INSTRUCTOR: JOHN SPURRIER, 420I LECONTE, 777-5072

TEXT: CLASSICAL AND MODERN REGRESSION WITH APPLICATIONS
RAYMOND H. MYERS, DUXBURY PRESS

OFFICE HOURS: I am generally available for questions whenever I am in my office. Feel free to make an appointment, if you prefer. Formal office hours are 2:00-3:00 M-Th. If you can not find me, leave your name and number with my secretary in the Department office.

GRADING: EXAMS (200 POINTS) There will be two exams prior to the final exam. The first exam will be on Thursday, October 1.

HOMEWORK (100 POINTS) Homework is a vital part of your learning experience in this course. You are to receive NO assistance on the problems. You may discuss them with the instructor. Some of the problems will be easy and some will be hard. You may use the computer to solve the homework unless the problem specifically states otherwise. Homework is due on the ~~Monday~~ Tuesday after it is assigned. Late papers will be accepted only under extreme circumstances.

FINAL EXAM (150 POINTS) The final is scheduled for Friday December 18, 9:00 AM - Noon

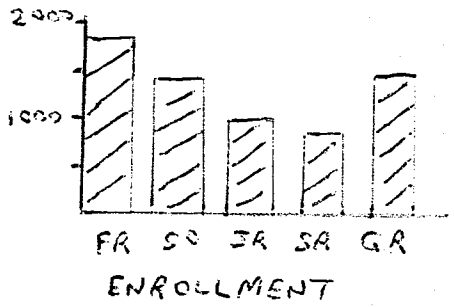
GENERAL COMMENTS: In graduate school, you will not necessarily understand everything during the lecture. Look over your notes after class and try to understand every line. If you can not, then contact me. We will be building constantly on previous lectures. You do not want to fall behind.

We will begin with a review of some concepts of applied statistics. Most of you will have seen at least some of the material. We will start in the text in about four weeks. Until that time, you will want to read the class notes and refer to previous texts that you have studied. If you do not have a previous text, please see me.

TOPIC 1 - DESCRIPTIVE STATISTICS

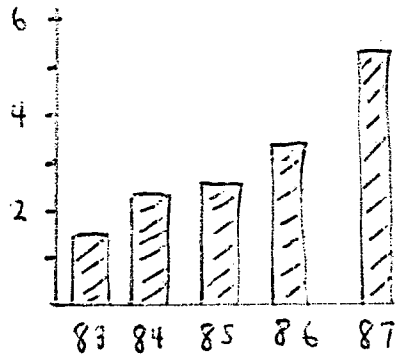
The field of statistics can be divided into two parts, descriptive and inferential. DESCRIPTIVE STATISTICS are techniques for providing a summary of a data set. In medium to large data sets, it is impossible to comprehend all of the data simultaneously. These techniques include tables, charts, graphs, diagrams, and summary statistics such as the sample mean and sample variance. The aim of descriptive statistics is to provide the user with a summary of the data that can be absorbed at a glance without inflicting any pain. The best descriptive techniques provide the summary without losing any of the original information. For example, we will see that the

stem-leaf diagram gives a graphical display of the data such that the user can recover all of the original data points. If one samples from the normal distribution, then the sample mean and variance contain all of the pertinent information.



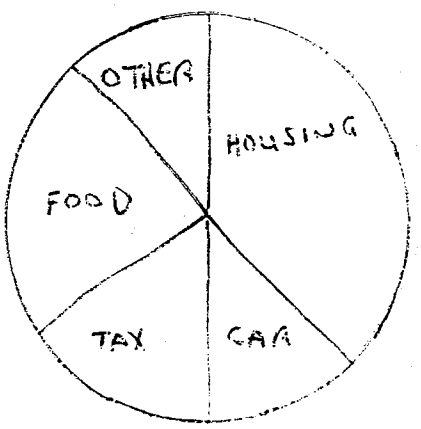
ENROLLMENT

Descriptive statistics that are popular for presenting information to a nontechnical audience are the BAR CHART and the PIE CHART. The bar chart presents a bar representing the frequency or relative frequency of each value of the variable of interest. Bar charts are also used at times to represent a trend in a variable over time. In this way, the bar chart is an alternative graphical presentation of a plot of the variable versus time. A pie chart is an effective way of presenting the proportion of the data that falls into each of a small number of categories. The angle of a slice of pie is 360 degrees times the proportion of items falling into that category. These charts particularly effective if they are done in color.



SALES IN MILLIONS

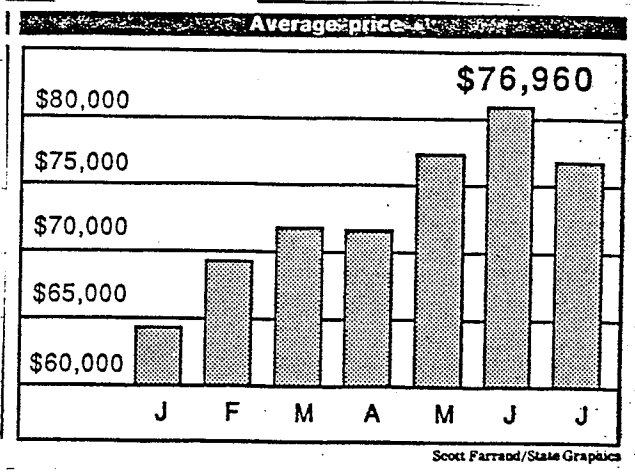
A descriptive statistic for summarizing numerical data is the histogram. To construct a histogram, one divides the real line into intervals of equal width. The histogram is a bar chart showing the frequency or relative frequency of each interval. When scaled properly, the histogram estimates the probability density function for continuous variables (the probability mass function for discrete variables) in the population from which the sample was taken.



FAMILY BUDGET

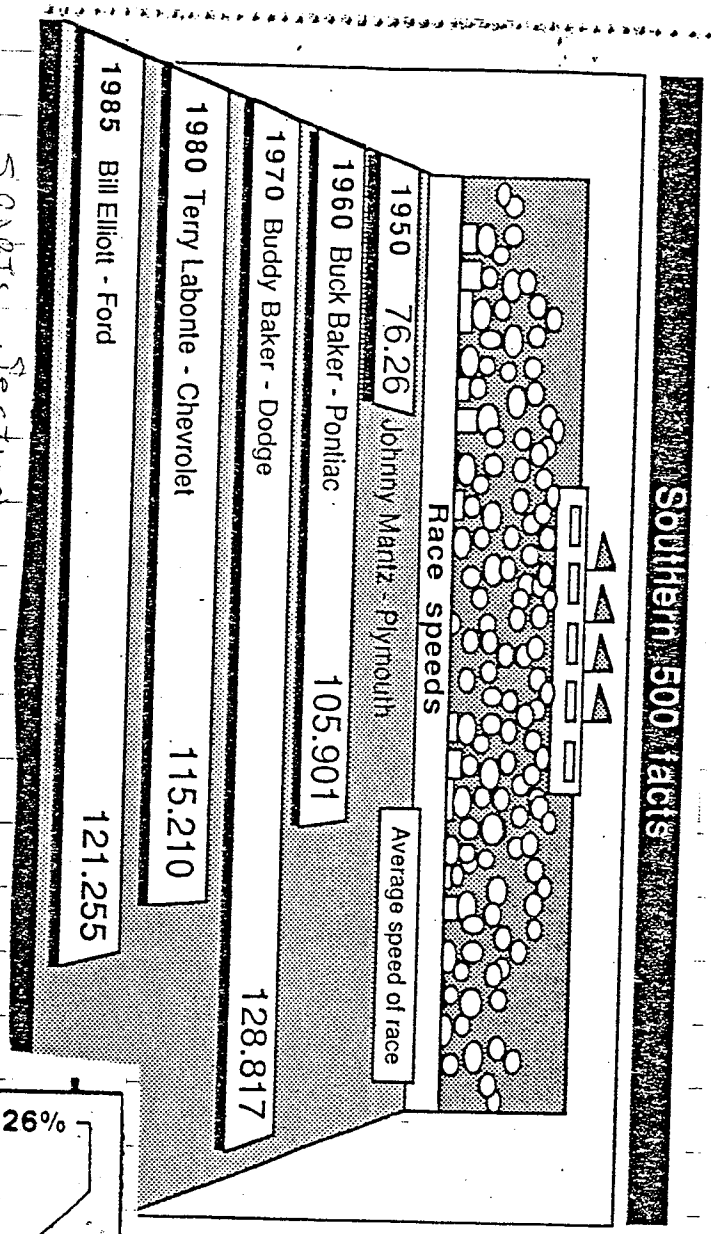
Histograms have two basic weaknesses. First, the choice of the intervals is arbitrary. With some data sets, a different selection

Columbia housing update



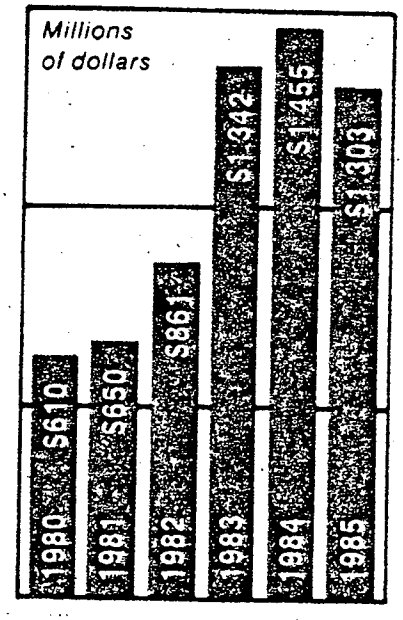
Sports Section

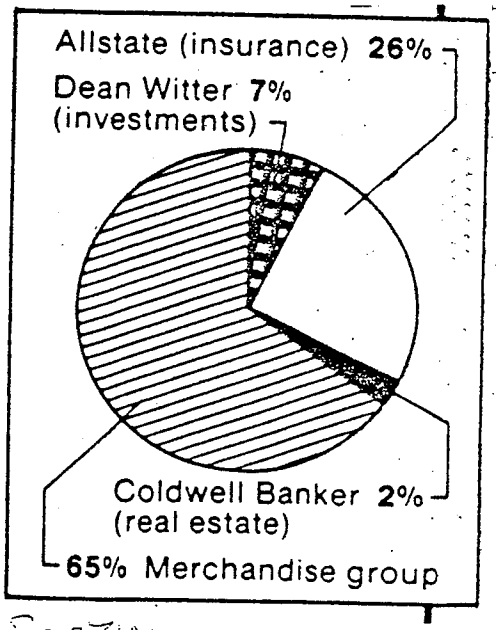
Housing Section improper goes to bars
A → June appears dld



sloppy - 70-80 # 80-85

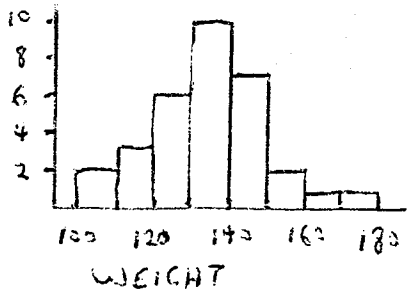
Sears, Roebuck & Co.





Business Section

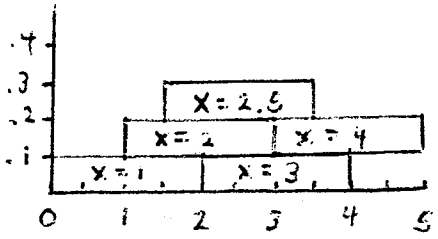
GRAPHICS FROM THE STATE 8-30 & 8-31, 1986

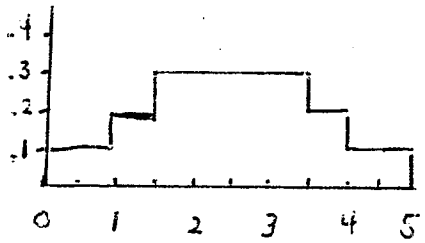


of intervals gives a drastically different picture. Second, the histogram loses information. By looking at a histogram, you can determine the number of observations falling in each interval, but you can not determine the precise values of those observations.

Three alternatives to histograms have become popular in the last several years. The first of these is known as DENSITY ESTIMATION. One such method, kernel density estimation, computes for each real number x, [1/(2h)] times the proportion of the sample falling within h units of x. This has the effect of representing each observation X by a rectangle of width 2h and height [1/(2hn)] centered at X, where n is the sample size. Applying pressure from the top to the stacks of rectangles yields the density estimate. One can uses shapes other than rectangles which have area 1/n. The choice of h is arbitrary. Statistical theory suggests using $h = n^{-4/5}$. For more information about density estimation, see Nonparametric Probability Density Estimation by Richard A. Tapia and James R. Thompson.

n=5 h=1
DATA 1, 2, 2.5, 3, 4





KERNEL DENSITY ESTIMATE

A second alternative to histograms is the STEM-LEAF diagram. This technique is described in detail in Exploratory Data Analysis by John W. Tukey. In the stem-leaf diagram each data point is represented by a gross measurement, the stem, and a fine measurement, the leaf. The following example, taken from Tukey's book, illustrates the concept. The data set consists of the height of the heighest point in each of the 50 states. The unit of measurement is 100 feet.

* graphs are best for discrete data

```
 0*|34588        Fla,Del,La,Miss,RI
 1 |236788
 2 |003448
 3 |24556
 4 |01489
 5 |03347
 6 |367
 7 |2
 8 |8
 9 |
10*|
11 |2
12 |678
13 |12588
14 |455           Colo,Wash,Calif
15 |
16 |
17 |
18 |
19 |
20*|3             Alaska
```

The numbers to the left of the vertical line are the stems and those to the right are the leaves. The stem 0 has 5 leaves 3,4,5,8 & 8. This indicates that five states have highest heights ≥ 0 and < 10 hundred feet units (1000 feet). Specifically, they are 03,04,05,08, and 08 in terms of hundred feet units. Between 10 and 20 hundred feet there are six states, etc. At times, it is informative to label the extreme points. If you rotate the stem-leaf diagram 90 degrees, you get a histogram. However unlike the usual histogram, you can recover the data values.

The third alternative to the histogram is also found in Tukey's book. It is the BOX PLOT. In order to discuss the box plot, we need to first consider some sample statistics. We use sample statistics to estimate or describe some feature of the population from which the sample is taken. One such feature is the CENTER of the population. Unfortunately, the concept of center is not uniquely defined. We can think of the center as being the POPULATION MEAN, the simple average of all values in the population. We can also think of the center as being the POPULATION MEDIAN, a point such that at least half of the items have values less than or equal to the value and at least half of the items have values greater than or equal to the value. The population mean and median are not always equal. Further complicating the problem is the fact that one can create several other reasonable measures of the center.

A wide variety of sample statistics can be used to estimate the population center. The first of these is the SAMPLE MEAN, the simple average of the sample items, $\bar{x}$. In many situations this is an excellent estimator of the population center. By studying its form, we can gain insight into the situations where it does not perform well. Since it is the simple average, its value can be dominated by one or a small number of observations that are very extreme relative to the other observations. Thus, if one is sampling from a population containing a small proportion of very extreme observations, then the sample mean is not very reliable as an estimate of the population center.

A second drawback of the sample mean in some experiments is that we must observe all data points to compute it. In many survival experiments we observe n items until $t < n$ items fail. The variable of interest is the time until failure. In this case it is not possible to compute the sample mean. This type of data is one example of CENSORED DATA.

An alternative estimator of the population center is the SAMPLE MEDIAN. To compute the sample median, one orders the data from smallest to largest. If n is odd, the median is the middle value. If n is even, the median is the average of the two middlemost values. This algorithm is consistent with the definition of the population median. That is, at least half of the sample observations are less than or equal to the sample median and at least half are greater than or equal to the median.

The median is not effected by extreme data points. It can often be computed with censored data. However, when one is sampling from a population where extreme values are not present, it tends to be a poorer estimator of the center of the population than is the sample mean.

