1 Introduction

Reading: SW Chapter 1

What is statistics/biostatistics/biometry?

Examples of medical and research problems:

1. A couple is deciding whether or not to have a child, because of the existence of certain diseases within the family. With present understanding of genetics, they are told that the probability that a child of theirs having this defect is 0.01. What might they want to do? How would the type of disease affect this? What if the probability is 0.50? What other factors besides probabilities and the type of disease would be pertinent?

2. Research question: Is HPV (human papilloma virus) a risk factor for cervical dysplasia? How does one approach answering this question? One possibility: Becker et al. (1994) conducted a case-control study. The women in the study were patients at UNM clinics. The 175 cases were women, aged 18-40, who had cervical dysplasia. The 308 controls were women aged 18-40 who did not have cervical dysplasia. Each women was classified as positive or negative, depending on the presence of HPV. The data collected from the study are summarized below.

HPV Outcome	Cases	Controls
Positive	164	130
Negative	11	178
Sample size	175	308

   The results can be summarized in a number of ways. The proportion positive among cases is $164/175 = 0.94$. the proportion positive among controls is $130/308 = 0.42$. This gives an odds ratio of $164 * 178/(11 * 130) = 20.4$. Do these results indicate that HPV is a risk factor for cervical dysplasia?

3. Research question: Is a new drug more effective in treating an illness than a previously used drug? How to approach this question? One possibility: conduct a clinical trial (Phase II) with one treatment group where all patients receive the new drug. The old drug has an assumed cure rate obtained from repeated use of this treatment.

   Outcomes and conclusions: Assume old drug cures 70%. If 9 people out of 10 with the illness were cured with new treatment, then what would you conclude? If 6 were cured? If 90 out of a sample of 100?

   Alternative possibility: conduct a clinical trial (Phase III) with two groups (new treatment, old treatment), and randomize patients to the two groups.

   Other possible outcomes of interest: reduction in fever, pain, itching of skin rash in 24-hour period (quantify reduction), reduction in tumor size.

SO WHAT DOES STATISTICS LEND TO THESE PROBLEMS?

1. **What is statistics?**

   - Statistics is concerned with the STUDY, DESCRIPTION, and MANAGEMENT of variability.
   - There are many ways to define statistics, but common components in the definitions are: variation; uncertainty; inference.
   - Biostatistics is the subset of statistics that is concerned with applications in biological/medical areas.

2. **What should you get out of an introductory course in biostatistics?**

   - Understand basic statistical concepts
   - Be able to read papers in your field and understand the statistical results, and, hopefully, the statistical methods that were used.
   - Be able to determine appropriate statistical methods to use and implement them – in simple analyses.
   - Be able to determine when you can't do something and seek out help from a statistician.

ASPECTS OF STATISTICS THAT WE WILL BE CONCERNED WITH

- Descriptive statistics and exploratory data analysis: ways to describe data using graphical displays and numerical summaries
- Basic ideas of probability as a means of quantifying uncertainty
- Statistical inference: wish to draw some conclusions from data, based on hypothesis testing and estimation methods.

Types of data/situations we will examine:

- data on one continuous variable (one, two and multiple samples)
- discrete data (single sample and two-way tables, including logistic regression)
- data on two or more continuous variables (linear regression and correlation, and survival analysis)

2 Descriptive Statistics

Reading: SW Chapter 2, Sections 1-6

A natural first step towards answering a research question is for the experimenter to design a study or experiment to collect data from the **population** (i.e. collection of individuals) of interest. In most studies, data are collected on only a subset or **sample** from the **population**. Typically, a number of different characteristics or **variables** are measured on each selected individual.

Once the data are collected, we should summarize the information **graphically** and **numerically**. The actual methods used to summarize data depend on the types of variables that were recorded.

Quantitative versus Qualitative

Simply, a quantitative variable is a variable expressed by a quantity, while a qualitative variable is expressed by a quality (i.e. **categorical**).

Examples:

- number of pregnancies (quantitative)
- eye color (qualitative or categorical)
- age (quantitative)
- ethnic group (qualitative or categorical)

Discrete versus Continuous:

Variables that are expressed numerically can be further subdivided into **discrete** and **continuous** variables. A discrete or counting variable is a variable that takes on a finite or countably infinite number of values, while a continuous variable is a variable that assumes any of the values in at least one interval of the real number line.

Examples:

- number of pregnancies (discrete)
- age (continuous)
- city population size (discrete)
- proportion of population who are HIV+ (continuous)

Nominal versus Ordinal:

Categorical variables are **ordinal** if the order of the categories is meaningful and are **nominal** if the order is unimportant.

Examples:

- stage of cancer: in situ, local, regional, distant (ordinal)
- ethnic group (nominal)

Notes:

1. Continuous variables often have a well-defined measurement scale. For example, time in seconds, temperature in degrees Celsius. However, the scale is often not unique. With continuous variables you should always define the unit of measurement.
2. Discrete variables can be constructed from continuous variables. For example, age is a continuous variable, but the variable X defined by $X = 1$ if age is less than 40, otherwise $X = 2$ is a discrete variable that has been created by categorizing age. Note X is ordinal.
3. A qualitative variable can be **coded** to have numerical values. For example, if the variable is eye color, we might define $X = 1$ if person has blue eyes, $X = 0$ otherwise.
4. A discrete variable that has only two possible values is called **binary**. The variable X above is binary.
5. We are limited in our ability to measure continuous variables. Furthermore, many discrete variables can be analyzed with methods for continuous variables, provided the discrete variables are "close-enough" to being continuous. For example, if scores on a psychological test can take on integer values from 1 to 50, then the score variable is discrete. However, if a sample distribution of the scores contains many of the possible values, then it may be possible to use methods for continuous data for analyzing the discrete data.

REMARK: We commonly use capital letters, say X and Y, to identify variables. This is useful mathematical shorthand that is not intended to confuse you.

Summarizing and Displaying Numerical Data

Suppose we have a sample of n individuals, and we measure each individual's response on one quantitative characteristic, say height, weight, or systolic blood pressure. For notational simplicity, the collected measurements are denoted by $Y_1,\ Y_2, ..., Y_n$, where n is the **sample size**. The order in which the measurements are assigned to the place-holders $Y_1,\ Y_2, ..., Y_n$ is irrelevant.

Two standard numerical summary measures are the **sample mean** $\overline{Y}$ and the **sample standard deviation** s. A numerical summary measure is called a **statistic**, so both the sample mean and standard deviation are statistics.

The sample mean is a measure of **central location**, or a measure of a **typical value** for the data set. The standard deviation is a measure of **spread** in the data set. These summary statistics might be familiar to you. Let us consider a simple example to show you how to compute them. Suppose we have a sample of $n = 8$ children with weights (in pounds): 5, 9, 12, 30, 14, 18, 32, 40. Then

$$\overline{Y} = \frac{\sum_i Y_i}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$
$$= \frac{5 + 9 + 12 + 30 + 14 + 18 + 32 + 40}{8} = \frac{160}{8} = 20.$$

The sample standard deviation is the square root of the sample variance given by the formula:

$$s^2 = \frac{\sum_i (Y_i - \overline{Y})^2}{n-1} = \frac{(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + \cdots + (Y_k - \overline{Y})^2}{n-1}.$$

For hand calculations, it is common to create a **table** from which s is computed, as below:

Data	Deviation	Squared Deviation
5	5-20 = -15	(-15)^2 = 225
9	9-20 = -11	(-11)^2 = 121
12	12-20 = - 8	(- 8)^2 = 64
14	14-20 = - 6	(- 6)^2 = 36
18	18-20 = - 2	(- 2)^2 = 4
30	30-20 = 10	10^2 = 100
32	32-20 = 12	12^2 = 144
40	40-20 = 20	20^2 = 400

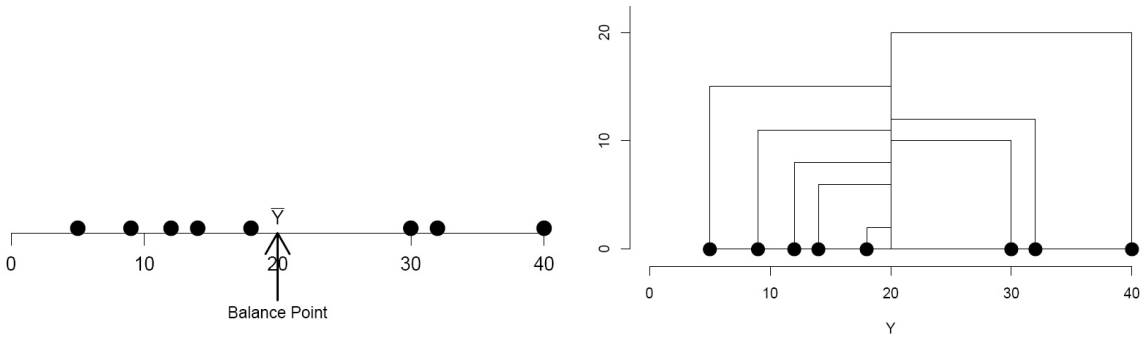
The sample variance is obtained by adding the entries in the last column and dividing by $n-1$:

$$s^2 = \frac{225 + 121 + 64 + 36 + 4 + 100 + 144 + 400}{8-1} = \frac{1094}{7} = 156.3.$$

Thus, $s = \sqrt{s^2} = 12.5$. Summary statistics have well-defined units of measurement, for example, $\overline{Y} = 20lb$, $s^2 = 156.3lb^2$, and $s = 12.5lb$. The standard deviation is often used instead of s^2 as a measure of spread because s is measured in the same units as the data.

REMARK: If the divisor for s^2 was n instead of $n-1$, then the variance would be the average squared deviation observations are from the center of the data as measured by the mean.

The following graphs should help you to see some physical meaning of the sample mean and variance. If the data values were placed on a "massless" ruler, the balance point would be the mean (20). The variance is basically the "average" (remember n-1 instead of n) of the total areas of all the squares obtained when squares are formed by joining each value to the mean. In both cases think about the implication of unusual values (**outliers**). What happens to the balance point if the 40 were a 400 instead of a 40? What happens to the squares?



The **sample median** M is an alternative measure of central location. The measure of spread reported along with M is the **interquartile range**, $IQR = Q_3 - Q_1$, where Q_1 and Q_3 are the **first** and **third** quartiles of the data set, respectively. To calculate the median and interquartile range, order the data from lowest to highest values, all repeated values included. The ordered weights are

$$5 \;\; 9 \;\; 12 \;\; 14 \;\; 18 \;\; 30 \;\; 32 \;\; 40.$$

The median M is the value located at the half-way point of the ordered string. There is an even number of observations, so M is defined to be half-way between the two middle values, 14 and 18. That is, $M = .5(14 + 18) = 16lb$. To get the quartiles, break the data into the lower half: 5 9 12 14, and the upper half: 18 30 32 and 40. Then

$$Q_1 = \text{first quartile} = \text{median of lower half of data} = \tfrac{9+12}{2} = 10.5lb,$$

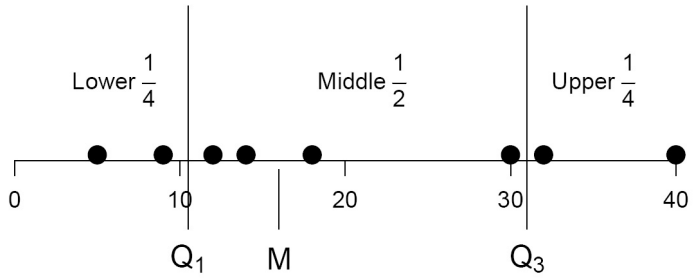
and

$$Q_3 = \text{third quartile} = \text{median of upper half of data} = .5(30{+}32) = 31lb.$$

The interquartile range is

$$IQR = Q_3 - Q_1 = 31 - 10.5 = 20.5lb.$$

The quartiles, with M being the second quartile, break the data set roughly into fourths. The first quartile is also called the 25^{th} percentile, whereas the median and third quartiles are the 50^{th} and 75^{th} percentiles, respectively.. The IQR is the **range** for the middle half of the data.



Suppose we omit the largest observation from the weight data:

5 9 12 14 18 30 32.

How do M and IQR change? With an odd number of observations, there is a unique middle observation in the ordered string which is M. Here $M = 14lb$. It is unclear which half the median should fall into, so M is placed into both the lower and upper halves of the data. The lower half is 5 9 12 14, and the upper half is 14 18 30 32. With this convention, $Q_1 = .5(9 + 12) = 10.5$ and $Q_3 = .5(18 + 30) = 24$, giving $IQR = 24 - 10.5 = 13.5(lb)$.

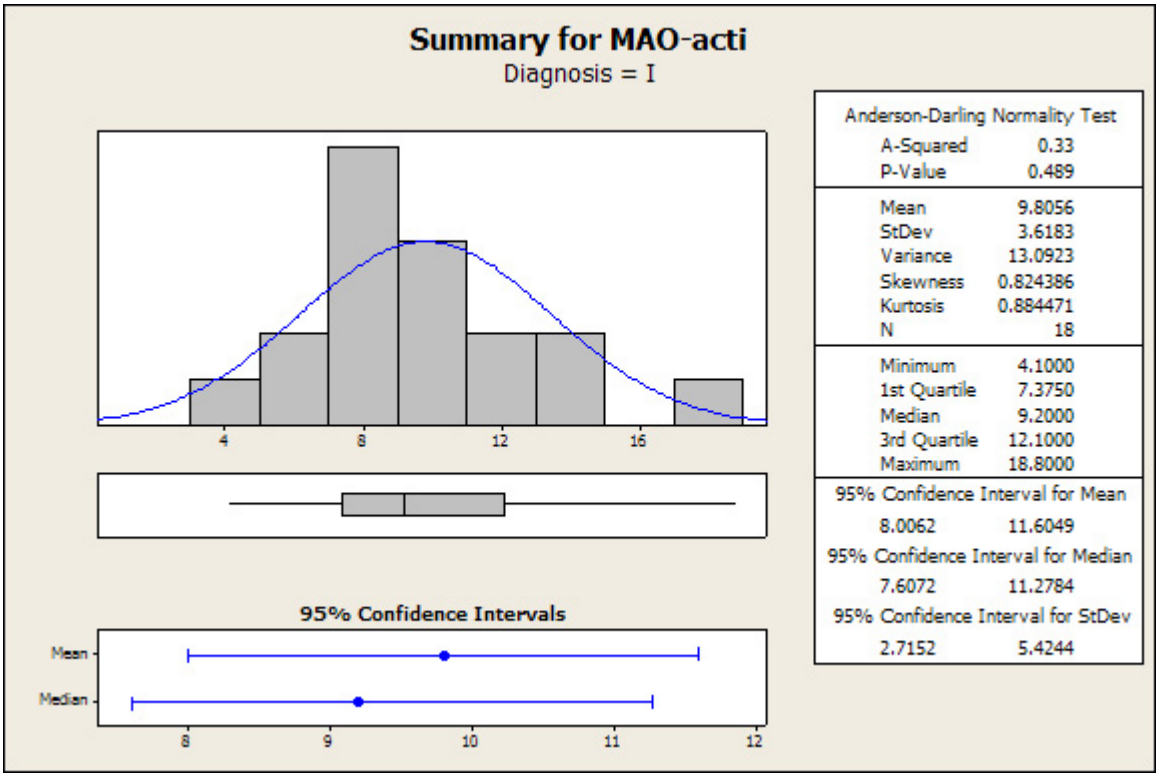
If you look at the data set with all eight observations, there actually are many numbers that split the data set in half, so the median is not uniquely defined, although "everybody" agrees to use the average of the two middle values. With quartiles there is the same ambiguity but no such universal agreement on what to do about it, however, so Minitab will give slightly different values for Q_1 and Q_3 than we just calculated, and other packages will report even different values. This has no practical implication (all the values are "correct") but it can appear confusing.

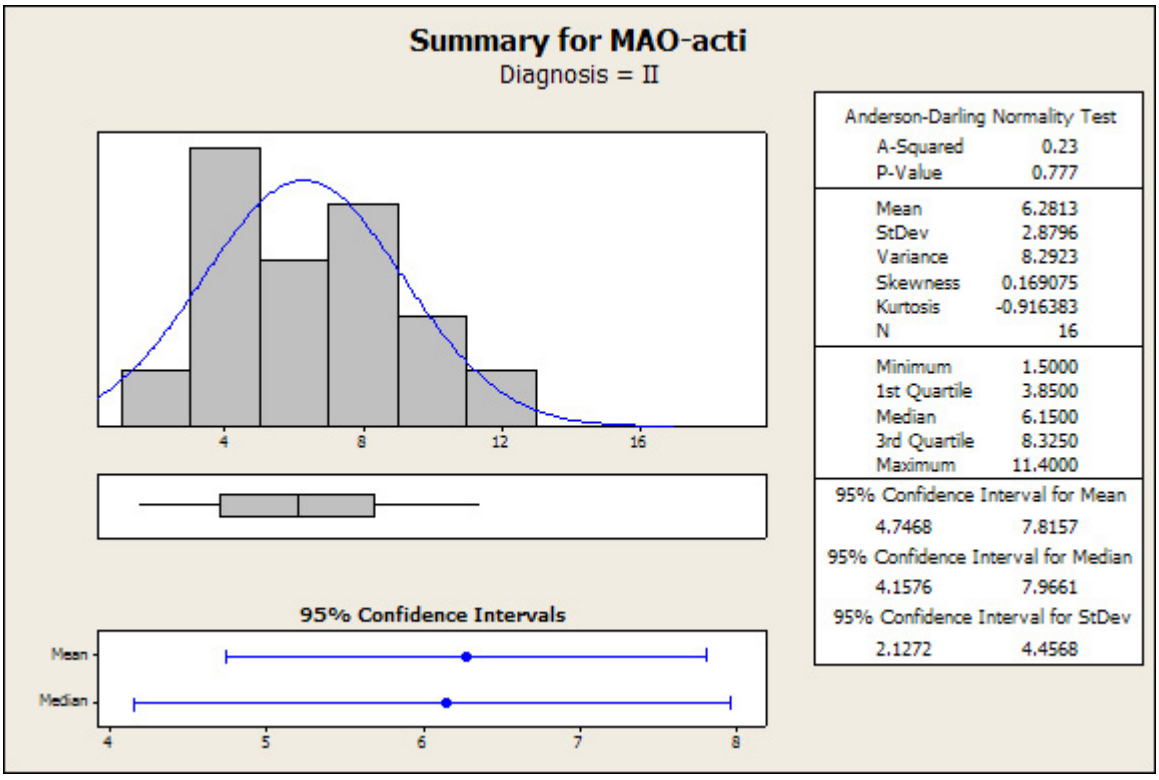
Minitab Implementation

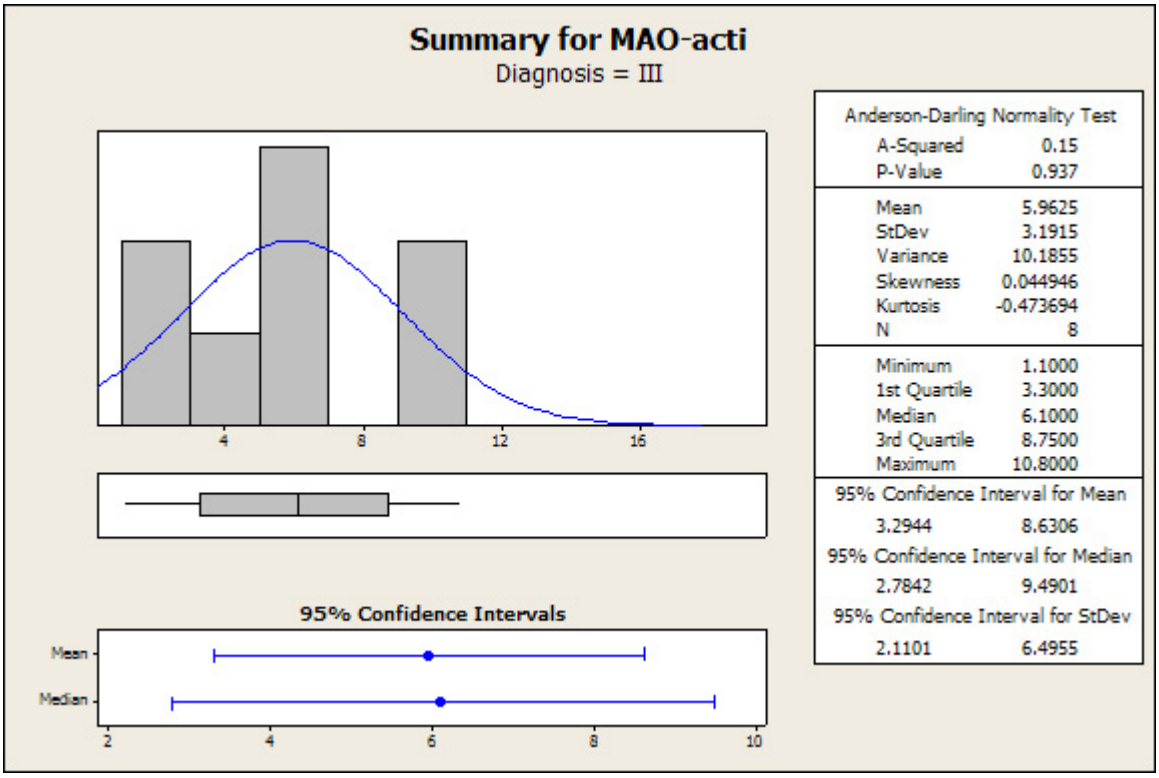
Minitab will automatically compute the summaries we have discussed, and others. Erik will show you how to do this in LAB. Following are numerical and graphical summaries for the data in Example 1.4 pages 3-4 of SW. Monoamine oxidase (MAO) activity expressed as nmol benzylaldehyde product per 108 platelets per hour was measured on schizophrenic patients of three different diagnoses. The data are on the CD in the back of SW.

The first display is simple descriptive statistics, the graphs are an enhancement of the simple descriptive statistics. Eric will show you how to obtain both, and how to import into a program like WORD. Let us discuss the output.

```
Descriptive Statistics: MAO-acti

Variable  Diagnosis   N  N*   Mean  SE Mean  StDev  Minimum     Q1 Median
MAO-acti  I          18   0  9.806    0.853  3.618    4.100  7.375  9.200
          II         16   2  6.281    0.720  2.880    1.500  3.850  6.150
          III         8  10   5.96     1.13   3.19     1.10   3.30   6.10

Variable  Diagnosis      Q3  Maximum
MAO-acti  I          12.100   18.800
          II          8.325   11.400
          III          8.75    10.80
```


Summary for MAO-acti
Diagnosis = I
4
8
12
16
95% Confidence Intervals
Mean
Median
8
9
10
11
12
Anderson-Darling Normality Test
A-Squared 0.33
P-Value 0.489
Mean 9.8056
StDev 3.6183
Variance 13.0923
Skewness 0.824386
Kurtosis 0.884471
N 18
Minimum 4.1000
1st Quartile 7.3750
Median 9.2000
3rd Quartile 12.1000
Maximum 18.8000
95% Confidence Interval for Mean
8.0062 11.6049
95% Confidence Interval for Median
7.6072 11.2784
95% Confidence Interval for StDev
2.7152 5.4244


Summary for MAO-acti
Diagnosis = II
4
8
12
16
95% Confidence Intervals
Mean
Median
4
5
6
7
8
Anderson-Darling Normality Test
A-Squared 0.23
P-Value 0.777
Mean 6.2813
StDev 2.8796
Variance 8.2923
Skewness 0.169075
Kurtosis -0.916383
N 16
Minimum 1.5000
1st Quartile 3.8500
Median 6.1500
3rd Quartile 8.3250
Maximum 11.4000
95% Confidence Interval for Mean
4.7468 7.8157
95% Confidence Interval for Median
4.1576 7.9661
95% Confidence Interval for StDev
2.1272 4.4568



Mean versus Median

Although the mean is the most commonly used measure of central location, it (and the standard deviation) is very sensitive to the presence of extreme observations, sometimes called **outliers**. The median and interquartile range are more **robust** (less sensitive) to the presence of outliers.

For example, the following data are the incomes in 1000 dollar units for a sample of 12 retired couples: 7, 1110, 7, 5, 8, 12, 0, 5, 2, 2, 46, 7. The sample has two extreme outliers at 46 and 1110. For these data $\overline{Y} = 100.9$ and $s = 318$, whereas $M = 7$ and $IQR = 8.3$. If we hold out the two outliers, then $\overline{Y} = 5.5$ and $s = 3.8$, whereas $M = 6$ and $IQR = 5.25$.

The mean and median often have similar values in data sets without outliers, so in such a case it does not matter much which one is used as the **typical value**. This issue is important, however, in data sets with extreme outliers. In such instances, the median is often more reasonable. For example, is $\overline{Y} = 100.9$ a reasonable measure for a typical income in this sample, given that the second largest income is only 46?

Further Points That Will Emphasized in Class:

1. I will mention another summary measure, the **coefficient of variation**: $CV = 100\% * s/\overline{Y}$.

2. I will briefly discuss how the mean and standard deviation change if the units are changed. For example, what happens in the weight problem if I change units from pounds to ounces?

3. The size of the standard deviation depends on the units of measure. We often use s to compare spreads from different samples measured on the same attribute.

3 Graphical Displays of Data

Reading: SW Chapter 2, Sections 1-6

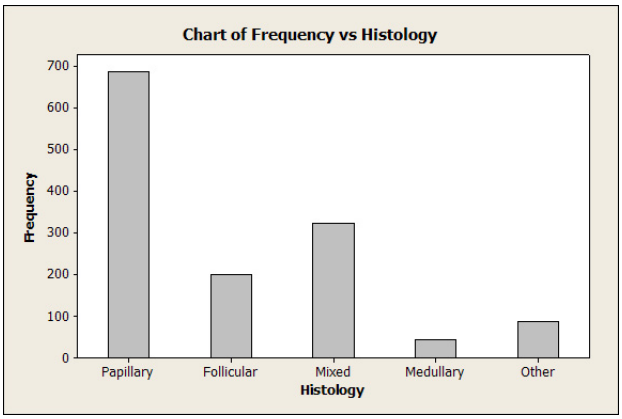
Summarizing and Displaying Qualitative Data

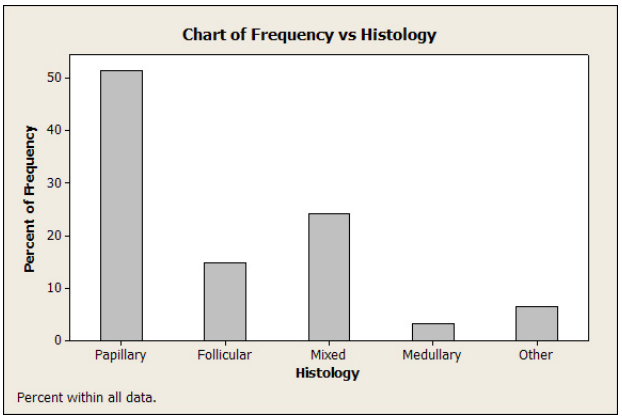
The data below are from a study of thyroid cancer, using NMTR data. The investigators looked at all thyroid cancer cases diagnosed among NM residents between 1/1/69 and 12/31/91. A small percentage of cases were omitted (those that weren't first primary; those without more than 60 days of follow-up without another diagnosis of cancer), leaving 1338 cases of thyroid cancer.

