.305) compared to H_2O(118.765), use NaOH in the manufacturing process.

The time levels(Factor C) had an overall p-value of 0.0002. See Appendix A.3.2 on page 1247. Time is a significant factor in the experiment. Time level 16 resulted in the highest yield. Using the Tukey pairwise comparison test, time level 16 is statistically different from levels 1 thru 7. At time

level 8, the yields become insignificant when compared with level 16. See Appendix A.3.4 on page 1249. Thus, use time level 8(40 minutes) in the manufacturing process.

11.2.1 Residual Analysis

The purpose of perfoming a residual analysis is the verify the normality assumption of the model. Since the residuals were correlated, they had to be transformed to make them uncorrelated. The transformation involved creating the 18×18 matrix $\widehat{V}$ such that $\widehat{V} = \widehat{\sigma}^2 \widehat{\rho}^{|i-j|}$, where $i = 1, 2, ..., 18$; $j = 1, 2, ..., 18$. Then, the following transformation matrix is derived:

$$\widehat{\Lambda}^{-1/2} = \left(I \otimes \widehat{V} \right)^{-1/2}.$$

$\otimes$ is the direct product of the 18×18 identity matrix I and $\widehat{V}$ and produces a 324×324 matrix. The transformed residuals are obtained by multiplying $\widehat{\Lambda}^{-1/2}$ by the residuals from PROC MIXED. The transformation matrix was created in PROC IML.

The hypotheses tests are: H_0 : uncorrelated residuals are normally distributed, vs H_1 : uncorrelated residuals are not normally distributed. Looking at Appendix A.3.5 on page 1250, the p-value of the Wilkins test for normality is 0.0001. Thus, reject H_0. The residuals do not come from a normal population. Looking at Appendix A.3.5 again, many of the extreme residuals came from the A2B2 treatment(ground yeast and NaOH).

11.3 Normality Remedy

Since many of the extreme residuals came from the A2B2 block, one or more of the replications from that block should removed. By trial and error, it was decided to remove the third replication. This still leaves two replications to estimate the mean response of the A2B2 treatment. Upon removing the third replication, the correlated residuals became normally distributed(Appendix A.4.6 on page 1258). However, the interaction[2] between Factor A(disruption) and Factor B(digestion) is now significant. The interaction between Factor B(digestion) and Factor C(time) is now significant. See Appendix A.4.1 on page 1252.

[2] As in the statistical sense: the level of one factor affects the mean response of another factor.

11.4 Re-Analysis of the Data

Currently, the model equation is similar to that of equation 2. The mean interaction responses are being estimated by :

$$\widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot}$$

The mean interaction responses should be estimated by:

$$\bar{y}_{ij\cdot} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk\cdot}$$

which corresponds to the following model $y_{ijk} = \mu_{ij} + \epsilon_{ijk}$. PROC MIXED was run again with just the interaction terms. The interaction means appear on pages 1254 and 1255 in Appendix A.4.3.

A Tukey pairwise comparison test(Appendix A.4.4 on page 1256) was run on the AB interaction. The highest mean occurred at the ground yeast and NaOH combination. The mean is 332.56, and the mean is statistically different from the other various combinations of each level of Factor A and Factor B. Thus, ground yeast in sodium hydroxide should be used in the manufacturing process.

A Tukey pairwise comparison test(Appendix A.4.5 on page 1257) was run on the BC interaction. The highest mean being 306.28 occurred at the NaOH level of Factor B and level 13 of Factor C. Holding NaOH constant, level 13 of Factor C is not statistically different from levels 9 thru 12 but is statistically different from levels 1 thru 8. The combination NaOH and level 9 of Factor C is statistically different from the H_2O level of Factor B and level 17 of Factor C. The yeast should be exposed to NaOH for 45 minutes.

11.5 Conclusions

1. Which disruption method(Factor A) is the most effective? Holding the digestion levels constant, there is a statistical significance and difference among the levels of the disruption methods at the 95% confidence level. Recommendation: ground yeast.

2. Which digestion method(Factor B) is the most effective? Holding the disruption levels constant, there is a statistical significance and difference between H_2O and NaOH at the 95% confidence level. Recommendation: NaOH(sodium hydroxide).

3. Which of the factors or combination of factors produced the fastest rate of extraction? Omitted.

4. What is the optimal extraction time? Holding the digestion levels constant, there is a statistical significance and difference between the various levels of time. Recommendation: 45 minutes.