It is possible to define classes of estimators which are compromises between the sample mean and the sample median. The TRIMMED MEAN computes the simple average of the sample after one deletes a fixed number of the smallest and largest observations. The name is derived from the fact that we are trimming extreme observations from the sample. With no trimming, you get the sample

mean. With complete trimming, you get the sample median. Between these two extremes, you get a new compromise estimator.

A related estimator is the WINSORIZED MEAN. To compute the Winsorized mean, one first trims the sample. Then all trimmed small values are replaced by the smallest untrimmed value and all trimmed large values are replaced by the largest untrimmed value. The Winsorized mean is the simple average of the adjusted data set.

As an example, consider the ordered data set of size n=10:

1.3, 2.7, 2.9, 4.1, 5.6, 6.2, 6.6, 7.8, 8.9, 12.7

The sample mean is 58.8/10 = 5.88. The sample median is (5.6+6.2)/2 = 5.9. If we trim 20% on each side, the trimmed mean is (2.9 + 4.1 + 5.6 + 6.2 + 6.6 + 7.8)/6 = 33.2/6 = 5.533. If we Winsorized 20% of each side, the Winsorized mean is (2.9 + 2.9 + 2.9 + 4.1 + 5.6 + 6.2 + 6.6 + 7.8 + 7.8 + 7.8)/10 = 54.6/10 = 5.46. Note that for the sample we have several estimates of the center. Without knowing more about the population there is no way to know which is the best estimate. Statistical theory gives us a way to choose among estimators.

At times we wish to estimate a point of the distribution other than the center. This is done by using PERCENTILES. The 30th percentile is a number such that at least 30% of the data is less than or equal to the number and at least 70% of the data is greater than or equal to the number. For the above data set the 30th percentile is 3.5. Similar definitions hold for the other percentiles. The median is the 50th percentile. Percentiles are useful in reliability theory and guarantee times. For example, in studying electronic components, you may wish to know the time such that 95% of the components survive past that time. This is the 5th percentile. At times, percentiles are combined to estimate the center of the population. The average of the 25th and 75th percentiles is an estimate of the center.

[2.9, 4.1] ⇒ interval

PROBLEMS

The following data set gives the number of revolutions until failure in millions of 25 ball bearings:

7.3, 5.7, 8.1, 9.2, 6.3, 4.2, 7.2, 5.8, 6.9, 9.1, 5.4, 7.8, 5.2, 4.3, 6.2, 7.9, 6.4, 7.4, 7.0, 2.8, 4.3, 8.1, 7.3, 5.8, 5.0

1. Compute the sample mean, the sample median, the trimmed mean with 20% trimming on each tail, the Winsorized mean adjusting for 20% in each tail, the 20th percentile, and the 90th percentile.
2. Construct a stem-leaf diagram.
3. Construct a kernel density estimate with h=0.5.

In addition to estimating the center of the distribution or other measures of location. It is important to describe the amount of VARIABILITY from the center of the population. As with the center, there are a number of ways that we can measure variability. The simplest measure of variability is the RANGE, which is the difference between the highest score and the lowest score. From its definition, we can see that the range is very sensitive to extreme observations. It is also very inefficient in most settings except for very small sample sizes. For these reasons, it is seldom used in practice, although in some situations there are theoretical reasons for using it to describe variability. Its main advantage is ease of computation.

The most popular measures of variability are the VARIANCE and its companion the STANDARD DEVIATION. The POPULATION VARIANCE, denoted by σ^2, measures the average squared distance between population values and the population mean, μ. Suppose the population has N members with values $X_1,\ldots,X_N$. The population mean and variance are given by

$$\mu = \Sigma X_i / N \quad \text{and} \quad \sigma^2 = \Sigma (X_i - \mu)^2 / N.$$

The population standard deviation, denoted by σ, is the nonnegative square root of the population variance. The POPULATION STANDARD DEVIATION is expressed in the original units of the measurements. That is, if the measurement is in inches, the standard deviation is expressed in inches. The variance does not have this property.

The SAMPLE VARIANCE, denoted by S^2, estimates the population variance. For a sample of size n, $x_1,\ldots,x_n$, the sample variance is

$$S^2 = \Sigma (x_i - \bar{x})^2 / (n - 1).$$

Some authors use the divisor n rather than (n-1) in the definition of the sample variance. While there are some reasons for doing this, the use of n causes S^2 on the average to underestimate σ^2. The

$S^2 = \frac{\Sigma x^2 - n\bar{x}^2}{(n-1)}$ alternate form - time efficient, error inefficient

sample standard deviation, S, is the nonnegative square root of the sample variance.

The above formula for the sample variance gives us good intuition, but it is not very convenient for hand or calculator calculation. By expanding the square, one can rewrite the formula as

$$S^2 = [(\Sigma x_i^2) - (\Sigma x_i)^2/n\ ] \ /\ (n-1).$$

When using a computer both algorithms have advantages. The second formula requires only one pass through the data, while the first requires one pass to compute $\bar{x}$ and a second pass to compute the sum of squares. The second formula, however, is much more sensitive to roundoff error than the first.

Like the mean, the variance and standard deviation are sensitive to a small number of extreme observations. In fact, the effect is more dramatic because we are squaring the extreme departure. In situations where you do not encounter extreme observations the variance and standard deviation are excellent measures of variability.

There are, of course, other measures of variability. One method of making the estimator less sensitive to extreme observations is to compute the average absolute deviation from the sample mean

$$\Sigma\ |x_i - \bar{x}|\ /\ n.$$

One can also compute the median of the absolute deviations. Also the deviations can be measured from the sample median, trimmed mean, etc. rather than from the sample mean.

A different approach to measuring variability is to take the difference of two percentiles. The most popular of these is the INTERQUARTILE RANGE, the difference between the 75th percentile and the 25th percentile. This measure is called the H-SPREAD by Tukey. He calls the 25th and 75th percentiles the hinges of the sample.

EXAMPLE Let us again use the sample of size 10 found on page 6.

1.3, 2.7, 2.9, 4.1, 5.6, 6.2, 6.6, 7.8, 8.9, 12.7

We have previously seen that $\bar{x} = 5.88$ and that the median is 5.9. The sample range is 12.7 - 1.3 = 11.4. The sample variance

$S^2 = \{[(1.3)^2+(2.7)^2+\ldots+(12.7)^2] - [58.8^2/10]\} \ /\ (10-1) = 11.4618$

This can also be computed

$S^2 = [(1.3-5.88)^2+\ldots+(12.7-5.88)^2] \ /\ (10-1) = 11.4618$

The sample standard deviation is $(11.4618)^{1/2} = 3.3855$

The average absolute deviation from the mean is
[|1.3-5.88| +...+ |12.7-5.88|] / 10 = 2.56
The median absolute deviation from the mean is (1.92+2.98)/2 = 2.45
The 25th percentile is 2.8, the 75th percentile is ~~8.35~~ 7.8, and the H-spread is 8.35 - 2.8 = 5.55.

Notice that unlike the measures of center, the measures of variability are not directly comparable.

We are now ready to describe the third alternative to the histogram which is the BOX PLOT. Box plots are also known as box and whisker plots. The box plot is a schematic representation of a data set. It can be arranged with a vertical or horizontal orientation. We will illustrate it here with a horizontal orientation to save space. Construction of the box plot begins by drawing vertical lines to represent the 25th and 75th percentiles. These vertical lines are then connected by two horizontal lines to form a rectangle, box. A vertical line is drawn across the box at the 50th percentile, median. A horizontal line, whisker, is drawn outside the box from the 25th percentile line covering all data points falling up to 1.5 H-spreads less than the 25th percentile. The process is repeated for the upper tail. Data points falling more than 1.5 H-spreads but less than 3 H-spreads outside the box are known as "outside" values. Points falling 3 or more H-spreads away from the box are known as "far out" values. Outside values are denoted by small circles and far out values are denoted by circles with dots in the middle. Some authors use different notation for marking these values. The sample mean is denoted by a "+" sign.

Outside and far out values demand special attention. They may represent mistakes due to measurement or data entry errors. They may also indicate situations that are much different from the rest. For example, if the data set represents monthly sales for each of a companies 100 account representatives, then outside and far out values in the upper tail may suggest that these representatives have positive techniques that could be used be the other representatives. In the lower tail, extreme values may suggest the need for more training or motivation.

Let us now construct a box plot for the highest peak data given in the stem leaf diagram on page 4. Straightforward calculations yield that the 25th percentile is 20, the median is 46, the 75th percentile is 112, and the mean is 61.64. The H-spread is 112 - 20 = 92. The 25th percentile minus 1.5 H-spreads is 20 - 1.5(92) = -118.0. The 75th percentile plus 1.5 H-spreads is 112 + 1.5(92) = 250.05. Thus, the lower whisker runs from 20 and covers all data points down to -118.0. Because our smallest data point is 3, the whisker runs from 3 to 20. The upper whisker runs from 112 and covers all data points up to 250.0. Because our largest data point is 203, the whisker runs from 112 to 203. In our example, there are no outside or far out values. The box plot is given below:

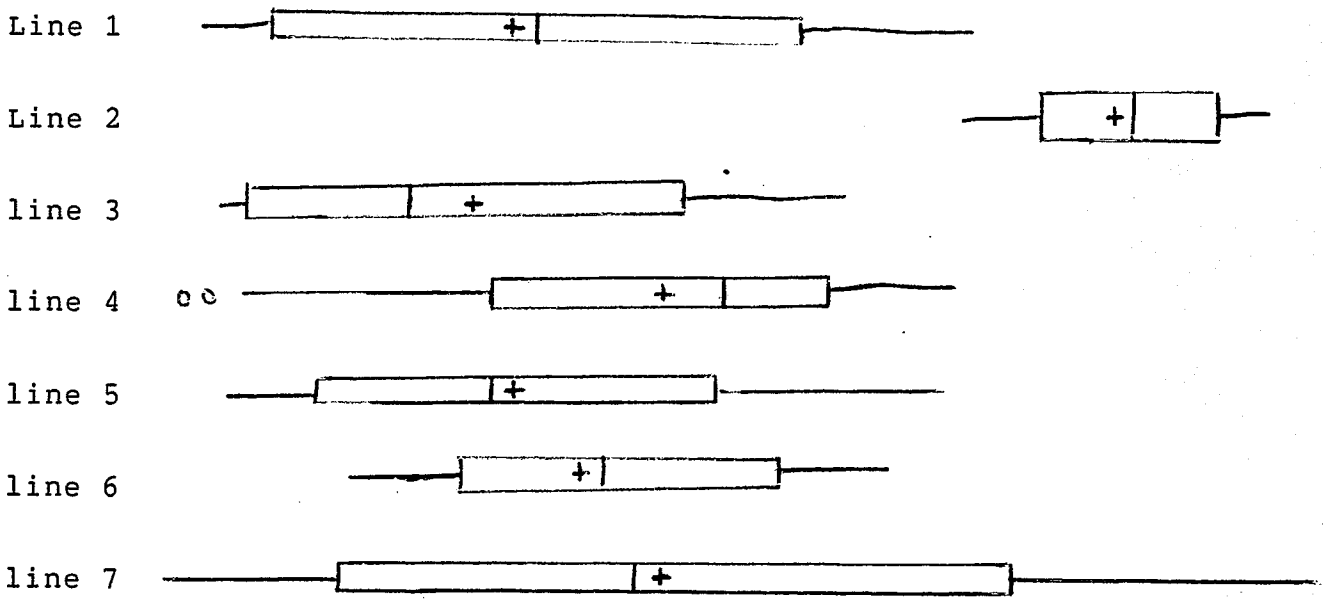
0	0	0	0	0	1	1	1	1	1	2	2
0	2	4	6	8	0	2	4	6	8	0	2
0	0	0	0	0	0	0	0	0	0	0	0

The facts that the median is much closer to the 25th percentile than the 75 percentile and that the right whisker is much longer than the left suggests that the data set has a long right tail and a short left tail.

Notice that the box plot does not preserve all of the information. We can not recreate all of the data points as we could with the stem-leaf diagram. For the peak data the box plot gives us no clue that the data set is bimodal (has two peaks). Thus, the box plot should be thought of as a quick summary of the data. It provides a fair amount of information, but we will generally want to take a closer look.

Box plots are particularly effective for comparing several data sets. This might arise when one takes a sample of items produced on several production lines or for displaying the size distributions of samples of an animal species collected at different times throughout the year.

Breaking Strength (pounds/square inch) For 7 Production Lines



Notice how easily one can compare the distributions. Line 2 tends to produce items having larger breaking strenghs and less variability than the other lines. Line 4 displays some outside observations on the low side. Line 7 displays large variability.

PROBLEMS

4. Construct by hand a box plot for the data in the problem set on page 7.

TOPIC 2 - INTRODUCTION TO SAS

SAS is a statistical computing package which is extremely helpful in data analysis and in report writing. We will begin our study of SAS by taking a simple program and looking at the printout that it generates. Our program will analyze the following data set:

ID	SEX	WEIGHT	HEIGHT	AGE	ID	SEX	WEIGHT	HEIGHT	AGE
1	F	119	66	17	21	M	164	70	21
2	F	130	63	19	22	F	163	73	19
3	F	140	65	21	23	M	186	69	21
4	F	163	68	23	24	F	125	65	19
5	F	201	66	21	25	F	133	64	21
6	M	156	68	22	26	M	177	68	22
7	F	121	65	23	27	F	119	63	23
8	M	183	70	22	28	M	135	66	17
9	F	95	57	18	29	F	154	68	19
10	F	145	66	20	30	F	121	67	20
11	F	128	65	21	31	F	106	60	23
12	M	168	70	24	32	M	154	70	18
13	F	134	68	23	33	F	178	68	19
14	M	215	72	21	34	M	188	74	20
15	M	195	72	22	35	M	212	72	19
16	F	145	68	19	36	F	169	71	20
17	M	159	69	22	37	M	152	66	18
18	F	128	62	21	38	F	121	65	19
19	M	175	67	24	39	M	166	70	21
20	M	230	75	20	40	F	155	64	20

We will first list the program and then study the effect of each line of the program. Each line in the listing of the computer program represents a separate card image. The entries in lower case letters are supplied by the user.

```
//jobname JOB (I5400015),'yourname',CLASS=Q,USER=userid,
// PASSWORD=password,MSGCLASS=A
/*ROUTE PRINT MVS.R43
$JOB SAS I5400015,yourname
DATA ONE;
INPUT ID 1-2 SEX $ 4 WEIGHT 6-8 HEIGHT 10-11 AGE 13-14;
CARDS;
 1 F 119 66 17
  .
  .
  .
40 F 155 64 20
PROC PRINT;
PROC CHART;VBAR HEIGHT/MIDPOINTS=58 62 66 70 74;
PROC FREQ;TABLES AGE;
PROC FREQ;TABLES SEX*AGE;
PROC CHART;PIE SEX;
PROC MEANS;
PROC SORT;BY SEX;
PROC UNIVARIATE PLOT;VAR HEIGHT;
PROC PLOT;PLOT WEIGHT*HEIGHT=SEX;
PROC SPLOT;VAR WEIGHT;CLASSES SEX;
$FINISH
/*
//
```

PROC MEANS; BY SEX; ⇒

NOTE: The above program is designed to produce a hard copy printout in Remote 43 on the first floor of Leconte. If you wish to have the output returned to your terminal replace the third line by /*ROUTE PRINT VM.userid

EXPLANATION OF THE PRINTOUT

We will go through the program line by line and explain the effects on the printout.

The first three lines are JOB CONTROL CARDS. They alert the computer that you wish to run a SAS program, that the cost should be billed to account I5400015, and what your name is. The version of SAS that we are using is known as FASTBATCH. It has the advantage that your program has high priority for running on the computer. The disadvantages are that the amount of time and the number of pages of printout are limited. Also you can not read data from mass storage or tapes under FASTBATCH. In analyzing real data sets, you may not be able to use FASTBATCH. In that case the JOB CONTROL CARDS are somewhat different.

DATA ONE;

This statement tells SAS to create a data set named ONE.