A **frequency distribution** for a categorical variable gives the counts or frequency with which the values occur in the various categories. The frequency distribution for histologic type is given below. The **relative** frequency distribution gives the proportion (i.e number of cases divided by sample size) or percentage (proportion times 100%) of cases in each histologic category.

Histology	Frequency	Relative Frequency	Percentage
Papillary	687	$687/1338 = 0.51$	51%
Follicular	199	$199/1338 = 0.15$	15%
Mixed	323	$323/1338 = 0.24$	24%
Medullary	43	$43/1338 = 0.03$	3%
Other	86	$86/1338 = 0.06$	6%
Total	1338	0.99(1.00)	99% (100%)

The frequency distribution is usually summarized graphically via a **bar graph**, sometimes called a **bar chart**. The next page give frequency and relative frequency distributions generated by **Minitab**. Erik will show you how to do this in LAB.





The information conveyed is the same in both graphs. The graph of percentages has real advantages when comparing two groups with much different sample sizes, however.

Example: SW pages 12, 14 - colors of Poinsettia.

Graphical Summaries of Numerical Data

There are four (actually, there a many more) graphical summaries of primary interest: the **histogram**, the **dotplot**, the **stem and leaf** display, and the **boxplot**. Each of these is easy to generate in **Minitab**. Our goal with a graphical summary is to see patterns in the data. We want to see what values are typical, how spread out are the values, where do the values tend to cluster, and what (if any) big deviations from the overall patterns are present. Sometimes one summary is better than another for a particular data set.

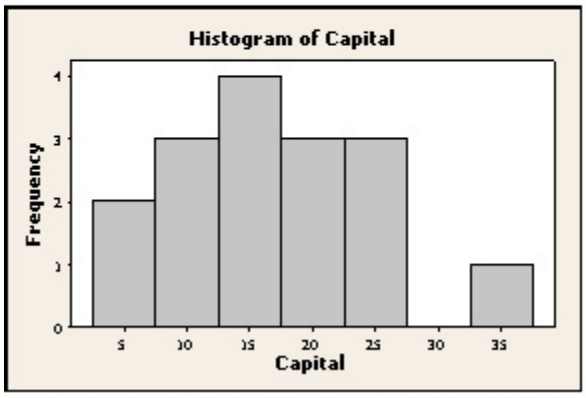
Histogram

The **histogram** breaks the range of data into several equal width intervals, and counts the number (or proportion, or percentage) of observations in each interval. The histogram can be viewed as a **grouped frequency distribution** for a continuous variable. Here is the "help" entry from Minitab describing histograms:

Histograms

Graph > Histogram

Use to examine the shape and spread of sample data. Histograms divide sample values into many intervals called bins. Bars represent the number of observations falling within each bin (its frequency). In the histogram below, for example, there are two observations with values between 2.5 and 7.5, three observations with values between 7.5 and 12.5, and so on.



Observations that fall exactly on an interval boundary are included in the interval to the right (or left, if the last bin).

Why is it reasonable to group measurements whereas with categorical data we computed the number of observations with each distinct data value?

Most texts, including SW, discuss the choice of intervals. We will use **Minitab** for our calculations, which usually does quite a good job of choosing the intervals for us. We already saw histograms of MAO levels in the previous section.

The real strength of histograms is showing where data values tend to cluster. Their real weakness is that the choice of intervals (bins) can be arbitrary, and the apparent clustering can depend considerably on the choice of bins. Histograms work pretty well with larger data sets, where the choice of bins usually has little effect; for smaller data sets, dotplots or stem and leaf displays usually are a much better choice.

Dotplot

Where histograms try to condense the data into relatively few bins, dotplots present a similar picture but emphasize the distinct values. Dotplots are particularly good at comparing different data sets, especially smaller data sets. One big advantage is that you usually see all the data, so no information is lost in the dotplot. The biggest disadvantage is that it gets pretty "noisy" for large data sets.

Here is the "help" entry from Minitab describing dotplots:

Dotplots

Graph > Dotplot

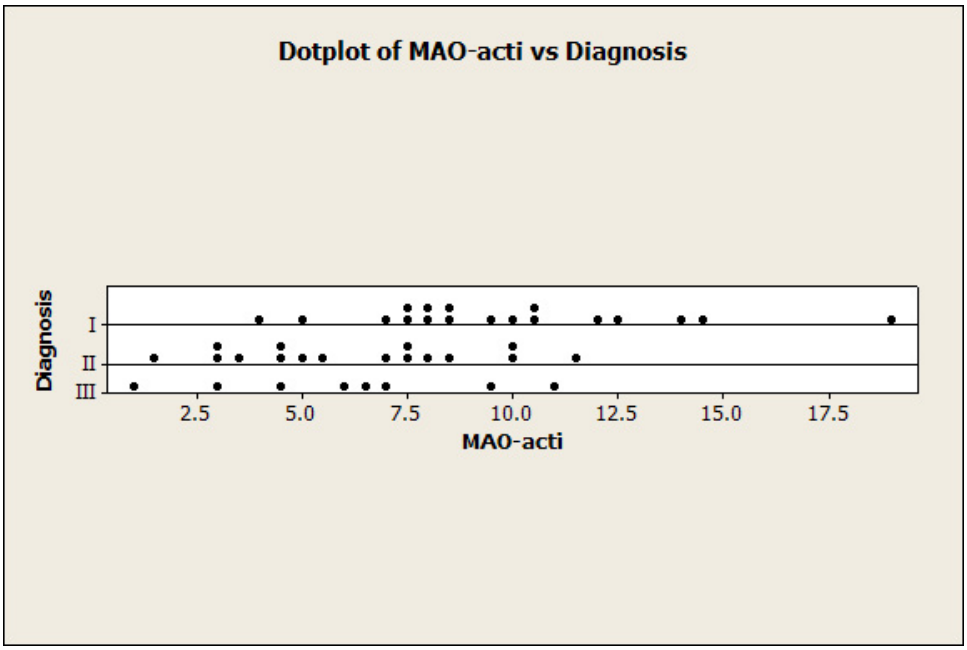
Use to assess and compare distributions by plotting the values along a number line. Dotplots are especially useful for comparing distributions.

The x-axis for a dotplot is divided into many small intervals, or bins. Data values falling within each bin are represented by dots.

If possible, Minitab displays a dot for each observation. Otherwise, a dot represents multiple observations with a footnote indicating the maximum number of observations represented by each dot.

Note Dotplots can only be brushed when dots represent individual observations.

Earlier we looked at histograms of MAO activity levels for schizophrenic patients of three different diagnoses. The dotplots for the three data sets make comparisons quite easy. Isn't it a lot easier to see the nature of differences here than using the three histograms in the previous section?



Stem and Leaf Display

A **stem and leaf** display defines intervals for a grouped frequency distribution using the base 10 number system. Intervals are generated by selecting an appropriate number of lead digits for the

data values to be the stem. The remaining digits comprise the leaf. Following is Minitab's "help" entry for the Stem and Leaf:

Stem-and-Leaf

Graph > Stem-and-Leaf
Stat > EDA > Stem-and-Leaf
***Character Graphs* > Stem-and-Leaf**

Use to examine the shape and spread of sample data. Minitab displays a stem-and-leaf plot in the Session window. The plot is similar to a histogram on its side, however, instead of bars, digits from the actual data values indicate the frequency of each bin (row).

Below is a stem-and-leaf plot for a data set with the following five values: 3, 4, 8, 8, and 10.

```
Stem-and-leaf of C1  N  = 5
Leaf Unit = 1.0

 1   0  3
 2   0  4
 2   0
(2)  0  88
 1   1  0
```

The display has three columns:

- The leaves (right) – Each value in the leaf column represents a digit from one observation. The "leaf unit" (declared above the plot) specifies which digit is used. In the example, the leaf unit is 1.0. Thus, the leaf value for an observation of 8 is 8 while the leaf value for an observation of 10 is 0.
- The stem (middle) – The stem value represents the digit immediately to the left of the leaf digit. In the example, the stem value of 0 indicates that the leaves in that row are from observations with values greater than or equal to zero, but less than 10. The stem value of 1 indicates observations greater than or equal to 10, but less than 20.
- Counts (left) – If the median value for the sample is included in a row, the count for that row is enclosed in parentheses. The values for rows above and below the median are cumulative. The count for a row above the median represents the total count for that row and the rows above it. The value for a row below the median represents the total count for that row and the rows below it.

  In the example, the median for the sample is 8, so the count for the fourth row is enclosed in parentheses. The count for the second row represents the total number of observations in the first two rows.

Look carefully at the display – how would the example above change if the numbers were 30, 40, 80, 80 and 100 instead of 3, 4, 8, 8, and 10. Try it and confirm the display looks the same with one important difference. Following is the stem and leaf for the MAO activity levels of Diagnosis I patients.

```
Stem-and-Leaf Display: MAO-acti

Stem-and-leaf of MAO-acti  group = 1    N  = 18
Leaf Unit = 1.0

 2   0  45
 7   0  67777
(4)  0  8899
 7   1  001
 4   1  2
 3   1  44
 1   1
 1   1  8
```

Let's examine this display, and make sure we can pick out what the actual numbers are. Look at the original values (from SW). Is Minitab rounding numbers or just truncating excess digits? SW would have you put larger numbers on top. That would seem more conventional, except stem

and leaf displays almost always are done Minitab's way with the larger numbers on the bottom. There is a good reason for this – if you turn the graph 90 degrees counterclockwise, you end up with a regular histogram (what are the bins?)

The stem and leaf was invaluable for "paper and pencil" data analysis. It is very quick to do by hand, and it has the advantage of keeping the original data right on the display. It also sorts the data (puts them in order), which allows quick calculation of medians and quartiles. I find the dotplot a better tool, often, when summarizing small to moderate-sized data sets on the computer. The stem and leaf is harder to use for comparing several groups, but still is more common in practice than dotplots.

Erik will show you how to generate stem and leaf displays in **Minitab**, and a few of the options.

Example

Two stem and leaf displays for a data set on age at death for SIDS cases in Washington state are given below. The first is for the data recorded in days, the second for the data recorded in weeks. Note that the maximum value is 307 days, or 43.9 weeks.

```
Stem-and-Leaf Display: SIDS days

Stem-and-leaf of SIDS days  N  = 78
Leaf Unit = 10


 9    0  222222333
 18   0  444444555
 31   0  6666666667777
(16)  0  8888888889999999
 31   1  00000111111
 20   1  22333
 15   1  4455
 11   1  6777
 7    1  88
 5    2  0
 4    2  23
 2    2
 2    2  7
 1    2
 1    3  0

Stem-and-Leaf Display: SIDS weeks

Stem-and-leaf of SIDS weeks  N  = 78
Leaf Unit = 1.0


 8    0  33334444
 27   0  5666666778889999999
(22)  1  01111111122222333444
 29   1  55556666677889
 15   2  0112344
 8    2  5669
 4    3  23
 2    3  9
 1    4  3
```

The structure of the two stem and leaf displays is slightly different. In particular, the days display corresponds to a histogram with intervals of width 20 (confirm this!). The weeks display corresponds to a histogram with intervals of width 5 (confirm!). Minitab does give you some control over interval widths, but usually makes the right choice by default.

Boxplots

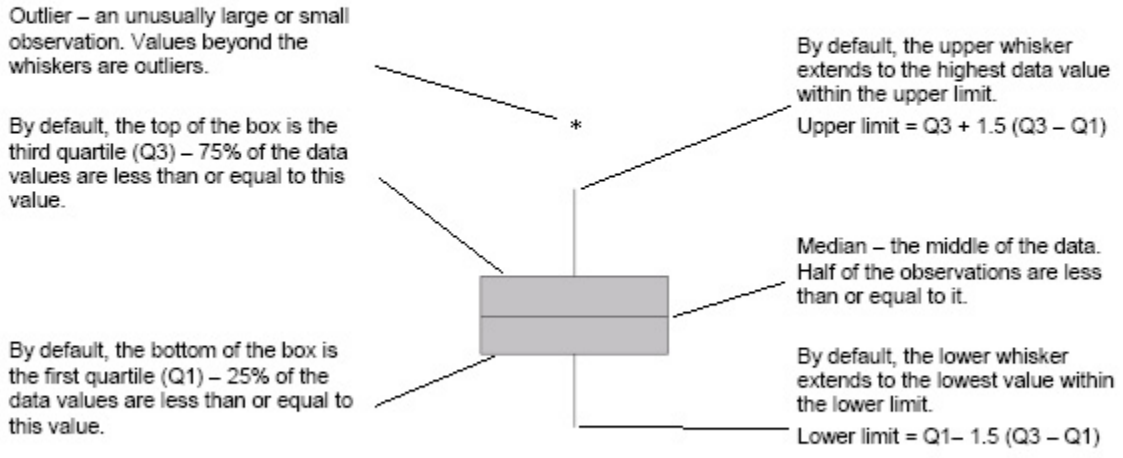
Boxplots have become probably the most useful of all the graphical displays of numerical data. I can go weeks without computing histograms, dotplots, or stem and leaf displays, but I usually compute several boxplots per week. They succinctly summarize central location (average), spread and shape of the data, and highlight outliers while permitting simple comparison of many data sets at once. Following is the Minitab "help" description of boxplots.

Boxplots

Graph > Boxplot

Stat > EDA > Boxplot

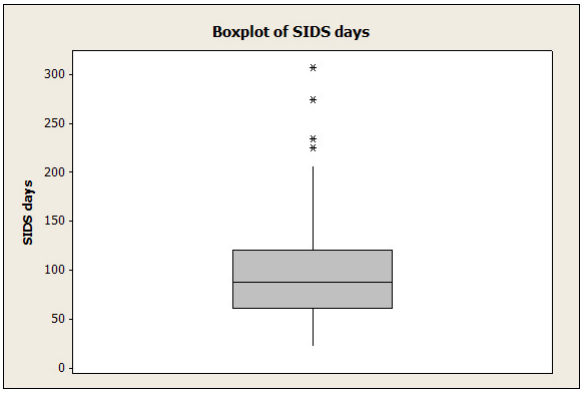
Use boxplots (also called box-and-whisker plots) to assess and compare sample distributions. The figure below illustrates the components of a default boxplot.

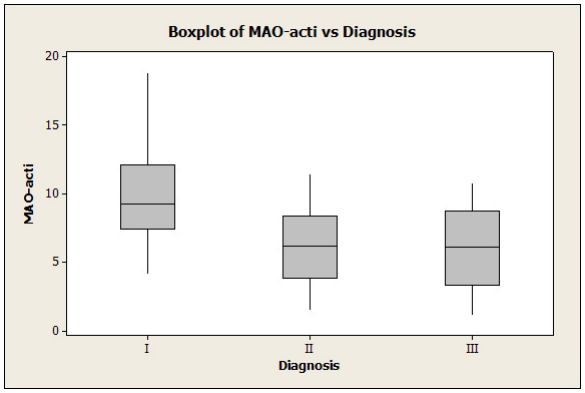


Note By default, Minitab uses the quartile method for calculating box endpoints. To change the method for a specific graph to hinge or percentile, use Editor > Edit Interquartile Range Box > Options. To change the method for all future boxplots, use Tools > Options > Individual Graphs > Boxplots.

Lots of elementary texts make the boxplots simpler by connecting the whiskers to the extremes of the data; this keeps them from highlighting outliers and, in my opinion, erases substantial utility of the boxplot. Minitab will allow you to compute those neutered boxplots, but you should not. The box part of the boxplot is Q_1, M, and Q_3, a range containing half the data. The whiskers connect the box to the extremes of "normal" looking data, and anything more extreme is plotted separately (and importantly) as an outlier. Relative distance of the quartiles from the median, and relative length of the whiskers tells us a lot about the shape of the data (we will explore that below). Several packages, including Minitab, allow you to clutter the boxplot with a lot of other features, but I usually prefer not to.

Boxplots of the SIDS and MAO data sets are below. Let's pick out important features.





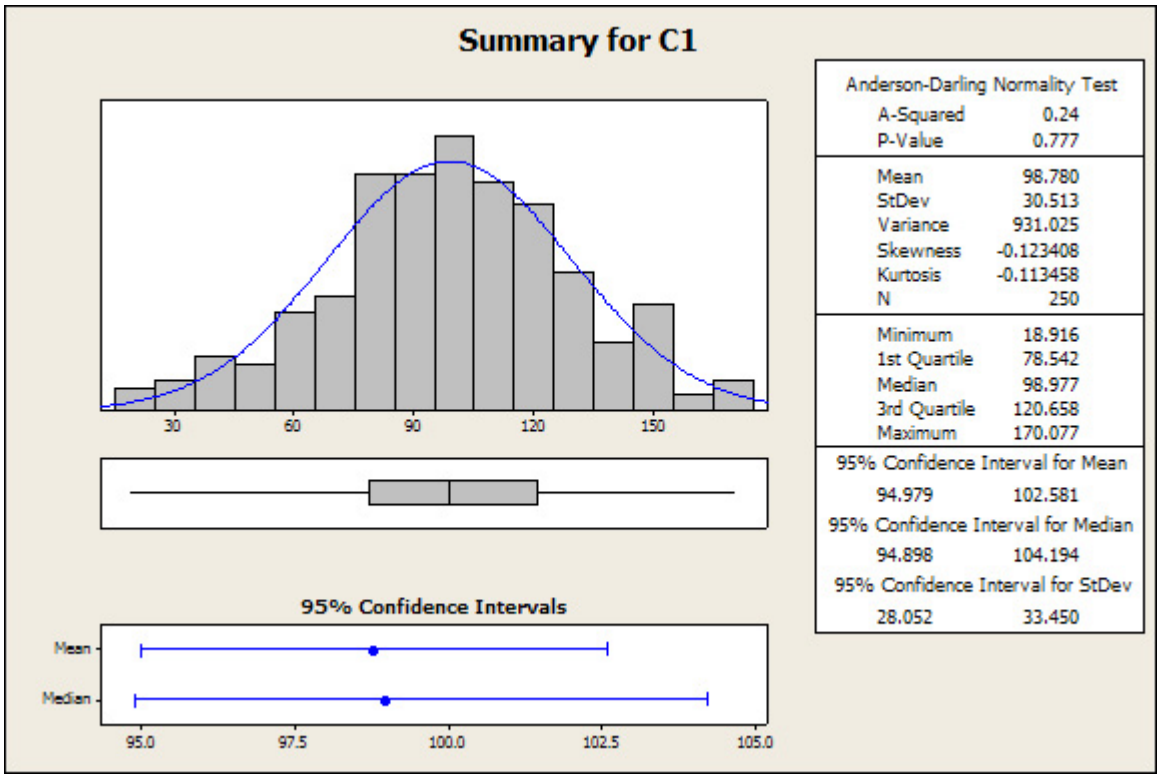
Interpretation of Graphical Displays for Numerical Data

In many studies, the data are viewed as a subset or **sample** from a larger collection of observations or individuals under study, called the **population**. A primary goal of many statistical analyses is to generalize the information in the sample to **infer** something about the population. For this generalization to be possible, the sample must reflect the basic patterns of the population. There are several ways to collect data to ensure that the sample reflects the basic properties of the population, but the simplest approach, by far, is to take a random or "representative" sample from the population. A **random sample** has the property that every possible sample of a given size has the same chance of being the sample eventually selected. Random sampling eliminates any systematic biases associated with the selected observations, so the information in the sample should accurately reflect features of the population. The process of sampling introduces random variation or random errors associated with summaries. Statistical tools are used to calibrate the size of the errors.

Whether we are looking at a histogram (or stem and leaf, or dotplot) from a sample, or are conceptualizing the histogram generated by the population data, we can imagine approximating the "envelope" around the display with a smooth curve. The smooth curve that approximates the population histogram is called the **population frequency curve**. Statistical methods for inference about a population usually make assumptions about the shape of the population frequency curve. A common assumption is that the population has a normal frequency curve. In practice, the observed data are used to assess the reasonableness of this assumption. In particular, a sample display should resemble a population display, provided the collected data are a random or representative sample from the population. Several common shapes for frequency distributions are given below, along with the statistical terms used to describe them.

The first display is **unimodal** (one peak), **symmetric** and **bell-shaped**. This is the prototypical normal curve. The boxplot (laid on its side for this display) shows strong evidence of symmetry: the median is about halfway between the first and third quartiles, and the tail lengths are roughly equal. The boxplot is calibrated in such a way that 7 of every 1000 observations are outliers (more than $1.5(Q_3 - Q_1)$ from the quartiles) in samples from a population with a normal frequency curve. Only 2 out of every 1 million observations are extreme outliers (more than $3(Q_3 - Q_1)$ from the quartiles). We do not have any outliers here out of 250 observations, but we certainly could have

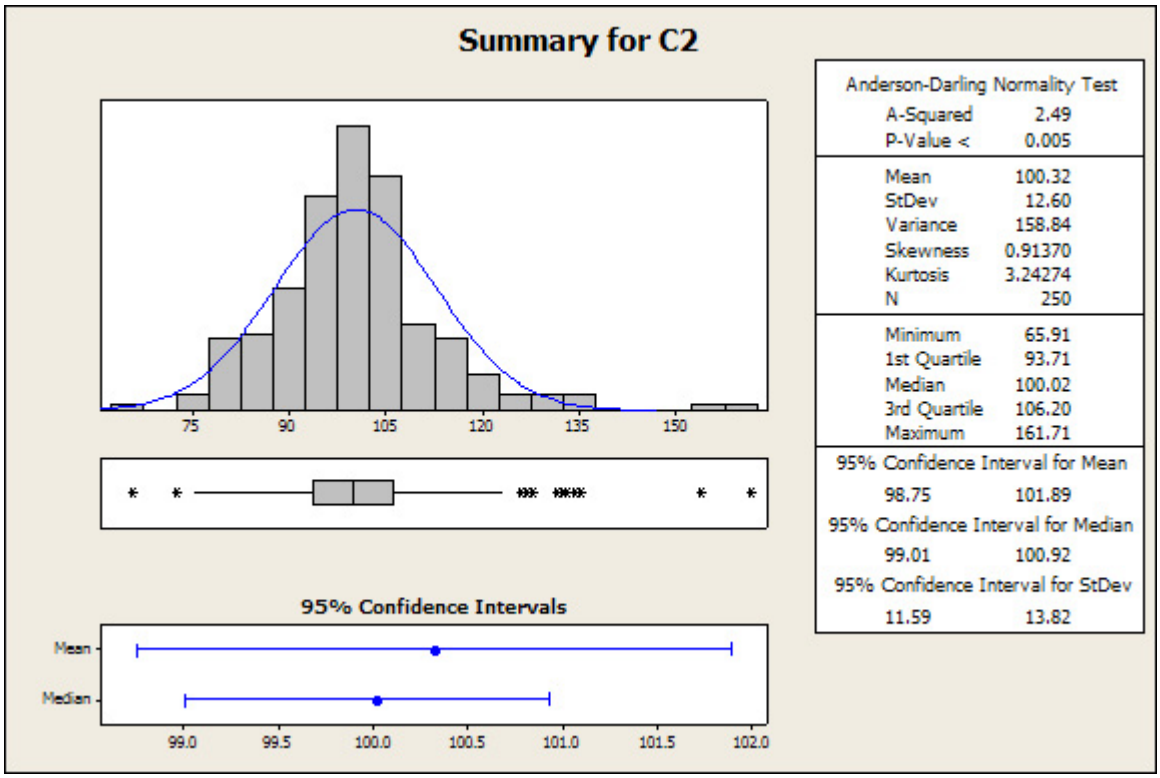
some without indicating nonnormality. If a sample of 30 observations contains 4 outliers, two of which are extreme, would it be reasonable to assume the population from which the data were collected has a normal frequency curve? Probably not.




```
Stem-and-Leaf Display: C1

Stem-and-leaf of C1   N   = 250
Leaf Unit = 1.0

  1     1    8
  5     2    1378
  9     3    3379
  17    4    11223567
  25    5    13455789
  38    6    2222444458899
  65    7    11222223345555567777888889
  98    8    000011112233445555666666678888889
(32)    9    11112223344455555555667888899999
  120   10   000123334444444455566667788889
  90    11   001111122233344455566689 9
  64    12   00011111122234444555556 89
  39    13   001112344466779
  24    14   011366677778
  12    15   001133
  6     16   04669
  1     17   0
```

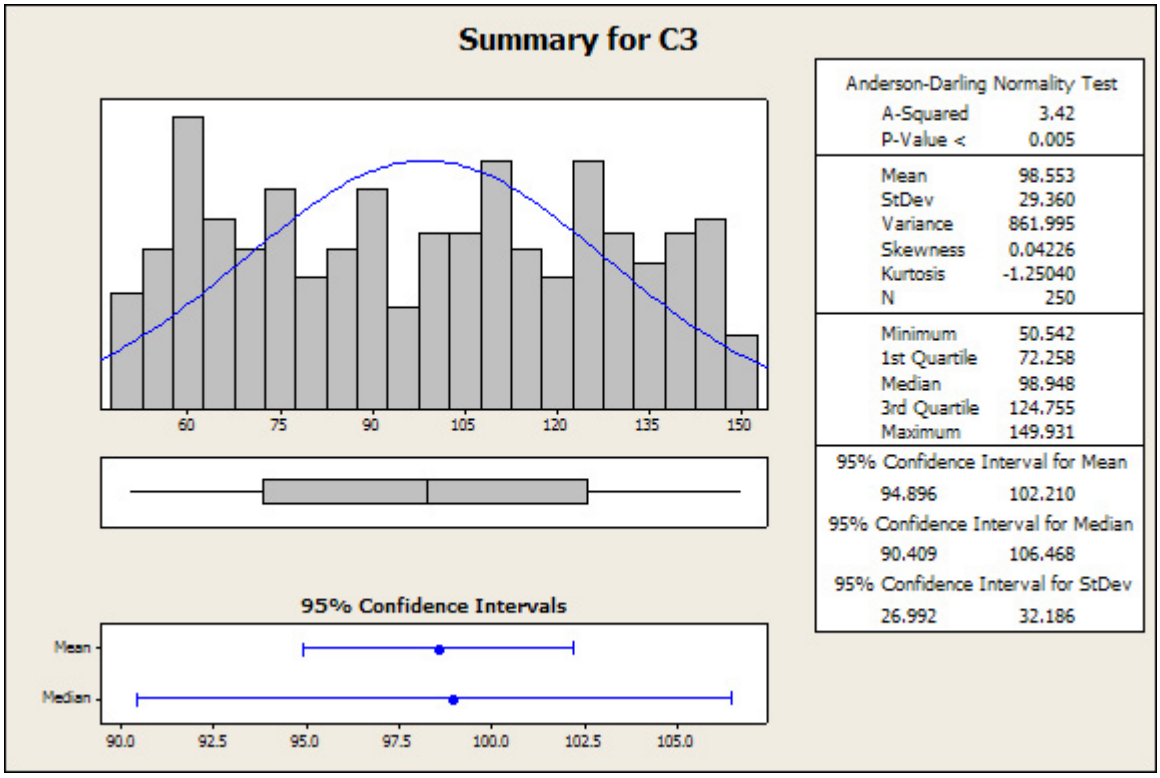
The boxplot is better at highlighting outliers than are other displays. The histogram and stem and leaf displays below appear to have the same basic shape as a normal curve (unimodal, symmetric). However, the boxplot shows that we have a dozen outliers in a sample of 250 observations.