5. Are either of the two extraction methods more effective than the control? Holding NaOH constant, ground yeast results in a statistically significant higher mean response than either unground yeast or yeast treated with liquid nitrogen. Holding H_2O constant, there is no statistical difference in the various levels of the extraction methods.

Appendix A

A.1 SAS Source Code

```
OPTIONS LS = 72;
DATA YEAST;
INFILE 'A:\DATA.TXT';
INPUT DISRUPT $ DIGEST $ Y1-Y18;


* PRINT THE DATA;


PROC PRINT;
VAR DISRUPT DIGEST Y1-Y18;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'RAW DATA OF 1,3-BETA-GLUCAN';
RUN;


* REPEATED MEASURES MODEL;

PROC GLM;
CLASS DISRUPT DIGEST;
MODEL Y1-Y18 = DIGEST | DISRUPT/NOUNI;
REPEATED TIME 18 (5 10 15 20 25 30 35 40 45 50
                  55 60 65 70 75 80 85 90)/PRINTE;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'GLM ANALYSIS OF THE DATA';
RUN;
```

```
* RUN THE SPHERICITY TEST;

PROC GLM;
CLASS DISRUPT DIGEST;
MODEL Y1-Y6 = DIGEST | DISRUPT/NOUNI;
REPEATED TIME 6/PRINTE;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'GLM ANALYSIS OF THE DATA/SPHERICITY TEST';
RUN;



* CREATE A NEW DATA SET TO USE WITH PROC MIXED;

DATA NEW1; SET YEAST;
ARRAY T{18} Y1-Y18;
YEAST+1;
DO TIME = 1 TO 18;
   Y=T{TIME};
OUTPUT;
END;
DROP Y1-Y18;


* MODEL THE CORRELATION STRUCTURE;

PROC MIXED DATA = NEW1 METHOD = REML;
CLASS YEAST DIGEST DISRUPT TIME;
MODEL Y = DIGEST | DISRUPT | TIME /CHISQ PREDICTED;
REPEATED /TYPE = AR(1) SUBJECT = YEAST R;
LSMEANS DISRUPT DIGEST TIME/ADJUST=TUKEY;
MAKE 'PREDICTED' OUT=NEW2 noprint;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'DATA ANALYSIS WITH PROC MIXED';
RUN;


* UN-CORRELATE THE RESIDUALS;

PROC IML;

        START XFORM;
```

```
        USE NEW2;
        READ ALL;

        I18 = I(18);
        HATV = J(18, 18, 0);
        SIGMAH = J(324, 324, 0);
        U = J(324, 1, 0);

        DO I = 1 TO 18 BY 1;
           DO J = 1 TO 18 BY 1;
                HATV[I,J] = 1362.326*(0.8922)**(ABS(I-J));
           END;
        END;

        SIGMAH = I18 @ HATV;
        SIGMAH = SQRT(SIGMAH);
        SIGMAH = INV(SIGMAH);
        U = SIGMAH*RESID;

        CREATE NEW3 VAR{U};
        APPEND;
        SHOW CONTENTS;

FINISH XFORM;
RUN XFORM;


PROC UNIVARIATE DATA = NEW3 PLOT NORMAL;
VAR U;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'UN-CORRELATED RESIDUAL ANALYSIS';
RUN;

PROC UNIVARIATE DATA = NEW2 PLOT NORMAL;
VAR RESID;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'CORRELATED RESIDUAL ANALYSIS';
RUN;

* REMOVE THE 3-RD REPLICATION FROM THE A2B2 BLOCK;

DATA YEAST; SET YEAST;
```

```
IF _N_ = 18 THEN DELETE;


* PRINT THE DATA;


PROC PRINT DATA = YEAST;
VAR DISRUPT DIGEST Y1-Y18;
RUN;




DATA FINAL; SET YEAST;

ARRAY T{18} Y1-Y18;
YEAST+1;
DO TIME = 1 TO 18;
   Y=T{TIME};
OUTPUT;
END;
DROP Y1-Y18;
RUN;

RUN;

* MODEL THE REDUCED DATA SET;

PROC MIXED DATA = FINAL METHOD = REML;
CLASS YEAST DISRUPT DIGEST TIME;
MODEL Y = DISRUPT | DIGEST | TIME /CHISQ PREDICTED S;
*LSMEANS DISRUPT | DIGEST | TIME /ADJUST=TUKEY;
MAKE 'PREDICTED' OUT=NEW4 noprint;
REPEATED /TYPE = AR(1) SUBJECT = YEAST R;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'ANALYSIS ON THE REDUCED DATA SET';
RUN;




*MODEL THE INTERACTION TERMS ONLY;
```

```
PROC MIXED DATA = FINAL METHOD = REML;
CLASS YEAST DISRUPT DIGEST TIME;
MODEL Y = DISRUPT*DIGEST DIGEST*TIME /CHISQ PREDICTED S;
LSMEANS DISRUPT*DIGEST DIGEST*TIME /ADJUST=TUKEY;
MAKE 'PREDICTED' OUT=NEW4 noprint;
REPEATED /TYPE = AR(1) SUBJECT = YEAST R;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'ANALYSIS OF INTERACTION TERMS ONLY';
RUN;

* UN-CORRELATE THE RESIDUALS;

PROC IML;

        START XFORM;
        USE NEW4;
        READ ALL;

        I18 = I(18);
        HATV = J(17, 18, 0);
        SIGMAH = J(306, 306, 0);
        TEMP = J(306, 324, 0);
        U = J(324, 1, 0);

        DO I = 1 TO 17 BY 1;
           DO J = 1 TO 18 BY 1;
                HATV[I,J] = 556.43*(0.727)**(ABS(I-J));
           END;
        END;

        TEMP = I18 @ HATV;

        * THE LAST 18 COLUMNS ARE JUNK;
        * OMIT THE LAST 18 COLUMNS;

        DO I = 1 TO 306 BY 1;
          DO J = 1 TO 306 BY 1;
             SIGMAH[I,J] = TEMP[I,J];
          END;
        END;
```

```
        SIGMAH = SQRT(SIGMAH);
        SIGMAH = GINV(SIGMAH);
        U = SIGMAH*RESID;

        CREATE NEW5 VAR{U};
        APPEND;
        SHOW CONTENTS;

FINISH XFORM;
RUN XFORM;



* TRANSFORMED RESIDUAL ANALYSIS;

PROC UNIVARIATE DATA=NEW5 PLOT NORMAL;
VAR U;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'UN-CORRELATED RESIDUAL ANALYSIS';
RUN;


PROC UNIVARIATE DATA=NEW4 PLOT NORMAL;
VAR RESID;
TITLE 'IMMUDYNE PROJECT';
TITLE2 'CORRELATED RESIDUAL ANALYSIS';
RUN;
```