INPUT ID 1-2 SEX $ 4 WEIGHT 6-8 HEIGHT 10-11 AGE 13-14;

The statement informs the computer that the data will be entered such that the ID value will be in colums 1-2, the value of SEX will be in column 4, the value of WEIGHT will be in columns 6-8, etc. The $ following SEX indicates that character rather than numeric data will be entered for the variable SEX. We have used formatted input. If we had used unformatted input the statement would have been

```
INPUT ID SEX $ WEIGHT HEIGHT AGE;
```

With unformatted input the data can be placed in any column as long as the values are in proper order and separated by at least one space.

At times we wish to create new variables or change the variable listed on the INPUT card. For example, our data has weight in pounds. We might wish to convert the value to kilograms. This is done by putting the following card after the INPUT card:

```
WEIGHT = WEIGHT/2.2046;
```

A list of possible control statements is given in the USER'S GUIDE.

CARDS;

This statement informs the computer that the data will be read in on cards (or card images). The data follows the CARDS statement.

```
 1 F 119 66 17
      .
      .
      .
40 F 155 64 20
```

Using our input statement we have one data card per individual. Thus, we have 40 data cards. The first data card tells the computer that the first observation has ID=1, SEX=F, WEIGHT=119, HEIGHT=66 and AGE=17. After the last data card is read, our data set ONE is formed with forty observations. The rest of the cards perform operations on this data set.

PROC PRINT;
The PRINT procedure prints the data set that we have just created. It is important to check this printing carefully to make sure that the data in the computer is what we think it should be. It also serves as a written record of the data set.

PROC CHART;VBAR HEIGHT/MIDPOINTS=58 62 66 70 74;
We are using the CHART procedure to produce a histogram for the variable HEIGHT. The word VBAR tells SAS that we want vertical bars in our histogram (HBAR for horizontal bars). We are specifying that the midpoints of the classes should be 58 62 66 70 74. If we leave this off, SAS will select classes for us.

PROC FREQ;TABLES AGE;
This statement gives a frequency table for the variable AGE. Note that individual values of AGE are used and not classes.

PROC FREQ;TABLES SEX*AGE;
Here the FREQUENCY procedure is used to give a cross-classification of the sample by age and sex.

PROC CHART;PIE SEX;
A pie chart is presented for the variable SEX.

PROC MEANS;
The MEANS procedure produces a set of descriptive statistics for all numeric variables. Note in our data set SEX in not a numeric variable. The procedure serves two purposes. First it gives a handy set of descriptive statistics. Second it serves as a check for errors in the data set. The user should look at the minimum and maximum values of each variable to see if they are reasonable.

PROC SORT;BY SEX;
The SORT procedure sorts the data set by the variable SEX. Thus our data set is arranged such that all females are listed first followed by all males. The following statement would sort the data by sex first and that by age within sex.
PROC SORT;BY SEX AGE;
The SORT procedure produces no printout.

PROC MEANS;BY SEX;
The MEANS procedure is run separately for females and males. Note the data must be sorted by SEX to use this option.

PROC UNIVARIATE PLOT;VAR HEIGHT;
The UNIVARIATE procedure produces a vast set of descriptive statistics summarizing the variable HEIGHT in our data set. The word PLOT yields the stem-leaf diagram, the box plot, and the normal probability plot.

PROC PLOT;PLOT WEIGHT*HEIGHT=SEX;
The PLOT procedure produces scatterplots of two variables. Here we are asking for a plot of WEIGHT versus HEIGHT. We are using the value of SEX as the printing symbol. If we had left =SEX off the plot request a different type of symbol would have been used. Under

that system the letter 'A' would represent one observation, the letter 'B' would represent two observations, etc.

PROC SPLOT;VAR WEIGHT;CLASSES SEX;
The SPLOT procedure produces separate box plots on the variable weight for each sex. Again, the data set had to be sorted by SEX to use this command.

$FINISH
/*
//

So long computer. I am done with you for now. It has been fun. Don't call me. I'll call you.

PROBLEM

Thurs 5. The following data set has diameter in millimeters (D) and the breaking strengh in pounds (B) of 20 samples of monofiliment fishing line:

item	D	B	item	D	B	item	D	B
1	13	10	8	12	11	15	19	21
2	8	6	9	12	14	16	11	15
3	5	1	10	15	20	17	10	13
4	9	12	11	14	20	18	14	18
5	11	10	12	15	18	19	20	24
6	13	9	13	10	12	20	16	18
7	8	4	14	8	7			

a)Write a SAS program which plots B versus D and which gives box plots for the variables B and D.
b)Based on the plot, describe the relationship between B and D.
c)Is it reasonable to assume that reducing the variability in D will reduce the variability in B?

TOPIC 3 - ELEMENTS OF PROBABILITY

In most of the mathematical sciences we work with deductive logic. That is, we make a set of definitions and assumptions and see what theorems we can deduce from them. If our assumptions are valid and our method of proof is correct, then the theorems are unquestionably correct.

In developing the theory of statistics, there is a great deal of theorem proving - deductive logic. However, the basic nature of statistics involves inductive and not deductive. The experimenter wishes to make an inference about a population of interest. It is not possible due to time or money constraints to make a measurement on every item in the population. Therefore, measurements are made on only a sample of items drawn from the population. We then generalize from the results of the sample to the entire population - inductive logic. We must realize that inductive logic is subject to error. That is, some of our inferences about the population based on the sample results will be wrong.

Much of the study of statistics involves quantitifying the probability that our inference will be wrong and in designing experiments in such a way as to limit the probability of error to some "acceptable" level. What is "acceptable" often depends on the amount of money available for the study and the ramifications of a wrong decision. In order to quantify and limit these probabilities, we need to first formally define some concepts of probability theory.

In the probability context, an EXPERIMENT is any process which yields an outcome or observation. In statistics we are interested in RANDOM EXPERIMENTS. That is, an experiment where the outcome can not be predicted in advance with certainty. The set of all possible outcomes is known as the SAMPLE SPACE. The sample space is denoted by the symbol Ω. Any particular outcome is called a SAMPLE POINT. Any collection of sample points is known as an EVENT. We generally denote events by capital letters from the first of the alphabet. Two events are MUTUALLY EXCLUSIVE if they contain no sample points in common. The event made up of no sample points is called the NULL EVENT. It is denoted by ϕ.

EXAMPLE A left-handed wino rolls a die. His sober friend records the number of spots showing. The possible outcomes are 1,2,3,4,5, and 6. Thus, $\Omega = \{1,2,3,4,5,6\}$. There are $2^6 = 64$ distinct events formed by all possibilities of including or excluding the integers $1,\ldots,6$. The events $A = \{1,3,5\}$ and $B = \{2,4,6\}$ are mutually exclusive.

We are now ready to define a probability model. A PROBABILITY MODEL is a set function defined on the collection of events such that

1. For each event A in Ω, $1 \geq P(A) \geq 0$
2. $P(\Omega) = 1$
3. For any collection of mutually exclusive events $A_1, A_2, A_3, \ldots$
$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$$

It is important to note that in any random experiment, there are an infinite number of set functions that will satisfy these rules. That is, there are an infinite number of probability models.

EXAMPLE In the roll the die experiment, we can assign any nonnegative numbers which sum to 1 to the events {1}, {2}, {3}, {4},

{5}, {6}. Any such assignment satisfies the above rules. Some assignment patterns follow naturally from assumptions that we might consider making. For example, if we assume that the die is perfectly balanced, then it is natural to assign probability 1/6 to each of the sample points. However, if we assume that the die is weighted such that on each roll it yields 1 spot, then the probability model which assign probability 1 to the sample point {1} and 0 to all others is natural.

We would like to choose P so that for every event A, P(A) is the proportion of the time that A would occur if we performed the experiment a very large number of times. In general we do not have an opportunity to perform the experiment a large number of times. Thus, we must use our best judgment to assign the P function. Much of statistical inference involves choosing among competing probability models.

A RANDOM VARIABLE is a numerical value assigned to every sample point in Ω. Thus, it is a function from Ω to the real line. Because we can not predict with certainty which sample point will occur, we also can not predict the value of the random variable which will occur.

EXAMPLE

A student is selected at random from a population of students. The random variable of interest is height in inches. Note that it is possible to define other random variables in this experiment. We could have measured weight, age, head circumference, IQ, or any of several other measurements.

EXAMPLE

A manufactured product is selected at random from all similar items produced during the day. It is placed in use and the time until failure is measured.

EXAMPLE

A bird is selected at random from a particular species. The variable X is set equal to 0 if the bird is female and 1 if the bird is male. Note that in the strict sense, using M or F would not define a random variable.

Random variables are generally denoted by capital letters at the end of the alphabet. The probability model on the sample space induces a probability model on the random variable. Thus, we speak about the probability that the random variable X takes on a value in the subset A of the real line. We define the CUMULATIVE DISTRIBUTION FUNCTION (or distribution function) by $F(x) = P(X \leq x)$ for all x.

The most common types of random variables are either discrete or continuous. A DISCRETE random variable has a range of possible values consisting of a countable set of points on the real line, for example the integers or the positive integers. A CONTINUOUS random variable has a range of possible values of the entire real line or some combination of nondegenerate intervals on the real line. Discrete random variables tend to arise when one counts and continuous variables arise when one measures. Note that all measuring devices are discrete since they can only be read with finite precision. However, the quantity that they measure, such as length, is continuous. In this case, we consider the random variable to be continuous.

With discrete random variables, probability is assigned to the possible values of the random variable through the PROBABILITY MASS FUNCTION f(x). Thus $f(x) = P(X = x)$. One computes $P(X \in A)$ by summing f(x) over all $x \in A$. Thus, one computes F(z) by summing f(x) over all $x \leq z$. Let us consider some examples of families of discrete probability distributions. In each case one gets a particular discrete distribution by specifying values of the parameter(s).

EXAMPLE - DISCRETE UNIFORM ON 1,...,n

$f(x) = 1/n$ for $x = 1,\ldots,n$ and 0 otherwise.

In this distribution each of the numbers from 1,...,n are equally likely to occur. The probability distribution of the number of spots showing in the roll of a balanced die is discrete uniform with $n = 6$.

EXAMPLE - BINOMIAL n,p

$f(x) = \binom{n}{x}p^x(1-p)^{n-x}$ for $x = 0,1,\ldots,n$ and 0 otherwise, where $\binom{n}{x} = n!/[x!(n-x)!]$ and $0 \leq p \leq 1$. The binomial distribution is the probability distribution of the number of successes in a binomial experiment. That is an experiment where there are n independent

trials each of which ends in either a success or failure and the probability of success is constant at p in all trials.

EXAMPLE - GEOMETRIC p

$f(x) = (1-p)^x p$ for $x = 0,1,2,\ldots$ and 0 otherwise for $0 < p \leq 1$. The geometric distribution is the probability distribution of the number of failures before the first success in an experiment where there are independent trials each of which ends in success or failure, the probability of success is constant at p in all trials, and the trials continue until one success is achieved.

EXAMPLE - NEGATIVE BINOMIAL k,p

$f(x) = \binom{x+k-1}{x}(1-p)^x p^k$ for $x = 0,1,2,\ldots$ and 0 otherwise for $0 < p \leq 1$, k is a positive integer. The negative binomial distribution is a generalization of the last example where the trials continue until there have been k successes.

EXAMPLE - HYPERGEOMETRIC DISTRIBUTION n,a,b

$f(x) = \binom{a}{x}\binom{b}{n-x}/\binom{a+b}{n}$ for $x=\max(0,n-b),\ldots,\min(n,a)$ and 0 otherwise, where a and b are nonnegative integers and n is a positive integer $\leq a+b$. The hypergeometric distribution is the distribution of the number of successes in n draws without replacement from a finite population containing a successes and b failures.

EXAMPLE - POISSON λ

$f(x) = \lambda^x e^{-\lambda} / x!$ for $x=0,1,2,\ldots$ and 0 otherwise, where $\lambda \geq 0$. The Poisson distribution is the probability distribution of the number of occurrences of an event over a fixed time period where the probability of a single occurence of the event in any short time interval $(t,t+\Delta)$ is proportional to Δ, the probability of two or more occurrences in such a time interval is negligible, and the occurrence or nonoccurrence of events in nonoverlapping time intervals are independent.

With continuous random variables, we can not assign probabilities to each of the possible values, there are an uncountable number of them. Instead, we assign probability to intervals via the distribution function F. Thus, $P(a < X \leq b) = F(b) - F(a)$. If F(x) is differentiable with respect to x, then its derivative is known as the PROBABILITY DENSITY FUNCTION. The density function is denoted by f(x). Thus $P(a < X \leq b) = \int_a^b f(x)dx$ provided the density exists.

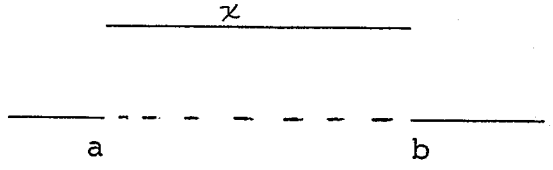
Also, $F(z) = \int_{-\infty}^{z} f(x)dx$. With continuous random variables the probability that X equals any particular point on the real line is said to be zero. Thus, $P(a < X \leq b) = P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b)$. It is f(x) that the kernel density estimator approximates.

It is convenient to describe continuous distributions by their density functions. They resemble histograms. Let us consider some examples of families of continuous distributions which possess density functions. In each case one gets a particular continuous distribution by specifying a value for the parameter(s).

EXAMPLE - UNIFORM (a,b)

$f(x) = 1/(b-a)$ for $a < x < b$ and 0 otherwise, where $a < b$.

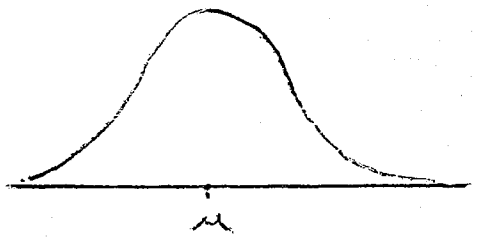
In this probability model all subintervals of (a,b) having equal length have equal probability.



EXAMPLE - NORMAL (μ, σ^2)

$\frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

$f(x) = [1/(2\pi\sigma^2)^{1/2}]\exp[-(1/2)(x-\mu)^2/\sigma^2]$ for $-\infty < x < \infty$, where μ is any real number and $\sigma^2 > 0$. This density function gives us the familiar bell shaped curve which is centered at μ and is symmetric about μ. The parameter μ is the location parameter and we will see later that it is the expected value or mean of the distribution.



The parameter σ^2 effects the spread of the curve. We will see later that σ^2 is the variance of the distribution. The parameter σ is known as a scale parameter for this distribution. The distribution function for the normal distribution does not have a closed form. It has to be approximated using numerical integration. Consequently, most statistics book publish tables for the normal distribution with $\mu = 0$ and $\sigma^2 = 1$. This distribution is known as the STANDARD NORMAL DISTRIBUTION. These tables vary some in format and one has to be careful to see what probability the number in the table represents. Some give $P(-\infty < X \leq a)$, while others give

$P(0 < X \leq a)$. For normal distibutions other than the standard normal, one makes the transformation $Z = (X-\mu)/\sigma$ prior to using the table. For a normal distribution with $\mu = 0.5$ and $\sigma = 2$, $P(X \leq 2) = P[Z = (X - 0.5)/2 \leq (2 - 0.5)/2 = 0.75] = 0.7734$.

OPTIONAL HOMEWORK [To be done if you need practice on the normal tables. Recall that $P(Z < -c) = P(Z > c)$]

Let Z have a standard normal distribution. Show that $P(Z \leq 2.1) = 0.9821$, $P(0 < Z \leq 1.46) = 0.4279$, $P(-1.64 < Z < 1.64) = 0.8990$, $P(Z < -1.96) = 0.0250$.

Let X have a normal distribution with $\mu = 0.5$ and $\sigma = 2$. Show that $P(X \leq 2.1) = 0.7881$, $P(0 < X \leq 1.46) = 0.2831$, $P(-1.64 < X < 1.64) = 0.5734$, $P(X < -1.96) = 0.1093$.