We would only expect about two outliers in 250 observations when sampling from a population with a normal frequency curve. The frequency curve is best described as unimodal, symmetric, and **heavy-tailed**.




```
Stem-and-Leaf Display: C2

Stem-and-leaf of C2  N  = 250
Leaf Unit = 1.0

 1    6   5
 11   7   2578899999
 45   8   00011222333333344567777778888899999
 124  9   000000111222222333333344444455555555566666666666677777777888888889+
(84)  10  000000000000000000111111111112222223333333334444444444455555566666+
 42   11  00000011122222333345556689
 16   12  000113567
 7    13  12345
 2    14
 2    15  3
 1    16  1
```

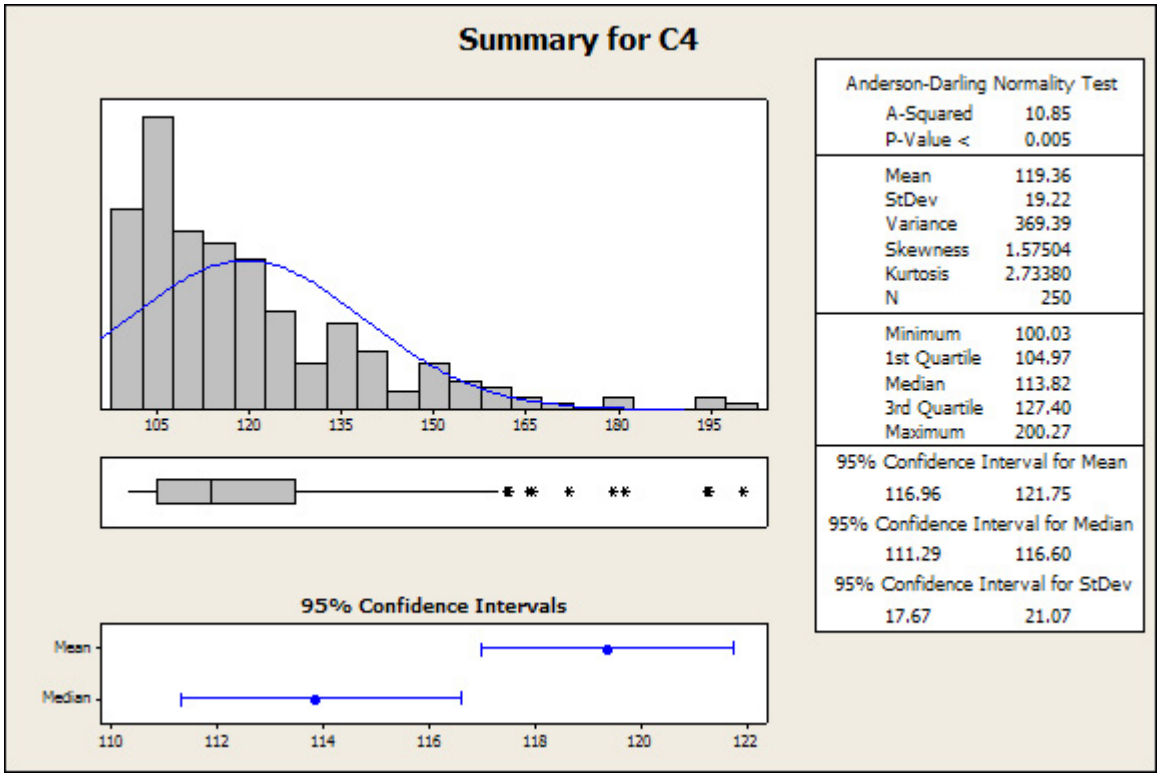
Not all symmetric distributions are mound-shaped, as the display below suggests. The boxplot shows symmetry, but the tails of the distribution are shorter (lighter) than in the normal distribution. Note that the distance between quartiles is roughly constant here.




```
Stem-and-Leaf Display: C3

Stem-and-leaf of C3  N  = 250
Leaf Unit = 1.0

 29    5   00111122334555666677777899999
 56    6   000001111223334566666777889
 82    7   000111233344455556667889
 108   8   111222223445566677788888889
(18)   9   001113334466788889
 124   10  000001223455566667788899999
 97    11  00000111223334566668899
 73    12  00113334444445555666778899
 48    13  000001112333444567778889999
 22    14  0001244455566666777999
```

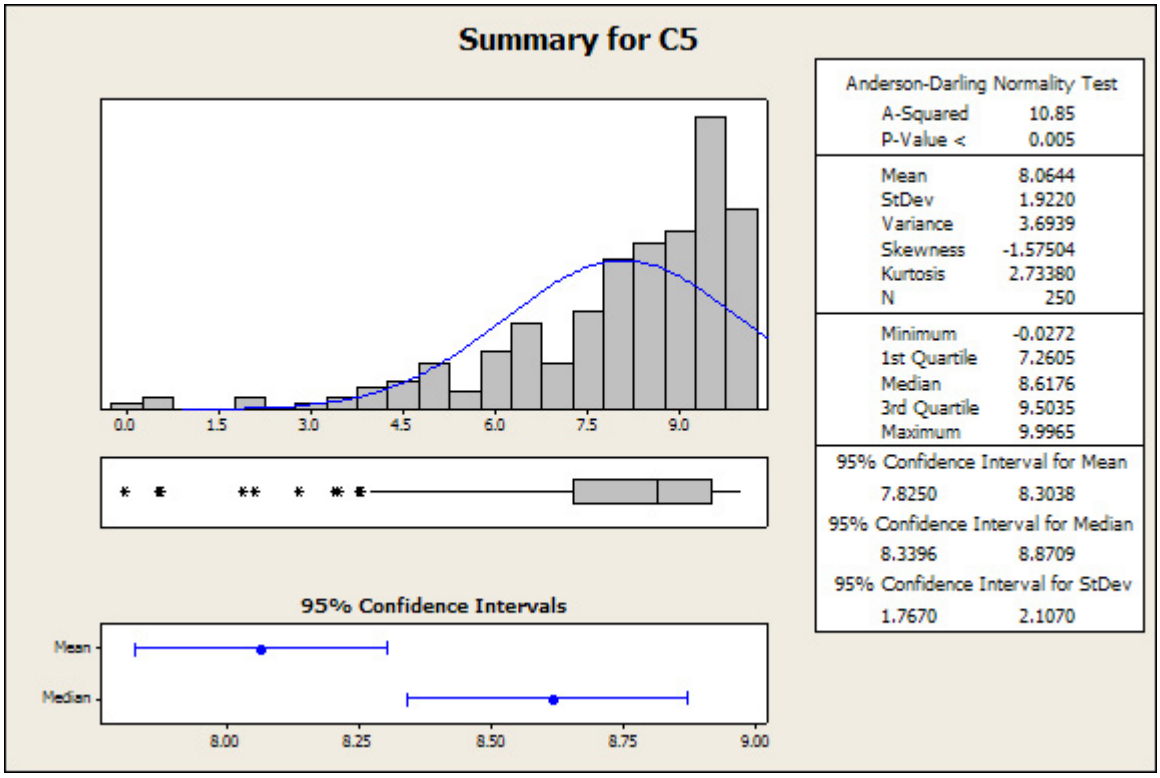
The mean and median are identical in a population with a (exact) symmetric frequency curve. The histogram and stem and leaf displays for a sample selected from a symmetric population will tend to be fairly symmetric. Further, the sample means and medians will likely be close.

The distribution below is unimodal, and asymmetric or **skewed**. The distribution is said to be **skewed to the right**, or upper end, because the right tail is much longer than the left tail. The boxplot also shows the skewness - the region between the minimum observation and the median contains half the data in less than 1/5 the range of values. In addition, the upper tail contains several outliers.



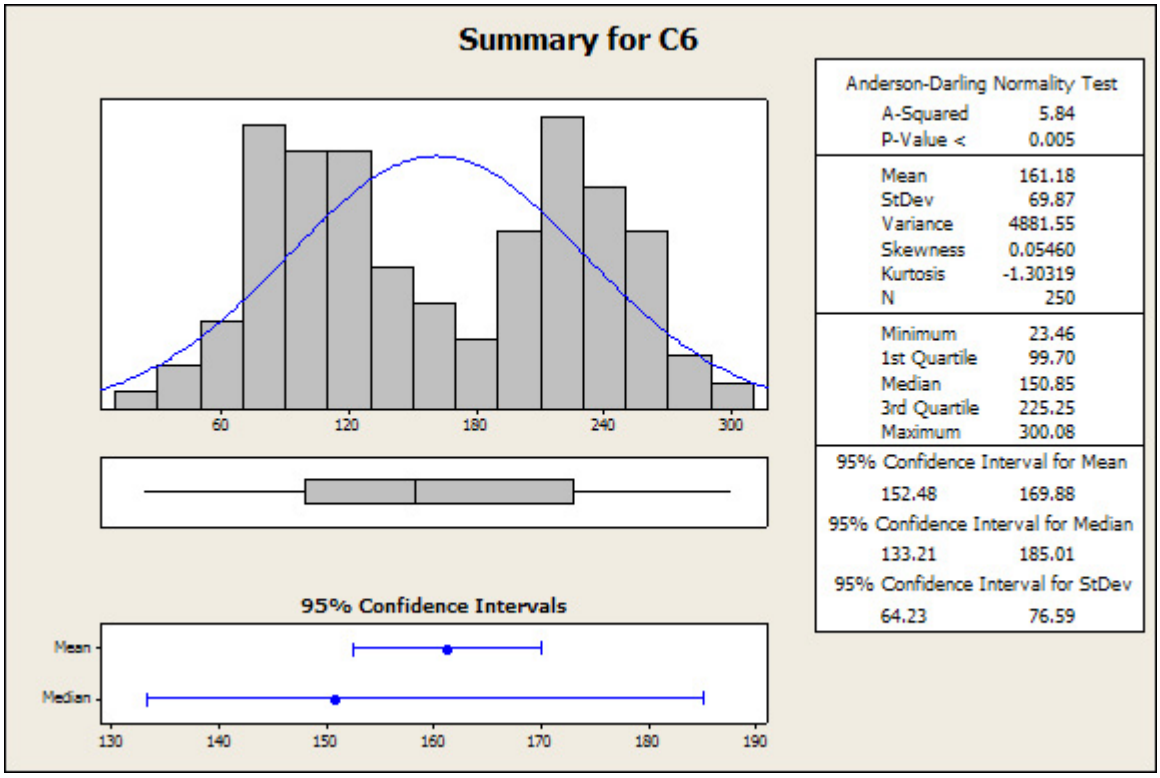

```
Stem-and-Leaf Display: C4

Stem-and-leaf of C4  N  = 250
Leaf Unit = 1.0

 105   10  000000000000000000000001111111111222222222333333333333334444444555+
(57)   11  000111112222222333333344444444455666667777777788888899999
 88    12  000000011223344445556666677788
 57    13  0112333445555666678889
 34    14  00112234489
 23    15  00111235556
 12    16  001256
 6     17  19
 4     18  1
 3     19  44
 1     20  0
```

The distribution below is unimodal and **skewed to the left**. The two examples show that extremely skewed distributions often contain outliers in the longer tail of the distribution.




```
Stem-and-Leaf Display: C5

Stem-and-leaf of C5  N  = 250
Leaf Unit = 0.10

 3    0   055
 4    1   8
 6    2   08
 12   3   347899
 23   4   34446788899
 34   5   01556778899
 57   6   011122333344445566678889
 88   7   11222233333444555566778899999999
(57)  8   000001111112222222233333445555555556666666777777788888999
 105  9   00000000111111111122222223333333444444444445555555666666666666677+
 0    10
```

Not all distributions are unimodal. The distribution below has two modes or peaks, and is said to be **bimodal**. Distributions with three or more peaks are called **multi-modal**.




```
Stem-and-Leaf Display: C6

Stem-and-leaf of C6  N  = 250
Leaf Unit = 10

  4     0  2233
 12     0  44455555
 32     0  66666777777777777777
 64     0  88888888888888888999999999999999
 95     1  00000000000001111111111111111
 115    1  22222222222333333333
(15)    1  444444445555555
 120    1  6666677777
 110    1  88899999999
 99     2  000000000001111111111111111
 71     2  22222222222222222333333333333333
 38     2  4444444445555555555555
```

The boxplot and histogram or stem and leaf display (or dotplot) are used **together** to describe the distribution. The boxplot does not provide information about modality - it only tells you about skewness and the presence of outliers.

As noted earlier, many statistical methods assume the population frequency curve is normal. Small deviations from normality usually do not dramatically influence the operating characteristics of these methods. We worry most when the deviations from normality are severe, such as extreme skewness or heavy tails containing multiple outliers.

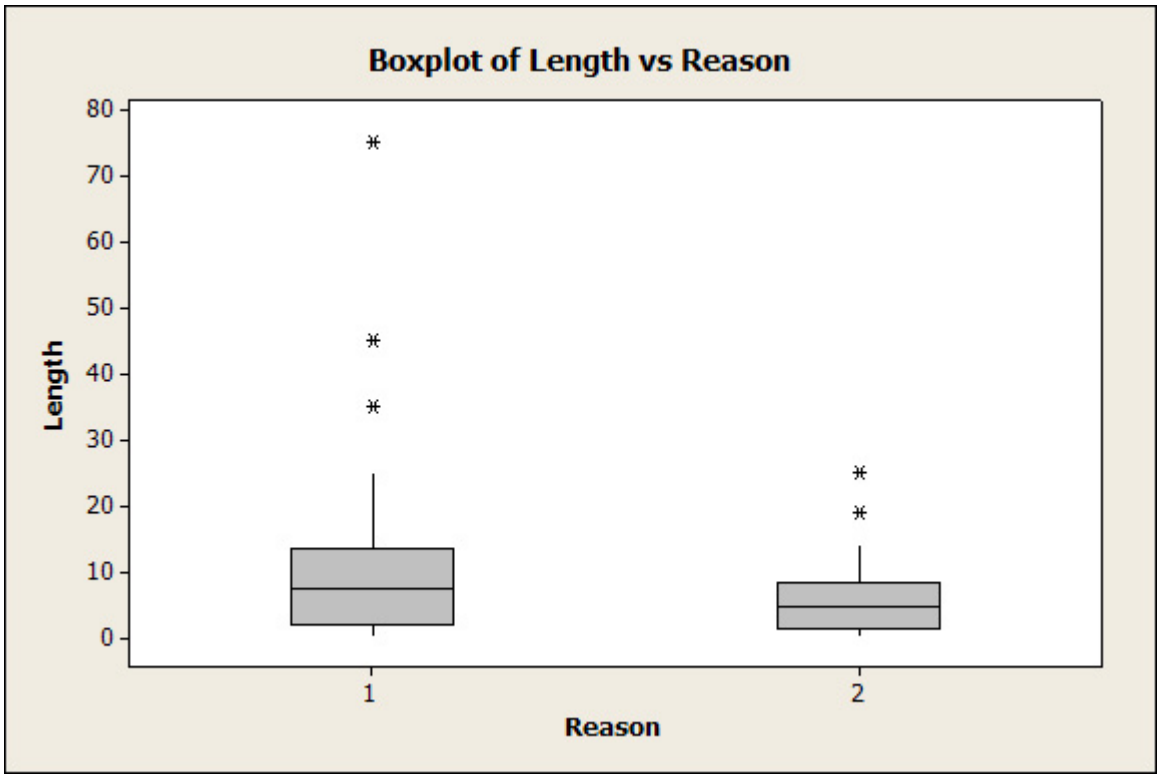
Interpretations for Examples

The **MAO** samples are fairly symmetric, unimodal (?), and have no outliers. The distributions do not deviate substantially from normality. The various measures of central location ($\overline{Y}$, M) are fairly close, which is common with reasonably symmetric distributions containing no outliers.

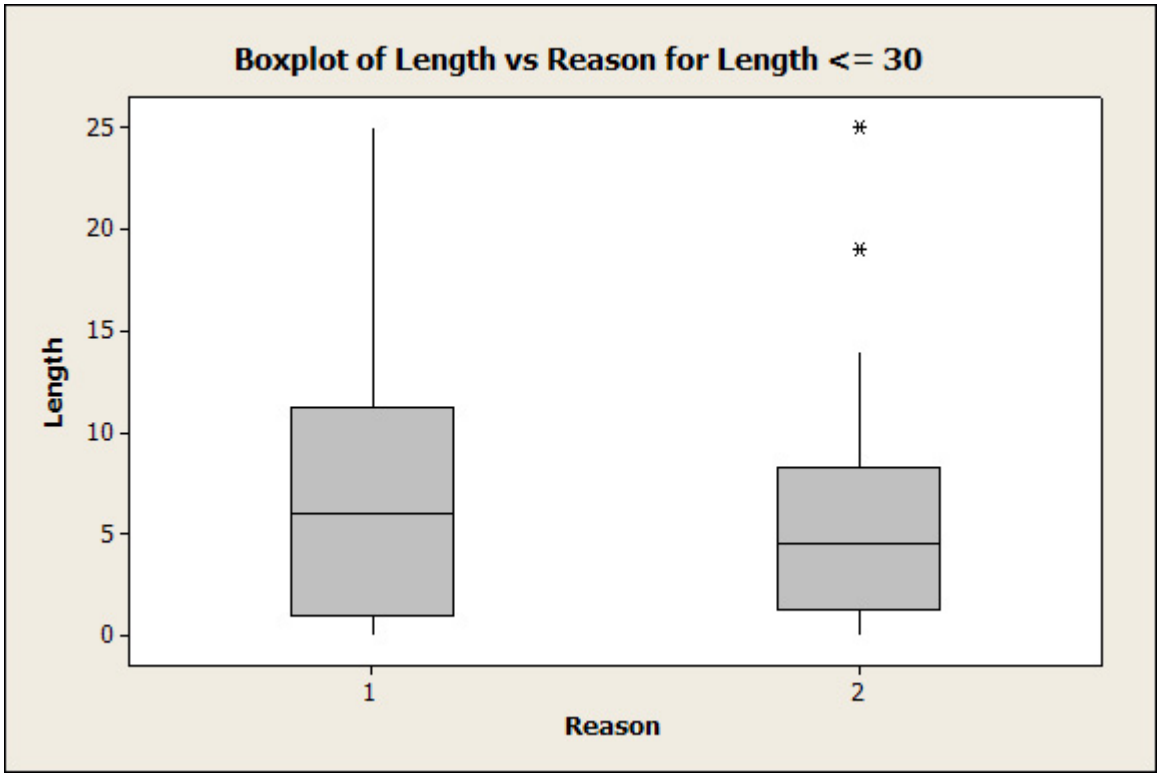
The **SIDS** sample is unimodal, and skewed to the right due to the presence of four outliers in the upper tail. Although not given, we expect the mean to be noticeably higher than the median (Why?). A normality assumption here is unrealistic.

Example: Length of Stay in a Psychiatric Unit

Data on all 58 persons committed voluntarily to the acute psychiatric unit of a health care canter in Wisconsin during the first six months of a year are stored in the worksheet HCC that installs with Minitab. Two of the variables are Length (of stay, in number of days), and Reason (for discharge, 1=normal, 2=other). It is of interest to see if the length of stay differs for the two types of discharge. The main parts of the boxplots comparing the groups are rather compressed (and not very useful) because outliers are using up all the scale.



The solution in a case like this is to zoom in using the Data Options in the boxplot display. In this case let us exclude rows where Length > 30. Now we have a little more basis for comparison.



Does it look like there is a really large difference between the groups? What would you say about the shape of the distributions? Does it look like these are normally distributed values?

Examine the descriptive statistics. What is a reasonable summary here and what probably is pretty distorted? What is your summary of the data based upon the boxplots and numerical summaries?

Descriptive Statistics: Length

Variable	Reason	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Length	1	42	0	11.55	2.25	14.60	0.00	1.75	7.50	13.50
	2	16	0	6.44	1.78	7.11	0.00	1.25	4.50	8.25

Variable	Reason	Maximum
Length	1	75.00
	2	25.00

4 Basics of Probability

Most of this material is covered quite nicely in SW, so I plan to stick with the text very closely.

Populations and Samples

- SW Section 2.8. covers populations.
- SW Section 3.2 covers simple random sampling (SRS). We will use Minitab to illustrate random sampling for Example 3.1 p.74
- Bias in sampling is illustrated in Examples 3.2, and 3.3. The other examples are worth studying also.
- Sampling human populations often involves stratification and/or clustering of individuals into groups. We'll look at this, but for now just note that it is more complicated.

Probability

We will focus on the relative frequency interpretation of probability on p. 80, and the examples.

Probability Rules (Section 3.5) is really a huge topic, requiring more time than the value it adds to your understanding. While it is on the syllabus, we are going to skip it. The material on probability trees is more accessible and gets you most of what you need in probability calculation.

Probability Trees

SW Section 3.4. Trees provide a device for organizing probability calculations. Let us examine in detail examples 15 and 17, and do problem 3.9. We want to get the terms *sensitivity* and *specificity* out of this section, and understand how little information a test may have even with very good values of both.

Density Curves

SW Section 3.6. These are basically histograms of populations standardized to have and area of 1 under them, so that area can be related to sampling probability. We will do some simulations in Minitab to see how histograms of huge sets of numbers can look like smooth curves. We will cover Example 3.30.

Random Variables

SW Section 3.7. All we really want to cover is the definition. This is a mathematical model for a population. Populations have means (μ) and standard deviations (σ), and the idea is identical for random variables. We will skip that part of this section.

Binomial Distribution

SW Section 3.8. This is our model for binary outcomes. We really want to understand well the *Independent-Trials Model* on p. 103. We will use Minitab to do the calculation for Example 3.45, and we want to understand how the model breaks down for Example 3.50.

The Normal Distribution

SW Sections 4.1-4.3. We will just get a good start on this and continue next time. We need to see how to use a table like Table 3 p. 675-6, although we will see how to get the answers more easily out of Minitab. A great deal of what we do in statistics involves normal distribution calculations, and while those usually are done within software we need to understand what is being done behind the scenes.

5 Probability, Sampling Distributions, Central Limit Theorem

As with last week, most of this material is covered quite nicely in SW, so I plan to stick with the text very closely. We will do a quite a bit of computer work to accompany this material, both during lecture and the lab.

Random Variables

SW Section 3.7.

This is our model for sampling from a population. If we sample (from either a categorical or quantitative) population, we write Y = the value obtained. If we sample n values from a population, the values obtained are $Y_1, Y_2, \ldots, Y_n$. The population we sampled from has a mean μ and standard deviation σ (we'll force even categorical variables into such a structure) – those are the mean and standard deviation of the random variable as well. Don't worry about the more mathematical treatment in SW.

Binomial Distribution

SW Section 3.8.

Eric covered the Independent-Trials Model with you in lab. The binomial distribution lays out probabilities for all the possible numbers of successes in n independent trials with probability p of Success each trial. This is a new population with a mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$. The model is important, but don't worry about all the formulae in SW. Minitab does a great job of calculating probabilities when needed.

Normal Distribution

SW Sections 4.1-3.

Eric also covered this in lab. We want to revisit Figure 4.7, the standard normal Z, and the Standardization Formula $Z = \frac{Y-\mu}{\sigma}$ on p. 124. The figures on p. 125 are a valuable working guide. We will work a couple of examples, including Minitab calculations.

Normal distributions pop up in many more situations than you would expect. We need to be able to use them.

Assessing Normality

SW Section 4.4.

Eric will cover this in the lab. The normal probability plot is a widely used tool. SW do not talk about box plots here, but those also serve as valuable tools. We will be making the assumption many times that we sampled from a normal population. The assumption really does matter, so we need methods to assess it.

Sampling Distribution of $\overline{Y}$

SW Section 5.3.

If we randomly sample (SRS) n values $Y_1, Y_2, \ldots, Y_n$ from a quantitative population and calculate $\overline{Y}$, then $\overline{Y}$ depends upon the random sample we drew — if we drew another random sample

we would get a different value of $\overline{Y}$. This means $\overline{Y}$ is a random variable, i.e. it is a single random number sampled from *some* population.

From what population is $\overline{Y}$ drawn? It most certainly is not the same as the population $Y_1, Y_2, \ldots, Y_n$ come from (the possible values may not even be the same). It is a new population called the ***sampling distribution*** of $\overline{Y}$. I'll spare you any derivations and just cite some results.

Mean and Standard Deviation of $\overline{Y}$

If the population $Y_1, Y_2, \ldots, Y_n$ are sampled from has mean μ and standard deviation σ, the sampling distribution of $\overline{Y}$ has mean $\mu_{\overline{Y}} = \mu$ and standard deviation $\sigma_{\overline{Y}} = \sigma/\sqrt{n}$. On average $\overline{Y}$ values come out the same place (μ) as the Y_i values, but they tend to be closer to μ than are the individual Y_i, since the standard deviation is smaller.

Shape of Sampling Distribution of $\overline{Y}$

This is the part with a lot of mathematics behind it. There are two cases when we can treat $\overline{Y}$ as if it was sampled from a Normal Distribution (and we know how to use normal distributions!):

1. If the population $Y_1, Y_2, \ldots, Y_n$ were sampled from is normal, no matter how small n is,
2. if n is large, almost no matter what the shape of the population from which $Y_1, Y_2, \ldots, Y_n$ were sampled.

We cannot say what the shape is for small n unless we originally sampled from a normal distribution. This is why we worry so much about assessing normality with boxplots and normal probability plots. Part 2 is the **Central Limit Theorem**.

Let us go over Examples 5.9 and 5.10. We will do a few simulations in Minitab to demonstrate the preceding results.

Sampling Distribution of $\hat{p}$

SW Sections 5.2, 5.5

This is how we really use the Independent-Trials Model, and the way we think of binary response variables. We now randomly sample n individuals from a population where every value is either a S or F (just generic labels). The proportion of S's in the population is p, the proportion of S's in the sample is $\hat{p}$. Again, $\hat{p}$ is a random variable since it depends upon the random sample, so it has a sampling distribution. What population is $\hat{p}$ sampled from?

The amazing result is that if n is large, we can assume $\hat{p}$ was drawn from a *Normal* population with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. For this to hold we need $np \geq 5$ and $n(1-p) \geq 5$.

We will use Minitab to demonstrate this, and do a few calculations.

6 Estimation in the One-Sample Situation

SW Chapter 6

Standard Errors and the $t-$Distribution

We need to add one more small complication to the sampling distribution of $\overline{Y}$. What we saw last time, and in SW Chapter 5, is that if Y_1, Y_2, ..., Y_n is a random sample from a normal population and that population has mean μ and standard deviation σ, then $\overline{Y}$ looks like it is a single number randomly selected from a normal distribution also with mean $\mu_{\overline{Y}} = \mu$ but with standard deviation $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$. We get from this that $\frac{\overline{Y}-\mu}{\sigma_{\overline{Y}}} = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} = Z$ is a standard normal random variable, and we can use the table on the inside front cover of SW to compute probabilities involving $\overline{Y}$.

Unfortunately, in the context in which we need to use this result, we would need to know σ in order to apply the result. We are sampling from a population in order to find out something about it, so almost certainly we do not know what σ is. What works well is to estimate the population standard deviation σ with sample standard deviation S calculated from the random sample. Our best guesses of the population mean and standard deviation μ and σ are the corresponding sample values $\overline{Y}$ and S. While μ and σ are constants (we do not know the actual values, but they are constants), $\overline{Y}$ and S depend upon the actual sample randomly selected from the population. If we repeated the experiment and drew a second random sample of n observations, we would get different values for $\overline{Y}$ and S, which is to say $\overline{Y}$ and S are random variables.