A.2 Repeated Measures Analysis using PROC GLM

A.2.1 Selected Partial Correlation Coefficients

General Linear Models Procedure
Repeated Measures Analysis of Variance

Partial Correlation Coefficients from the Error SS&CP Matrix / Prob > |r|

DF = 12	Y1	Y2	Y3	Y4	Y5	Y6
Y1	1.000000	0.805055	0.684215	0.797613	0.739900	0.779293
	0.0001	0.0009	0.0099	0.0011	0.0038	0.0017
Y2	0.805055	1.000000	0.910334	0.959974	0.967580	0.951462
	0.0009	0.0001	0.0001	0.0001	0.0001	0.0001
Y3	0.684215	0.910334	1.000000	0.924955	0.945084	0.932897
	0.0099	0.0001	0.0001	0.0001	0.0001	0.0001
Y4	0.797613	0.959974	0.924955	1.000000	0.974034	0.931556
	0.0011	0.0001	0.0001	0.0001	0.0001	0.0001
Y5	0.739900	0.967580	0.945084	0.974034	1.000000	0.952571
	0.0038	0.0001	0.0001	0.0001	0.0001	0.0001
Y6	0.779293	0.951462	0.932897	0.931556	0.952571	1.000000
	0.0017	0.0001	0.0001	0.0001	0.0001	0.0001

Highly correlated. So were the other 12 variables.

Test for Sphericity: Mauchly's Criterion = 0.0066855
Chisquare Approximation = 50.57899 with 14 df Prob > Chisquare = 0.0000

$H_0 : \sigma^2 I$
$H_1 : \Lambda.$
$\Rightarrow$ reject H_0.