EXAMPLE - GAMMA α, β

$f(x) = \{1/[\Gamma(\alpha)\beta^{\alpha}]\}x^{\alpha-1}\exp(-x/\beta)$ for $0 < x < \infty$ and 0 otherwise, where α and $\beta > 0$. The symbol $\Gamma(a)$ denotes the gamma function evaluated at α. $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}\exp(-y)dy$ for $\alpha \geq 0$. . The parameter α is known as a shape parameter since changing its value changes the basis shape of the density function. The parameter β is a scale parameter. The gamma family of distributions is one of several used to model the lifetimes of manufactured parts. There are two important special cases. The EXPONENTIAL DISTRIBUTION ($\alpha = 1$) has the property that the probability that an item fails by time t+s given that it has survived until t does not depend upon t. The CHI SQUARE DISTRIBUTION WITH υ (nu) DEGREES OF FREEDOM ($\alpha = \upsilon/2$, $\beta = 2$) plays an important role in making inferences about the variance of a normal distribution.

We will address other important continuous distributions as the need arises. For more information about probability distributions see a textbook in mathematical statistics such as Introduction to Mathematical Statistics by Robert V. Hogg and Allen T. Craig or Statistical Theory by Bernard W. Lindgren. Another excellent source of facts about probability distributions is a series of four volumes entitled Distributions in Statistics by Norman Johnson and Samuel Kotz.

It is convenient to summarize probability distributions with descriptive measures. This is generally done using the concept of

expected values. For a discrete random variable X, the EXPECTED VALUE or MEAN, denoted by $E(X)$ or μ_X, is defined by $E(X) = \Sigma x f(x)$ provided the sum exists, where the summation is over all possible values of X.

EXAMPLE

For the discrete uniform distribution $E(X) = \Sigma_{x=1}^{n} x\,(1/n) = (1/n)\Sigma x = (1/n)n(n+1)/2 = (n+1)/2$.

EXAMPLE

Consider the discrete distribution with mass 1/2, 1/4, 1/8, 1/16, ... at the points 2, 4, 8, 16, ..., respectively. In this case $\Sigma\, x\, f(x) = \Sigma(1)$ where the summation is over an infinite set of points. Thus, $E(X)$ does not exist.

We can also speak of the expectation of a function of X. $E(g(X)) = \Sigma g(x)f(x)$ provided the summation exists, where the summation is over all possible values of X. Of particular interest are functions of the type $g(X)=X^r$. $E(X^r)$ is known as the rth MOMENT of X. The VARIANCE of a random variable is defined by $\sigma^2 = E(X^2)-E(X)^2$.

For continuous random variables with a density $f(x)$, $E(X)$ is defined by $E(X) = \int_{-\infty}^{\infty} x f(x)dx$ provided the integral exists.

EXAMPLE

For the continuous uniform distribution, $E(X) = \int_{-\infty}^{\infty} x f(x)dx = \int_a^b x[1/(b-a)]dx = [1/(b-a)](b^2/2 - a^2/2) = (a+b)/2$.

EXAMPLE

For the Cauchy distribution $f(x)=[\sigma/\pi]/[\sigma^2+(x-\mu)^2]$ for all x. In this case $\int_{-\infty}^{\infty} x f(x)dx$ does not exist. Thus, the Cauchy distribution does not have a mean.

The following table gives the mean and variance for some of the probability distributions that we have considered in the examples.

DISTRIBUTION	MEAN	VARIANCE
discrete uniform	$(n+1)/2$	$(n+1)(n-1)/12$
binomial	np	$np(1-p)$
Poisson	λ	λ
continuous uniform	$(a+b)/2$	$(b-a)^2/12$
normal	μ	σ^2
gamma	$\alpha\beta$	$\alpha\beta^2$
Cauchy	none	none

We can also describe a probability distribution in terms of probability points. That is the value of x such that $F(x)=p$ for a constant $0 < p < 1$.

HOMEWORK

6. Graph the density function of the gamma distribution with the following (α,β) values: (1,1), (1,2), (2,1) and (3,1). Put all graphs on the same sheet of graph paper or computer plot.
7. Graph the density functions of the Cauchy distribution with $\mu=0$ and $\sigma=1$ and the standard normal distribution on the same sheet.

An important fact about expectation is that it is a linear operator. That is, if $X_1,\ldots,X_n$ are random variables with expectations $\mu_1,\ldots,\mu_n$ and $a_0,a_1,\ldots,a_n$ are constants then $E[a_0+a_1X_1+\ldots+a_nX_n] = a_0+a_1\mu_1+\ldots a_n\mu_n$. It follows from this fact that if $X_1,\ldots,X_n$ each have mean μ then $\bar{X}$ also has mean μ.

TOPIC 4 - SAMPLING

We make our inferences about a population by taking measurements on a sample of size n and generalizing to the population. If our sample is not representative of the population then there is no reason to believe that our inference will be correct. Unfortunately, it is often difficult to look at a particular sample and know whether or not it is representative. This takes a considerable knowledge of the population of interest. If we knew that much about the population, we could probably answer all of our questions without taking a sample. Moreover, a sample that you think is representative, may not appear to be representative to another judge.

frag

Since we will never be able to agree that we have a representative sample. Our attention centers on techniques of drawing samples in a scientific fashion. Such sampling techniques allow one to compute expected values and standard deviations of the random variables which are used to estimate the parameters of interest such as the population mean or the proportion of individuals that have a characteristic of interest. Scientific samples do not insure that indivdual samples are representative. They can however in most cases insure that the expected value of the estimator is equal to the quantity that we are trying to estimate.

In order to be a scientific sample, the sampling process has to satisfy the following:

1. It is possible to define the set of distinct possible samples $S_1, S_2, S_3, \ldots, S_v$.
2. Each possible sample S_i has assigned to it a known probability π_i of being selected.
3. Using a random process in which each possible sample S_i has probability π_i of being selected, one of the samples is selected.
4. A well stated rule for computing the desired estimator for every possible sample is given.

The above rules define a random experiment where the random variable of interest is the estimator. Since we know how likely each sample is to occur, it is possible to compute the expected value and the standard deviation of the estimator in terms of the population values. If we do not know how likely each sample is to occur, it is not possible to make these calculations.

It order to satisfy the above rules, we generally have to have a list of the members of the population. This list is called a FRAME. Once we have a frame, it is conceptually simple to list all possible samples. In practice, it would be a tedious job make such a list. Fortunately, we only need to be able to make the list. It is not necessary to actually do it. It is often a difficult task to construct a frame. For example, if the population of interest is the residents of Columbia, then an up-to-date frame does not exist. However, one could find a list of all addresses or a list of all phone numbers.

Once one has the frame, the assignment of the probabilities, π_i, is generally determined by the method of sampling that one is using. The simplest scientific sampling technique is SIMPLE RANDOM SAMPLING. A sample of size n is a simple random sample if all possible samples of size n have an equal chance of being selected. Note that this is a stronger statement than saying that each item in the population has an equal chance of being selected.

Prior to selecting the sample, it has to be determined whether an item will be allowed to appear in the sample more than once. If we allow an item to appear more than once, our sample is a SAMPLE WITH REPLACEMENT. If we do not allow an item to appear more than once, our sample is a SAMPLE WITHOUT REPLACEMENT. Sampling without replacement has the advantages of producing estimators with smaller standard deviations than sampling with replacement and of not bothering an individual more than once. Sampling with replacement has the advantage that we do not have to check for duplicates in our sample.

w/ replacement has larger variance

If we sample with replacement, there are N^n possible samples. In simple random sampling with replacement, $\pi_i = 1/N^n$. If we sample without replacement, there are $\binom{N}{n}$ possible samples. In simple random sampling without replacement, $\pi_i = 1/\binom{N}{n}$.

It follows from the definition that simple random sampling is free of selection bias provided that the frame is accurate. Because all possible samples are equally likely to be picked, no subgroup has an unfair weighting in the selection process. Note that this is not true, if our sample is the first n people that we see on the street. However, we are not assured of a representative sample. If we took a simple random sample of 200 University of South Carolina students, it is possible although unlikely that all of them would have the last name Smith. In sampling with replacement it is possible although extremely unlikely that we would select the same person all 200 times.

With simple random sampling, the random process of selecting the sample is generally done by using a TABLE OF RANDOM NUMBERS or by generating random numbers on the computer. The process will be explained first for the table of random numbers. A table of random

Table I Random Numbers

COLUMN / ROW	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	53342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953
26	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
27	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
28	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253
29	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
31	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
32	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
33	69011	65795	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
34	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547

(*continued*)

numbers is a table of numbers, often 5 digit numbers, generated by a computer to approximate a simple random sample of all 5 digit numbers. The steps for sampling with replacement are the following:

1. Assign each sampling unit a unique identification number (e.g. number the frame from 1 to N)
2. Pick a random starting point in the random number table.
3. If the random number corresponds to an identification number in the population, select that item in your sample. If not, skip that random number and go to the next one.
4. Continue down the column until a sample of size n is selected.
5. If the bottom of the column is reached, move to the top of the next column.

If one is sampling without replacement then step 3 is modified so that one checks to see if that identification number is already in the sample.

To use the computer in place of a table of random numbers, then, one generates a string of random numbers that approximates a sample from the uniform distribution on (0,1). The selected identification number is then the greatest integer less than or equal to NU+1, where U is the uniform random variable.

9/15 PROBLEM 8. It is desired to take a simple random sample without replacement of size 10 from the households with surname Hampton in the Metropolitan Columbia area. The frame consists of the Hampton listings in the white pages of the Greater Columbia telephone book. Start in row 5 column 9 of the table of random numbers and use the first two digits.

a) What is your sample?

b) Comment on the appropriateness of the frame.

After one selects the sample, data is collected on each item in the sample. We are generally interested in estimating the population mean μ, the population total τ, or the population proportion P. The population total is the sum of the measurement of interest for all members of the population. That is, $\tau = N\mu$. The population proportion is the number of individuals in the population having the characteristic of interest divided by N.

For simple random sampling our point estimator for μ is $\bar{X}$. For τ, our point estimator is $N\bar{X}$. For estimating P our estimator is $\hat{P}$ = the number of individuals in the sample having the characteristic of interest divided by n.

These estimators are random variables and as such have means and standard deviations. These estimators have expected values equal to the population values that they estimate. That is $E(\bar{X})=\mu$, $E(N\bar{X}) = \tau$, and $E(\hat{P}) = P$. Estimators whose expected values equal the population values that they estimate are called UNBIASED ESTIMATORS. The standard deviation of an estimator is known as its STANDARD ERROR. The standard error of these estimators differ for sampling with replacement and sampling without replacement.

For sampling without replacement, the estimated standard error of $\bar{X}$ is $[(s^2/n)(N-n)/N]^{1/2}$. If we multiply this quantity by N, we get the estimated standard error of $N\bar{X}$. The estimated standard error of $\hat{P}$ is $\{[\hat{p}(1-\hat{p})/(n-1)](N-n)/N\}^{1/2}$. $\frac{N-n}{N}$ is finite correction factor

For sampling with replacement, the estimated standard error of $\bar{X}$ is $[s^2/n]^{1/2}$. Again, we multiply by N to get the estimated standard error of $N\bar{X}$. The estimated standard error of $\hat{P}$ is $[\hat{p}(1-\hat{p})/n]^{1/2}$. Note that the standard errors for sampling with replacement are larger than those for sampling without replacement.

We can be approximately 95% confident that the unknown population value falls within two estimated standard errors of the point estimate. That is, when using simple random sampling without replacement, $\mu \in \bar{x} \pm 2\,[(s^2/n)(N-n)/N]^{1/2}$ for approximately 95% of all possible samples.

Point estimates are generally meaningless without interval

Simple random sampling is the simplest approach to scientific sampling. There are numerous other techniques. The text Sampling Techniques by William G. Cochran is a classic reference to these ideas. We will consider one other sampling method in this course. That method is STRATIFIED SAMPLING.

In stratified sampling we divide our population into K nonoverlapping subgroups called strata. We then draw a simple random sample of size n_i from the N_i members of the ith stratum, i=1,...,k. Note that in stratified sampling we are ruling out many of the samples that were possible with simple random sampling. We are

N - population size
Strata - 1 .. k
N_i - size of the i^{th} Stratum
n_i - sample size from N_i stratum

forcing our sample to have a set number of items from each stratum. In simple random sampling, the number from each stratum is random.

Stratified sampling offers two advantages over simple random sampling.

1. Stratified sampling insures that we can make estimates for specific subpopulations. In simple random sampling, a subpopulation may not occur in the sample or it may occur so seldom that subpopulation estimates are unreliable. It may be important to have an estimate of a proportion on a statewide basis and to also have estimates on a countywide basis or for the set of urban counties and the set of rural counties.

2. If the strata can be constructed so that there is less variability in the variable of interest within the strata than in the population as a whole, then the standard errors of our estimators are smaller for stratified sampling than for simple random sampling. In many cases the reduction in the standard error due to stratification can be tremendous. It is sometimes difficult to know what variable on which to stratify. Note that it can not be the variable of interest, since we must stratify before we make our measurements. It must be a variable which is known for all members of the population and ideally it is highly related to the variable of interest. For example, if we wished to estimate the number of acres in South Carolina planted in corn, we might wish to stratify based on the farm size.

Stratification can but not always would reduce variability in estimate & weight

The estimators used with stratified sampling are pieced together from simple random sampling estimators computed for the strata. For example, the population total is the sum of the strata total. Thus, we separately estimate the total for each strata and then add the results. In symbols we write the stratified total estimate as

$$\Sigma_{i=1}^{K} N_i \bar{X}_i,$$

where $\bar{X}_i$ is the mean of the items drawn from the ith stratum. The stratified estimator of the population mean is the estimator of the total divided by N. That is,

$$\bar{X}_{st} = \Sigma (N_i/N) \bar{X}_i. \quad = \sum_{i=1}^{K} \frac{N_i \bar{X}_i}{N}$$

For the stratified estimator of the population proportion, we replace $\bar{X}_i$ by $\hat{p}_i$. That is,

$$\hat{p}_{st} = \Sigma (N_i/N) \hat{p}_i.$$

$N_i \bar{X}_i$ - estimate of total for i^{th} stratum

To compute the estimated standard error of $\bar{X}_{st}$, we first compute the sample variance, s_i^2, for the items from the ith stratum, i = 1,...,k. The estimated standard error of the stratified estimator of the mean when sampling without replacement is

$$\{\Sigma_{i=1}^{K}(N_i/N)^2\,[(s_i^2/n_i)(N_i-n_i)/N_i]\}^{1/2}.$$

The estimated standard error of the stratified estimator of the total when sampling without replacement is the same formula multiplied by N. The estimated standard error of the stratified estimator of the population proportion when sampling without replacement is

$$\{\Sigma_{i=1}^{K}(N_i/N)^2\,[\hat{p}_i(1-\hat{p}_i)/(n_i-1)]\,[(N_i-n_i)/N_i]\}^{1/2}.$$

In each case the estimated standard error come from piecing together the results of the independent simple random samples. We utilize the fact that for independent random variables $\mathrm{Var}(\Sigma a_i X_i)=\Sigma a_i^2 \mathrm{Var}(X_i)$.

EXAMPLE The city of Ferndale has two residential areas. The first area with 1000 households is comprised of typically lower income families. The second area with 200 households is comprised of typically upper income families. It is of interest to estimate the mean household income in Ferndale. It is decided to use stratified sampling with household income determined for 100 households in the first area and 20 in the second area. The data is summarized as

AREA	N_i	n_i	$\bar{x}_i$	s_i^2
1	1000	100	11,376	7,595,536
2	200	20	65,682	374,577,320

The stratified estimate of the mean is

$$(1000/1200)\ 11376 + (200/1200)\ 65682 = 20427.$$

The estimated standard error of the estimate is

$$\{(1000/1200)^2\,[(7595536/100)\,(1000-100)/1000] + (200/1200)^2\,[(374577320/20)\,(200-20)/200]\}^{1/2} = 718.$$

Thus, we are approximately 95% confident that the mean household income is between $20,427 \pm 2(718)$. That is ($18,991,$21,863). The total income for the 1200 household is estimated to be 1200(20,427)= $24,512,400 and the estimated standard error of the estimate is 1200(718)=$861,600. We are approximately 95% confident that the total income is in the interval $24,512,400 $\pm$ 2(861,600).