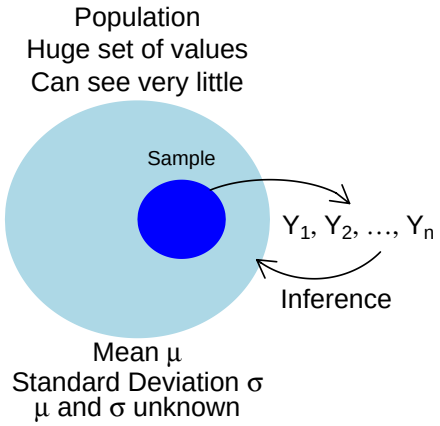
If we are going to estimate σ with S, then of course we would estimate $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$ with $\frac{S}{\sqrt{n}}$. That is exactly what we do, and we give this quantity the name **Standard Error of** $\overline{Y}$, $SE_{\overline{Y}}$. Instead of standardizing $\overline{Y}$ with the expression $\frac{\overline{Y}-\mu}{\sigma_{\overline{Y}}}$ we use the new expression $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$. Using $SE_{\overline{Y}}$ in the denominator introduces extra variability, though, so this no longer looks like a random number that came from a Z distribution. Provided our assumptions are correct (random sampling from a normally distributed population), then $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$ looks like a single random number randomly selected from a Student's $t-$distribution with $n-1$ degrees of freedom (df). The amount of extra variability introduced depends upon the sample size n; if n is very small it is a lot, but by the time n is 30 or so, there is very little difference from a Z, and in fact df $= \infty$ makes the $t-$ and Z distributions the same. SW on p. 187 show how the distribution compares to the normal – it doesn't look like a big difference, but the probability statements we can make are different enough to matter.

Table 4 p. 677 in SW is a standard table of the $t-$distribution. It is organized differently from the Normal Table, since it gives areas under the curve across the top and lets you look up the "critical" values that generate those areas, while the Z table gives you critical values across the side and top and lets you look up areas. We will go through some examples of reading this table during the lecture.

Inference for a Population Mean

Suppose that you have identified a population of interest where individuals are measured on a single quantitative characteristic, say, weight, height or IQ. You select a random or representative sample from the population with the goal of estimating the (unknown) **population mean** value, identified by μ.

This is a standard problem in statistical inference, and the first inferential problem that we will tackle. For notational convenience, identify the measurements on the sample as Y_1, Y_2, ..., Y_n, where n is the sample size. Given the data, our best guess, or estimate, of μ is the sample mean:



$\bar{Y} = \frac{\sum_i Y_i}{n} = \frac{Y_1+Y_2+\cdots+Y_n}{n}$.

There are two main methods that are used for inferences on μ: **confidence intervals** (CI) and **hypothesis tests**. The standard CI and test procedures are based on the sample mean and the sample standard deviation, denoted by s. We will consider CIs in this lecture, and hypothesis tests in the next lecture.

Let's apply the results of the preceding section, and then lay out the mechanics of the procedure. The main idea behind a CI is this: $\overline{Y}$ should be a pretty good guess as to what μ is, but while μ is a constant (we don't know the value, though), $\overline{Y}$ is a random variable (every possible sample gives a different value), so most assuredly $\overline{Y} \neq \mu$. Still, $\overline{Y}$ should not be too far from μ, but how far away from μ do we think $\overline{Y}$ could be? As a specific example, suppose we randomly sample $n = 9$ values from a normal population and get $\overline{Y} = 22$ and $S = 6$. What could μ be?

To answer such a question, apply the $t-$distribution. $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$ looks like a single random number sampled from a $t-$distribution with 8 df, so it *should have* come out somewhere in the middle of that distribution. The middle 95% of that distribution is between -2.306 and 2.306 (from the table). So, we *had* a 95% chance that $\frac{\overline{Y}-\mu}{SE_{\overline{Y}}}$ would fall in that range. Substituting the actual values of $\overline{Y}$ and S we obtained, we are 95% confident that $\frac{22-\mu}{6/\sqrt{9}} = \frac{22-\mu}{2}$ is between -2.306 and 2.306, or equivalently we are 95% confident that $22-\mu$ is between -4.612 and 4.612. This says that μ should be within 4.612 of 22, or in the range $22-4.612$ to $22+4.612$, i.e. between 17.388 and 26.612. We still do not know what μ is, but to have gotten data like this μ must be somewhere between 17.388 and 26.612.

The interval $17.388 \leq \mu \leq 26.612$ is referred to as a 95% confidence interval for μ. It is improper to say there is a 95% chance that μ is in that range: If it is in that range, say 25, there is a 100% chance it is in that range, while if it is not in that range, say 30, there is a 0% chance it is in that range. The 95% refers to how often using this technique works (like a lifetime batting average) - this interval either worked in capturing μ or it did not work, and we cannot know which is true.

Mechanics of a CI for μ

A CI for μ is a range of plausible values for the unknown population mean μ, based on the observed data. To compute a CI for μ:

1. Specify the **confidence coefficient**, which is a number between 0 and 100%, in the form $100(1-\alpha)\%$. Solve for α.

2. Compute the $t-$critical value: $t_{crit} = t_{.5\alpha}$ such that the area under the $t-$ curve ($df = n-1$) to the right of t_{crit} is $.5\alpha$.

3. The desired CI has lower and upper endpoints given by $L = \bar{Y} - t_{crit}SE_{\overline{Y}}$ and $U = \bar{Y} + t_{crit}SE_{\overline{Y}}$, respectively, where $SE_{\overline{Y}} = s/\sqrt{n}$ is the standard error of the sample mean. The CI is often written in the form $\bar{Y} \pm t_{crit}SE_{\overline{Y}}$.

In practice, the confidence coefficient is large, say 95% or 99%, which correspond to $\alpha = .05$ and .01, respectively. The value of α expressed as a percent is known as the **error rate** of the CI.

The CI is determined once the confidence coefficient is specified and the data are collected. Prior to collecting the data, the interval is unknown and is viewed as random because it will depend on the actual sample selected. Different samples give different CIs. The "confidence" in, say, the 95% CI (which has a 5% error rate) can be interpreted as follows. If you repeatedly sample the population and construct 95% CIs for μ, then 95% of the intervals will contain μ, whereas 5% will not. The interval you construct from your data will either cover μ, or it will not.

The length of the CI

$$U - L = 2t_{crit}SE_{\overline{Y}}$$

depends on the accuracy of our estimate $\overline{Y}$ of μ, as measured by $SE_{\overline{Y}} = s/\sqrt{n}$ the standard error of $\overline{Y}$. Less precise estimates of μ lead to wider intervals for a given level of confidence.

Assumptions for Procedures

I described the classical CI. The procedure is based on the assumptions that the data are a random sample from the population of interest, and that the population frequency curve is normal. The population frequency curve can be viewed as a "smoothed histogram" created from the population data.

The normality assumption can be checked using a stem-and-leaf display, a boxplot, or a normal scores plot of the sample data (probably the more the better).

Example

Let us go through a hand-calculation of a CI, using Minitab to generate summary data. I will then show you how the CI is generated automatically in Minitab. The ages (in years) at first transplant for a sample of 11 heart transplant patients are as follows: 54 42 51 54 49 56 33 58 54 64 49.

```
Data Display

 AgeTran
    54   42   51   54   49   56   33   58   54   64   49

 Stem-and-Leaf Display: AgeTran
```

```
Stem-and-leaf of AgeTran  N  = 11
Leaf Unit = 1.0

 1   3  3
 1   3
 2   4  2
 4   4  99
(4)  5  1444
 3   5  68
 1   6  4
```

```
Descriptive Statistics: AgeTran

Variable   N  N*   Mean  SE Mean  StDev  Minimum    Q1  Median      Q3 Maximum
AgeTran   11   0  51.27     2.49   8.26    33.00 49.00   54.00  56.00   64.00
```

The summaries for the data are: $n = 11$, $\overline{Y} = 51.27$, and $s = 8.26$ so that $SE_{\overline{Y}} = 8.26/\sqrt{11} = 2.4904$. The degrees of freedom are $df = 11 - 1 = 10$.

A necessary first step in every problem is to define the population parameter in question. Here, let

$$\mu = \text{mean age at time of first transplant for population of patients.}$$

Let us calculate a 95% CI for μ. The degrees of freedom are $df = 11 - 1 = 10$. For a 95% CI $\alpha = .05$, so we need to find $t_{crit} = t_{.025} = 2.228$.

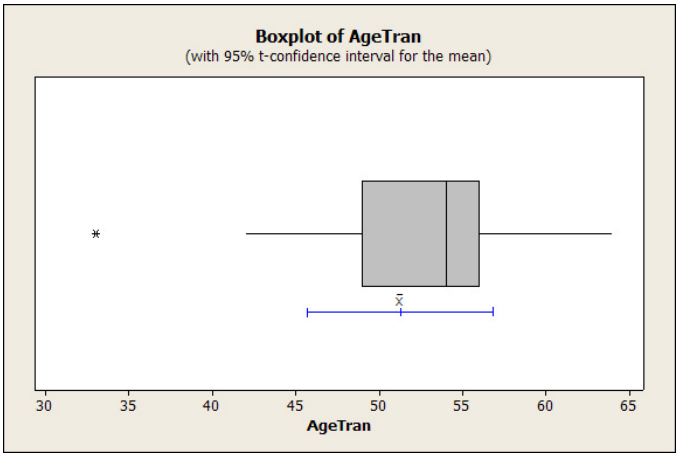
Now $t_{crit} SE_{\overline{Y}} = 2.228 * 2.4904 = 5.55$. The lower limit on the CI is $L = 51.27 - 5.55 = 45.72$. The upper endpoint is $U = 51.27 + 5.55 = 56.82$.

I insist that the results of every CI be summarized in words. For example, I am 95% confident that the population mean age at first transplant is between 45.7 and 56.8 years (rounding off to 1 decimal place).

Minitab does all this very easily. Follow the menu path `Stat > Basic Statistics > 1-Sample t` (be careful that you don't select the `1-Sample Z` — it will treat S as if it is actually σ and give you incorrect bounds). Under `Options...` select Confidence Level of 95 (the default) and Alternative: not equal (we will understand that next week). Under `Graphs` check Boxplot. Do not check `Summarized data` or `Perform hypothesis test`. You get the following results:

```
One-Sample T: AgeTran

Variable   N     Mean   StDev  SE Mean         95% CI
AgeTran   11  51.2727  8.2594   2.4903  (45.7240, 56.8215)
```

We might be a little concerned about the outlier and the possible skewness indicated in the boxplot below, since that could be evidence we did not sample from a normal distribution. It will be worth trying one of the nonparametric procedures we will learn about later, since the assumption of normality is not made there.



The Effect of α on a Two-Sided CI

A two-sided $100(1-\alpha)\%$ CI for μ is given by $\overline{Y} \pm t_{crit} s/\sqrt{n}$. The CI is centered at $\overline{Y}$ and has length $2t_{crit}s/\sqrt{n}$. The confidence coefficient $100(1-\alpha)\%$ is *increased* by *decreasing* α, which increases t_{crit}. That is, increasing the confidence coefficient makes the CI wider. This is sensible: to increase your confidence that the interval captures μ you must pinpoint μ with less precision by making the CI wider. For example, a 95% CI is wider than a 90% CI.

SW Example 6.9 page 192: Let us compute a 90% and a 95% CI by hand.

Note: For large n the Central Limit Theorem gives us the ability to treat $\frac{\overline{Y}-\mu}{\sigma_{\overline{Y}}}$ as a Z random variable even without sampling from a normal distribution. Some texts would suggest using the `1-Sample Z` procedure in this case (although that still begs the issue of not knowing σ). In practice what we do about large n is to worry a little less about lack of normality in the population we sampled from (outliers and extreme skewness are still problems, just slightly different ones), but continue to use the t-procedures. Remember for large n we get large df, and for large df there is little difference between Z and t.

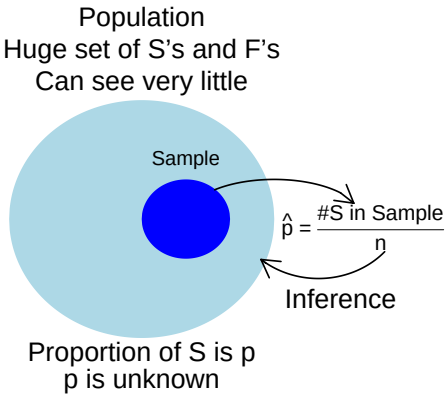
Inference for a Population Proportion

Assume that you are interested in estimating the proportion p of individuals in a population with a certain characteristic or attribute based on a random or representative sample of size n from the population. The **sample proportion** $\hat{p}$ =(# with attribute in the sample)/n is the best guess for p based on the data.

This is the simplest **categorical data** problem. Each response falls into one of two exclusive and exhaustive categories, called success and failure. Individuals with the attribute of interest are in the success category. The rest fall into the failure category. Knowledge of the **population proportion** p of successes characterizes the distribution across both categories because the population proportion of failures is $1-p$.

As an aside, note that the probability that a randomly selected individual has the attribute of interest is the population proportion p with the attribute, so the terms population proportion and probability can be used interchangeably with random sampling.

The diagram of this is very similar to the earlier one. Note that a random sample of size n now becomes just a set of S's and F's.



A CI for p

The derivation of the CI follows the same basic ideas as before, except we do not have the idea of df since we are considering n as large ($np \geq 5$ and $n(1-p) \geq 5$). $\hat{p}$ is a random variable (it almost surely is not p), and it looks like a single number randomly selected from a normal distribution with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, so $\frac{\hat{p}-p}{\sigma_{\hat{p}}}$ looks like a Z. We have the same problem as before – to use this as we wish, we need to compute the denominator, but we need to know p to compute it. We estimate it instead, and call the estimated standard deviation of $\hat{p}$ the standard error of $\hat{p}$, $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Everything proceeds as before.

A two-sided CI for p is a range of plausible values for the unknown population proportion p, based on the observed data. To compute a two-sided CI for p:

1. Specify the confidence level as the percent $100(1-\alpha)\%$ and solve for the error rate α of the CI.

2. Compute $z_{crit} = z_{.5\alpha}$ (i.e. area under the standard normal curve to the right of z_{crit} is $.5\alpha$.)

3. The $100(1-\alpha)\%$ CI for p has endpoints $L = \hat{p} - z_{crit}SE$ and $U = \hat{p} + z_{crit}SE$, respectively, where the "CI standard error" is
$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$
The CI is often written as $\hat{p} \pm z_{crit}SE$.

The CI is determined once the confidence level is specified and the data are collected. Prior to collecting data, the CI is unknown and can be viewed as random because it will depend on the

actual sample selected. Different samples give different CIs. The "confidence" in, say, the 95% CI (which has a .05 or 5% error rate) can be interpreted as follows. If you repeatedly sample the population and construct 95% CIs for p, then 95% of the intervals will contain p, whereas 5% (the error rate) will not. The CI you get from your data either covers p, or it does not.

The length of the CI

$$U - L = 2z_{crit}SE$$

depends on the accuracy of the estimate $\hat{p}$, as measured by the standard error SE. For a given $\hat{p}$, this standard error decreases as the sample size n increases, yielding a narrower CI. For a fixed sample size, this standard error is maximized at $\hat{p} = .5$, and decreases as $\hat{p}$ moves towards either 0 or 1. In essence, sample proportions near 0 or 1 give narrower CIs for p. However, the normal approximation used in the CI construction is less reliable for extreme values of $\hat{p}$.

Example:

The 1983 Tylenol poisoning episode highlighted the desirability of using tamper-resistant packaging. The article "Tamper Resistant Packaging: Is it Really?" (Packaging Engineering, June 1983) reported the results of a survey on consumer attitudes towards tamper-resistant packaging. A sample of 270 consumers was asked the question: "Would you be willing to pay extra for tamper resistant packaging?" The number of yes respondents was 189. Construct a 95% CI for the proportion p of all consumers who were willing in 1983 to pay extra for such packaging.

Here $n = 270$ and $\hat{p} = 189/270 = .700$. The critical value for a 95% CI for p is $z_{.025} = 1.96$. The CI standard error is given by

$$SE = \sqrt{\frac{.7 * .3}{270}} = .028,$$

so $z_{crit}SE = 1.96 * .028 = .055$. The 95% CI for p is $.700 \pm .055$. You are 95% confident that the proportion of consumers willing to pay extra for better packaging is between .645 and .755. (How much extra?).

Appropriateness of the CI

The standard CI is based on a large sample standard normal approximation to

$$z = \frac{\hat{p} - p}{SE}.$$

A simple rule of thumb requires $np \geq 5$ and $n(1-p) \geq 5$ for the method to be suitable. The population proportion p is unknown so you should use $\hat{p}$ in these formulae to check the suitability of the CI. Given that $n\hat{p}$ and $n(1-\hat{p})$ are the observed numbers of successes and failures, you should have at least 5 of each to apply the large sample CI.

In the packaging example, $n\hat{p} = 270*(.700) = 189$ (the number who support the new packaging) and $n(1-\hat{p}) = 270*(.300) = 81$ (the number who oppose) both exceed 5. The normal approximation is appropriate here.

More Accurate Confidence Intervals

Large sample CIs for p should be interpreted with caution in small sized samples because the true error rate usually exceeds the assumed (nominal) value. For example, an assumed 95% CI, with a nominal error rate of 5%, may be only an 80% CI, with a 20% error rate. The large sample CIs are usually overly optimistic (i.e. too narrow) when the sample size is too small to use the normal approximation.

SW use the following method, originally suggested by Alan Agresti, for a 95% CI. The standard method computes the sample proportion as $\hat{p} = y/n$ where y is the number of individuals in the sample with the characteristic of interest, and n is the sample size. Agresti suggested estimating the proportion with $\tilde{p} = (y+2)/(n+4)$, with a standard error of

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}},$$

and using the "usual interval" with these new summaries: $\tilde{p} \pm 1.96SE$. This appears odd, but just amounts to adding two successes and two failures to the observed data, and then computing the standard CI.

This adjustment has little effect when n is large and $\hat{p}$ is not close to either 0 or 1, as in the Tylenol example. Let us do examples using SW's proposed CI.

SW Examples 6.16 and 6.17, page 208-9

Minitab Implementation

A CI for p can be obtained in Minitab from summary data from the menu path `Stat > Basic Statistics > 1 Proportion`, check `Summarized data`, enter `Number of trials` (n) and `Number of events` (# Successes), click `Options`, enter `Confidence level` in percent (95.0 usually), ignore `Test proportion` for now, select `Alternative: not equal`, and check `Use test and interval based on normal distribution`.

The above choices produce a CI based upon $\hat{p}$. In order to use SW's CI based on $\tilde{p}$, add 4 to n and 2 to # Successes. Finally, to get the best interval (arguably *the* correct one), **do not** check `Use test and interval based on normal distribution`. This third choice produces what is known as an exact interval – it is a lot harder to explain how we get it (I'll indicate where it comes from next week), but the confidence level and error rate are correct and not subject to approximation like the other two intervals. Minitab is a little unique in providing this. Let us examine Minitab results from two examples:

The Tylenol Example:

Using $\hat{p}$:

```
Sample    X    N  Sample p         95% CI         Z-Value  P-Value
1       189  270  0.700000  (0.645339, 0.754661)     6.57    0.000
```

Using $\tilde{p}$:

```
Sample    X    N  Sample p         95% CI         Z-Value  P-Value
1       191  274  0.697080  (0.642670, 0.751490)     6.52    0.000
```

Using exact interval:

```
                                                    Exact
Sample    X    N  Sample p         95% CI         P-Value
1       189  270  0.700000  (0.641500, 0.754047)    0.000
```

Ignore the Z-Value and P-Value entries for now. You can see that the intervals all agree for any practical interpretation.

Example 6.17 p. 209 of SW

Using $\hat{p}$:

```
 Sample  X   N  Sample p    CI    Z-Value  P-Value
 1       0  11  0.000000  (*, *)    -3.32    0.001

 * NOTE * The normal approximation may be inaccurate for small samples.
```

Using $\tilde{p}$:

```
Sample  X   N  Sample p         95% CI         Z-Value  P-Value
1       2  15  0.133333  (0.000000, 0.305361)    -2.84    0.005

* NOTE * The normal approximation may be inaccurate for small samples.
```

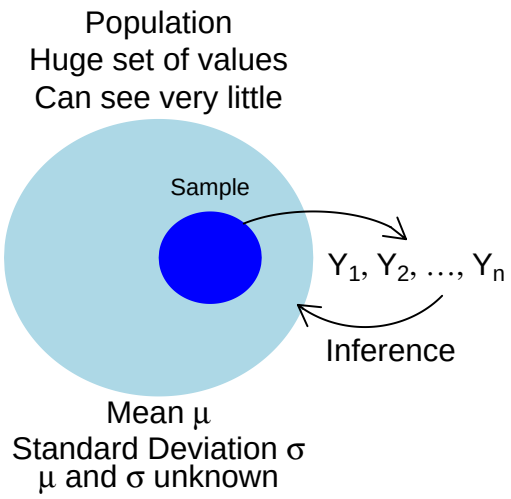
Using exact interval:

```
                                                  Exact
Sample  X   N  Sample p         95% CI         P-Value
1       0  11  0.000000  (0.000000, 0.238404)    0.001
```

The only one of these I would trust is the exact one. The one based on $\tilde{p}$ is surprisingly informative, though. Minitab's warning on the other two should not be ignored.

7 Hypothesis Testing in the One-Sample Situation

Suppose that you have identified a population with the goal of estimating the (unknown) **population mean** value, identified by μ. You select a random or representative sample from the population where, for notational convenience, the sample measurements are identified as $Y_1,\ Y_2,\ ...,\ Y_n$, where n is the sample size.



Given the data, our best guess, or estimate, of μ is the sample mean:

$$\bar{Y} = \frac{\sum_i Y_i}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

There are two main methods for inferences on μ: **confidence intervals** (CI) and **hypothesis tests**. The standard CI and test procedures are based on $\overline{Y}$ and s, the sample standard deviation. I discussed CIs in the last lecture.

Hypothesis Test for μ

Suppose you are interested in checking whether the population mean μ is equal to some prespecified value, say μ_0. This question can be formulated as a two-sided hypothesis test, where you are trying to decide which of two contradictory claims or hypotheses about μ is more reasonable given the observed data. The **null hypothesis**, or the hypothesis under test, is $H_0 : \mu = \mu_0$, whereas the **alternative hypothesis** is $H_A : \mu \neq \mu_0$.

I will explore the ideas behind hypothesis testing later. At this point, I focus on the mechanics behind the test. The steps in carrying out the test are:

1. Set up the null and alternative hypotheses: $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$, where μ_0 is specified by the context of the problem.

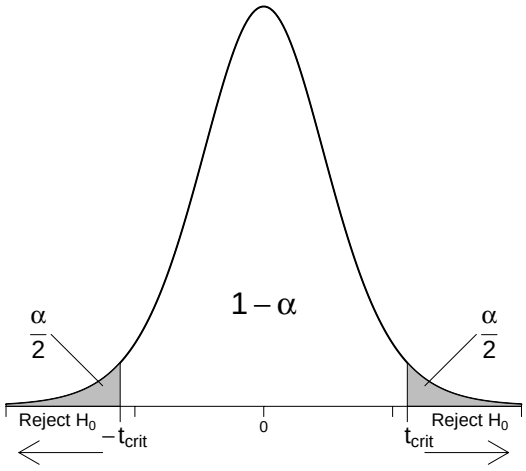
2. Choose the **size** or **significance level** of the test, denoted by α. In practice, α is set to a small value, say, .01 or .05, but theoretically can be any value between 0 and 1.

3. Compute the critical value t_{crit} from the $t-$distribution table with degrees of freedom $df = n - 1$. In terms of percentiles, $t_{crit} = t_{.5\alpha}$.

4. Compute the **test statistic**
$$t_s = \frac{\bar{X} - \mu_0}{SE},$$
where $SE = s/\sqrt{n}$ is the standard error.

5. **Reject** H_0 in favor of H_A (i.e. decide that H_0 is false, based on the data) if $|t_s| > t_{crit}$. Otherwise, do not reject H_0. An equivalent rule is to Reject H_0 if $t_s < -t_{crit}$ or if $t_s > t_{crit}$. I sometimes call the test statistic t_{obs} to emphasize that the computed value depends on the observed data.

The process is represented graphically below. The area under the $t-$probability curve outside $\pm t_{crit}$ is the size of the test, α. One-half α is the area in each tail. You reject H_0 in favor of H_A only if the test statistic is outside $\pm t_{crit}$.



Assumptions for Procedures

I described the classical $t-$test, which assumes that the data are a random sample from the population and that the population frequency curve is normal. The population frequency curve can be

viewed as a "smoothed histogram" created from the population data. You assess the reasonableness of the normality assumption using a stem-and-leaf, histogram, and a boxplot of the sample data. The stem-and-leaf and histogram should resemble a normal curve.

The $t-$test is known as a small sample procedure. For large samples, researchers sometimes use a $z-$test, which is a minor modification of the $t-$method. For the $z-$test, replace t_{crit} with a critical value z_{crit} from a standard normal table. The $z-$critical value can be obtained from the $t-$table using the $df = \infty$ row. The z-test does not require normality, but does require that the sample size n is large. In practice, most researchers just use the $t-$test whether or not n is large – it makes little difference since z and t are very close when n is large.

Example: Age at First Transplant

The ages (in years) at first transplant for a sample of 11 heart transplant patients are as follows: 54 42 51 54 49 56 33 58 54 64 49. The summaries for these data are: $n = 11, \overline{Y} = 51.27$, and $s = 8.26$. Test the hypothesis that the mean age at first transplant is 50. Use $\alpha = .05$. Also, find a 95% CI for the mean age at first transplant.

A good (necessary) first step is to define the population parameter in question, and to write down hypotheses symbolically. These steps help to avoid confusion. Let

$$\mu = \text{mean age at time of first transplant for population of patients.}$$

You are interested in testing $H_0 : \mu = 50$ against $H_A : \mu \neq 50$, so $\mu_0 = 50$.

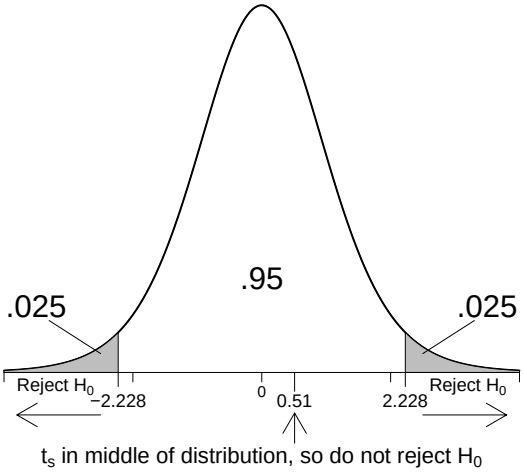
The degrees of freedom are $df = 11 - 1 = 10$. The critical value for a 5% test is $t_{crit} = t_{.025} = 2.228$. (Note $.5\alpha = .5 * .05 = .025$). The same critical value is used with the 95% CI.