A.2.2 Hypotheses Tests

General Linear Models Procedure
Class Level Information

Class	Levels	Values
DISRUPT	3	GROUND LN2 UNGROUND
DIGEST	2	H2O NAOH

Number of observations in data set = 18

General Linear Models Procedure
Repeated Measures Analysis of Variance
Repeated Measures Level Information

Dependent Variable	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
Level of TIME	5	10	15	20	25	30	35	40

Dependent Variable	Y9	Y10	Y11	Y12	Y13	Y14	Y15	Y16
Level of TIME	45	50	55	60	65	70	75	80

Dependent Variable	Y17	Y18
Level of TIME	85	90

General Linear Models Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DIGEST	1	1510092.93175021	1510092.93175021	68.46	0.0001
DISRUPT	2	93660.91046843	46830.45523421	2.12	0.1624
DISRUPT*DIGEST	2	73718.67333538	36859.33666769	1.67	0.2290
Error	12	264697.51683458	22058.12640288		

General Linear Models Procedure
Repeated Measures Analysis of Variance
Univariate Tests of Hypotheses for Within Subject Effects

Source: TIME

DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
17	148365.0545959	8727.3561527	38.55	0.0001	0.0001	0.0001

Source: TIME*DIGEST

DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
17	37511.9907485	2206.5876911	9.75	0.0001	0.0001	0.0001

Source: TIME*DISRUPT

DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
34	6918.3772001	203.4816824	0.90	0.6324	0.5285	0.5851

Source: TIME*DISRUPT*DIGEST

DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
34	6568.0519730	193.1779992	0.85	0.7019	0.5660	0.6397

Source: Error(TIME)

DF	Type III SS	Mean Square
204	46181.3002023	226.3789226

Greenhouse-Geisser Epsilon = 0.2480
Huynh-Feldt Epsilon = 0.5585

A.3 Repeated Measures Analysis using PROC MIXED

A.3.1 Estimates of ρ and σ^2

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
DIAG AR(1)	0.00065491	0.89220563	0.02574286	34.66	0.0001
Residual	1.00000000	1362.3260943	318.03926192	4.28	0.0001

Model Fitting Information for Y

Description	Value
Observations	324.0000
Variance Estimate	1362.326
Standard Deviation Estimate	36.9097
REML Log Likelihood	-983.088
Akaike's Information Criterion	-985.088
Schwarz's Bayesian Criterion	-988.463
-2 REML Log Likelihood	1966.176
Null Model LRT Chi-Square	336.1816
Null Model LRT DF	1.0000
Null Model LRT P-Value	0.0000

A.3.2 Tests of Hypotheses

Tests of Fixed Effects

Source	NDF	DDF	Type III ChiSq	Type III F	Pr > ChiSq	Pr > F
DIGEST	1	12	109.56	109.56	0.0001	0.0001
DISRUPT	2	12	6.80	3.40	0.0334	0.0677
DIGEST*DISRUPT	2	12	5.35	2.67	0.0690	0.1095
TIME	17	204	48.67	2.86	0.0001	0.0002
DIGEST*TIME	17	204	26.86	1.58	0.0601	0.0718
DISRUPT*TIME	34	204	33.03	0.97	0.5148	0.5187
DIGEST*DISRUPT*TIME	34	204	34.86	1.03	0.4270	0.4374

A.3.3 Least Squares Means

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
DISRUPT GROUND	207.13302148	11.29689462	12	18.34	0.0001
DISRUPT LN2	188.41610907	11.29689462	12	16.68	0.0001
DISRUPT UNGROUND	165.55494593	11.29689462	12	14.65	0.0001
DIGEST H2O	118.76478185	9.22387583	12	12.88	0.0001
DIGEST NAOH	255.30460247	9.22387583	12	27.68	0.0001
TIME 1	147.26801167	8.69970017	204	16.93	0.0001
TIME 2	152.62652444	8.69970017	204	17.54	0.0001
TIME 3	158.16350278	8.69970017	204	18.18	0.0001
TIME 4	164.35957389	8.69970017	204	18.89	0.0001
TIME 5	171.29198056	8.69970017	204	19.69	0.0001
TIME 6	176.52320667	8.69970017	204	20.29	0.0001
TIME 7	179.44056111	8.69970017	204	20.63	0.0001
TIME 8	182.58687500	8.69970017	204	20.99	0.0001
TIME 9	185.44511056	8.69970017	204	21.32	0.0001
TIME 10	187.36433667	8.69970017	204	21.54	0.0001
TIME 11	196.97766333	8.69970017	204	22.64	0.0001
TIME 12	200.23405278	8.69970017	204	23.02	0.0001
TIME 13	210.38874167	8.69970017	204	24.18	0.0001
TIME 14	210.22912889	8.69970017	204	24.17	0.0001
TIME 15	206.52012778	8.69970017	204	23.74	0.0001
TIME 16	214.74531333	8.69970017	204	24.68	0.0001
TIME 17	214.07752778	8.69970017	204	24.61	0.0001
TIME 18	208.38222000	8.69970017	204	23.95	0.0001