In this example we sampled the same proportion of observations from each stratum. Since the variance is larger in stratum 2, it

would have been advantageous to take a larger proportion of the items from stratum 2 than from stratum 1. See Cochran for futher details.

In this example, there is a big advantage in using stratified sampling over simple random sampling. If we had been using simple random sampling and had obtained the same sample, our data would have yielded $\bar{x} = 20427$ and $s^2 = 479,170,490$. Thus, our point estimate of the mean would have been the same, but the estimated standard error would have been $\{(479170490/120)[(1200-120)/1200]\}^{1/2}=1896$. Thus, effective stratification has reduced the estimated standard error of the mean from \$1896 to \$718.

9/22 ⟹ PROBLEM 9. In this problem provide enough detail so that I can check your work. I need to check how you selected your sample and your calculations.

a) Using the table of random numbers, draw a simple random sample without replacement of size 6 from the population listed on page 12. Give an approximate 95% confidence interval for the mean weight.

b) Repeat part a using a stratified random sample without replacement where you stratify based on sex and select 3 individuals from each stratum.

c) Comment of the difference in the answers in parts a and b. Would you expect similar differences if the variable of interest had been age?

TOPIC 5 - SAMPLING DISTRIBUTIONS

In this section we wish to state facts about the probability distribution of $\hat{p}$, $\bar{X}$, and S^2. Let us begin with $\bar{X}$.

Let $X_1,\ldots,X_n$ be independent outcomes of a random variable having distribution function F. These outcomes may come from independent trials of a experiment such as measuring the resistance of capacitors being produced on an assembly line or they may be survey results with the sampling done with replacement. We are interested in the probability distribution of $\bar{X}$. The probability distribution of $\bar{X}$ depends on F. It follows from the laws of expectation that if F has a finite mean μ then $E(\bar{X}) = \mu$. Futhermore, if F has a finite variance σ^2, then $Var(\bar{X}) = \sigma^2/n$. Thus, the variability of $\bar{X}$ decreases with n.

We can say more about the distribution of $\bar{X}$ for large n provided that the variance of F exists. In this case the distribution of $Z = n^{1/2}(\bar{X} - \mu)/\sigma$ approaches the standard normal as n approaches ∞. This result is known as the CENTRAL LIMIT THEOREM. Thus provided F has a finite variance the distribution of $\bar{X}$ is approximately normal with mean μ and variance σ^2/n. The precision of this approximation depends upon F and improves as n increases.

The display on the following page illustrates the central limit theorem. The top row of figures depict density functions corresponding to four different distribution functions. These figures give us the density function of $\bar{X}$ when n=1. The second row of figures depict the density function of $\bar{X}$ based on samples of size two from the same four distributions. The third and fourth row of figures depict the density function of $\bar{X}$ based on samples of size five and thirty for the same four distributions. Notice how different the four figures are in the top row and how similar they are in the bottom row.

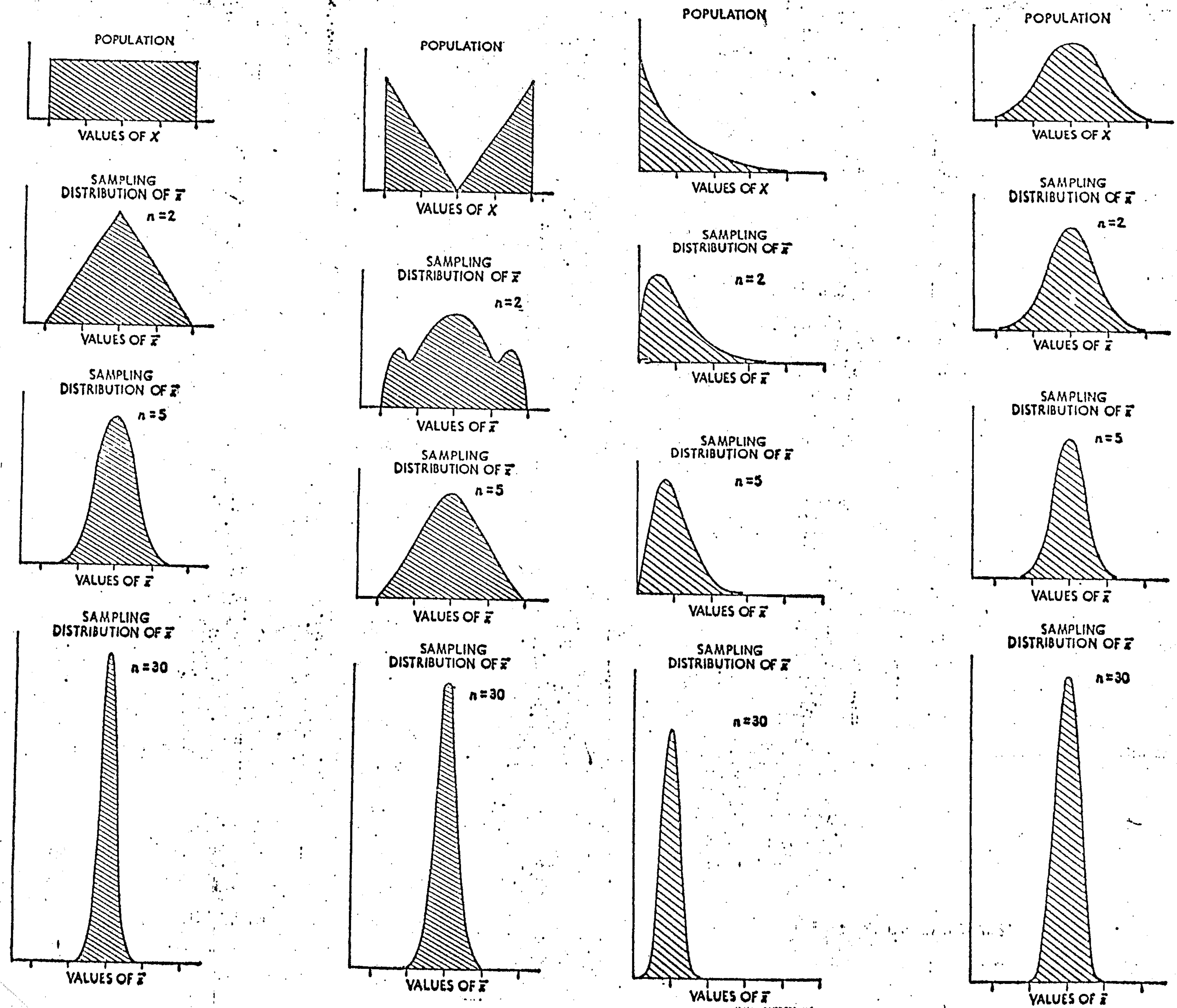
If F is a normal distribution, then the distribution of $\bar{X}$ is normal with mean μ and variance σ^2/n. That is, the approximation is exact for all n.

If F is a normal distribution, then the random variable $(n-1)S^2/\sigma^2$ has a chi square with n-1 degrees of freedom. Recall from page 24 that the chi square distribution is a special case of the Gamma family of distributions. It follows from page 23, that the expected value of a chi square random variable equals the number of degrees of freedom. Thus, $E(S^2) = \sigma^2$. A table of the probability points of chi square distributions is found in most elementary statistics books.

If F is not a normal distribution, then $(n-1)S^2/\sigma^2$ does not have chi square distribution. However, if F has a finite variance σ^2, then $E(S^2) = \sigma^2$. Thus, S^2 is an unbiased estimator of σ^2.

In making inferences about μ, we will be interested in the random variable $T = n^{1/2}(\bar{X} - \mu)/S$. This is the variable Z that we discussed in the central limit theorem with σ replaced by its estimator S. If F is a normal distribution, the random variable T has the STUDENT'S T DISTRIBUTION WITH n-1 DEGREES OF FREEDOM. The name student refers to the fact that the originator of this distribution W. S. Gosset

THE EFFECT OF SAMPLE SIZE AND SHAPE OF UNIVERSE ON THE DISTRIBUTION OF MEANS OF RANDOM SAMPLES



published has results under the pen-name Student. The T distribution is continuous and is symmetric about zero. The density function of the t distribution with ν degrees of freedom is given by

$$f(x) = \{\Gamma[(\nu+1)/2]/[(\pi\nu)^{1/2}\Gamma(\nu/2)]\}\ (1+x^2/\nu)^{-(\nu+1)/2}.$$

As ν approaches ∞, the T distribution approaches the standard normal. Most elementary statistics books provide tables of the probability points of the T distribution.

If F is not a normal distribution, then the variable T does not have the T distribution. However, if F has a finite variance then the distribution of the variable T still approaches the standard normal as n approaches ∞.

In a binomial experiment we perform n independent trials each having probability of success p. The total number of successes X has a binomial distribution with parameters n and p. The expected value and variance of X are np and np(1-p), respectively. An estimator of p is $\hat{p} = X/n$. From the laws of expectation, the expected value and variance of $\hat{p}$ are p and p(1-p)/n, respectively. Thus, the variance of $\hat{p}$ decreases with n.

If we write $X = \Sigma_{i=1}^{n} X_i$, where $X_i = 1$, if the ith trial is a success and 0, otherwise. We see that $\hat{p}$ can be thought of as a sample mean of the X_i's. Thus for $0 < p < 1$, by the central limit theorem the distribution of $(\hat{p}-p)/[p(1-p)/n]^{1/2}$ approaches the standard normal as n approaches ∞. Equivalently, for large n, the binomial distribution can be approximated by the normal distribution with mean np and variance np(1-p). The approximation improves as n increases and is better for values of p close to 1/2. For extremely small values of p the distribution of X can be approximated by a Poisson distribution with parameter $\lambda = np$. Poisson approx to Binomial

Approximating a discrete binomial distribution by a continuous normal distribution causes some technical difficulties. This arises from the fact that the binomial distribution assigns probabilities to integer values and the probability of integer values under the normal distribution is zero. To correct for this we generally approximate the probability that the binomial variable X equal an integer x by the probability that the normal variable falls in the interval $x \pm 1/2$.

$$E(X^2) - (E(X_i))^2 =$$

$$p - p^2 = p(1-p)$$

For example with n=20 and p=1/2, we approximate P(X=10) by the probability that a normal random variable with mean 20(1/2)=10 and variance 20(1/2)(1/2)=5 lies between 9.5 and 10.5. Similarly, we approximate P(X<10) by the probability that the same normal distribution is less than 9.5.

In making inferences about p, we will be interested in the random variable $(\hat{p} - p)/[\hat{p}(1-\hat{p})/n]^{1/2}$. It can be shown that for $0 < p < 1$, the distribution of this variable approaches the standard normal as n approaches ∞.

One example of a binomial experiment comes from sampling with replacement and determining if each sampled item has a characteristic of interest. In this case p is the population proportion. Note that when we sample without replacement, the distribution of X is hypergeometric. If n is small relative to N, then the binomial distribution provides a good approximation to the hypergeometric.

TOPIC 6 - HYPOTHESIS TESTING

Hypothesis testing is a major branch of statistical inference. To understand the role of hypothesis testing in science, we will begin by reviewing the SCIENTIFIC METHOD. The scientific method was formulated by Francis Bacon in the early 1600s.

1. State a hypothesis
   The hypothesis is a statement or conjecture about the state of nature or reality. The hypothesis is made based on the knowledge of science.
2. Perform an experiment
   The experiment is designed to show whether or not the hypothesis is true.
3. Make a conclusion about your hypothesis.
4. Formulate a new hypothesis
   The new hypothesis is based on the past knowledge of science plus that gathered through the current experiment.

We see that hypotheses are a central part of the scientific investigation. We should also note a role of the statistician throughout the process. The statistician can help in formulating the hypotheses, designing the experiment, and in analyzing the experimental data.

We formulate STATISTICAL HYPOTHESES by making statements about the population parameters or the distribution functions. Some examples of statistical hypotheses are the following:

The mean weight μ of the population of males is 160 pounds.

The proportion of undergraduate students who have smoked marijuana is 0.6.

The mean lifetime of GE light bulbs is greater than the mean lifetime of Sylvania light bulbs.

We see a number of advertizing claims that are in the form of hypotheses. Some examples are:

Nothing is more effective for fighting athlete's foot than Desenex.

STP Gas Treatment improves your gas mileage.

Schlitz drinkers are better lovers.

These hypotheses may or may not be correct.

The procedure of hypothesis testing is that we state two hypotheses, the NULL HYPOTHESIS H_0 and the ALTERNATIVE HYPOTHESIS H_1. The hypothesis that the experimenter wishes to prove is stated as the alternative hypothesis. The null hypothesis is the contradiction of the alternative hypothesis and generally contains $=$, $\leq$, or $\geq$. The method of proof in hypothesis testing is to try to prove the alternative hypothesis by showing beyond reasonable doubt that the null hypothesis is false. That is, the burden of proof is on the experimenter. If we are unable to show beyond a reasonable doubt that the null hypothesis is false, then our conclusion is that we can not reject the null hypothesis. Note that we do not prove that the null hypothesis is true.

We have a perfect analogy to the criminal court system where the accused is presumed innocent until proven guilty beyond a reasonable doubt. If there is reasonable doubt, then the accused is released. He is not required to prove that he is innocent.

Notice that we can make two types of errors. We can make a TYPE I ERROR of rejecting H_0 when in fact H_0 is true or we can make a TYPE II ERROR OF failing to reject H_0 when H_0 is false. We try to control the probability of each of these errors.

We refer to the probability of a type I error as the SIGNIFICANCE LEVEL of the hypothesis test. We denote the significance level by α.

due 9/24

11. A sample of 1200 registered nurses shows that 580 of them are employed by hospitals. Can we conclude at the 5% level that less than 50% of the population of nurses are employed by hospitals?

The decision whether or not to reject H_0 depends upon the choice of α. The smaller the value of α is, the stronger the evidence must be in support of H_1 before we reject H_0. The choice of α is subjective. It is very possible that different investigators might wish to use different values of α. Because of these facts it is convenient to report the results of a test of hypotheses in terms of the P-VALUE, also known as the OBSERVED LEVEL OF SIGNIFICANCE. The p-value has two equivalent definitions.

Interpreting

1. The p-value is the smallest value of α such that we would reject the null hypothesis.

Computing

2. The p-value is the probability under the null hypothesis of observing a result as extreme or more extreme in support of the alternative.

interpretation

The first definition is better for interpreting the meaning of a reported p-value. If the p-value = 0.0322, we know immediately that if $\alpha \geq 0.0322$ we reject H_0 and if $\alpha < 0.0322$ we fail to reject H_0.

computation

The second definition is better for computation of the p-value. In our example test of H_0: $\mu = 1.5$ versus H_1: $\mu \neq 1.5$, the value of the test statistic was -1.199. For the two sided alternative, large positive and large negative values of T support H_1. Thus,

$$\text{p-value} = P(T \leq -1.199 \text{ or } T \geq 1.199 | \mu=1.5).$$

Since the distribution of T is symmetric under H_0, we have

$$\text{p-value} = 2\ P(T \geq 1.199 \mid \mu = 1.5).$$

From the table for the t distribution with 19 degrees of freedom, we bound $P(T \geq 1.199 | \mu=1.5)$ between 0.10 and 0.20. Thus, the p-value is between 0.20 and 0.40. Hence, our result is not that unusual if H_0 is true.

For the one sided alternative H_1: $\sigma^2 > 0.01$, large values of χ^2 support H_1. Thus, the p-value is $P(\chi^2 \geq 104.927 | \sigma^2=0.01)$. The p-value can be bounded by looking at a table of the chi square distribution with 19 degrees of freedom. The p-value is less than 0.001. Thus, if H_0 is true, we have observed a most unusual event.

due 9/24 PROBLEM

12. An experiment with n = 13 yields a computed t statistic of 0.873. Report the p-values for the following alternatives: $\mu > \mu_0$, $\mu < \mu_0$, $\mu \neq \mu_0$.