Let us first look at the CI calculation. Here $SE = s/\sqrt{n} = 8.26/\sqrt{11} = 2.4904$ and $t_{crit} * SE = 2.228 * 2.4904 = 5.55$. The lower limit on the CI is $51.27 - 5.55 = 45.72$. The upper endpoint is $51.27 + 5.55 = 56.82$. Thus, you are 95% confident that the population mean age at first transplant is between 45.7 and 56.8 years (rounding to 1 decimal place).

For the test,

$$t = \frac{\bar{X} - \mu_0}{SE} = \frac{51.27 - 50}{2.4904} = 0.51.$$

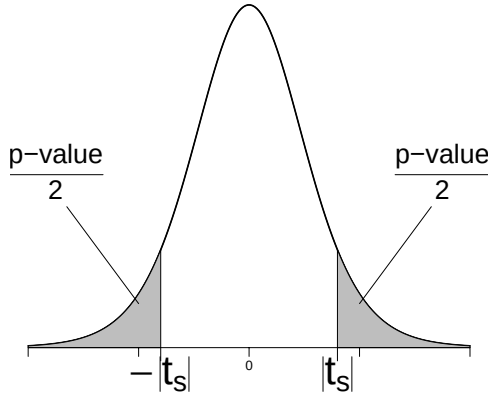
Since $t_{crit} = 2.228$, we do not reject H_0 using a 5% test. Note the placement of t relative to t_{crit} in the picture below. The results of the hypothesis test should not be surprising, since the CI tells you that 50 is a plausible value for the population mean age at transplant. Note: All you can say is that the data *could have* come from a distribution with a mean of 50 – this is not convincing evidence that μ actually *is* 50.



P-values

The **p-value**, or **observed significance level** for the test, provides a measure of plausibility for H_0. Smaller values of the p-value imply that H_0 is less plausible. To compute the p-value for a two-sided test, you

1. Compute the test statistic t_s as above.
2. Evaluate the area under the $t-$probability curve (with $df = n - 1$) outside $\pm|t_s|$.



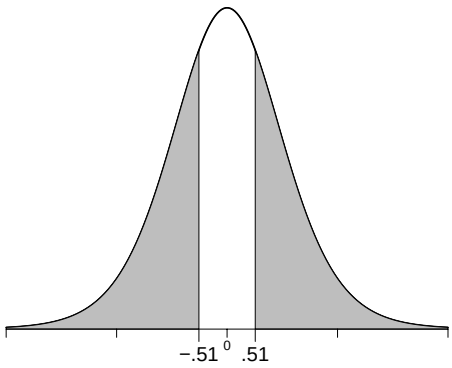
In the picture above, the p-value is the total shaded area, or twice the area in either tail. You can only get bounds on the p-value using SW's $t-$table.

Most, if not all, statistical packages, including **Minitab**, summarize hypothesis tests with a p-value, rather than a decision (i.e reject or not reject at a given α level). You can make a decision to reject or not reject H_0 for a size α test based on the p-value as follows - reject H_0 if the p-value is less than or equal to α. This decision is identical to that obtained following the formal rejection procedure given earlier. The reason for this is that the p-value can be interpreted as the smallest value you can set the size of the test and still reject H_0 given the observed data.

There are a lot of terms to keep straight here. α and t_{crit} are constants we choose (actually, one determines the other so we really only choose one, usually α) to set how rigorous evidence against H_0 needs to be. t_s and the p-value (again, one determines the other) are random variables because they are calculated from the random sample. They are the evidence against H_0.

Example: Age at First Transplant

The picture below is used to calculate the p-value. Using SW's table, all we know is that the p-value is greater than .40. (Why?) The exact p-value for the test (generated with JMP-in) is 0.62. For a 5% test, the p-value indicates that you would not reject H_0 (because $.62 > .05$).



Minitab output for the heart transplant problem is given below. Let us look at the output and find all of the summaries we computed. Also, look at the graphical summaries to assess whether the $t-$test and CI are reasonable here.

COMMENTS:

1. The data were entered into the worksheet as a single column (C1) that was labelled **agetran**.
2. To display the data follow the sequence Data > Display Data, and fill in the dialog box.
3. To get the stem and leaf display, follow the sequence Graph > Stem and Leaf ..., then fill in the dialog box.

4. To get a one-sample t-test and CI follow the sequence: STAT > BASIC STATISTICS > 1-sample t... . In the dialog box, select the column to analyze (C1). For the test, you need to check the box for Perform Hypothesis Test and specify the null mean (i.e. μ_0) and the type of test (by clicking on OPTIONS): not equal gives a two-sided test (default), less than gives a lower one-sided test, and greater than gives an upper one-sided test. The results of the test are reported as a p-value. We have only discussed two-sided tests up to now. Click on the Graphs button and select Boxplot of data.

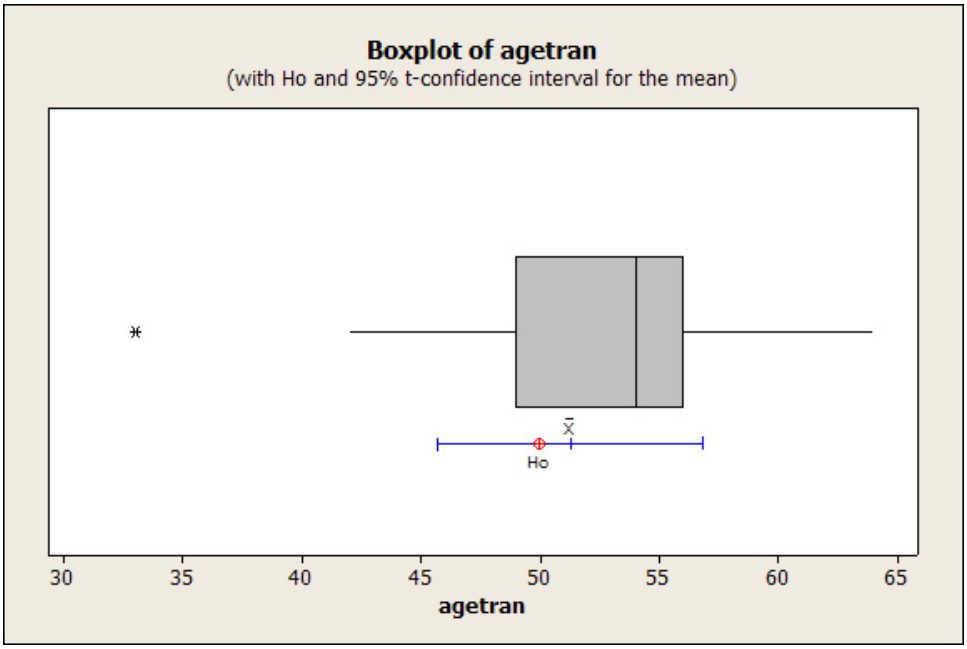
5. I would also follow Stat > Basic Statistics > Display Descriptive Statistics to get a few more summary statistics. The default from the test is a bit limited.

6. If you ask for a test, you will get a corresponding CI. The CI level is set by clicking on Option in the dialog box. If you want a CI but not a test, do not check Perform Hypothesis Test in the main dialog box. A 95% CI is the default.

7. The boxplot will include a CI for the mean.

8. The plots generated with Stat > Basic Statistics > Graphical Summary include a CI for the population mean.

```
Data Display

agetran
   33   42   49   49   51   54   54   54   56   58   64

Stem-and-Leaf Display: agetran

Stem-and-leaf of agetran  N  = 11
Leaf Unit = 1.0

 1   3  3
 1   3
 2   4  2
 4   4  99
(4)  5  1444
 3   5  68
 1   6  4

One-Sample T: agetran

Test of mu = 50 vs not = 50


Variable   N     Mean   StDev  SE Mean         95% CI           T      P
agetran   11  51.2727  8.2594   2.4903  (45.7240, 56.8215)  0.51  0.620

Descriptive Statistics: agetran

Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
agetran   11   0  51.27     2.49   8.26    33.00  49.00   54.00  56.00    64.00
```



Example: Meteorites

One theory of the formation of the solar system states that all solar system meteorites have the same evolutionary history and thus have the same cooling rates. By a delicate analysis based on measurements of phosphide crystal widths and phosphide-nickel content, the cooling rates, in degrees Celsius per million years, were determined for samples taken from meteorites named in the accompanying table after the places they were found.

Suppose that a hypothesis of solar evolution predicted a mean cooling rate of $\mu = .54$ degrees per million year for the Tocopilla meteorite. Do the observed cooling rates support this hypothesis? Test at the 5% level. The boxplot and stem and leaf display (given below) show good symmetry. The assumption of a normal distribution of observations basic to the $t-$test appears to be realistic.

Meteorite	Cooling rates
Walker County	0.69 0.23 0.10 0.03 0.56 0.10 0.01 0.02 0.04 0.22
Uwet	0.21 0.25 0.16 0.23 0.47 1.20 0.29 1.10 0.16
Tocopilla	5.60 2.70 6.20 2.90 1.50 4.00 4.30 3.00 3.60 2.40 6.70 3.80

Let

$$\mu = \text{mean cooling rate over all pieces of the Tocopilla meteorite.}$$

To answer the question of interest, we consider the test of $H_0 : \mu = .54$ against $H_A : \mu \neq .54$. I will explain later why these are the natural hypotheses here. Let us go carry out the test, compute the p-value, and calculate a 95% CI for μ. The sample summaries are $n = 12$, $\overline{Y} = 3.892$, $s = 1.583$. The standard error is $SE_{\overline{Y}} = s/\sqrt{n} = 0.457$.

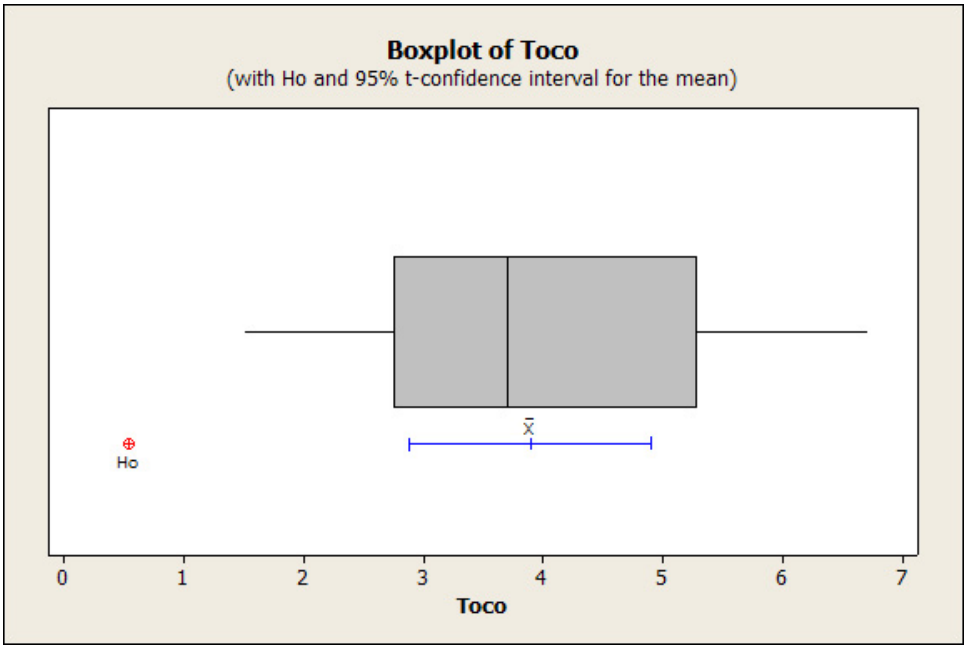
Minitab output for this problem is given below. For a 5% test (i.e. $\alpha = .05$), you would reject H_0 in favor of H_A because the $p - value \leq .05$. The data strongly suggest that $\mu \neq .54$. The 95% CI

says that you are 95% confident that the population mean cooling rate for the Tocopilla meteorite is between 2.89 and 4.90 degrees per million years. Note that the CI gives us a means to assess how different μ is from the hypothesized value of .54.

COMMENTS:

1. The data were entered as a single column in the worksheet, and labelled **Toco**.
2. Remember that you need to specify the null value for the mean (i.e. .54) in the 1-sample t dialog box!
3. I generated a boxplot within the 1-sample t dialog box. A 95% CI for the mean cooling rate is superimposed on the plots.

```
Data Display

Toco
   5.6   2.7   6.2   2.9   1.5   4.0   4.3   3.0   3.6   2.4   6.7   3.8

Stem-and-Leaf Display: Toco

Stem-and-leaf of Toco  N  = 12
Leaf Unit = 0.10

 1   1  5
 2   2  4
 4   2  79
 5   3  0
(2)  3  68
 5   4  03
 3   4
 3   5
 3   5  6
 2   6  2
 1   6  7

One-Sample T: Toco

Test of mu = 0.54 vs not = 0.54


Variable   N     Mean    StDev  SE Mean         95% CI           T      P
Toco      12  3.89167  1.58255  0.45684  (2.88616, 4.89717)  7.34  0.000

Descriptive Statistics: Toco

Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median    Q3  Maximum
Toco      12   0  3.892    0.457  1.583    1.500 2.750   3.700  5.275    6.700
```



The Mechanics of Setting up an Hypothesis Test

SW Section 7.10

When setting up a test you should imagine you are the researcher conducting the experiment. In many studies, the researcher wishes to establish that there has been a change from the **status quo**, or that they have developed a method that produces a **change** (possibly in a specified direction) in the typical response. The researcher sets H_0 to be the **status quo** and H_A to be the **research hypothesis** - the claim the researcher wishes to make. In some studies you define the hypotheses so that H_A is the **take action** hypothesis - rejecting H_0 in favor of H_A leads one to take a radical action.

Some perspective on testing is gained by understanding the mechanics behind the tests. An hypothesis test is a decision process in the face of uncertainty. You are given data and asked which of two contradictory claims about a population parameter, say μ, is more reasonable. Two decisions are possible, but whether you make the correct decision depends on the true state of nature which is unknown to you.

Decision	If H_0 true	If H_A true
Reject H_0 in favor of H_A	Type I error	correct decision
Do not Reject [accept] H_0	correct decision	Type II error

For a given problem, only one of these errors is possible. For example, if H_0 is true you can make a Type I error but not a Type II error. Any reasonable decision rule based on the data that tells us when to reject H_0 and when to not reject H_0 will have a certain probability of making a Type I error if H_0 is true, and a corresponding probability of making a Type II error if H_0 is false and H_A is true. For a given decision rule, define

$$\alpha = \text{Prob(Reject } H_0 \text{ given } H_0 \text{ is true)} = \text{Prob(Type I error)}$$

and

$$\beta = \text{Prob(Do not reject } H_0 \text{ when } H_A \text{ true)} = \text{Prob(Type II error)}.$$

The mathematics behind hypothesis tests allows you to prespecify or control α. For a given α, the tests we use (typically) have the smallest possible value of β. Given the researcher can control α, you set up the hypotheses so that committing a Type I error is more serious than committing a Type II error. The magnitude of α, also called the **size** or **level** of the test, should depend on the seriousness of a Type I error in the given problem. The more serious the consequences of a Type I error, the smaller α should be. In practice α is often set to .10, .05, or .01, with $\alpha = .05$ being the scientific standard. By setting α to be a small value, you reject H_0 in favor of H_A only if the data **convincingly indicate** that H_0 is false.

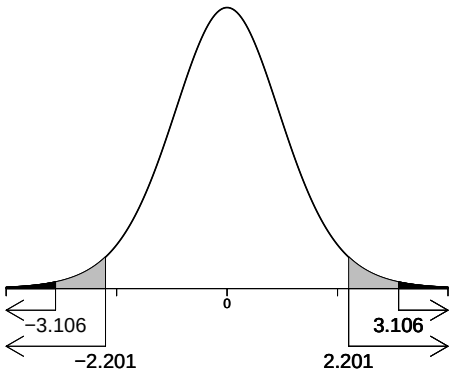
Let us piece together these ideas for the meteorite problem. Evolutionary history predicts $\mu = .54$. A scientist examining the validity of the theory is trying to decide whether $\mu = .54$ or $\mu \neq .54$. Good scientific practice dictates that rejecting another's claim when it is true is more serious than not being able to reject it when it is false. This is consistent with defining $H_0 : \mu = .54$ (the status quo) and $H_A : \mu \neq .54$. To convince yourself, note that the implications of a Type I error would be to claim the evolutionary theory is false when it is true, whereas a Type II error would correspond to not being able to refute the evolutionary theory when it is false. With this setup, the scientist will refute the theory only if the data overwhelmingly suggest that it is false.

The Effect of α on the Rejection Region of a Two-Sided Test

For a size α test, you reject $H_0 : \mu = \mu_0$ if

$$t_s = \frac{\overline{Y} - \mu_0}{SE}$$

satisfies $|t_s| > t_{crit}$.



Rejection Regions for .05 and .01 level tests

The critical value is computed so that the area under the $t-$probability curve (with $df = n-1$) outside $\pm t_{crit}$ is α, with $.5\alpha$ in each tail. Reducing α makes t_{crit} larger. That is, reducing the size of the test makes rejecting H_0 harder because the rejection region is smaller. A pictorial representation is given above for the Tocopilla data, where $\mu_0 = 0.54$, $n = 12$ and $df = 11$. Note that $t_{crit} = 2.201$ and 3.106 for $\alpha = 0.05$ and 0.01, respectively.

The mathematics behind the test presumes that H_0 is true. Given the data, you use

$$t_s = \frac{\bar{Y} - \mu_0}{SE}$$

to measure how far $\overline{Y}$ is from μ_0, relative to the spread in the data given by SE. For t_s to be in the rejection region, $\overline{Y}$ must be significantly above or below μ_0, relative to the spread in the data. To see this, note that rejection rule can be expressed as: **Reject** H_0 if

$$\overline{Y} < \mu_0 - t_{crit}SE \quad \text{or} \quad \overline{Y} > \mu_0 + t_{crit}SE.$$

The rejection rule is sensible because $\overline{Y}$ is our best guess for μ. You would reject $H_0 : \mu = \mu_0$ only if $\overline{Y}$ is so far from μ_0 that you would question the reasonableness of assuming $\mu = \mu_0$. How far $\overline{Y}$ must be from μ_0 before you reject H_0 depends on α (i.e. how willing you are to reject H_0 if it is true), and on the value of SE. For a given sample, reducing α forces $\overline{Y}$ to be further from μ_0 before you reject H_0. For a given value of α and s, increasing n allows smaller differences between $\overline{Y}$ and μ_0 to be **statistically significant** (i.e. lead to rejecting H_0). In problems where small differences between $\overline{Y}$ and μ_0 lead to rejecting H_0, you need to consider whether the observed differences are important.

In essence, the $t-$ distribution provides an objective way to calibrate whether the observed $\overline{Y}$ is typical of what sample means look like when sampling from a normal population where H_0 is true. If all other assumptions are satisfied, and $\overline{Y}$ is inordinately far from μ_0, then our only recourse is to conclude that H_0 must be incorrect.

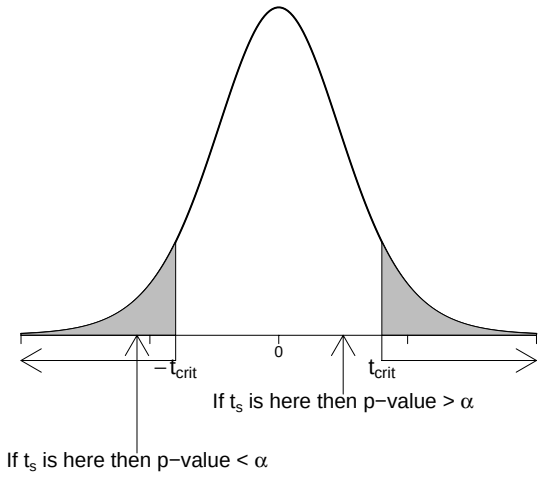
Two-Sided Tests, CI and P-Values

An important relationship among two-sided tests of $H_0 : \mu = \mu_0$, CI, and p-values is that

$$\text{size } \alpha \text{ test rejects } H_0 \Leftrightarrow 100(1-\alpha)\% \text{ CI does not contain } \mu_0 \Leftrightarrow p - value \leq \alpha.$$

$$\text{size } \alpha \text{ test does not reject } H_0 \Leftrightarrow 100(1-\alpha)\% \text{ CI contains } \mu_0 \Leftrightarrow p - value > \alpha.$$

For example, an $\alpha = .05$ test rejects $H_0 \Leftrightarrow 95\%$ CI does not contain $\mu_0 \Leftrightarrow p - value \leq .05$. The picture above illustrates the connection between p-values and rejection regions.



Either a CI or a test can be used to decide the plausibility of the claim that $\mu = \mu_0$. Typically, you use the test to answer the question **is there a difference?** If so, you use the CI to assess **how much of a difference exists**. I believe that scientists place too much emphasis on hypothesis testing. See the discussion below.

Statistical Versus Practical Significance

Suppose in the Tocopilla meteorite example, you rejected $H_0 : \mu = .54$ at the 5% level and found a 95% two-sided CI for μ to be .55 to .58. Although you have sufficient evidence to conclude that the population mean cooling rate μ differs from that suggested by evolutionary theory, the range of plausible values for μ is small and contains only values close to .54. Although you have shown statistical significance here, you need to ask ourselves whether the actual difference between μ and .54 is large enough to be important. The answer to such questions is always problem specific.

Design Issues and Power

An experiment may not be sensitive enough to pick up true differences. For example, in the Tocopilla meteorite example, suppose the true mean cooling rate is $\mu = 1.00$. To have a 50% chance of rejecting $H_0 : \mu = .54$, you would need about $n = 30$ observations. If the true mean is $\mu = .75$, you would need about 140 observations to have a 50% chance of rejecting H_0. In general, the smaller the difference between the true and hypothesized mean (relative to the spread in the population), the more data that is needed to reject H_0. If you have prior information on the expected difference between the true and hypothesized mean, you can design an experiment appropriately by choosing the sample size required to likely reject H_0.

The **power** of a test is the probability of rejecting H_0 when it is false. Equivalently,

power = 1 - Prob(not rejecting H_0 when it is false) = 1- Prob(Type II error).

For a given sample size, the tests I have discussed have maximum power (or smallest probability of a Type II error) among all tests with fixed size α. However, the actual power may be small, so sample size calculations, as briefly highlighted above, are important prior to collecting data. See your local statistician.

One-Sided Tests on μ

There are many studies where a one-sided test is appropriate. The two common scenarios are the **lower one-sided test** $H_0 : \mu = \mu_0$ (or $\mu \geq \mu_0$) versus $H_A : \mu < \mu_0$ and the **upper one-sided test** $H_0 : \mu = \mu_0$ (or $\mu \leq \mu_0$) versus $H_A : \mu > \mu_0$. Regardless of the alternative hypothesis, the tests are based on the t-statistic:

$$t_s = \frac{\overline{Y} - \mu_0}{SE}.$$

For the **upper one-sided test**

1. Compute the critical value t_{crit} such that the area under the t-curve to the **right** of t_{crit} is the desired size α, that is $t_{crit} = t_\alpha$.
2. Reject H_0 if and only if $t_s \geq t_{crit}$.
3. The p-value for the test is the area under the $t-$curve to the **right** of the test statistic t_s.

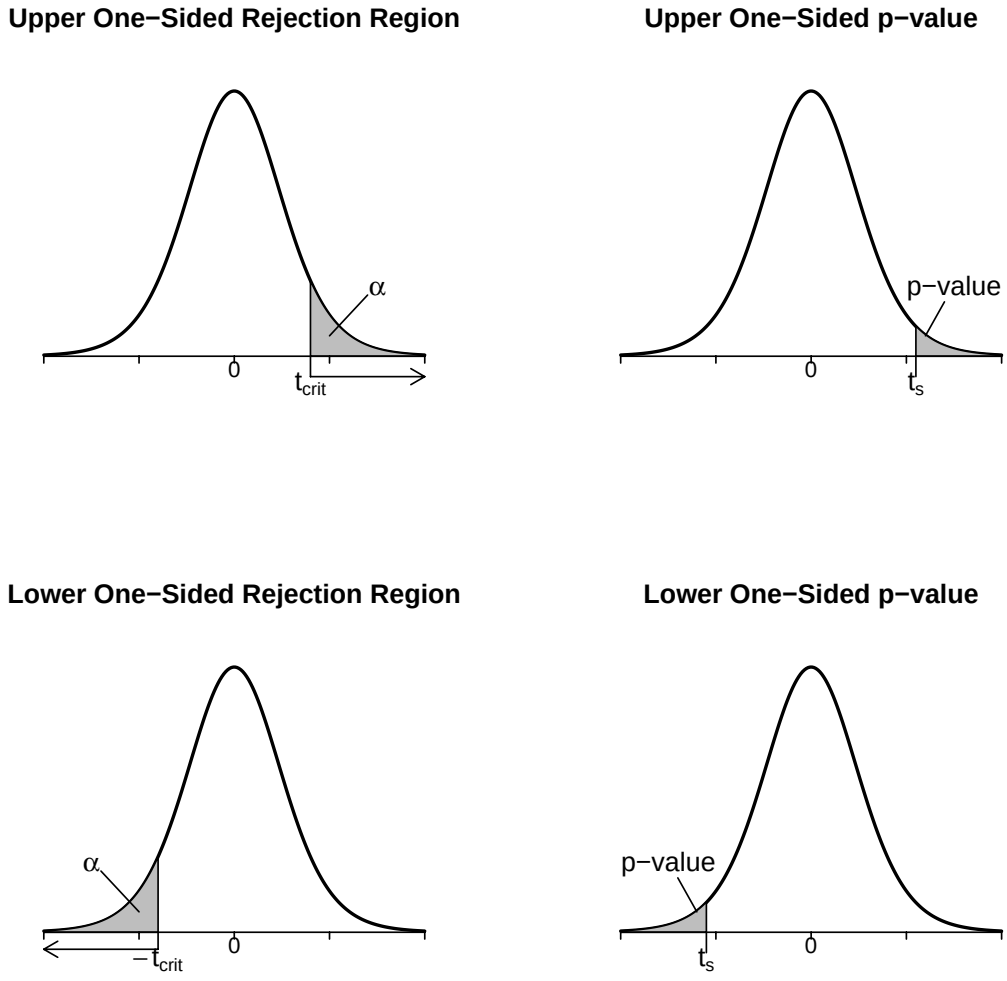
The **upper one-sided test** uses the **upper tail** of the $t-$ distribution for a rejection region. The p-value calculation reflects the form of the rejection region. You will reject H_0 only for large positive values of t_s which require $\overline{Y}$ to be significantly greater than μ_0. Does this make sense?

For the **lower one-sided test**

1. Compute the critical value t_{crit} such that the area under the t-curve to the **right** of t_{crit} is the desired size α, that is $t_{crit} = t_{alpha}$.
2. Reject H_0 if and only if $t_s \leq -t_{crit}$.
3. The p-value for the test is the area under the $t-$curve to the **left** of the test statistic t_s.

The **lower one-sided test** uses the **lower tail** of the $t-$ distribution for a rejection region. The calculation of the rejection region in terms of $-t_{crit}$ is awkward but is necessary for hand calculations because SW only give upper tail percentiles. Note that here you will reject H_0 only for large negative values of t_s which require $\overline{Y}$ to be significantly less than μ_0.

Pictures of the rejection region and the p-value evaluation for a lower one-sided test are given on the next page. As with two-sided tests, the p-value can be used to decide between rejecting or not rejecting H_0 for a test with a given size α.