A.3.4 Tukey Pairwise Comparisons

Means

Level 1	Level 2	Difference	Std Error	DDF	T	Pr > |T|	Adjustment	Adj P
GROUND	LN2	18.71691241	15.97622158	12	1.17	0.2641	Tukey	0.4913
GROUND	UNGROUND	41.57807556	15.97622158	12	2.60	0.0231	Tukey	0.0561
LN2	UNGROUND	22.86116315	15.97622158	12	1.43	0.1780	Tukey	0.3568
H2O	NAOH	-136.5398206	13.04453030	12	-10.47	0.0001	Tukey	0.0000

Level 1	Level 2	Difference	Std Error	DDF	T	Pr > |T|	Adjustment	Adj P
TIME 1	TIME 16	-67.47730167	11.13625000	204	-6.06	0.0001	Tukey-Kramer	0.0000
TIME 2	TIME 16	-62.11878889	10.98686818	204	-5.65	0.0001	Tukey-Kramer	0.0000
TIME 3	TIME 16	-56.58181056	10.81698679	204	-5.23	0.0001	Tukey-Kramer	0.0001
TIME 4	TIME 16	-50.38573944	10.62335239	204	-4.74	0.0001	Tukey-Kramer	0.0005
TIME 5	TIME 16	-43.45333278	10.40204033	204	-4.18	0.0001	Tukey-Kramer	0.0055
TIME 6	TIME 16	-38.22210667	10.14825521	204	-3.77	0.0002	Tukey-Kramer	0.0237
TIME 7	TIME 16	-35.30475222	9.85604464	204	-3.58	0.0004	Tukey-Kramer	0.0431
TIME 8	TIME 16	-32.15843833	9.51787324	204	-3.38	0.0009	Tukey-Kramer	0.0792
TIME 9	TIME 16	-29.30020278	9.12395985	204	-3.21	0.0015	Tukey-Kramer	0.1252
TIME 10	TIME 16	-27.38097667	8.66118801	204	-3.16	0.0018	Tukey-Kramer	0.1424
TIME 11	TIME 16	-17.76765000	8.11118518	204	-2.19	0.0296	Tukey-Kramer	0.7575
TIME 12	TIME 16	-14.51126056	7.44660629	204	-1.95	0.0527	Tukey-Kramer	0.8891
TIME 13	TIME 16	-4.35657167	6.62294373	204	-0.66	0.5114	Tukey-Kramer	1.0000
TIME 14	TIME 16	-4.51618444	5.55650223	204	-0.81	0.4173	Tukey-Kramer	1.0000
TIME 15	TIME 16	-8.22518556	4.03940434	204	-2.04	0.0430	Tukey-Kramer	0.8479

A.3.5 Residual Analysis

```
                        Univariate Procedure

Variable=U

                 Moments

N               324  Sum Wgts        324
Mean              0  Sum               0
Std Dev    5.519142  Variance   30.46093
Skewness   0.117602  Kurtosis   9.038323
USS        9838.879  CSS        9838.879
CV                .  Std Mean   0.306619
T:Mean=0          0  Pr>|T|       1.0000
Num ^= 0        324  Num > 0         170
M(Sign)           8  Pr>=|M|      0.4047
Sgn Rank        337  Pr>=|S|      0.8421

      W:Normal 0.907035 Pr<W 0.0001
```

H_0 : Normally distributed.
H_1 : Not.
$\Rightarrow$ reject H_0.

```
              Quantiles(Def=5)

100% Max  34.20967       99%  16.31307
75% Q3     2.3683        95%  7.481788
50% Med   0.12592        90%  5.313722
25% Q1   -2.38169        10%  -5.36281
0% Min  -27.3134         5%  -7.21337
1%  -15.9949
Range     61.52305
Q3-Q1     4.749986
                Mode      -27.3134
```