13. If the p-value = 0.008, would H_0 be rejected at $\alpha = 0.01$?

TOPIC 7 - INTERVAL ESTIMATION

We have discussed estimating a population parameter by computing a point estimator. The point estimate is our best guess as to the value of the parameter. However, the point estimator is a random variable, and we know that it is very unlikely that the point estimator equals the true value of the parameter. With point estimation, we know our estimate is most likely wrong, and we hope that it is close to the true value.

In order to establish a reliable estimate of the parameter we must allow for the error in the estimation process. That is, we state that the unknown parameter falls in a specified interval on the real line rather than saying it equals a specific point. By taking this approach, we can calculate or approximate the probability that we are correct. By adjusting the width of the interval we can achieve any degree of reliability that we wish. However, increasing the reliability increases the width of the interval. An interval estimator that has probability $1 - \alpha$ of covering the value of the unknown parameter is known as a $(1-\alpha)100\%$ CONFIDENCE INTERVAL for the parameter. Confidence intervals can be derived using two essentially equivalent approaches. Depending on the setting, one approach is sometimes more intuitive than the other.

The first approach for constructing confidence intervals is known as the PIVOTAL method. In the pivotal method, we find a random variable which is a function of the sample values and the unknown parameter of interest. It can not be a function of other parameters. This random variable is known as the pivotal. The probability distribution of the pivotal must not depend on the unknown parameters. We write either an exact or approximate probability statement involving the pivotal. We then use algebraic manipulations

to rewrite the probability statement so that the parameter of interest is isolated and the other terms are functions only of the sample. The functions of the sample serve as the endpoints of our interval estimate.

EXAMPLE

If $X_1,\ldots,X_n$ is a random sample from a population having a normal distribution, then $T = (n)^{1/2}(\bar{X}-\mu)/S$ has a T distribution with n-1 degrees of freedom. The variable T is our pivotal for μ. Our probability statement is

$$1-\alpha = P(-t_{\alpha/2} < T < t_{\alpha/2}) = P(-t_{\alpha/2} < (n)^{1/2}(\bar{X}-\mu)/S < t_{\alpha/2})$$

Rewriting within the probability statement yields

$$1-\alpha = P(\bar{X} - t_{\alpha/2}S/n^{1/2} < \mu < \bar{X} + t_{\alpha/2}S/n^{1/2}).$$

Thus, we are $(1-\alpha)100\%$ confident that μ lies in the interval

$$\bar{X} \pm t_{\alpha/2}\ S\ /\ n^{1/2}.$$

Decreasing α, increases our degree of confidence and the value of $t_{\alpha/2}$, which increases the width of the interval.

For our example data, we had $\bar{X} = 1.437$, $S = 0.235$, and $n=20$. If we use $\alpha=0.05$, then with 19 degrees of freedom $t_{\alpha/2} = 2.0930$. Thus, we are 95% confident that the true value of μ lies in the interval $1.437 \pm 2.0930(0.235)/(20)^{1/2}$. Evaluating the interval yields 1.437 ± 0.110 or (1.327,1.547).

At times it is of interest to bound μ only on one side. This is accomplished by making the probability statement about the pivotal involve a one sided inequality. For example,

$$1-\alpha = P(T < t_\alpha) = P[n^{1/2}(\bar{X}-\mu)/S < t_\alpha].$$

Rewriting the inequality yields

$$1-\alpha = P[\mu > \bar{X} - t_\alpha\ S\ /\ n^{1/2}].$$

Thus, we are $(1-\alpha)100\%$ confident that μ exceeds $\bar{X} - t_\alpha S/n^{1/2}$. For our example data, $t_{.05} = 1.729$. Thus, we are 95% confident that $\mu > 1.437 - 1.729(0.235)/(20)^{1/2} = 1.346$. Notice that the one sided interval gives us a tighter lower bound than the two sided interval. The price of this tighter lower bound is that our upper bound is ∞.

The second approach to constructing confidence intervals is known as HYPOTHESIS TEST INVERSION method. When we reject the null hypothesis H_0: $\theta=\theta_0$ in favor of the alternative H_1: $\theta\neq\theta_0$, we are concluding that θ does not equal θ_0. That is, θ_0 can be ruled out as

a possible value of θ. In constructing a confidence interval, we wish to include all values of the parameter that we can not rule out. Thus, the hypothesis test inversion method for constructing a two sided $(1-\alpha)100\%$ confidence interval for a parameter is to include all values θ_0 such that we can not reject H_0: $\theta=\theta_0$ in favor of H_1: $\theta\neq\theta_0$ with a significance level of α.

In our example, we could not reject H_0: $\mu=1.5$ in favor of H_1: $\mu\neq1.5$ at $\alpha=0.05$. Thus, the point 1.5 falls within the 95% two sided confidence interval for μ. In order to conveniently use the hypothesis test inversion method, we need an algorithm to compute the endpoints of the confidence interval. In the case of confidence intervals for the mean of a normal distribution the algorithm follows the same steps as the pivotal method. We are looking for the set of values of μ_0 such that $-t_{\alpha/2} < n^{1/2}(\bar{X}-\mu_0)/S < t_{\alpha/2}$. The resulting confidence interval, $\bar{X}\pm t_{\alpha/2}S/n^{1/2}$ agrees with that found through the pivotal method.

One sided confidence intervals can be found by inverting tests of hypotheses with one sided alternatives. For example, to bound θ from below, we invert the test of H_0: $\theta \leq \theta_0$ versus H_1: $\theta > \theta_0$ with significance level α. The logic of this is that if we reject H_0, we are concluding that $\theta > \theta_0$, Thus, we can rule out all values of $\theta \leq \theta_0$.

In the case of confidence intervals for the mean of a normal distribution the algorithm for finding the endpoint of the one sided interval again follows the same steps as the pivotal method. We are looking for the set of μ_0 such that $n^{1/2}(\bar{X}-\mu_0)/S < t_\alpha$. The resulting confidence interval $(\bar{X} - t_\alpha S/n^{1/2},\infty)$ agrees with that found through the pivotal method.

In the case of the mean of the normal distribution, the two methods yield the same confidence intervals. The approximate confidence intervals for the binomial parameter P do not agree although the differences are quite small. The pivotal variable is

$$Z = (\hat{P}-P)/[\hat{P}(1-\hat{P})/n]^{1/2}$$

which is approximately distributed as the standard normal. The resulting approximate 95% confidence interval for P is

$$\hat{P} \pm 1.96[\hat{P}(1-\hat{P})/n]^{1/2}.$$

For the hypothesis test inversion method, we wish to find the set of values of P_0 such that $-1.96 < (\hat{P}-P_0)/[P_0(1-P_0)/n]^{1/2} < 1.96$. The endpoints are the solutions for P_0 of the equations

$$\hat{P}-P_0 = \pm 1.96[P_0(1-P_0)/n]^{1/2}.$$

The pivotal method yields a simpler solution and is generally used.

In the case of the variance of a normal distribution the two methods agree. The pivotal variable is $(n-1)S^2/\sigma^2$ which has a chi square distribution with n-1 degrees of freedom. The two sided probability statement is $P(\chi^2_{1-\alpha/2} < (n-1)S^2/\sigma^2 < \chi^2_{\alpha/2}) = 1-\alpha$. Rewriting the statement yields

$$P[(n-1)S^2/\chi^2_{\alpha/2} < \sigma^2 < (n-1)S^2/\chi^2_{1-\alpha/2}] = 1-\alpha.$$

Thus, the $(1-\alpha)100\%$ two sided confidence interval for σ^2 is

$$((n-1)S^2/\chi^2_{\alpha/2},\ (n-1)S^2/\chi^2_{1-\alpha/2}).$$

For our example, n=20, thus we have 19 degrees of freedom. For a 95% two sided confidence interval, we find $\chi^2_{.975}=8.91$ and $\chi^2_{.025}=32.9$. The sample variance was $0.235^2 = 0.0552$. Thus, the 95% confidence interval for σ^2 is (19(0.0552)/32.9, 19(0.0552)/8.91) which simplifies to (0.0319, 0.1178).

not symmetric around point due to χ^2 nonsymmetric

PROBLEMS

done 9/24

14. Using the data from problem 10 on page 40, find
a) a 95% confidence interval for the mean reaction time,
b) a 95% lower confidence bound on the mean reaction time,
c) a 95% confidence interval for the population variance,
d) a 95% confidence interval for the population standard deviation.

15. Using the data from problem 11 on page 41, find
a) a 95% confidence interval for the population proportion,
b) a 99% confidence interval for the population proportion.

TOPIC 8 - COMPARING TWO TREATMENTS (NORMAL THEORY)

A major role of the applied statistician is to assist the experimenter in designing the experiment so that the experimenter gets the most possible information from the data. To do this the statistician must be involved in the project from the beginning. We will introduce the concept of EXPERIMENTAL DESIGN through the problem

A study is designed to determine if a weight loss plan is more effective for obese men than for obese women. The variable of interest is the percentage change in body weight over a three month period. That is,

100 * (begin weight - end weight) / begin weight.

Positive scores indicate weight loss and negative scores indicate weight gain. A sample of $n_1 = 20$ women yielded $\bar{X}_1 = 6.32$ and $S_1^2 = 4.39$. A sample of $n_2 = 18$ men yielded $\bar{X}_2 = 9.14$ and $S_2^2 = 11.34$. We wish to test H_0: $\mu_1 \geq \mu_2$ versus H_1: $\mu_1 < \mu_2$ at $\alpha = 0.05$. Large negative T support H_1. Let us compute k and ν.

$$k = (4.39/20) / [(4.39/20) + (11.34/18)] = 0.2584$$

$$\nu = 1 / \{[(0.2584)^2/(20-1)] + [(1-0.2584)^2/(18-1)]\} = 27.88$$

Looking in the t table for 27.88 degrees of freedom we find that we will reject H_0 if $T \leq -1.701$. Our test statistic is

$$T = [(6.32-9.14)-0]/[(4.39/20)+(11.34/18)]^{1/2} = -3.060.$$

Thus, we reject H_0 at $\alpha=0.05$ and conclude that the plan provides men a mean percentage weight loss which is statistically significantly larger than women. The p-value is between 0.001 and 0.0025.

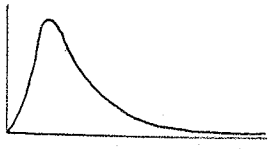
A 95% confidence interval for $\mu_1-\mu_2$ is

$$(6.32-9.14) \pm 2.048\ [(4.39/20) + (11.34/18)]^{1/2}$$

$$-2.82 \pm 1.89$$

Thus men on average lose between 0.93 and 4.71 more percentage points of weight than women do.

You may now be asking "How do I know if $\sigma_1^2 = \sigma_2^2$?". Note that this is a statistical hypothesis that can be tested. In order to test this hypothesis we need to introduce the F DISTRIBUTION. Let X have a chi square distribution with ν_1 degrees of freedom and Y have a chi square distribution with ν_2 degrees of freedom. If X and Y are independent, then $F = (X/\nu_1)/(Y/\nu_2)$ has an F distribution with ν_1 and ν_2 degrees of freedom. ν_1 is called the numerator degrees of freedom and ν_2 is called the denominator degrees of freedom. If $\nu_2 > 2$, then E(F) exists and equals $\nu_2/(\nu_2-2)$. If $\nu_2 > 4$, then Var(F) exists and equals $2\nu_2^2(\nu_1+\nu_2-2)/[\nu_1(\nu_2-2)^2(\nu_2-4)]$. The distribution is bounded below by 0 and is unbounded above. These distributions have a long right tail. It is of interest to note that 1/F has an F distribution with ν_2 and ν_1 degrees of freedom. Probability points for the F



F-distribution

Table C3

distribution are generally tabled for only the upper tail. That is, in terms of points F_α such that $P(F \geq F_\alpha) = \alpha$ for $\alpha = 0.10, 0.05, 0.025$, and 0.01. To get a lower tail point, we look up the upper tail point for the F distribution with the roles of the numerator and denominator degrees of freedom reversed and then invert the result. For example, if $\nu_1 = 4$ and $\nu_2 = 10$, the fact that $P(F > 4.47) = 0.025$ can be determined directly from the F table. To find the point such that the probability of being less than or equal to the point is 0.025, $F_{.975}$, we look at the F table with 10 and 4 degrees of freedom and $\alpha = 0.025$ and find the value 8.84 and then take the inverse $F_{.975}=1/8.84 = 0.113$. Thus, with $\nu_1=4$ and $\nu_2=10$, $P(F \leq 0.113)=0.025$.

We know that $(n_1-1)S_1^2/\sigma_1^2$ has a chi square distribution with n_1-1 degrees of freedom and $(n_2-1)S_2^2/\sigma_2^2$ has a chi square distribution with n_2-1 degrees of freedom provided that we are sampling from populations with normal distributions. In the independent sample design $(S_1^2/\sigma_1^2) / (S_2^2/\sigma_2^2)$ has an F distribution with n_1-1 and n_2-1 degrees of freedom.

Under the assumption that $\sigma_1^2=\sigma_2^2$, the statistic $F=S_1^2/S_2^2$ has an F distribution with n_1-1 and n_2-1 degrees of freedom. Values of $F > 1$ suggest $\sigma_1^2 > \sigma_2^2$ and values of $F < 1$ suggest the reverse. To test H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ at level α, we reject H_0 if the test statistic F is $\leq$ the lower $\alpha/2$ point or if $F \geq$ the upper $\alpha/2$ point of the F distribution with n_1-1 and n_2-1 degrees of freedom.

We can get a $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2 by inverting the probability statement

$$P[F_{1-\alpha/2} < (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2) < F_{\alpha/2}] = 1-\alpha.$$

12/14 ⇒ PROBLEM 16. Show that the above probability statement yields the following $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2:

$$([S_1^2/S_2^2]/F_{\alpha/2},\ [S_1^2/S_2^2]/F_{1-\alpha/2}).$$

EXAMPLE

Suppose that two independent samples yield the following results: $n_1=8$, $\bar{X}_1=43.5$, $S_1^2=142.5$ and $n_2=11$, $\bar{X}_2=34.6$, $S_2^2=65.4$. We wish to test H_0: $\sigma_1^2=\sigma_2^2$ versus H_1: $\sigma_1^2\neq\sigma_2^2$ at $\alpha=0.05$. With 7 and 10 degrees of freedom, we find $F_{.025}=3.95$ and $F_{.975}=1/4.76=0.210$. Thus, we

reject H_0 if $F \geq 3.95$ or $F \leq 0.210$. Because $F = 142.5/65.4 = 2.18$, we can not reject H_0 at $\alpha = 0.05$.

A 95% confidence interval for σ_1^2/σ_2^2 is $(2.18/3.95, 2.18/0.210)$ which simplifies to $(0.552, 10.38)$.

sample variances are very unstable, which leads to large CI

PROBLEMS

17. Repeat the test of hypotheses in the example with $\alpha = 0.10$. What is the p-value?

18. Follow the logic at the bottom of page 40 to derive the power of the test of equality of variances. What can you say about the power of the $\alpha = 0.05$ test in the example if in fact $\sigma_1^2 = 10\sigma_2^2$.

Note that the confidence interval in the example is quite wide. We need large sample sizes to get narrow intervals or to get high power in tests of hypotheses.

It is at times of interest to use the one sided alternative H_1: $\sigma_1^2 > \sigma_2^2$. This could be of interest in industrial settings where we might be trying to prove that method 2 produces less variability than method 1. In this case we would reject H_0 if $F \geq F_\alpha$.

It is also possible to test H_0: $\sigma_1^2/\sigma_2^2 = c$. In that case our test statistic is $F = (S_1^2/S_2^2)/c$. The rejection region is unchanged.