Example: Weights of canned tomatoes

A consumer group suspects that the average weight of canned tomatoes being produced by a large cannery is less than the advertised weight of 20 ounces. To check their conjecture, the group purchases 14 cans of the canner's tomatoes from various grocery stores. The weights of the contents of the cans to the nearest half ounce were as follows: 20.5, 18.5, 20.0, 19.5, 19.5, 21.0, 17.5, 22.5, 20.0, 19.5, 18.5, 20.0, 18.0, 20.5. Do the data confirm the group's suspicions? Test at the 5% level.

Let $\mu =$ the population mean weight for advertised 20 ounce cans of tomatoes produced by the cannery. The company claims that $\mu = 20$, but the consumer group believes that $\mu < 20$. Hence the consumer group wishes to test $H_0 : \mu = 20$ (or $\mu \geq 20$) against $H_A : \mu < 20$. The consumer group will reject H_0 only if the data overwhelmingly suggest that H_0 is false.

You should assess the normality assumption prior to performing the $t-$test. The stem and leaf display and the boxplot suggest that the distribution might be slightly skewed to the left. However, the skewness is not severe and no outliers are present, so the normality assumption is not unreasonable.

Minitab output for the problem is given below. Let us do a hand calculation using the summarized data. The sample size, mean, and standard deviation are 14, 19.679, and 1.295, respectively. The standard error is $SE_{\overline{Y}} = s/\sqrt{n} = .346$. We see that the sample mean is less than 20. But is it sufficiently less than 20 for us to be willing to publicly refute the canner's claim? Let us carry out the test, first using the rejection region approach, and then by evaluating a p-value.

The test statistic is

$$t_s = \frac{\overline{Y} - \mu_0}{SE_{\overline{Y}}} = \frac{19.679 - 20}{.346} = -.93$$

The critical value for a 5% one-sided test is $t_{.05} = 1.771$, so we reject H_0 if $t_s < -1.771$ (you can get that value from Minitab or from the table). The test statistic is not in the rejection region. Using the t-table, the p-value is between .15 and .20. I will draw a picture to illustrate the critical region and p-value calculation. The exact p-value from Minitab is .185, which exceeds .05.

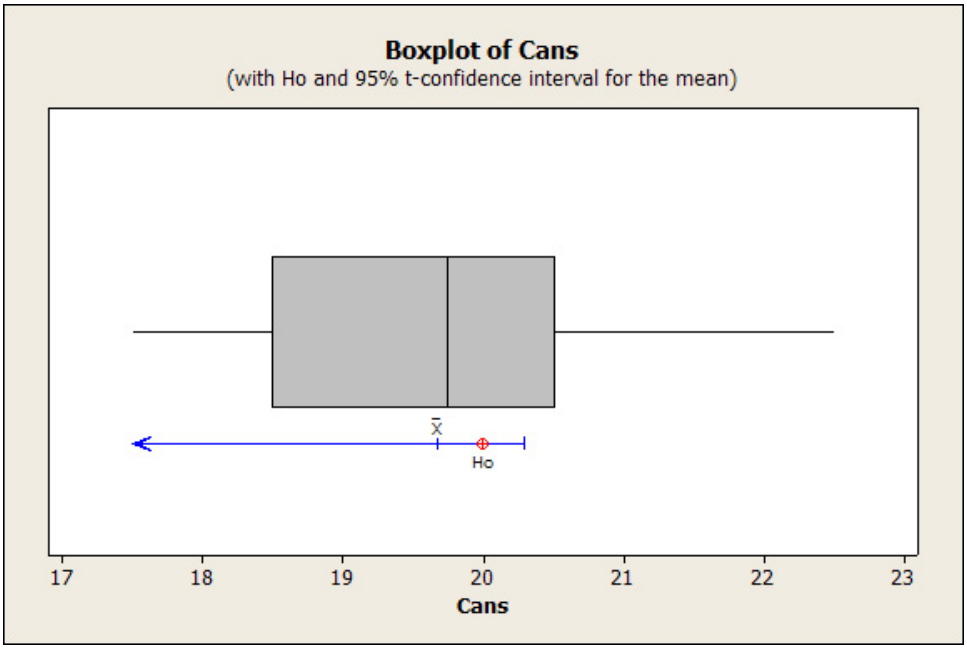
Both approaches lead to the conclusion that we do not have sufficient evidence to reject H_0. That is, we do not have sufficient evidence to question the accuracy of the canner's claim. If you did reject H_0, is there something about how the data were recorded that might make you uncomfortable about your conclusions?

COMMENTS:

1. The data are entered into the first column of the worksheet, which was labelled **cans**.
2. You need to remember to specify the lower one-sided test as an option in the 1 sample t-test dialog box.

```
Descriptive Statistics: Cans

Variable   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Cans      14   0  19.679    0.346  1.295   17.500  18.500  19.750  20.500

Variable  Maximum
Cans       22.500

Stem-and-Leaf Display: Cans

Stem-and-leaf of Cans  N  = 14
Leaf Unit = 0.10


 1  17  5
 2  18  0
 4  18  55
 4  19
 7  19  555
 7  20  000
 4  20  55
 2  21  0
 1  21
 1  22
 1  22  5
```

```
One-Sample T: Cans

Test of mu = 20 vs < 20

                                                    95%
                                                  Upper
Variable   N     Mean   StDev  SE Mean    Bound      T      P
Cans      14  19.6786  1.2951   0.3461  20.2915  -0.93  0.185
```



How should you couple a one-sided test with a CI procedure? For a **lower one-sided test**, you are interested only in an **upper bound** on μ. Similarly, with an **upper one-sided test** you are interested in a **lower bound** on μ. Computing these type of bounds maintains the consistency between tests and CI procedures. The general formulas for lower and upper $100(1-\alpha)\%$ confidence bounds on μ are given by

$$\overline{Y} - t_{crit} SE_{\overline{Y}} \qquad \text{and} \qquad \overline{Y} + t_{crit} SE_{\overline{Y}}$$

respectively, where $t_{crit} = t_\alpha$.

In the cannery problem, to get an upper 95% bound on μ, the critical value is the same as we used for the one-sided 5% test: $t_{.05} = 1.771$. The upper bound on μ is

$$\overline{Y} + t_{.05} SE_{\overline{Y}} = 19.679 + 1.771 * .346 = 19.679 + .613 = 20.292.$$

Thus, you are 95% confident that the population mean weight of the canner's 20oz cans of tomatoes is less than or equal to 20.29. As expected, this interval covers 20.

If you are doing a one-sided test in Minitab, it will generate the correct one-sided bound. That is, a lower one-sided test will generate an upper bound, whereas an upper one-sided test generates

a lower bound. If you only wish to compute a one-sided bound without doing a test, you need to specify the direction of the alternative which gives the type of bound you need. An upper bound was generated by Minitab as part of the test we performed earlier. The result agrees with the hand calculation.

Quite a few packages, including only slightly older versions of Minitab, do not directly compute one-sided bounds so you have to fudge a bit. In the cannery problem, to get an upper 95% bound on μ, you take the upper limit from a 90% two-sided confidence limit on μ. The rationale for this is that with the 90% two-sided CI, μ will fall above the upper limit 5% of the time and fall below the lower limit 5% of the time. Thus, you are 95% confident that μ falls below the upper limit of this interval, which gives us our one-sided bound. Here, you are 95% confident that the population mean weight of the canner's 20oz cans of tomatoes is less than or equal to 20.29, which agrees with our hand calculation.

```
One-Sample T: Cans

Variable   N     Mean   StDev  SE Mean         90% CI
Cans      14  19.6786  1.2951   0.3461  (19.0656, 20.2915)
```

The same logic applies if you want to generalize the one-sided confidence bounds to arbitrary confidence levels and to lower one-sided bounds - always double the error rate of the desired one-sided bound to get the error rate of the required two-sided interval! For example, if you want a lower 99% bound on μ (with a 1% error rate), use the lower limit on the 98% two-sided CI for μ (which has a 2% error rate).

Two-Sided Hypothesis Test for p

Suppose you are interested in whether the population proportion p is equal to a prespecified value, say p_0. This question can be formulated as a two-sided hypothesis test. To carry out the test:

1. Define the null hypothesis $H_0 : p = p_0$ and alternative hypothesis $H_A : p \neq p_0$.
2. Choose the size or significance level of the test, denoted by α.
3. Using the standard normal probability table, find the critical value z_{crit} such that the areas under the normal curve to the left and right of z_{crit} are $1 - .5\alpha$ and $.5\alpha$, respectively. That is, $z_{crit} = z_{.5\alpha}$.
4. Compute the test statistic (often to be labeled z_{obs})

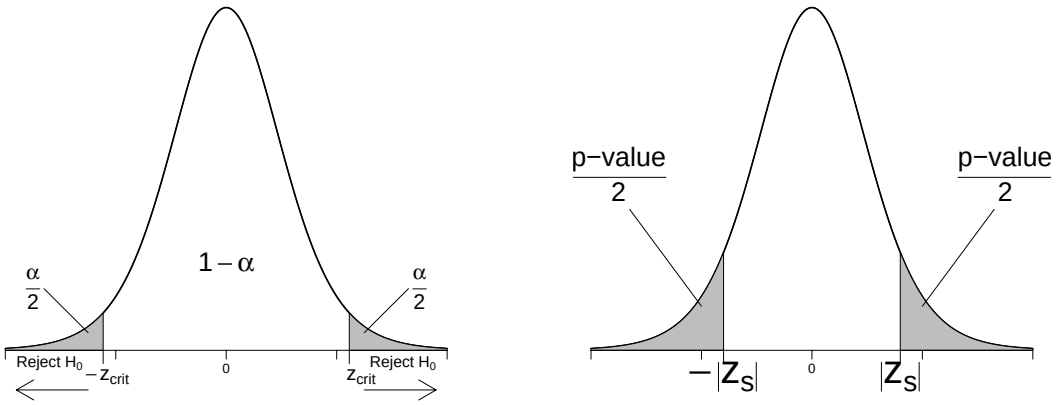
$$z_s = \frac{\hat{p} - p_0}{SE},$$

where the "test standard error" is

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}}.$$

5. Reject H_0 in favor of H_A if $|z_{obs}| \geq z_{crit}$. Otherwise, do not reject H_0.

The rejection rule is easily understood visually. The area under the normal curve outside $\pm z_{crit}$ is the size α of the test. One-half of α is the area in each tail. You reject H_0 in favor of H_A if the test statistic exceeds $\pm z_{crit}$. This occurs when $\hat{p}$ is significantly different from p_0, as measured by the standardized distance z_{obs} between $\hat{p}$ and p_0.



The P-Value for a Two-Sided Test

To compute the p-value (not to be confused with the value of p!) for a two-sided test:

1. Compute the test statistic z_s.

2. Evaluate the area under the normal probability curve outside $\pm z_s$.

Given the picture above with $z_{obs} > 0$, the p-value is the shaded area under the curve, or twice the area in either tail.

Recall that the null hypothesis for a size α test is rejected if and only if the p-value is less than or equal to α.

Example (Emissions data) Each car in the target population (L.A. county) either has been tampered with (a success) or has not been tampered with (a failure). Let p = the proportion of cars in L.A. county with tampered emissions control devices. You want to test $H_0 : p = .15$ against $H_A : p \neq .15$ (here $p_0 = .15$). The critical value for a two-sided test of size $\alpha = .05$ is $z_{crit} = 1.96$.

The data are a sample of $n = 200$ cars. The sample proportion of cars that have been tampered with is $\hat{p} = 21/200 = .105$. The test statistic is

$$z_s = \frac{.105 - .15}{.025} = -1.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{.15 * .85}{200}} = .025.$$

Given that $|z_s| = 1.78 < 1.96$, you have insufficient evidence to reject H_0 at the 5% level. That is, you have insufficient evidence to conclude that the proportion of cars in L.A. county that have been tampered with differs from the statewide proportion.

This decision is reinforced by the p-value calculation. The p-value is the area under the standard normal curve outside ± 1.78. This is about $2 * .0375 = .075$, which exceeds the test size of .05.

REMARK: It is important to recognize that the mechanics of the test on proportions is similar to tests on means, except we use a different test statistic and a different probability table for critical values.

Appropriateness of Test

The z-test is based on a large sample normal approximation, which works better for a given sample size when p_0 is closer to .5. The sample size needed for an accurate approximation increases dramatically the closer p_0 gets to 0 or to 1. Unfortunately, there is no universal agreement as to when the sample size n is "large enough" to apply the test. A simple rule of thumb says that the test is appropriate when $np_0(1 - p_0) \geq 5$.

In the emissions example, $np_0(1 - p_0) = 200 * (.15) * (.85) = 25.5$ exceeds 5, so the normal approximation is appropriate.

Minitab Implementation

This is done precisely as in constructing CIs, covered last week. Follow `Stat > Basic Statistics > 1 Proportion` and enter summarized data. We are using the normal approximation for these calculations. You need to enter p_0 and make the test two-sided under `Options`.

```
Test and CI for One Proportion

Test of p = 0.15 vs p not = 0.15

Sample   X    N  Sample p         95% CI         Z-Value  P-Value
1       21  200  0.105000  (0.062515, 0.147485)    -1.78    0.075
```

My own preference for this particular problem would be to use the exact procedure. What we are doing here is the most common practice, however, and does fit better with procedures we do later. You should confirm that the exact procedure (not using the normal approximation) makes no difference here (because the normal approximation is appropriate).

One-Sided Tests and One-Sided Confidence Bounds

For one-sided tests on proportions, we follow the same general approach adopted with tests on means, except using a different test statistic and table for evaluation of critical values.

For an upper one-sided test $H_0 : p = p_0$ (or $p \leq p_0$) versus $H_A : p > p_0$, you reject H_0 when $\hat{p}$ is significantly greater than p_0, as measured by test statistic

$$z_s = \frac{\hat{p} - p_0}{SE}.$$

In particular, you reject H_0 when $z_s \geq z_{crit}$, where the area under the standard normal curve to the right of z_{crit} is α, the size of the test. That is $z_{crit} = z_\alpha$. The p-value calculation reflects the form of the rejection region, so the p-value for an upper one-sided test is the area under the $z-$curve to the right of z_s. The graphs on page 51 of the notes illustrated all this for the $t-$statistic; the picture here is the same except we now are using a z.

The lower tail of the normal distribution is used for the lower one-sided test $H_0 : p = p_0$ (or $p \geq p_0$) versus $H_A : p < p_0$. Thus, the p-value for this test is the area under the $z-$curve to the left of z_s. Similarly, you reject H_0 when $z_s \leq -z_{crit}$, where z_{crit} is the same critical value used for the upper one-sided test of size α.

Lower and upper one-sided $100(1-\alpha)\%$ confidence bounds for p are

$$\hat{p} - z_{crit}SE \quad \text{and} \quad \hat{p} + z_{crit}SE,$$

respectively, where $z_{crit} = z_\alpha$ is the critical value for a one-sided test of size α and $SE = \sqrt{\hat{p}(1-\hat{p})/n}$ is the "confidence interval" standard error. Recall that upper bounds are used in conjunction with lower one-sided tests and lower bounds are used with upper one-sided tests.

These are large sample tests and confidence bounds, so check whether n is large enough to apply these methods.

Example An article in the April 6, 1983 edition of *The Los Angeles Times* reported on a study of 53 learning impaired youngsters at the Massachusetts General Hospital. The right side of the brain was found to be larger than the left side in 22 of the children. The proportion of the general population with brains having larger right sides is known to be .25. Do the data provide strong evidence for concluding, as the article claims, that the proportion of learning impaired youngsters with brains having larger right sides exceeds the proportion in the general population?

I will answer this question by computing a p-value for a one-sided test. Let p be the population proportion of learning disabled children with brains having larger right sides. I am interested in testing $H_0 : p = .25$ against $H_A : p > .25$ (here $p_0 = .25$).

The proportion of children sampled with brains having larger right sides is $\hat{p} = 22/53 = .415$. The test statistic is

$$z_s = \frac{.415 - .25}{.0595} = 2.78,$$

where the test standard error satisfies

$$SE = \sqrt{\frac{.25 * .75}{53}} = .0595.$$

The p-value for an upper one-sided test is the area under the standard normal curve to the right of 2.78, which is approximately .003. I would reject H_0 in favor of H_A using any of the standard test levels, say .05 or .01. The newspaper's claim is reasonable.

A sensible next step in the analysis would be to compute a lower confidence bound for p. For illustration, consider a 95% bound. The CI standard error is

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.415 * .585}{53}} = .0677.$$

The critical value for a one-sided 5% test is $z_{crit} = 1.645$, so a lower 95% bound on p is $.415 - 1.645 * .0677 = .304$. I am 95% confident that the population proportion of learning disabled children with brains having larger right sides is at least .304. Values of p smaller than .304 are not plausible.

You should verify that the sample size is sufficiently large to use the approximate methods in this example.

Minitab does this one sample procedure very easily, and it makes no real difference if you use the normal approximation or the exact procedure (what does that say about the normal approximation?).

```
Test of p = 0.25 vs p > 0.25


                                95%
                              Lower
Sample   X   N  Sample p     Bound  Z-Value  P-Value
1       22  53  0.415094  0.303766     2.78    0.003

Test and CI for One Proportion

Test of p = 0.25 vs p > 0.25


                                95%
                              Lower    Exact
Sample   X   N  Sample p     Bound  P-Value
1       22  53  0.415094  0.300302    0.006
```

8 Two-Sample Inferences for Means

SW Chapters 7 and 9

Comparing Two Sets of Measurements

Suppose you have collected data on one variable from two (independent) samples and you are interested in "comparing" the samples. What tools are good to use?

Example: Head Breadths

In this analysis, we will compare a physical feature of modern day Englishmen with the corresponding feature of some of their ancient countrymen. The Celts were a vigorous race of people who once populated parts of England. It is not entirely clear whether they simply died out or merged with other people who were the ancestors of those who live in England today. A goal of this study might be to shed some light on possible genetic links between the two groups.

The study is based on the comparison of maximum head breadths (in millimeters) made on unearthed Celtic skulls and on a number of skulls of modern-day Englishmen. The data are given below. We have a sample of 18 Englishmen and an independent sample of 16 Celtic skulls.

Row	ENGLISH	CELTS
1	141	133
2	148	138
3	132	130
4	138	138
5	154	134
6	142	127
7	150	128
8	146	138
9	155	136
10	158	131
11	150	126
12	140	120
13	147	124
14	148	132
15	144	132
16	150	125
17	149	
18	145	

What features of these data would we likely be interested in comparing? The centers of the distributions, the spreads within each distribution, the distributional shapes, etc.

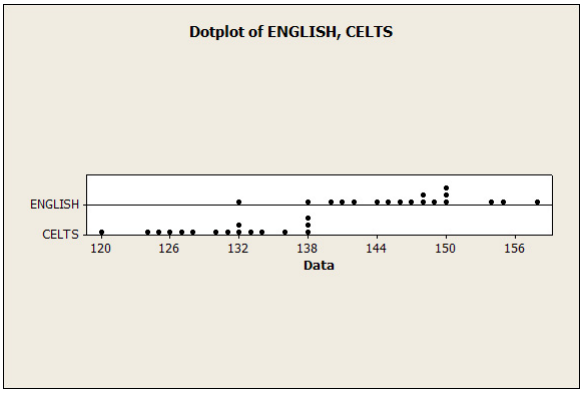
These data can be analyzed in Minitab as either STACKED data (1 column containing both samples, with a separate column of labels or **subscripts** to distinguish the samples) or UNSTACKED (2 columns, 1 for each sample). The form of subsequent Minitab commands will depend on which data mode is used. It is often more natural to enter UNSTACKED data, but with large data bases STACKED data is the norm (for reasons that I will explain verbally). It is easy to create STACKED data from UNSTACKED data and vice-versa. Graphical comparisons usually require the plots for the two groups to have the same scale, which is easiest to control when the data are STACKED.

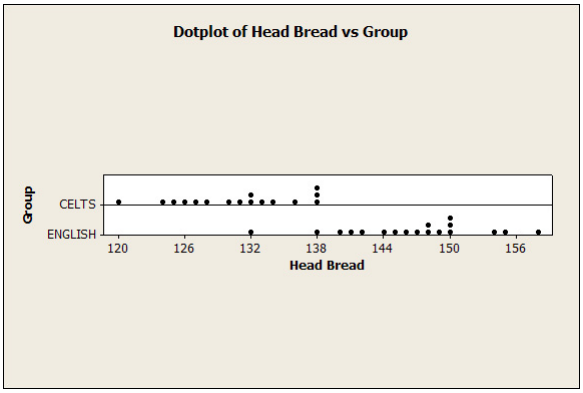
The head breadth data was entered as two separate columns, c1 and c2 (i.e. UNSTACKED). To STACK the data, follow: Data > Stack > Columns. In the dialog box, specify that you wish to stack the English and Celt columns, putting the results in c3, and storing the subscripts in c4. The output below shows the data in the worksheet after stacking the two columns.

```
Data Display

                        Head
Row  ENGLISH  CELTS     Bread  Group
  1      141    133       141  ENGLISH
  2      148    138       148  ENGLISH
  3      132    130       132  ENGLISH
  4      138    138       138  ENGLISH
  5      154    134       154  ENGLISH
  6      142    127       142  ENGLISH
  7      150    128       150  ENGLISH
  8      146    138       146  ENGLISH
  9      155    136       155  ENGLISH
 10      158    131       158  ENGLISH
 11      150    126       150  ENGLISH
 12      140    120       140  ENGLISH
 13      147    124       147  ENGLISH
 14      148    132       148  ENGLISH
 15      144    132       144  ENGLISH
 16      150    125       150  ENGLISH
 17      149              149  ENGLISH
 18      145              145  ENGLISH
 19                       133  CELTS
 20                       138  CELTS
 21                       130  CELTS
 22                       138  CELTS
 23                       134  CELTS
 24                       127  CELTS
 25                       128  CELTS
 26                       138  CELTS
 27                       136  CELTS
 28                       131  CELTS
 29                       126  CELTS
 30                       120  CELTS
 31                       124  CELTS
 32                       132  CELTS
 33                       132  CELTS
 34                       125  CELTS
```

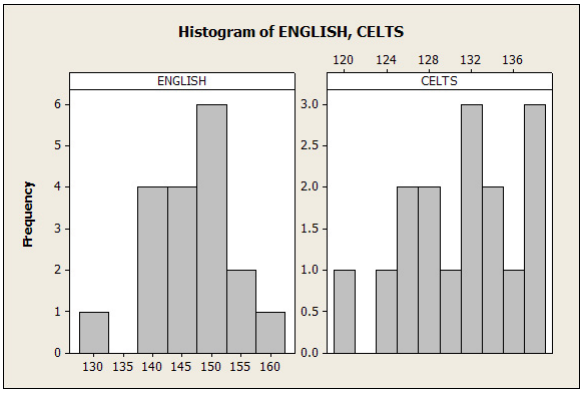
Plotting head breadth data:

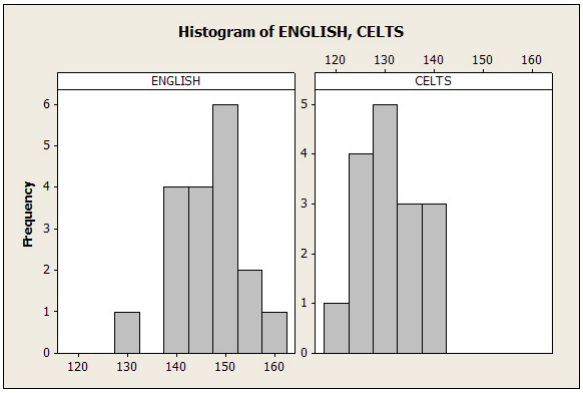
1. A dotplot with the same scale for both samples is obtained from the UNSTACKED data by selecting Multiple Y's with the Simple option, and then choosing C1 and C2 to plot. For the STACKED data, choose One Y With Groups, select c3 as the plotting variable and c4 as the Categorical variable for grouping. There are minor differences in the display generated – I prefer the Stacked data form. In the following, the Unstacked form is on the left, the stacked form on the right.



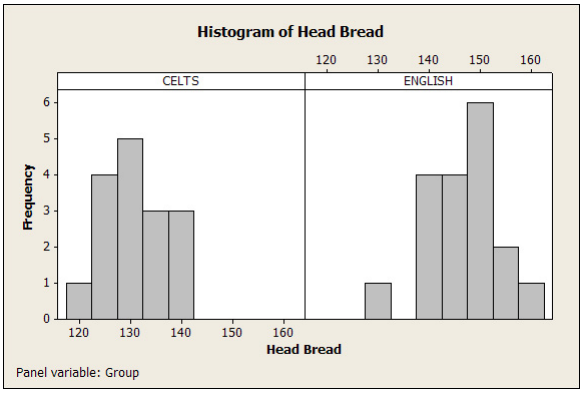


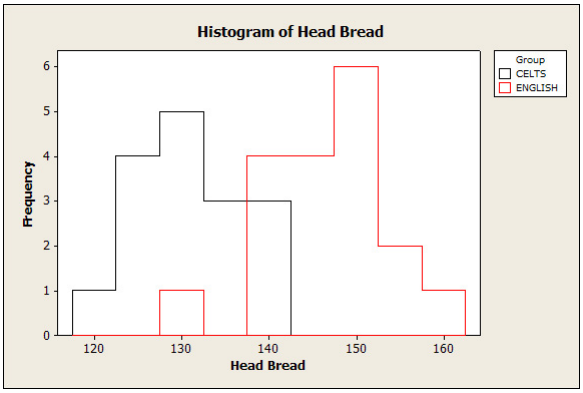
2. Histograms are hard to compare unless you make the scale and actual bins the same for both. Click on *Multiple Graphs* and check *In separate panels of the same graph.* That puts the two graphs next to each other. The left graph below is the unstacked form with only that option. Next check *Same X, including same bins* so you have some basis of comparison. The right graph below uses that option. Why is that one clearly preferable?





The stacked form is more straightforward (left graph below). Click on Multiple Graphs and define a By Variable. The Histogram With Outline and Groups is an interesting variant (right graph below).