Extremes

Lowest	Obs	Highest	Obs
-27.3134(	303)	13.30007(	178)
-26.8752(	286)	16.31307(	285)
-20.3073(	305)	16.97437(	125)
-15.9949(	124)	23.61644(	287)
-15.9366(	176)	34.20967(	304)

Univariate Procedure

Variable=U

```
                    Histogram                        #      Boxplot
  32.5+*                                             1         *
      .
      .*                                             1         *
      .*                                             2         0
      .**                                            5         0
      .**********                                   28         |
   2.5+*********************************************  133      +--+--+
      .****************************************       118      +-----+
      .*********                                     25         |
      .**                                            6         0
      .*                                             2         0
      .*                                             1         *
 -27.5+*                                             2         *
       ----+----+----+----+----+----+----+----+----+
       * may represent up to 3 counts
```

A.4 Analysis with the Unbalanced Data Set

A.4.1 Tests of Hypotheses

Tests of Fixed Effects

Source	NDF	DDF	Type III ChiSq	Type III F	Pr > ChiSq	Pr > F
DIGEST	1	11	487.63	487.63	0.0001	0.0001
DISRUPT	2	11	54.25	27.13	0.0001	0.0001
DIGEST*DISRUPT	2	11	49.15	24.57	0.0001	0.0001
TIME	17	187	92.36	5.43	0.0001	0.0001
DIGEST*TIME	17	187	42.96	2.53	0.0005	0.0012
DISRUPT*TIME	34	187	38.88	1.14	0.2593	0.2824
DIGEST*DISRUPT*TIME	34	187	39.52	1.16	0.2368	0.2611

A.4.2 Estimates Fitting $y_{ijk} = \mu_{ij} + \epsilon_{ijk}$

Covariance Parameter Estimates (REML)

Cov Parm	Ratio	Estimate	Std Error	Z	Pr > |Z|
DIAG AR(1)	0.00130647	0.72696144	0.04489917	16.19	0.0001
Residual	1.00000000	556.43333739	89.18816630	6.24	0.0001

Model Fitting Information for Y

Description	Value
Observations	306.0000
Variance Estimate	556.4333
Standard Deviation Estimate	23.5888
REML Log Likelihood	-1164.41
Akaike's Information Criterion	-1166.41
Schwarz's Bayesian Criterion	-1169.99
-2 REML Log Likelihood	2328.813
Null Model LRT Chi-Square	170.7341
Null Model LRT DF	1.0000
Null Model LRT P-Value	0.0000

A.4.3 Least Squares Means of Interaction Terms Only

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
DISRUPT*DIGEST GROUND H2O	117.36752850	7.17953069	12	16.35	0.0001
DISRUPT*DIGEST GROUND NAOH	332.55973756	8.76639565	12	37.94	0.0001
DISRUPT*DIGEST LN2 H2O	126.93807250	7.17953069	12	17.68	0.0001
DISRUPT*DIGEST LN2 NAOH	253.82989140	7.19040521	12	35.30	0.0001
DISRUPT*DIGEST UNGROUND H2O	111.98874456	7.17953069	12	15.60	0.0001
DISRUPT*DIGEST UNGROUND NAOH	219.28976503	7.19040521	12	30.50	0.0001
DIGEST*TIME H2O 1	103.78572333	7.86294642	254	13.20	0.0001
DIGEST*TIME H2O 2	103.23852667	7.86294642	254	13.13	0.0001
DIGEST*TIME H2O 3	103.65090556	7.86294642	254	13.18	0.0001
DIGEST*TIME H2O 4	108.95637000	7.86294642	254	13.86	0.0001
DIGEST*TIME H2O 5	107.01341667	7.86294642	254	13.61	0.0001
DIGEST*TIME H2O 6	107.53683556	7.86294642	254	13.68	0.0001
DIGEST*TIME H2O 7	113.04025556	7.86294642	254	14.38	0.0001
DIGEST*TIME H2O 8	112.58853889	7.86294642	254	14.32	0.0001
DIGEST*TIME H2O 9	115.09454333	7.86294642	254	14.64	0.0001
DIGEST*TIME H2O 10	113.04847333	7.86294642	254	14.38	0.0001
DIGEST*TIME H2O 11	118.35393778	7.86294642	254	15.05	0.0001
DIGEST*TIME H2O 12	120.53481667	7.86294642	254	15.33	0.0001
DIGEST*TIME H2O 13	127.45018333	7.86294642	254	16.21	0.0001
DIGEST*TIME H2O 14	129.71826889	7.86294642	254	16.50	0.0001
DIGEST*TIME H2O 15	135.03167778	7.86294642	254	17.17	0.0001
DIGEST*TIME H2O 16	138.67174889	7.86294642	254	17.64	0.0001
DIGEST*TIME H2O 17	144.11995556	7.86294642	254	18.33	0.0001
DIGEST*TIME H2O 18	135.93189556	7.86294642	254	17.29	0.0001