The T tests comparing means do not depend heavily on the assumption of normal population when we have large sample sizes. However, the F test for comparing variances does depend heavily on the normality assumption even for large sample sizes. That is, if we sample from highly non-normal populations then the true probability of making a type I error may differ greatly from the nominal or stated α value even for large sample sizes. We say that T tests are ROBUST to moderate departures from normality for large sample sizes. The F test is not robust. The fact that the T tests are robust do not mean they are the best test in the face of non-normality. The branch of statistics known as NONPARAMETRIC STATISTICS deals with methods derived without assuming a parametric family of distributions such as the normal family. For more information about nonparametric statistics, see Nonparametric Statistical Methods by Myles Hollander and Douglas Wolfe.

In using the F test as a preliminary test to decide which version of the independent sample T test to use, it is best to let α be large such as 0.20 for small and medium sample sizes since falsely rejecting the null hypothesis of equal variances in this case would not be a very serious error.

10/14 ⇒ PROBLEMS

19. A study is designed to compare the effect of an experimental mixture of cement to the standard mixture. The measure of interest is the amount of force required to break a concrete post made from the cement. The data consists of breaking strengths for 20 posts of each type of cement. The data are summarized by

	Experimental	Standard
sample size	20	20
sample mean	325.6	291.8
sample variance	432.4	398.4

a. Give a 95% confidence interval for the ratio of the variance for the experimental to the variance of the standard.
b. Is there sufficient evidence to conclude at the 5% level that the experimental cement yields stronger concrete than the standard cement? What is the p-value?
c. Give a 95% confidence interval for the difference in means. How would you explain these results to someone who has never studied statistics?

20. In a paired comparison experiment the boys in Ms. Proctor's 3rd grade class are timed in running two 40 yard dashes, one wearing Nike shoes and one wearing Keds shoes. The data are:

BOY	JDS	GDW	DRE	WAQ	TYU	OSS	PPW	QQW	KDE	WAS	MCM	AAW	IYZ
NIKE	7.5	9.3	8.1	5.9	7.3	8.5	9.9	6.5	8.3	8.7	9.2	7.3	6.7
KEDS	7.8	9.2	8.7	5.4	7.0	7.9	9.4	6.8	7.9	8.8	9.5	7.7	6.2

a. Can we conclude at $\alpha = 0.05$ that there is a difference in the effects of the shoes on speed? What is the p-value?
b. What suggestions would you make regarding the order of running in the shoes, i.e. Nike first and then Keds? Why?
c. What suggestions would you make regarding the length of time between the two dashes? Why?

Continued ...

21. A study is designed to compare the average salt content in two brands of cereal. A random sample of 15 boxes of cereal G yields the following results: 530, 518, 525, 534, 536, 529, 532, 524, 527, 523, 519, 522, 529, 526, 531. A random sample of 15 boxes of cereal W yields 483, 492, 520, 510, 502, 487, 513, 490, 528, 499, 463, 541, 495, 513, 525. Can we conclude at the 5% level that the two cereals differ in average salt content? What is the p-value?

It is very easy to perform these tests of hypotheses in SAS. For the one sample problem, PROC UNIVARIATE includes the test statistic for testing H_0: $\mu=0$ and the p-value for the two-sided alternative. If you wish to test that the mean of a variable X is 23, you define a new variable Y=X-23 and then test that the mean of Y is 0. For one sided alternatives, you divide the two sided alternative in half if the point estimate supports the alternative and take 1 minus half the p-value for the two sided alternative if the point estimate does not support the alternative.

For the dependent sample t test, the SAS data set would contain the variables PAIR, X1, and X2. One then defines the new variable DIFF = X1-X2 and runs PROC UNIVARIATE on the variable DIFF. This yields a test of H_0: $\theta_1-\theta_2 = 0$ versus the two sided alternative. The interpretation of the p-value is the same as above.

For the independent sample problem, the SAS procedure is PROC TTEST. The SAS statement to run PROC TTEST is

```
PROC TTEST;CLASS variable1;VAR variable2;
```

This statement tells the computer to divide the data set into two groups based on the value of variable 1 and to analyze the data using the variable variable2. The analysis gives a test of equality of variances versus a two-sided alternative and both versions of the t-test for equality of means versus a two-sided alternative. Again, one can adjust the p-value for one-sided alternatives.

The significance level is our measure of how strong the evidence must be in support of H_1 in order to reject H_0. That is, the significance level measures the amount of doubt that we think is reasonable. With $\alpha = 0.05$, we are willing to falsely reject a true H_0 five percent of the time. With $\alpha = 0.01$, we are willing to reject a true null hypothesis one percent of the time. It is harder to reject H_0 with $\alpha = 0.01$ than with $\alpha = 0.05$.

We refer to the probability of a type II error as the OPERATING CHARACTERISTIC of the test. We denote the operating characteristic by β. It is at times convenient to talk about the probability of rejecting H_0. This probability, which equals $1-\beta$, is known as the POWER of the test. The value of β depends on the true value of the parameter under H_1. That is, the operating characteristic is a function of the parameter. The same is true for the power.

Ideally, we would like to have α and β both close to zero. The value of α is chosen by the experimenter. The value of β depends on the choice of α and on the sample size. For fixed α, β decreases as our sample size increases. Thus, in determining our sample size we need to consider the β function as well as our dollar costs.

EXAMPLE - TEST OF BINOMIAL PARAMETER (ONE-SIDED ALTERNATIVE)

An experimenter wishes to prove that an experimental type of chemotherapy has a probability greater than 0.20 of producing a desired effect on patients having a particular type of cancer. He wishes to use α approximately equal to 0.05 and to have β approximately equal to 0.10 if the true probibility is 0.40. He is considering an experiment with $n = 15$ patients.

The hypotheses are H_0: $P \leq 0.2$ and H_1: $P > 0.2$. Let X denote the number of patients exhibiting the desired effect. Large values of X support the alternative hypothesis. The probability distribution of X is binomial with parameters n and P. We desire α to be approximately 0.05, so we need to find a value c such that

$$P(X \geq c | P=0.2) \text{ is approximately } 0.05.$$

The test would then reject H_0 if $X \geq c$. The following table gives the binomial probabilities for n=15 and p=0.2 and 0.4:

	p=0.2		p=0.4	
c	P(X=c)	P(X≥c)	P(X=c)	P(X≥c)
0	0.0352	1.0000	0.0005	1.0000
1	0.1319	0.9648	0.0047	0.9995
2	0.2309	0.8329	0.0219	0.9948
3	0.2501	0.6020	0.0634	0.9729
4	0.1876	0.3519	0.1268	0.9095
5	0.1032	0.1643	0.1859	0.7827
6	0.0430	0.0611	0.2066	0.5968
7	0.0138	0.0181	0.1771	0.3902

We see from the third column that $P(X \geq 6 | P=0.2)=0.0611$. Thus, if we use the decision rule to reject H_0 if $X \geq 6$, then $\alpha = 0.0611$. We see from the last column that $P(X \geq 6 | P=0.4)=0.5968$. Thus, if we use this decision rule then the power of our test is 0.5968 at P=0.4. This yields a value of $\beta = 1-0.5968 = 0.4032$ which is larger than the experimenter wanted. There are two solutions 1) take a larger sample size or 2) keep this sample size and realize that if P=0.4 you are going to make the wrong decision with probability 0.4032. (β)

How large does n have to be to have the power at P=0.4 approximately equal to 0.9? To answer this question, we will use the normal approximation to the binomial distribution. Recall for large n that $Z = (X-np)/[np(1-p)]^{1/2}$ is approximately distributed as the standard normal distribution. Thus,

$$P\{[X-n(0.2)]/[n(0.2)(0.8)]^{1/2} \geq 1.645 \mid P=0.2\} \approx 0.05.$$

Rewriting yields

$$P\{X \geq n(0.2) + 1.645[n(0.2)(0.8)]^{1/2} \mid P=0.2\} \approx 0.05.$$

Thus, for large n we reject H_0 if $X \geq n(0.2) + 1.645[n(0.16)]^{1/2}$. The power of the test at P = 0.4 is

$$P\{X \geq n(0.2) + 1.645[n(0.16)]^{1/2} \mid P=0.4\}$$

which can be rewritten as

$$P\{[X-n(0.4)]/[n(0.4)(0.6)]^{1/2} \geq [n(0.2)+1.645(0.16n)^{1/2}-n(0.4)]/[n(0.4)(0.6)]^{1/2} \mid P=0.4\}$$

This can be approximated by

$$P\{Z \geq [n(-0.2)+1.645(0.16n)^{1/2}]/[0.24n]^{1/2}\}.$$

The problem is solved by setting the right hand side equal to -1.282 and solving for n, because $P(Z \geq -1.282) = 0.90$. Thus, we have

$$-0.2n + 1.645(0.16n)^{1/2} = -1.282(0.24n)^{1/2}.$$

⇐ see facing page

Simplifying yields $-0.2n = n^{1/2}[-1.286]$. Solving for n yields n = 41.3. Thus, to achieve a power of 0.9 at P=0.4, we must use approximately 41 patients in the study.

Let us now formally state the steps involved in a test of hypotheses.

1. Formally state the hypotheses that you wish to prove as the alternative hypothesis and its contradiction as the null hypothesis.
2. Choose the level of significance.
3. Choose a test statistic whose distribution is known under the point of equality in the null hypothesis.
4. Find the rejection region, the range of values of the test statistic such that you will reject H_0.
5. If possible, compute β for several values of the parameter under the alternative hypothesis. If these values of β are not acceptable, adjust the sample size.
6. Collect the data.
7. Compute the test statistic and make your decision.
8. Write the decision in terms of the original problem.

In our example, step 8 would result in one of the following two statements being written:

"At the 0.05 level of significance, the response rate was shown to be statistically significantly larger than 0.20."

"At the 0.05 level of significance, we could not reject the null hypothesis that the response rate was less than or equal to 0.20."

Let us now review the test statistics and decision rules for testing hypotheses about μ, σ^2, and P.

TESTS INVOLVING μ

Assumption: $X_1,\ldots,X_n$ is a random sample from a normal distribution.

Test statistic: $T = (n)^{1/2}(\bar{X} - \mu_0)/S$ $\qquad T = \sqrt{n}\,\frac{\bar{X}-\mu_0}{S}$

Hypotheses: H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$ where μ_0 is a constant.
H_0: $\mu \geq \mu_0$ versus H_1: $\mu < \mu_0$,
H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$.

Null distribution: If $\mu = \mu_0$, T has a t distribution with n-1 degrees of freedom.

Support for alternative: Positive T support $\mu > \mu_0$ and negative T support $\mu < \mu_0$.

Decision rule: Rejection region depends on alternative used. For alternative $\mu > \mu_0$, the rejection rule is reject H_0 if $T \geq t_\alpha$ where $P(T \geq t_\alpha | \mu=\mu_0)=\alpha$. For the alternative $\mu < \mu_0$, the rejection rule is to reject H_0 if $T \leq -t_\alpha$. For the alternative $\mu \neq \mu_0$, the rejection rule is to reject H_0 if $|T| \geq t_{\alpha/2}$.

TESTS INVOLVING σ^2

Assumption: $X_1,\ldots,X_n$ is a random sample from a normal distribution.

Test statistic: $\chi^2 = (n-1)S^2/{\sigma_0}^2$

Hypotheses: H_0: $\sigma^2 \leq {\sigma_0}^2$ versus H_1: $\sigma^2 > {\sigma_0}^2$, where ${\sigma_0}^2$ is a constant.

H_0: $\sigma^2 \geq {\sigma_0}^2$ versus H_1: $\sigma^2 < {\sigma_0}^2$.

H_0: $\sigma^2 = {\sigma_0}^2$ versus H_1: $\sigma^2 \neq {\sigma_0}^2$.

Null distribution: If $\sigma^2 = {\sigma_0}^2$, χ^2 has a chi square distribution with n-1 degrees of freedom.

Support for alternative: Large values of χ^2 support $\sigma^2 > {\sigma_0}^2$. Small values of χ^2 support $\sigma^2 < {\sigma_0}^2$.

Decision rule: Rejection region depends on alternative used. For alternative $\sigma^2 > {\sigma_0}^2$, the rejection rule is reject H_0 if $\chi^2 \geq \chi^2_\alpha$, where $P(\chi^2 \geq \chi^2_\alpha | \sigma^2 = {\sigma_0}^2) = \alpha$. For the alternative $\sigma^2 < {\sigma_0}^2$, reject H_0 if $\chi^2 \leq \chi^2_{1-\alpha}$. For the alternative $\sigma^2 \neq {\sigma_0}^2$, reject H_0, if $\chi^2 \geq \chi^2_{\alpha/2}$ or $\chi^2 \leq \chi^2_{1-\alpha/2}$.

TESTS INVOLVING P

Assumption: Binomial experiment.

Test statistic: X = number of successes in n trials

Hypotheses: H_0: $P \leq P_0$ versus H_1: $P > P_0$, where P_0 is a constant.

H_0: $P \geq P_0$ versus H_1: $P < P_0$.

H_0: $P = P_0$ versus H_1: $P \neq P_0$.

Distribution: X has a binomial distribution with parameters n and P. For large n, $Z=(x-np)/[np(1-p)]^{1/2}$ is approximately distributed as the standard normal.

Support for alternative: Large values of X (positive Z) support $P > P_0$. Small values of X (negative Z) support $P < P_0$.

Decision rule: Rejection region depends on alternative used. For

alternative $P>P_0$, reject H_0 for sufficiently large X (large positive Z). For alternative $P<P_0$, reject H_0 for sufficiently small X (large negative Z). For alternative $P\neq P_0$, reject H_0 for sufficiently large or small X (large positive or negative Z).

EXAMPLE

It is assumed that the diameter of a manufactured part follows a normal distribution. A buyer wishes to see if there is sufficient evidence to conclude that the mean diameter of items produced by the process differs from 1.500mm. The buyer selects a sample of 20 items from the production line and computes $\bar{X}$ = 1.437mm and S = 0.235mm.
The hypotheses are H_0: μ=1.500mm H_1: $\mu\neq$1.500mm
The test statistic is $T=(n)^{1/2}(\bar{X} - 1.500)/S$.
Under the null hypothesis the test statistic has a t distribution with 19 degrees of freedom. The value of $t_{.025}$ for 19 degrees of freedom is 2.0930. We will reject H_0 in favor of H_1 at the 5% level of significance if $|T|\geq 2.0930$.
The computed value of the statistic is

$$T = (20)^{1/2}(1.437-1.500)/0.235 = -1.199$$

We can not reject the null hypothesis at the 5% level. We do not have statistically significant evidence that the mean diameter differs from 1.500mm.

The buyer also wishes to investigate that the variance of the population of product diameters is not too large. He wishes to test at α=0.05 the null hypothesis that $\sigma^2\leq 0.0100$ versus the alternative that $\sigma^2>0.0100$. The test statistic is $\chi^2 = (19)S^2/0.0100$. We reject H_0 at the 5% level if $\chi^2 \geq 30.144$. Our data yields $\chi^2 = 19(0.235^2)/0.0100 = 104.927$. Hence, we reject H_0 at the 5% level and conclude that the variance is statistically significantly greater than 0.0100.

The power of tests of hypotheses involving μ is a function of μ/σ. That is, when testing H_0: μ = 10 versus H_1: $\mu \neq 10$ at α = 0.05 we can not compute the power of the test if μ = 12 without also specifying the value of σ. We can, however, compute the power for μ being one standard deviation greater than 10. The distribution of the test statistic $T=(n)^{1/2}(\bar{X} - \mu_0)/S$ when the null hypothesis is ~~true~~ false is a

noncentral t distribution. This distribution is discussed in detail in other courses. The sample size required to achieve a desired power can be found in Table A-12 of Introduction to Statistical Analysis (3rd Ed) by W. J. Dixon and F. J. Massey. For example, with a one sided alternative and $\alpha = 0.05$ the required sample size to achieve power equal to 0.9 for μ one σ away from μ_0 is 11 and for μ one half σ away from μ_0 is 37. Power can also be computed through the IMSL computer routines.