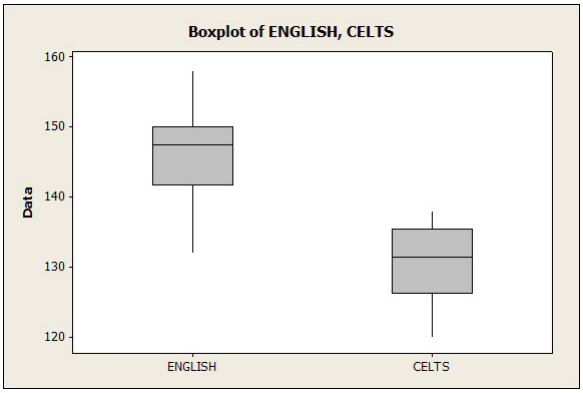
3. Stem-and-leaf displays in unstacked data can be pretty useless. The stems are not forced to match (just like with histograms). It is pretty hard to make quick comparisons with the following:

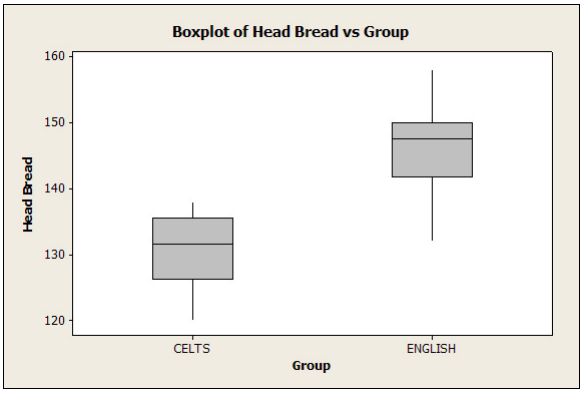
```
Stem-and-Leaf Display: ENGLISH, CELTS

Stem-and-leaf of ENGLISH  N  = 18
Leaf Unit = 1.0


 1   13  2
 2   13  8
 6   14  0124
(6)  14  567889
 6   15  0004
 2   15  58


Stem-and-leaf of CELTS  N  = 16
Leaf Unit = 1.0


 1  12  0
 1  12
 3  12  45
 5  12  67
 6  12  8
 8  13  01
 8  13  223
 5  13  4
 4  13  6
 3  13  888
```

Unfortunately, Minitab seems to be using an old routine for stem-and-leaf plots, and you cannot use stacked data with the Group variable we created. Minitab is wanting a numeric group variable in this case (their older routines always required numeric). Follow Data > Code > Text to Numeric in order to create a new variable in C5 with 1 for ENGLISH and 2 for CELTS. Now the stems at least match up:

```
Stem-and-Leaf Display: Head Bread

Stem-and-leaf of Head Bread  C5 = 1    N  = 18
Leaf Unit = 1.0


 1   13  2
 2   13  8
 6   14  0124
(6)  14  567889
 6   15  0004
 2   15  58


Stem-and-leaf of Head Bread  C5 = 2    N  = 16
Leaf Unit = 1.0


 2   12  04
 6   12  5678
(6)  13  012234
 4   13  6888
```

4. For boxplots, either Unstacked (Multiple Y's) or Stacked (One Y with Groups) works well. Again, I prefer the default from the stacked form, but it really doesn't matter much. Which is which below?





Many of the data summaries will work on either Unstacked or Stacked data. For the head breadth data, **descriptive statistics** output is given below, obtained from both the Stacked data (specifying data in c3 with c4 as a "by variable") and the Unstacked data (specifying data in separate columns c1 and c2).

```
Descriptive Statistics: ENGLISH, CELTS <<<<<<---------Unstacked

Variable   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
ENGLISH   18   0  146.50     1.50   6.38   132.00  141.75  147.50  150.00
CELTS     16   0  130.75     1.36   5.43   120.00  126.25  131.50  135.50

Variable  Maximum
ENGLISH    158.00
CELTS      138.00

Descriptive Statistics: Head Bread  <<<<<<---------Stacked

Variable    Group     N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median
Head Bread  CELTS    16   0  130.75     1.36   5.43   120.00  126.25  131.50
            ENGLISH  18   0  146.50     1.50   6.38   132.00  141.75  147.50

Variable    Group        Q3  Maximum
Head Bread  CELTS    135.50   138.00
            ENGLISH  150.00   158.00
```

Salient Features to Notice

The stem and leaf displays and boxplots indicate that the English and Celt samples are slightly skewed to the left. There are no outliers in either sample. It is not unreasonable to operationally assume that the population frequency curves (i.e. the histograms for the populations from which the samples were selected) for the English and Celtic head breadths are normal.

The sample means and medians are close to each other in each sample, which is not surprising given the near symmetry and the lack of outliers.

The data suggest that the typical modern English head breadth is greater than that for Celts. The data sets have comparable spreads, as measured by either the standard deviation or the IQR (you need to calculate IQR or ask for it in the above summaries).

Two-Sample Methods: Paired Versus Independent Samples

Suppose you have two populations of interest, say populations 1 and 2, and you are interested in comparing their (unknown) population means, μ_1 and μ_2. Inferences on the unknown population means are based on samples from each population. In practice, most problems fall into one of two categories.

1. **Independent samples**, where the sample taken from population 1 has no effect on which observations are selected from population 2, and vice versa. (SW Chapter 7)
2. **Paired** or dependent samples, where experimental units are paired based on factors related or unrelated to the variable measured. (SW Chapter 9)

The distinction between paired and independent samples is best mastered through a series of examples.

Example The English and Celt head breadth samples are independent

Example Suppose you are interested in whether the $CaCO_3$ (calcium carbonate) level in the Atrisco well field, which is the water source for Albuquerque, is changing over time. To answer this question, the $CaCO_3$ level was recorded at each of 15 wells at two time points. These data are paired. The two samples are the Times 1 and 2 observations.

Example To compare state incomes, a random sample of New Mexico households was selected, and an independent sample of Arizona households was obtained. It is reasonable to assume independent samples.

Example Suppose you are interested in whether the husband or wife is typically the heavier smoker among couples where both adults smoke. Data are collected on households. You measure the average number of cigarettes smoked by each husband and wife within the sample of households. These data are paired, i.e. you have selected husband wife pairs as the basis for the samples. It is reasonable to believe that the responses within a pair are related, or correlated.

Although the focus here will be on comparing population means, you should recognize that in paired samples you may also be interested, as in the problems above, in how observations compare within a pair. These goals need not agree, depending on the questions of interest. Note that with paired data, the sample sizes are equal, and equal to the number of pairs.

Two Independent Samples: CI and Test Using Pooled Variance

These methods assume that the populations have normal frequency curves, with equal population standard deviations, i.e. $\sigma_1 = \sigma_2$. Let $(n_1,\ \overline{Y}_1,\ s_1)$ and $(n_2,\ \overline{Y}_2,\ s_2)$ be the sample sizes, means and standard deviations from the two samples.

The standard CI for $\mu_1 - \mu_2$ is given by

$$\begin{aligned} Lower &= (\overline{Y}_1 - \overline{Y}_2) - t_{crit} SE_{\overline{Y}_1 - \overline{Y}_2} \\ Upper &= (\overline{Y}_1 - \overline{Y}_2) + t_{crit} SE_{\overline{Y}_1 - \overline{Y}_2} \end{aligned}$$

The t-statistic for testing $H_0 : \mu_1 - \mu_2 = 0 \quad (\mu_1 = \mu_2)$ against $H_A : \mu_1 - \mu_2 \neq 0 \quad (\mu_1 \neq \mu_2)$ is given by

$$t_s = \frac{\overline{Y}_1 - \overline{Y}_2}{SE_{\overline{Y}_1 - \overline{Y}_2}}.$$

The standard error of $\overline{Y}_1 - \overline{Y}_2$ used in both the CI and the test is given by

$$SE_{\overline{Y}_1 - \overline{Y}_2} = s_{pooled} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Here the **pooled variance estimator**,

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

is our best estimate of the common population variance. The pooled estimator of variance is a weighted average of the two sample variances, with more weight given to the larger sample. If $n_1 = n_2$ then s^2_{pooled} is the average of s_1^2 and s_2^2.

The critical value t_{crit} for CI and tests is obtained in usual way from a t-table with $df = n_1 + n_2 - 2$. For the test, follow the one-sample procedure, with the new t_s and t_{crit}.

The pooled CI and tests are sensitive to the normality and equal standard deviation assumptions. The observed data can be used to assess the reasonableness of these assumptions. You should look at boxplots and stem-and-leaf displays to assess normality, and should check whether $s_1 \approx s_2$ to assess the assumption $\sigma_1 = \sigma_2$. Formal tests of these assumptions will be discussed later.

Satterthwaite's Method

Satterthwaite's method assumes normality, but does not require equal population standard deviations. Satterthwaite's procedures are somewhat conservative, and adjust the SE and df to account for unequal population variances. Satterthwaite's method uses the same CI and test statistic formula, with a modified standard error:

$$SE_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

and degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Note that $df = n_1 + n_2 - 2$ when $n_1 = n_2$ and $s_1 = s_2$. The Satterthwaite and pooled variance procedures usually give similar results when $s_1 \approx s_2$.

SW use Satterthwaite's method for CI and tests, and only briefly touch upon the use of the pooled procedures. The *df* formula for Satterthwaite's method is fairly complex, so SW propose a conservative *df* formula that uses the minimum of $n_1 - 1$ and $n_2 - 1$ instead.

Examples: SW examples 7.7 and 7.8 pages 229-230.

Minitab does the pooled and Satterthwaite analyses, either on stacked or unstacked data. Follow the steps STAT > BASIC STATISTICS > 2 sample t. In the dialog box, specify the data to be analyzed, choose a CI level, and check if you wish to assume equal variances. The output will contain a p-value for a two-sided tests of equal population means and a CI for the difference in population means. If you check the box for assuming equal variances you will get the pooled method, otherwise the output is for Satterthwaite's method.

An important point to note: You can request individual values plots and side-by-side boxplots as an option in the dialog box - and the data need not be stacked.

Example: Head Breadths

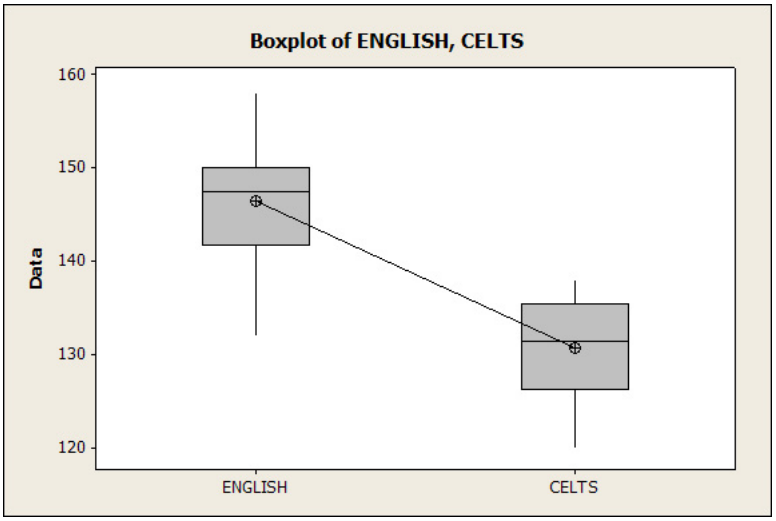
The English and Celts are independent samples. We looked at boxplots and stem and leaf displays, which suggested that the normality assumption for the t-test is reasonable. The Minitab output below shows the English and Celt sample standard deviations are fairly close, so the pooled and Satterthwaite results should be comparable. The pooled analysis is preferable here, but either is appropriate.

The form of the output will tell you which sample corresponds to population 1 and which corresponds to population 2. This should be clear from the dialog box if you use the UNSTACKED data, as I did. Here the CI tells us about the difference between the English and Celt population means, so I need to define μ_1 = population mean head breadths for all Englishmen and μ_2 = population mean head breadths for Celts.

```
Two-Sample T-Test and CI: ENGLISH, CELTS

Two-sample T for ENGLISH vs CELTS    <<<--------- Pooled

           N    Mean  StDev  SE Mean
ENGLISH   18  146.50   6.38      1.5
CELTS     16  130.75   5.43      1.4


Difference = mu (ENGLISH) - mu (CELTS)
Estimate for difference:  15.7500
95% CI for difference:  (11.5809, 19.9191)
T-Test of difference = 0 (vs not =): T-Value = 7.70  P-Value = 0.000  DF = 32
Both use Pooled StDev = 5.9569


Two-Sample T-Test and CI: ENGLISH, CELTS

Two-sample T for ENGLISH vs CELTS   <<<--------- Satterthwaite

           N    Mean  StDev  SE Mean
ENGLISH   18  146.50   6.38      1.5
CELTS     16  130.75   5.43      1.4
```

```
Difference = mu (ENGLISH) - mu (CELTS)
Estimate for difference:  15.7500
95% CI for difference:  (11.6158, 19.8842)
T-Test of difference = 0 (vs not =): T-Value = 7.77  P-Value = 0.000  DF = 31
```

The boxplot, asked for optionally, is nice here – it show means, and connects them to emphasize the analysis being done.



Remarks: The $T =$ entry on the T-TEST line is t_{obs}, whereas $P =$ is the p-value.

The pooled analysis strongly suggests that $H_0 : \mu_1 - \mu_2 = 0$ is false, given the 2-sided p-value of .0000. We are 95% confident that $\mu_1 - \mu_2$ is between 11.6 and 19.9 mm. That is, we are 95% confident that the population mean head breadth for Englishmen (μ_1) exceeds the population mean head breadth for Celts (μ_2) by between 11.6 and 19.9 mm.

The CI interpretation is made easier by recognizing that we concluded the population means are different, so the direction of difference must be consistent with that seen in the observed data, where the sample mean head breadth for Englishmen exceeds that for the Celts. Thus, the limits on the CI for $\mu_1 - \mu_2$ tells us how much larger the mean is for the English population (i.e. between 11.6 and 19.9 mm).

The interpretation of the analysis is always simplified if you specify the first sample in the dialog box (for an UNSTACKED analysis) to be the sample with the larger mean. Why?

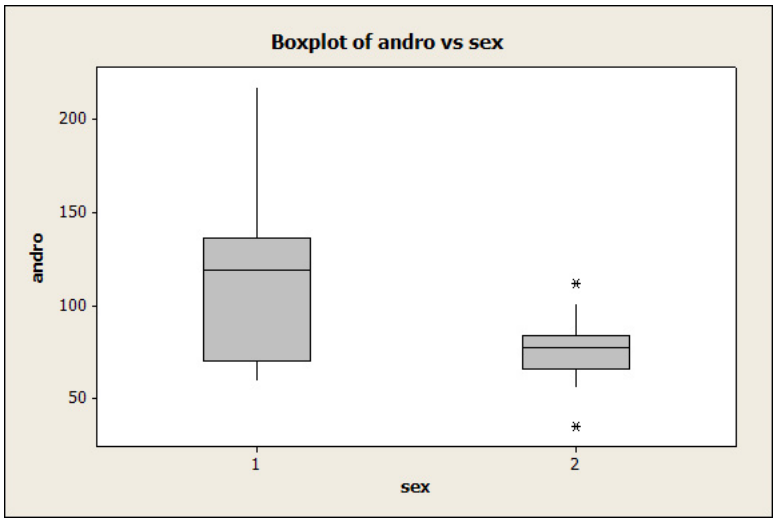
Example: Androstenedione Levels in Diabetics

The data consist of independent samples of diabetic men and women. For each individual, the scientist recorded their androstenedione level (a hormone - Mark McGwire's favorite dietary supplement). Let $\mu_1 =$ mean androstenedione level for the population of diabetic men, and $\mu_2 =$ mean androstenedione level for the population of diabetic women. We are interested in comparing the population means given the observed data.

The raw data and Minitab output is given below. The boxplots suggest that the distributions are reasonably symmetric. However, the normality assumption for the women is unreasonable due

to the presence of outliers. The equal population standard deviation assumption also appears unreasonable. The sample standard deviation for men is noticeably larger than the women's standard deviation, even with outliers in the women's sample.

I am more comfortable with the Satterthwaite analysis here than the pooled variance analysis. However, I would interpret all results cautiously, given the unreasonableness of the normality assumption.



Data Display

Row	men	women	andro	sex
1	217	84	217	1
2	123	87	123	1
3	80	77	80	1
4	140	84	140	1
5	115	73	115	1
6	135	66	135	1
7	59	70	59	1
8	126	35	126	1
9	70	77	70	1
10	63	73	63	1
11	147	56	147	1
12	122	112	122	1
13	108	56	108	1
14	70	84	70	1
15		80	84	2
16		101	87	2
17		66	77	2
18		84	84	2
19			73	2
20			66	2
21			70	2
22			35	2
23			77	2
24			73	2
25			56	2
26			112	2
27			56	2
28			84	2
29			80	2

```
 30                  101     2
 31                   66     2
 32                   84     2


Descriptive Statistics: men, women

Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
men       14   0  112.5     11.4   42.8     59.0   70.0   118.5  136.3    217.0
women     18   0  75.83     4.06  17.24    35.00  66.00   77.00  84.00   112.00


Stem-and-Leaf Display: andro

Stem-and-leaf of andro  sex = 1    N  = 14
Leaf Unit = 10


1  0  5
4  0  677
5  0  8
7  1  01
7  1  2223
3  1  44
1  1
1  1
1  2  1


Stem-and-leaf of andro  sex = 2    N  = 18
Leaf Unit = 10


 1   0  3
 3   0  55
(7)  0  6677777
 8   0  888888
 2   1  01
```

Using the Satterthwaite test, the data strongly suggest that the population mean androstenedione levels are different. In particular, the Welsh (Satterthwaite) p-value for testing $H_0 : \mu_1 - \mu_2 = 0$ is .008. The 95% Satterthwaite CI for $\mu_1 - \mu_2$ extends from 11.0 to 62.4, which implies that we are 95% confident that the population mean andro level for diabetic men exceeds that for diabetic women by at least 11.0 but by no more than 62.4.

As a comparison, let us examine the output for the pooled procedure. The p-value for the pooled t-test is .002, whereas the 95% confidence limits are 14.1 and 59.2. That is, we are 95% confident that the population mean andro level for men exceeds that for women by at least 14.1 but by no more than 59.2. These results are qualitatively similar to the Satterthwaite conclusions.

```
Two-Sample T-Test and CI: men, women

Two-sample T for men vs women

        N   Mean  StDev  SE Mean
men    14  112.5   42.8       11 women  18   75.8   17.2      4.1


Difference = mu (men) - mu (women)
Estimate for difference:  36.6667
95% CI for difference:  (10.9577, 62.3756)
T-Test of difference = 0 (vs not =): T-Value = 3.02  P-Value = 0.008  DF = 16
```

```
Two-Sample T-Test and CI: men, women

Two-sample T for men vs women

        N   Mean  StDev  SE Mean
men    14  112.5   42.8       11 women  18   75.8   17.2      4.1


Difference = mu (men) - mu (women) Estimate for difference:  36.6667
95% CI for difference:  (14.1124, 59.2210)
T-Test of difference = 0 (vs not =): T-Value = 3.32  P-Value = 0.002  DF = 30
Both use Pooled StDev = 30.9914
```

One-Sided Tests

SW discuss one-sided tests for two-sample problems, where the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$ but the alternative is directional, either $H_A : \mu_1 - \mu_2 < 0$ (i.e. $\mu_1 < \mu_2$) or $H_A : \mu_1 - \mu_2 > 0$ (i.e. $\mu_1 > \mu_2$). Once you understand the general form of rejection regions and p-values for one-sample tests, the one-sided two-sample tests do not pose any new problems. Use the $t-$ statistic, with the appropriate tail of the $t-$distribution to define critical values and p-values. One-sided two-sample tests are directly implemented in Minitab, by specifying the type of test in the dialog box. One-sided confidence bounds are given with the one-sided tests.

Paired Analysis

With paired data, inferences on $\mu_1 - \mu_2$ are based on the sample of differences within pairs. By taking differences within pairs, two dependent samples are transformed into one sample, which contains the relevant information for inferences on $\mu_d = \mu_1 - \mu_2$. To see this, suppose the observations within a pair are Y_1 and Y_2. Then within each pair, compute the difference $d = Y_1 - Y_2$. If the Y_1 data are from a population with mean μ_1 and the Y_2 data are from a population with mean μ_2, then the d's are a sample from a population with mean $\mu_d = \mu_1 - \mu_2$. Furthermore, if the sample of differences comes from a normal population, then we can use standard one sample techniques to test $\mu_d = 0$ (i.e. $\mu_1 = \mu_2$), and to get a CI for $\mu_d = \mu_1 - \mu_2$.

Let $\bar{d} = \overline{Y}_1 - \overline{Y}_2$ be the sample mean of the differences (which is also the mean difference), and let s_d be the sample standard deviation of the differences. The standard error of $\bar{d}$ is $SE_{\bar{d}} = s_d/\sqrt{n}$, where n is the number of pairs. The paired $t-$test (two-sided) CI for μ_d is given by $\bar{d} \pm t_{crit} SE_{\bar{d}}$. To test $H_0 : \mu_d = 0 \quad (\mu_1 = \mu_2)$ against $H_A : \mu_d \neq 0 \quad (\mu_1 \neq \mu_2)$, use

$$t_s = \frac{\bar{d} - 0}{SE_{\bar{d}}}$$

to compute a p-value as in a two-sided one-sample test. One-sided tests are evaluated in the usual way for one-sample tests on means.

A graphical analysis of paired data focuses on the **sample of differences**, and not on the original samples. In particular, the normality assumption is assessed on the sample of differences.

Minitab Analysis

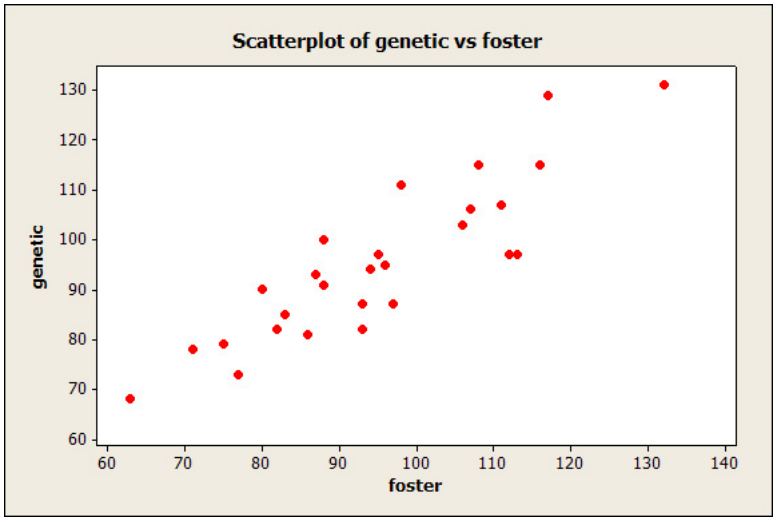
The most natural way to enter paired data is as two columns, one for each treatment group. At this point you can use the Minitab calculator to create a column of differences, and do the usual one-sample graphical and inferential analysis on this column of differences, or you can do the paired analysis directly without this intermediate step.

Example: Paired Analysis of Data on Twins

Burt (1966) presented data on IQ scores for identical twins that were raised apart, one by foster parents and one by the genetic parents. Assuming the data are a random sample of twin pairs, consider comparing the population mean IQs for twins raised at home to those raised by foster parents. Let μ_f=population mean IQ for twin raised by foster parents, and μ_g=population mean IQ for twin raised by genetic parents.

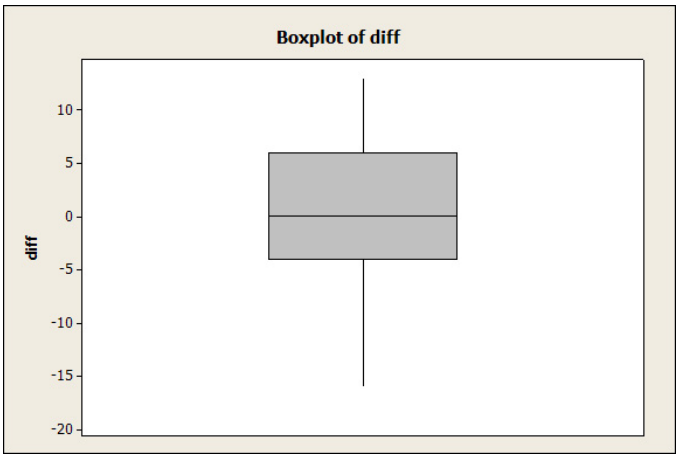
I created the data in the worksheet (c1=foster; c2=genetic), and computed the differences between the IQ scores for the children raised by the genetic and foster parents (c3=diff=genetic-foster). I also made a scatter plot of the genetic versus foster IQ scores.

```
Data Display

 Row  foster  genetic   diff

   1      82       82      0
   2      80       90     10
   3      88       91      3
   4     108      115      7
   5     116      115     -1
   6     117      129     12
   7     132      131     -1
   8      71       78      7
   9      75       79      4
  10      93       82    -11
  11      95       97      2
  12      88      100     12
  13     111      107     -4
  14      63       68      5
  15      77       73     -4
  16      86       81     -5
  17      83       85      2
  18      93       87     -6
  19      97       87    -10
  20      87       93      6
  21      94       94      0
  22      96       95     -1
  23     112       97    -15
  24     113       97    -16
  25     106      103     -3
  26     107      106     -1
  27      98      111     13
```



This plot of IQ scores shows that scores are related within pairs of twins. This is consistent with the need for a paired analysis.

Given the sample of differences, I created a boxplot and a stem and leaf display, neither which showed marked deviation from normality. The boxplot is centered at zero, so one would not be too surprised if the test result is insignificant.




```
Stem-and-Leaf Display: diff

Stem-and-leaf of diff      N   = 27
Leaf Unit = 1.0

   2    -1 65
   4    -1 10
   6    -0 65
  (8)   -0 44311110
  13     0 02234
   8     0 5677
   4     1 0223
```

Given the sample of differences, I generated a one-sample CI and test (i.e. STAT > BASIC STATISTICS > 1-sample t). The hypothesis under test is $\mu_d = \mu_g - \mu_f = 0$. The p-value for this test is large. We do not have sufficient evidence to claim that the population mean IQs for twins raised apart are different. This is consistent with the CI for μ_d given below, which covers zero.

```
One-Sample T: diff

Test of mu = 0 vs not = 0

Variable   N      Mean     StDev   SE Mean           95% CI            T      P
diff      27  0.185185  7.736214  1.488835  (-2.875159, 3.245529)  0.12  0.902
```

Alternatively, I can generate the test and CI directly from the raw data in two columns, following: STAT > BASIC STATISTICS > paired-t, and specifying genetic as the first sample and foster as the second. This gives the following output, which leads to identical conclusions to the earlier analysis. If you take this approach, you can get high quality graphics in addition to the test and CI.

You might ask why I tortured you by doing the first analysis, which required creating and analyzing the sample of differences, when the alternative and equivalent second analysis is so much easier. (A later topic deals with non-parametric analyses of paired data for which the differences must be first computed.)

```
Paired T-Test and CI: genetic, foster

Paired T for genetic - foster

             N      Mean     StDev   SE Mean
genetic     27   95.2963   15.7353    3.0283
foster      27   95.1111   16.0823    3.0950
Difference  27  0.185185  7.736214  1.488835

95% CI for mean difference: (-2.875159, 3.245529)
T-Test of mean difference = 0 (vs not = 0): T-Value = 0.12
P-Value = 0.902
```

Remark: I could have defined the difference to be the foster IQ score minus the genetic IQ score. How would this change the conclusions?

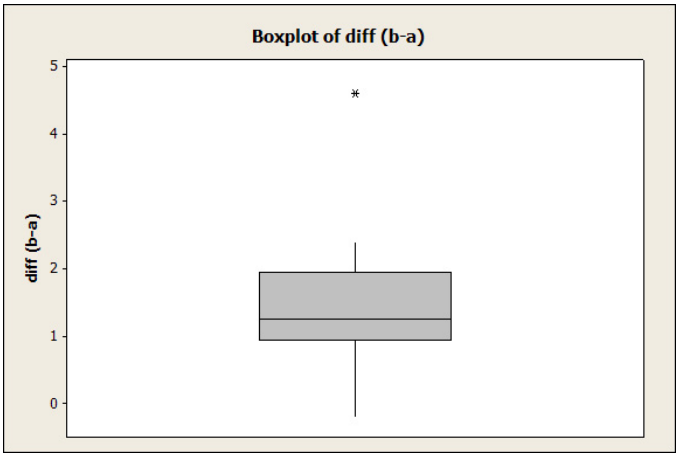
Example: Paired Comparisons of Two Sleep Remedies

The following data give the amount of sleep gained in hours from two sleep remedies, A and B, applied to 10 individuals who have trouble sleeping an adequate amount. Negative values imply sleep loss. In 9 of the 10 individuals, the sleep gain on B exceeded that on A.