Least Squares Means

Level	LSMEAN	Std Error	DDF	T	Pr > |T|
DIGEST*TIME NAOH 1	201.84637995	8.38168934	254	24.08	0.0001
DIGEST*TIME NAOH 2	212.45982995	8.38168934	254	25.35	0.0001
DIGEST*TIME NAOH 3	224.61384245	8.38168934	254	26.80	0.0001
DIGEST*TIME NAOH 4	229.83536745	8.38168934	254	27.42	0.0001
DIGEST*TIME NAOH 5	246.67845495	8.38168934	254	29.43	0.0001
DIGEST*TIME NAOH 6	259.42170495	8.38168934	254	30.95	0.0001
DIGEST*TIME NAOH 7	258.76501745	8.38168934	254	30.87	0.0001
DIGEST*TIME NAOH 8	265.25200495	8.38168934	254	31.65	0.0001
DIGEST*TIME NAOH 9	271.52601745	8.38168934	254	32.40	0.0001
DIGEST*TIME NAOH 10	276.61975495	8.38168934	254	33.00	0.0001
DIGEST*TIME NAOH 11	292.97327995	8.38168934	254	34.95	0.0001
DIGEST*TIME NAOH 12	293.85330495	8.38168934	254	35.06	0.0001
DIGEST*TIME NAOH 13	306.27707995	8.38168934	254	36.54	0.0001
DIGEST*TIME NAOH 14	302.39020495	8.38168934	254	36.08	0.0001
DIGEST*TIME NAOH 15	292.06071745	8.38168934	254	34.85	0.0001
DIGEST*TIME NAOH 16	305.15894245	8.38168934	254	36.41	0.0001
DIGEST*TIME NAOH 17	299.37300495	8.38168934	254	35.72	0.0001
DIGEST*TIME NAOH 18	294.97145495	8.38168934	254	35.19	0.0001

A.4.4 Tukey Pairwise Comparisons of the AB Interaction Terms

Differences of Least Squares Means

Level 1	Level 2	Difference	Std Error	DDF	T	Pr > |T|	Adjustment	Adj P
GROUND H2O	GROUND NAOH	-215.1922091	11.33116735	12	-18.99	0.0001	Tukey-Kramer	0.0000
GROUND H2O	LN2 H2O	-9.57054399	10.02950959	12	-0.95	0.3588	Tukey-Kramer	0.9239
GROUND H2O	LN2 NAOH	-136.4623629	10.16108203	12	-13.43	0.0001	Tukey-Kramer	0.0000
GROUND H2O	UNGROUND H2O	5.37878395	10.02950959	12	0.54	0.6016	Tukey-Kramer	0.9934
GROUND H2O	UNGROUND NAOH	-101.9222365	10.16108203	12	-10.03	0.0001	Tukey-Kramer	0.0000
GROUND NAOH	LN2 H2O	205.62166507	11.33116735	12	18.15	0.0001	Tukey-Kramer	0.0000
GROUND NAOH	LN2 NAOH	78.72984616	11.21333262	12	7.02	0.0001	Tukey-Kramer	0.0001
GROUND NAOH	UNGROUND H2O	220.57099301	11.33116735	12	19.47	0.0001	Tukey-Kramer	0.0000
GROUND NAOH	UNGROUND NAOH	113.26997253	11.21333262	12	10.10	0.0001	Tukey-Kramer	0.0000
LN2 H2O	LN2 NAOH	-126.8918189	10.16108203	12	-12.49	0.0001	Tukey-Kramer	0.0000
LN2 H2O	UNGROUND H2O	14.94932794	10.02950959	12	1.49	0.1619	Tukey-Kramer	0.6761
LN2 H2O	UNGROUND NAOH	-92.35169254	10.16108203	12	-9.09	0.0001	Tukey-Kramer	0.0000
LN2 NAOH	UNGROUND H2O	141.84114685	10.16108203	12	13.96	0.0001	Tukey-Kramer	0.0000
LN2 NAOH	UNGROUND NAOH	34.54012637	10.02950959	12	3.44	0.0049	Tukey-Kramer	0.0434
UNGROUND H2O	UNGROUND NAOH	-107.3010205	10.16108203	12	-10.56	0.0001	Tukey-Kramer	0.0000