The power of tests of hypotheses involving σ^2 is a function of σ^2/σ_0^2. Consider for example the one sided alternative $H_1: \sigma^2 > \sigma_0^2$. We reject H_0 if $(n-1)S^2/\sigma^2 > \chi^2_\alpha$. The power of the test is

$$P[(n-1)S^2/\sigma_0^2 \geq \chi^2_\alpha | \sigma^2] = P[(n-1)S^2/\sigma^2 \geq (\sigma_0^2/\sigma^2)\chi^2_\alpha | \sigma^2].$$

The left handside of the last inequality has a chi square distribution with n-1 degrees of freedom. Thus, the power is the probability that the chi square value is greater than or equal to a constant. This probability can be bounded from the chi square table or evaluated using the computer. In our example of testing H_0: $\sigma^2 \leq 0.01$ versus H_1: $\sigma^2 > 0.01$ with $\alpha = 0.05$, if $\sigma^2 = 0.04$, the power of the test is the probability that a chi square variable with 19 degrees of freedom is $\geq (0.01/0.04)30.144 = 7.536$. A chi square table bounds this probability between .99 and .995. Thus, for this value of σ^2 we are almost certain to reject H_0.

9/24 ⇒

PROBLEMS

10. A psychologist measures the reaction times in seconds of 20 individuals exposed to visual stimuli. The data in seconds are:

0.85 0.76 0.73 0.58 0.61 0.74 0.71 0.65 0.75 0.56
0.64 0.59 0.78 0.58 0.71 0.73 0.70 0.78 0.74 0.69

a) Is the assumption that the data follows a normal distribution unreasonable? Why? Regardless of your answer, assume normality for the rest of the problem.

b) The psychologist wishes to prove that the population mean differs from 0.70 seconds. Is there sufficient statistical evidence at $\alpha = 0.05$? Is there sufficient statistical evidence at $\alpha = 0.10$?

c) Test H_0: $\sigma^2 = 0.01$ versus H_1: $\sigma^2 \neq 0.01$ at $\alpha = 0.05$.

of comparing two treatments under the assumption of normal distributions.

We wish to compare treatments 1 and 2. For example, we may be comparing two types of fertilizers or two types of gasoline. Another possibility is that treatment 1 is a real treatment and treatment 2 is no treatment. In this case, we call treatment 2 a CONTROL. Through the use of a control, we separate the real effect of the treatment from other factors involved in conducting the experiment. For example in one study involving a certain type of headache pain, twenty percent of the subjects reported that their pain was deminished after a nurse gave them a PLACEBO, a treatment which has no effect. Thus, in that study it was important to separate the effect of a pain medicine from the effect of receiving something from the nurse. This could be accomplished by giving some subjects the pain medicine and other subjects the placebo in an identical fashion. It is important that the patients be randomly assigned to treatments and that in each case they be made to believe that the pill may be effective.

Generally the experimenter wishes to prove that one treatment is more effective than the other treatment. For sake of this discussion, we will suppose that larger is better. In this type of study, the treatments are applied to experimental units (subjects, plots, cars, etc.) and responses are measured. In such studies the experimenter has the choice of two different designs, INDEPENDENT samples and DEPENDENT samples. The second design is also known as PAIRED COMPARISONS.

In the independent sample design we randomly assign experimental units to the treatments to achieve the sample sizes that we desire. In the dependent sample design we divide the experimental units into subsets of size two so that within a subset the units are as much alike as possible. We then randomly select one unit from each pair to receive treatment 1 and the other unit from the pair receives treatment 2.

In order to decide which design to use, we must consider the SOURCES OF VARIATION for our response variable. There is possible variation due to the treatment effect. That is, treatments 1 and 2

may differ in their average effect on the response. It is this effect that we are wishing to establish. We estimate this effect by computing $\bar{X}_1 - \bar{X}_2$, the sample mean of the responses for treatment 1 minus the sample mean of responses for treatment 2.

There is also possible variation due to differences in experimental units. That is, variation in the responses among units receiving the same treatment. In the independent samples design, the variation in the experimental units shows up in the standard error of $\bar{X}_1 - \bar{X}_2$. Thus with the independent sample design, the more variation there is among experimental units, the larger the standard error. The power function of the test of hypotheses involving the treatment effect is a decreasing function of the standard error. Thus, the more variation there is among experimental units receiving the same treatment, the less likely we are to detect a true difference among the two treatments.

In the dependent sample design, we seek to control most of the variation among experimental units by pairing. It is our goal in pairing to make the experimental units within the pair as alike as possible. If we are successful, then within a pair there will be very little variation among experimental units. We then estimate the difference in the treatment effects separately for each pair by subtracting the response for treatment 2 from the response for treatment 1. The data analysis is then done on the difference scores. The standard error of $\bar{X}_1 - \bar{X}_2$ in the dependent sample design is not effected by variation among experimental units appearing in different pairs. In the dependent sample design we can not estimate the standard error as well as we can in the independent sample design since the taking of differences uses up degrees of freedom that could have been used to estimate the standard error.

The decision between the independent and dependent sample designs depends on our assessment of the variation among the experimental units. If this variation is large and we are able to pair units so that within a pair the units are much more alike than in the population as a whole, then the advantage of the dependent sample design outweighs the disadvantage. If the variation among experimental units is small or if we can not pair the units so that

within a pair the units are much more alike than in the population as a whole, then the disadvantage of the dependent sample design outweighs the advantage.

In some experiments we can measure both treatments on the same experimental unit. In this way the is no variation within the pair. This method can not be used if the treatment alters the experimental unit if there is a learning effect. If the treatments are given sequentially, it is important that half of the units get treatment 1 first and the other half get treatment 2 first.

EXAMPLE

The Army wishes to compare two brands of combat boot heels designed for male soldiers. The variable of interest is the amount of wear to the heel in six months of wear. The experimental units are individual soldiers. The experimenter realizes that there are many sources of variation among the soldiers with respect to boot heel wear. Some of these factors are how much the soldier weighs, how he walks, where he walks, how often we walks, how well he cleans foreign objects from the heel, etc. Thus, the experimenter anticipates a large amount of variability among the experimental units. Therefore, a dependent sample design is used where each soldier is his own control. That is each soldier in the study will wear one type of heel on his left shoe and the other type on his right shoe. The choice of which heel goes on the left foot is made at random. At the end of six months, the amount of wear on each heel is measured and the difference between the type 1 heel and the type 2 heel is recorded.

EXAMPLE

An investigator wishes to test the effectiveness of two different types of fertilizer on plant growth in a greenhouse experiment. The experimental unit is a clay pot filled with a prepared planting mixture. Seeds from a common source are planted in each pot. Temperature, light, and moisture are held constant for all pots. In this example the variation between experimental units is probably very small. There may be small environmental differences and there may be differences in the seeds. However, it is difficult to distinguish these differences before planting. Thus, the

investigator would use an independent sample design with random assignment of the clay pots to the fertilizers.

We have considered the situation of comparing two treatments which are applied to experimental units. Another situation that arises is comparing two populations. In this case we select random samples from each of the two populations and compare the results. The samples are selected independently. In this situation there is no choice in design. Our analysis is that of the independent sample design.

Let us now turn to inferences for the two designs. We will begin with the dependent sample design. Let θ_1 denote the effect of treatment 1 and θ_2 denote the effect of treatment 2. We wish to make inferences for $\theta = \theta_1 - \theta_2$. If $\theta > 0$ then treatment 1 gives a larger expected response than treatment 2. Suppose there are n pairs. Let X_{1i} denote the response for treatment 1 in the ith pair and X_{2i} denote the response for treatment 2, $i = 1,\ldots,n$. The difference score for the ith pair is $Y_i = X_{1i} - X_{2i}$. The difference Y_i estimates θ for each pair. Thus, we can think of $Y_1,\ldots,Y_n$ as being a sample from a population centered at θ. We make all of our inferences for θ based on this sample of differences. If we assume that the difference scores follow a normal distribution then the inferences are the same as those for μ in a sample from a single normal population. The power of the dependent sample t test is a function of $(\theta-\theta_0)/\sigma$, where σ is the standard deviation of the population of difference scores. The power is determined in the same manner as in the one sample t test.

EXAMPLE

An experiment is designed to test the effectiveness of a treated tape to discourage barnacles from attaching to the bottom of boats. In the experiment, eight panels are placed in a tidal creek at different locations. Half of the panel is covered with the tape and the other half is left uncovered, i.e. is a control. The measurement of interest is the surface area covered by barnacles at the end of one year. If the treatment is effective, the treatment-control difference will be negative.

i	X_{1i}	X_{2i}	Y_i
PANEL	TREATMENT	CONTROL	TREATMENT - CONTROL
1	13.7	24.8	-11.1
2	9.7	18.3	-8.6
3	12.4	21.6	-9.2
4	6.3	13.8	-7.5
5	10.4	22.6	-12.2
6	8.9	15.2	-6.3
7	17.2	31.5	-14.3
8	4.3	9.8	-5.5
AVERAGE	10.3625	19.7000	-9.3375

The analysis is done on the differences. The sample standard deviation of the difference scores is S = 3.0194. To test for the effectiveness of the treatment, we test H_0: $\theta \geq 0$ versus H_1: $\theta < 0$ at $\alpha = 0.05$. The T statistic is

$T = (8)^{1/2}(\bar{Y} - 0)/S = (8)^{1/2}(-9.3375-0)/3.0194 = -8.75.$

The rejection region is based on a T distribution with 8 - 1 = 7 degrees of freedom. We reject H_0 at $\alpha = 0.05$ if $T \leq -1.895$. Clearly, we reject H_0 and conclude that the treated tape is effective in reducing the barnacle coverage. The p-value for this test is less than 0.0005.

A 95% confidence interval for the effect of the treatment is

$$-9.3375 \pm 2.365(3.0194)/(8)^{1/2}$$

$$-9.3375 \pm 2.5247$$

Test 1

In the independent sample design, we have n_1 observations from treatment 1 with a sample mean of $\bar{X}_1$ and a sample variance of S_1^2 and n_2 observations from treatment 2 with a sample mean of $\bar{X}_2$ and a sample variance of S_2^2. The analysis for the independent design can follow one of two courses. The deciding factor is whether we are willing to assume that the variances are equal for the two groups.

We will first consider the analysis under the assumption that the variances are equal. Under our assumption, the two sample variances estimate the common variance σ^2. Thus, we can combine the estimates into a single estimate. This combined estimate is known as the POOLED ESTIMATE of the variance. It is given by

$$S_p^2 = [(n_1-1)S_1^2 + (n_2-1)S_2^2]/[n_1 + n_2 - 2].$$ ← note (-2) !!

If we are sampling from normal populations with equal variances, then $(n_1+n_2-2)S_p^2/\sigma^2$ has a chi-square distribution with n_1+n_2-2 degrees of

freedom. The difference in the sample means, $\bar{X}_1-\bar{X}_2$, has a normal distribution with mean $\mu_1 - \mu_2$ and variance $\sigma^2[(1/n_1)+(1/n_2)]$ and the difference in the sample means is independent of the pooled estimator of the variance. Thus, if we are sampling from normal populations with equal variances, the random variable

$$[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)] \;/\; \{S_p^2[(1/n_1)+(1/n_2)]\}^{1/2}$$

has a T distribution with n_1+n_2-2 degrees of freedom. This variable serves as the pivotal for confidence intervals for $\mu_1-\mu_2$, and upon substituting a hypothesized value for $\mu_1-\mu_2$, it serves as a test statistic.

The $(1-\alpha)100\%$ confidence interval for $\mu_1-\mu_2$ is given by

$$\bar{X}_1-\bar{X}_2 \pm t_{\alpha/2}\{S_p^2[(1/n_1)+(1/n_2)]\}^{1/2}$$

where the $t_{\alpha/2}$ is the probability point from the T distribution with n_1+n_2-2 degrees of freedom.

Tests of hypotheses involving differences of the means involve statements of the type $\mu_1-\mu_2=\theta_0$, where θ_0 is a specified constant. Of particular interest is the case of $\theta_0=0$ which corresponds to the statement $\mu_1 = \mu_2$. The test statistic for hypotheses of this type is

$$T = [(\bar{X}_1-\bar{X}_2)-\theta_0]/\{S_p^2[(1/n_1)+(1/n_2)]\}^{1/2}.$$

Positive values of T suggest $\mu_1 - \mu_2 > \theta_0$ and negative values of T suggest $\mu_1 - \mu_2 < \theta_0$. The power of this t test is a function of $[(\mu_1-\mu_2)-\theta_0]/\sigma$. Table A-12 of Dixon and Massey gives the sample size necessary to achieve a particular power. For example, if we are trying to prove $\mu_1 > \mu_2$ with $\alpha = 0.05$ and desire a power of 0.9 if $(\mu_1-\mu_2)/\sigma = 0.5$ then $n_1 = n_2 = 36$.

EXAMPLE

It is desired to know if there is a difference in the mean percentage sugar content of a particular species of orange grown in two different locations. We wish to test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$ at $\alpha = 0.05$. Samples of 10 oranges are selected from each location. The test statistic will have 18 degrees of freedom and we will reject H_0 if $|T| \geq 2.101$. The results for location 1 are summarized by $\bar{X}_1 = 5.62$ and $S_1^2 = 0.24$ and for location 2 are summarized by $\bar{X}_2 = 5.31$ and $S_2^2 = 0.32$. The pooled estimate of the variance is $S_p^2 = [(10-1)0.24 + (10-1)0.32]/[10+10-2] = 0.28$. The test statistic is

$$T = [(5.62-5.31)-0]/\{0.28[0.1+0.1]\}^{1/2} = 0.31/0.2366 = 1.310.$$

Hence, we can not reject H_0 at $\alpha=0.05$. The p-value is
$P(T > 1.310|\mu_1=\mu_2) + P(T < -1.310|\mu_1=\mu_2) = 2\ P(T > 1.310|\mu_1=\mu_2)$.
From the T table we can see that the p-value is slightly greater than 0.2.

A 95% confidence interval for $\mu_1-\mu_2$ is given by
$$(5.62-5.31)\pm 2.101[0.28(0.1+0.1)]^{1/2}$$
which yields the interval (-0.187,.807). Our conclusion is that the mean for location 1 exceeds the mean for location 2 by between -0.187% to 0.807%.

We will now consider the analysis when one is not willing to assume the two variances are equal. This is known as the BEHRENS-FISHER PROBLEM. Let σ_1^2 and σ_2^2 denote the variances associated with treatments 1 and 2, respectively. The difference of the two sample means, $\bar{X}_1-\bar{X}_2$, has a normal distribution with mean $\mu_1-\mu_2$ and variance $[(\sigma_1^2/n_1) + (\sigma_2^2/n_2)]$. It is natural to want to use the pivotal variable
$$T = [(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]/[(S_1^2/n_1)+(S_2^2/n_2)]^{1/2}.$$
Unfortunately, the distribution of T depends on the unknown variances. Several authors have approximated the distribution of T. One such approach says that the variable T is approximately distributed as a T distribution with ν degrees of freedom where ν is determined from the data by
$$\nu = 1 \Big/ \{[k^2/(n_1-1)] + [(1-k)^2/(n_2-1)]\},$$
where
$$k = (S_1^2/n_1) \Big/ [(S_1^2/n_1) + (S_2^2/n_2)].$$
The $(1-\alpha)100\%$ confidence interval for $\mu_1-\mu_2$ is
$$\bar{X}_1-\bar{X}_2\pm t_{\alpha/2}[(S_1^2/n_1)+(S_2^2/n_2)]^{1/2},$$
where $t_{\alpha/2}$ comes from the row of the t table with ν degrees of freedom. It will probably be necessary to interpolate between two rows of the table.

The test statistic for hypotheses of the type $\mu_1-\mu_2=\theta_0$ is
$$T = [(\bar{X}_1-\bar{X}_2)-\theta_0] \Big/ [(S_1^2/n_1)+(S_2^2/n_2)]^{1/2}.$$
The rejection region is found using the row of the t table with ν degrees of freedom.

EXAMPLE