A.4.5 Tukey Pairwise Comparisons of the BC Interaction Terms

Differences of Least Squares Means

Level 1	Level 2	Difference	Std Error	DDF	T	Pr > |T|	Adjustment	Adj P
NAOH 1	NAOH 13	-104.4307000	11.66524780	254	-8.95	0.0001	Tukey-Kramer	0.0000
NAOH 2	NAOH 13	-93.81725000	11.61636126	254	-8.08	0.0001	Tukey-Kramer	0.0000
NAOH 3	NAOH 13	-81.66323750	11.54877537	254	-7.07	0.0001	Tukey-Kramer	0.0000
NAOH 4	NAOH 13	-76.44171250	11.45515346	254	-6.67	0.0001	Tukey-Kramer	0.0000
NAOH 5	NAOH 13	-59.59862500	11.32510371	254	-5.26	0.0001	Tukey-Kramer	0.0002
NAOH 6	NAOH 13	-46.85537500	11.14372920	254	-4.20	0.0001	Tukey-Kramer	0.0163
NAOH 7	NAOH 13	-47.51206250	10.88929750	254	-4.36	0.0001	Tukey-Kramer	0.0089
NAOH 8	NAOH 13	-41.02507500	10.52926332	254	-3.90	0.0001	Tukey-Kramer	0.0481
NAOH 9	NAOH 13	-34.75106250	10.01287429	254	-3.47	0.0006	Tukey-Kramer	0.1703
NAOH 10	NAOH 13	-29.65732500	9.25558054	254	-3.20	0.0015	Tukey-Kramer	0.3206
NAOH 11	NAOH 13	-13.30380000	8.09897181	254	-1.64	0.1017	Tukey-Kramer	0.9997
NAOH 12	NAOH 13	-12.42377500	6.16294890	254	-2.02	0.0449	Tukey-Kramer	0.9889

A.4.6 Correlated Residual Analysis

```
                          Univariate Procedure

Variable=RESID         Residual

                Moments

N                306  Sum Wgts         306
Mean               0  Sum                0
Std Dev     21.06598  Variance    443.7754
Skewness    -0.17742  Kurtosis     0.34105
USS         135351.5  CSS         135351.5
CV                 .  Std Mean    1.204262
T:Mean=0           0  Pr>|T|       1.0000
Num ^= 0         306  Num > 0          157
M(Sign)            4  Pr>=|M|      0.6891
Sgn Rank       435.5  Pr>=|S|      0.7791
```

H_0 : Normally distributed.
W:Normal 0.990468 Pr<W 0.9528 } H_1 : Not.
$\Rightarrow$ accept H_0.

```
W:Normal    0.990468  Pr<W         0.9528

          Quantiles(Def=5)

100% Max  72.24006       99%  46.36316
75%  Q3   14.76201       95%  31.56615
50%  Med  0.280578       90%  24.28432
25%  Q1   -13.6608       10%  -27.2872
0%   Min   -68.681        5%  -34.3395
                           1%  -51.8734
Range      140.921
Q3-Q1     28.42282
Mode       -68.681
```

```
                 Extremes

Lowest    Obs     Highest    Obs
 -68.681(     303) 42.77665(     162)
-63.6405(     305) 46.36316(     284)
-59.9489(     306) 48.81099(     159)
-51.8734(     289) 48.99876(     278)
-48.0589(     124) 72.24006(     287)


                        Univariate Procedure

Variable=RESID         Residual

                    Histogram               #             Boxplot
          75+*                              1                0
            .                                
            .                                
            .***                            5                |
            .******                        12                |
            .******************            36                |
            .************************      48               +-----+
           5+****************************  55               *--+--*
            .************************      50               |     |
            .**********************        46               +-----+
            .**************                27                |
            .*********                     17                |
            .***                            5                |
            .*                              2                0
         -65+*                              2                0
             ----+----+----+----+----+---
             * may represent up to 2 counts
```

Index