Let μ_A = population mean sleep gain (among troubled sleepers) on remedy A, and μ_B = population mean sleep gain (among troubled sleepers) on remedy B. Consider testing $H_0 : \mu_B - \mu_A = 0$ or equivalently $\mu_d = 0$, where $\mu_d = \mu_B - \mu_A$.

The observed distribution of differences between B and A is slightly skewed to the right, with a single outlier in the upper tail. The normality assumption of the standard one-sample t-test and CI are suspect here. I will continue with the analysis.

```
Data Display

                  diff
Row     a     b  (b-a)
  1   0.7   1.9   1.2
  2  -1.6   0.8   2.4
  3  -0.2   1.1   1.3
  4  -1.2   0.1   1.3
  5   0.1  -0.1  -0.2
  6   3.4   4.4   1.0
  7   3.7   5.5   1.8
  8   0.8   1.6   0.8
  9   0.0   4.6   4.6
 10   2.0   3.0   1.0


 Stem-and-Leaf Display: diff (b-a)

 Stem-and-leaf of diff (b-a)  N  = 10
 Leaf Unit = 0.10


 1   -0  2
 2   0   8
(6)  1   002338
 2   2   4
 1   3
 1   4   6

 One-Sample T: diff (b-a)

 Test of mu = 0 vs not = 0

 Variable     N     Mean    StDev  SE Mean         95% CI           T      P
 diff (b-a)  10  1.52000  1.27174  0.40216  (0.61025, 2.42975)  3.78  0.004
```



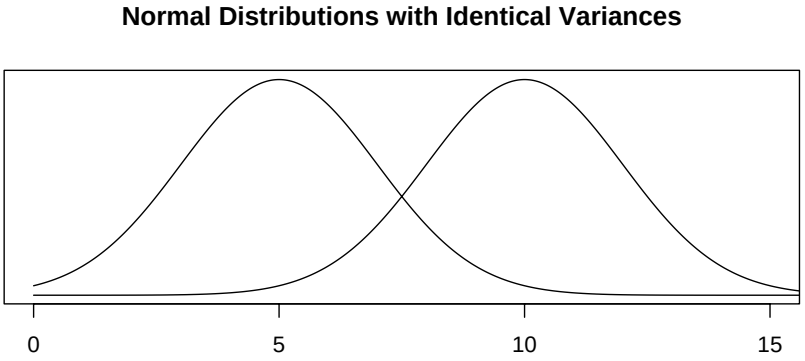
The p-value for testing H_0 is .004. We'd reject H_0 at the 5% or 1% level, and conclude that the population mean sleep gains on the remedies are different. We are 95% confident that μ_B exceeds μ_A by between .61 and 2.43 hours. Again, these results must be reported with caution, because the normality assumption is unreasonable. However, the presence of outliers tends to make the t-test and CI conservative, so we'd expect to find similar conclusions if we used the nonparametric methods discussed later.

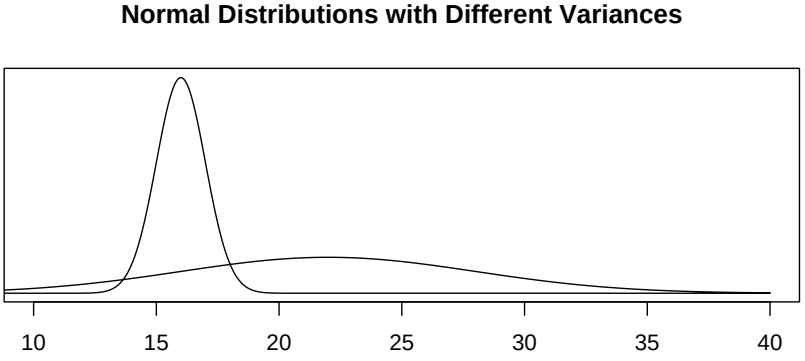
Query: In what order should the remedies be given to the patients?

Should You Compare Means?

The mean is the most common feature on which two distributions are compared. You should not, however, blindly apply the two-sample tests (paired or unpaired) without asking yourself whether the means are the relevant feature to compare. This issue is not a big concern when, as highlighted in the first graph below, the two (normal) populations have equal spreads or standard deviations. In such cases the difference between the two population means is equal to the difference between any fixed percentile for the two distributions, so the mean difference is a natural measure of difference.

Consider instead the hypothetical scenario depicted in the bottom pane below, where the population mean lifetimes using two distinct drugs for a fatal disease are $\mu_1 = 16$ months from time of diagnosis and $\mu_2 = 22$ months from time of diagnosis, respectively. The standard deviations under the two drugs are $\sigma_1 = 1$ and $\sigma_2 = 6$, respectively. The second drug has the higher mean lifetime, but at the expense of greater risk. For example, the first drug gives you a 97.7% chance of living at least 14 months, whereas the second drug only gives you a 90.8% chance of living at least 14 months. Which drug is best? It depends on what is important to you, a higher expected lifetime or a lower risk of dying early.





Nonparametric Procedures

Usually the biggest problems with assumptions of normality occur when we see extreme skewness and/or outliers. The first remedy most statisticians try in such cases is to transform the data using logs or another appropriate transformation to obtain approximate normality on the transformed scale. That often works well but does not handle nearly all problems. Nonparametric procedures are a set of methods designed as alternatives to procedures like t-tests and t-confidence intervals that can be applied even when sampling is not from a normal distribution. I will cover these in a very cursory fashion – this is actually a huge topic on its own.

Minitab implements some of the more popular methods if you follow the path `Stat > Nonparametrics`. The first three options are `1-Sample Sign`, `1-Sample Wilcoxon`, and `Mann-Whitney`. The Sign Test is an alternative to the 1-Sample t-test and makes no real assumption about the shape of the distribution sampled from; it focuses on the population *median* rather than the mean, however. The Wilcoxon Signed Rank test also is an alternative to the 1-Sample t-test; the only assumption about the distribution sampled from is that it is symmetric. The Mann-Whitney test is an alternative to the 2-Sample t-test. It focuses on differences in population medians, and assumes only that the two population distributions have the same general shape.

The Sign Test is pretty inefficient to use for data actually sampled from a normal distribution, but it protects against arbitrarily large outliers. The Wilcoxon and Mann-Whitney tests, if they are appropriate, are very efficient (just as powerful) relative to the t-test, and they also provide great protection against the bad effects of outliers.

Let's look at the Sign Test and Wilcoxon tests for the data on sleep remedies (paired data give rise to 1-Sample methods applied to the differences).

```
One-Sample T: diff (b-a)        <<<<<<<<< COMPARE WITH T

Test of mu = 0 vs not = 0

Variable     N     Mean    StDev  SE Mean         95% CI           T      P
diff (b-a)  10  1.52000  1.27174  0.40216  (0.61025, 2.42975)  3.78  0.004

Sign CI: diff (b-a)              <<<<<<<<< ASK FOR CI AND TEST SEPARATELY

Sign confidence interval for median

                                 Confidence
                          Achieved   Interval
             N  Median  Confidence  Lower  Upper  Position
diff (b-a)  10   1.250      0.8906  1.000  1.800         3
                            0.9500  0.932  2.005       NLI  <<-- USE THIS
                            0.9785  0.800  2.400         2

Sign Test for Median: diff (b-a)

Sign test of median =  0.00000 versus not = 0.00000

             N  Below  Equal  Above       P  Median
diff (b-a)  10      1      0      9  0.0215   1.250
```

```
Wilcoxon Signed Rank CI: diff (b-a)

                                          Confidence
               Estimated    Achieved      Interval
             N     Median  Confidence  Lower  Upper
diff (b-a)  10       1.30        94.7   0.80   2.70

Wilcoxon Signed Rank Test: diff (b-a)

Test of median = 0.000000 versus median not = 0.000000

                   N
                 for   Wilcoxon         Estimated
             N  Test  Statistic      P     Median
diff (b-a)  10    10       54.0  0.008      1.300
```

There is very little difference among these results. The sign test has the shortest CI (but it is for a population median, not mean). For real interpretation, though, your conclusions would not depend on which of these procedures you used. That at least makes you more comfortable if you go ahead and report the results of the t-test.

Let's go back to the androstenedione data set where we saw a problem with outliers. For purposes of illustration, we'll compare the Mann-Whitney to the 2-Sample t-test. Again, there is no real difference in a practical sense. I am uncomfortable with the Mann-Whitney here since the shapes do not really look the same.

```
Two-Sample T-Test and CI: men, women

Two-sample T for men vs women

        N   Mean  StDev  SE Mean
men    14  112.5   42.8       11
women  18   75.8   17.2      4.1


Difference = mu (men) - mu (women)
Estimate for difference:  36.6667
95% CI for difference:  (10.9577, 62.3756)
T-Test of difference = 0 (vs not =): T-Value = 3.02  P-Value = 0.008  DF = 16

Mann-Whitney Test and CI: men, women

        N  Median
men    14  118.50
women  18   77.00


Point estimate for ETA1-ETA2 is 38.00
95.4 Percent CI for ETA1-ETA2 is (3.99,56.01)
W = 293.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0185
The test is significant at 0.0183 (adjusted for ties)
```

Finally, to see that there really can be a difference, let's return to the income data from several lectures ago. The two large outliers pretty well destroy any meaning to the t-interval, but the sign-interval makes a lot of sense for a population median.

```
Data Display
```

```
Income
      7   1110      7      5      8     12      0      5      2      2     46
      7
```

```
One-Sample T: Income

Variable   N     Mean    StDev  SE Mean          95% CI
Income    12  100.917  318.008   91.801  (-101.136, 302.969)
```

```
Sign CI: Income

Sign confidence interval for median

                                 Confidence
                        Achieved    Interval
         N  Median  Confidence  Lower  Upper  Position
Income  12    7.00      0.8540   5.00   8.00         4
                        0.9500   2.79  10.95       NLI
                        0.9614   2.00  12.00         3

**** REMARK: NLI stands for non-linear interpolation
```

SW do discuss the Mann-Whitney test in Section 7.11.

9 One-Way Analysis of Variance

SW Chapter 11 - all sections except 6.

The one-way analysis of variance (**ANOVA**) is a generalization of the two sample $t-$test to $k \geq 2$ groups. Assume that the populations of interest have the following (unknown) population means and standard deviations:

	population 1	population 2	$\cdots$	population k
mean	μ_1	μ_2	$\cdots$	μ_k
std dev	σ_1	σ_2	$\cdots$	σ_k

A usual interest in ANOVA is whether $\mu_1 = \mu_2 = \cdots = \mu_k$. If not, then we wish to know which means differ, and by how much. To answer these questions we select samples from each of the k populations, leading to the following data summary:

	sample 1	sample 2	$\cdots$	sample k
size	n_1	n_2	$\cdots$	n_k
mean	$\overline{Y}_1$	$\overline{Y}_2$	$\cdots$	$\overline{Y}_k$
std dev	s_1	s_2	$\cdots$	s_k

A little more notation is needed for the discussion. Let Y_{ij} denote the j^{th} observation in the i^{th} sample and define the total sample size $n^* = n_1 + n_2 + \cdots + n_k$. Finally, let $\overline{\overline{Y}}$ be the average response over all samples (combined), that is

$$\overline{\overline{Y}} = \frac{\sum_{ij} Y_{ij}}{n^*} = \frac{\sum_i n_i \overline{Y}_i}{n^*}.$$

Note that $\overline{\overline{Y}}$ is *not* the average of the sample means, unless the samples sizes n_i are equal.

An $F-$statistic is used to test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against $H_A :$ not H_0. The assumptions needed for the standard ANOVA $F-$test are analogous to the independent two-sample $t-$test assumptions: (1) Independent random samples from each population. (2) The population frequency curves are normal. (3) The populations have equal standard deviations, $\sigma_1 = \sigma_2 = \cdots = \sigma_k$.

The $F-$test is computed from the ANOVA table, which breaks the spread in the combined data set into two components, or **Sums of Squares** (SS). The **Within SS**, often called the **Residual SS** or the **Error SS**, is the portion of the total spread due to variability *within* samples:

$$\text{SS(Within)} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 = \sum_{ij}(Y_{ij} - \overline{Y}_i)^2.$$

The **Between SS**, often called the Model SS, measures the spread between (actually among!) the sample means

$$\text{SS(Between)} = n_1(\overline{Y}_1 - \overline{\overline{Y}})^2 + n_2(\overline{Y}_2 - \overline{\overline{Y}})^2 + \cdots + n_k(\overline{Y}_k - \overline{\overline{Y}})^2 = \sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2,$$

weighted by the sample sizes. These two SS add to give

$$\text{SS(Total)} = \text{SS(Between)} + \text{SS(Within)} = \sum_{ij}(Y_{ij} - \overline{\overline{Y}})^2.$$

Each SS has its own degrees of freedom (*df*). The *df*(Between) is the number of groups minus one, $k-1$. The *df*(Within) is the total number of observations minus the number of groups: $(n_1-1)+(n_2-1)+\cdots(n_k-1) = n^*-k$. These two *df* add to give *df*(Total) $= (k-1)+(n^*-k) = n^*-1$.

The Sums of Squares and *df* are neatly arranged in a table, called the ANOVA table:

Source	*df*	SS	MS
Between Groups	$k-1$	$\sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2$	
Within Groups	n^*-k	$\sum_i (n_i-1)s_i^2$	
Total	n^*-1	$\sum_{ij}(Y_{ij}-\overline{\overline{Y}})^2$.	

The ANOVA table often gives a **Mean Squares** (MS) column, left blank here. The Mean Square for each source of variation is the corresponding SS divided by its *df*. The Mean Squares can be easily interpreted.

The MS(Within)

$$\frac{(n_1-1)s_1^2+(n_2-1)s_2^2+\cdots+(n_k-1)s_k^2}{n^*-k} = s^2_{pooled}$$

is a weighted average of the sample variances. The MS(Within) is known as the pooled estimator of variance, and estimates the assumed common population variance. If all the sample sizes are equal, the MS(Within) is the average sample variance. The MS(Within) is identical to the **pooled variance estimator** in a two-sample problem when $k=2$.

The MS(Between)

$$\frac{\sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2}{k-1}$$

is a measure of variability among the sample means. This MS is a multiple of the sample variance of $\overline{Y}_1,\ \overline{Y}_2, ..., \overline{Y}_k$ when all the sample sizes are equal.

The MS(Total)

$$\frac{\sum_{ij}(Y_{ij}-\overline{\overline{Y}})^2}{n^*-1}$$

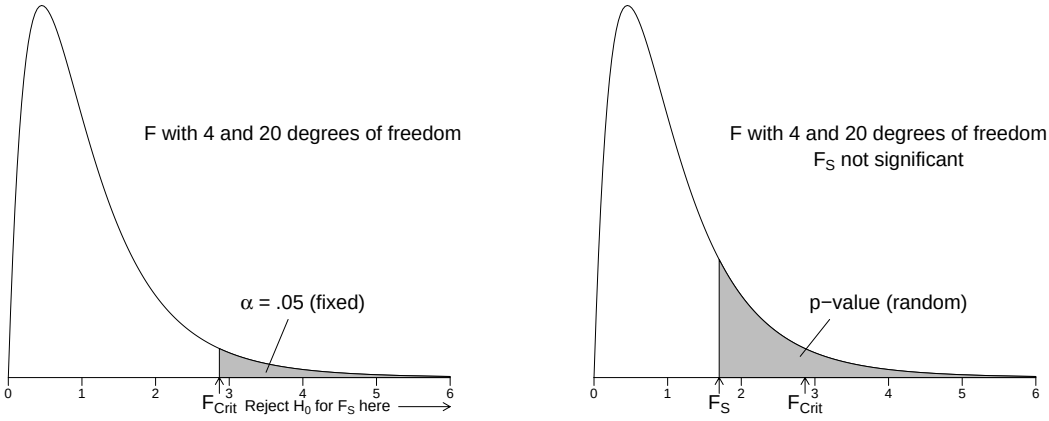
is the variance in the combined data set.

The decision on whether to reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is based on the ratio of the MS(Between) and the MS(Within):

$$F_s = \frac{MS(Between)}{MS(Within)}.$$

Large values of F_s indicate large variability among the sample means $\overline{Y}_1,\ \overline{Y}_2, ..., \overline{Y}_k$ relative to the spread of the data within samples. That is, large values of F_s suggest that H_0 is false.

Formally, for a size α test, reject H_0 if $F_s \geq F_{crit}$, where F_{crit} is the upper-α percentile from an F distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom n^*-k (i.e. the *df* for the numerators and denominators in the $F-$ratio.) The p-value for the test is the area under the $F-$ probability curve to the right of F_s:



For $k = 2$ the ANOVA $F-$ test is equivalent to the pooled two-sample $t-$test.

Minitab summarizes the ANOVA $F-$test with a p-value. The data can be either UNSTACKED or STACKED, but for multiple comparisons discussed later the data must be STACKED. To carry out the analysis, follow the sequence: STAT > ANOVA > ONE-WAY for STACKED data or ONE-WAY (unstacked) for UNSTACKED data. With STACKED data, you need to specify the **response variable** (i.e. the column containing the measurements to be analyzed) and the **factor** (i.e. the column with subscripts that identify the samples) in the dialog box. As with a two-sample analysis, high quality side-by-side boxplots and dotplots can be generated from the ANOVA dialog box. The command line syntax for ANOVA can be obtained from the on-line help, if you are interested.

Example: Comparison of Fats

During cooking, doughnuts absorb fat in various amounts. A scientist wished to learn whether the amount absorbed depends on the type of fat. For each of 4 fats, 6 batches of 24 doughnuts were prepared. The data are grams of fat absorbed per batch (minus 100).

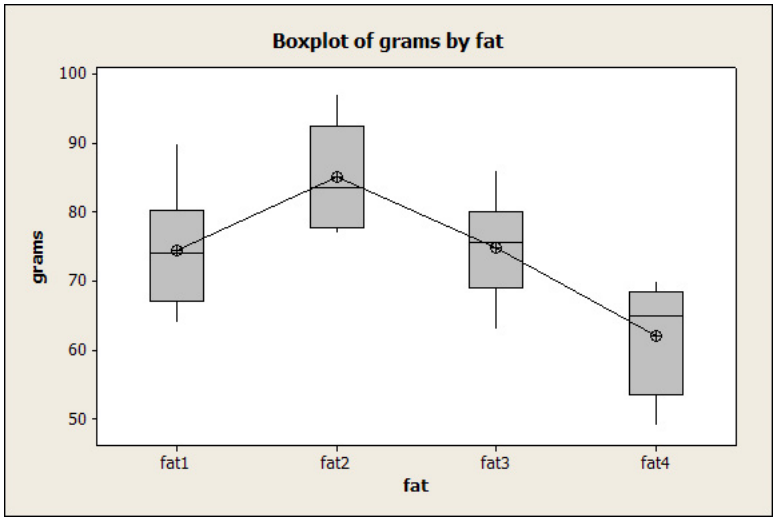
Let

$$\mu_i = \text{pop mean grams of fat } i \text{ absorbed per batch of 24 doughnuts (-100).}$$

The scientist wishes to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_A :$ not H_0. There is no strong evidence against normality here. Furthermore the sample standard deviations (see output below) are close. The standard ANOVA appears to be appropriate here.

The p-value for the $F-$test is .001. The scientist would reject H_0 at any of the usual test levels (i.e. .05 or .01). The data suggest that the population mean absorption rates differ across fats *in some way*. The $F-$test does not say *how* they differ.

```
Row  fat1  fat2  fat3  fat4
  1    64    78    75    55
  2    72    91    86    66
  3    68    97    78    49
  4    77    82    71    64
  5    90    85    63    70
  6    76    77    76    68
```




```
One-way ANOVA: grams versus fat

Source  DF      SS     MS     F      P
fat      3  1595.5  531.8  7.95  0.001
Error   20  1338.3   66.9
Total   23  2933.8

S = 8.180   R-Sq = 54.38%   R-Sq(adj) = 47.54%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean  StDev  -----+---------+---------+---------+----
fat1   6  74.500  9.028              (------*-----)
fat2   6  85.000  7.772                     (------*------)
fat3   6  74.833  7.627              (------*------)
fat4   6  62.000  8.222  (------*------)
                          -----+---------+---------+---------+----
                              60        70        80        90

Pooled StDev = 8.180

Fisher 95% Individual Confidence Intervals     <<<<<<<<<<< WILL EXPLAIN SOON
All Pairwise Comparisons among Levels of fat

Simultaneous confidence level = 80.83%

fat = fat1 subtracted from:

fat     Lower   Center   Upper  ------+---------+---------+---------+---
fat2    0.648   10.500  20.352                  (----*----)
fat3   -9.518    0.333  10.185            (----*----)
fat4  -22.352  -12.500  -2.648       (----*----)
                                 ------+---------+---------+---------+---
                                    -20         0        20        40
```

```
fat = fat2 subtracted from:

fat     Lower   Center    Upper  ------+---------+---------+---------+---
fat3  -20.018  -10.167   -0.315        (----*----)
fat4  -32.852  -23.000  -13.148  (----*---)
                                ------+---------+---------+---------+---
                                   -20         0        20        40

fat = fat3 subtracted from:

fat     Lower   Center   Upper  ------+---------+---------+---------+---
fat4  -22.685  -12.833  -2.982       (----*----)
                                ------+---------+---------+---------+---
                                    -20         0        20        40



Fisher 99.167% Individual Confidence Intervals  <<<<<<<<-- Bonferroni comparisons
All Pairwise Comparisons among Levels of fat

Simultaneous confidence level = 96.16%

fat = fat1 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat2   -3.325   10.500  24.325                (------*------)
fat3  -13.492    0.333  14.159           (------*------)
fat4  -26.325  -12.500   1.325      (------*------)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat2 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat3  -23.992  -10.167   3.659       (------*------)
fat4  -36.825  -23.000  -9.175  (------*-----)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat3 subtracted from:

fat     Lower   Center  Upper  --------+---------+---------+---------+-
fat4  -26.659  -12.833  0.992      (------*-----)
                               --------+---------+---------+---------+-
                                     -20         0        20        40
```

Multiple Comparison Methods: Fisher's Method

The ANOVA $F-$test checks whether all the population means are equal. **Multiple comparisons** are often used as a follow-up to a significant ANOVA $F-$test to determine which population means are different. I will discuss Fisher's, Bonferroni's and Tukey's methods for comparing all pairs of means. These approaches are implemented in **Minitab**.

Fisher's Least significant difference method **(LSD or FSD)** is a two-step process:

1. Carry out the ANOVA $F-$test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ at the α level. If H_0 is not rejected, stop and conclude that there is insufficient evidence to claim differences among population means. If H_0 is rejected, go to step 2.

2. Compare each pair of means using a pooled two sample $t-$test at the α level. Use s_{pooled} from the ANOVA table and $df = df$(Residual).

To see where the name LSD originated, consider the $t-$test of $H_0 : \mu_i = \mu_j$ (i.e. populations i and j have same mean). The $t-$statistic is

$$t_s = \frac{\overline{Y}_i - \overline{Y}_j}{s_{pooled}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}.$$

You reject H_0 if $|t_s| \geq t_{crit}$, or equivalently, if

$$|\overline{Y}_i - \overline{Y}_j| \geq t_{crit} s_{pooled}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

The minimum absolute difference between $\overline{Y}_i$ and $\overline{Y}_j$ needed to reject H_0 is the LSD, the quantity on the right hand side of this inequality. If all the sample sizes are equal $n_1 = n_2 = \cdots = n_k$ then the LSD is the same for each comparison:

$$LSD = t_{crit} s_{pooled}\sqrt{\frac{2}{n_1}},$$

where n_1 is the common sample size.

I will illustrate Fisher's method on the doughnut data, using $\alpha = .05$. At the first step, you reject the hypothesis that the population mean absorptions are equal because $p - value = .001$. At the second step, compare all pairs of fats at the 5% level. Here, $s_{pooled} = 8.18$ and $t_{crit} = 2.086$ for a two-sided test based on 20 *df* (the *df* for Residual SS). Each sample has six observations, so the LSD for each comparison is

$$LSD = 2.086 * 8.18 * \sqrt{\frac{2}{6}} = 9.85.$$

Any two sample means that differ by at least 9.85 in magnitude are **significantly different** at the 5% level.

An easy way to compare all pairs of fats is to order the samples by their sample means. The samples can then be grouped easily, noting that two fats are in the same group if the absolute difference between their sample means is smaller than the LSD.

Fats	Sample Mean
2	85.00
3	74.83
1	74.50
4	62.00

There are six comparisons of two fats. From this table, you can visually assess which sample means differ by at least the LSD=9.85, and which ones do not. For completeness, the table below summarizes each comparison:

Comparison	Absolute difference in means	Exceeds LSD?
Fats 2 and 3	10.17	Yes
2 and 1	10.50	Yes
2 and 4	23.00	Yes
Fats 3 and 1	0.33	No
3 and 4	12.83	Yes
Fats 1 and 4	12.50	Yes

The end product of the multiple comparisons is usually presented as a collection of **groups**, where a group is defined to be a set of populations with sample means that not significantly different from each other. Overlap among groups is common, and occurs when one or more populations appears in two or more groups. Any overlap requires a more careful interpretation of the analysis.

There are three groups for the doughnut data, with no overlap. Fat 2 is in a group by itself, and so is Fat 4. Fats 3 and 1 are in a group together. This information can be summarized by ordering the samples from lowest to highest average, and then connecting the fats in the same group using an underscore:

```
FAT 4    FAT 1    FAT 3     FAT 2
-----    ----------------   -----
```

The results of a multiple comparisons must be interpreted carefully. At the 5% level, you have sufficient evidence to conclude that the population mean absorption for Fat 2 exceeds the other population means, whereas the mean absorption for Fat 4 is smallest. However, there is insufficient evidence to conclude that the population mean absorptions for Fats 1 and 3 differ.

Be Careful with Interpreting Groups in Multiple Comparisons!

To see why you must be careful when interpreting groupings, suppose you obtain two groups in a three sample problem. One group has samples 1 and 3. The other group has samples 3 and 2:

```
1      3      2
----------
       ------------
```

This occurs, for example, when $|\overline{Y}_1 - \overline{Y}_2| \geq LSD$, but both $|\overline{Y}_1 - \overline{Y}_3|$ and $|\overline{Y}_3 - \overline{Y}_2|$ are less than the LSD. There is a tendency to conclude, and please try to avoid this line of attack, that populations 1 and 3 have the same mean, populations 2 and 3 have the same mean, but populations 1 and 2 have different means. This conclusion is illogical. The groupings imply that we have sufficient evidence to conclude that population means 1 and 2 are different, but insufficient evidence to conclude that population mean 3 differs from either of the other population means.

FSD Multiple Comparisons in Minitab

To get Fisher comparisons in **Minitab**, check on COMPARISONS in the one-way ANOVA dialog box. Then choose Fisher, with individual error rate = 5 to get the individual comparisons at the 5% level, as considered above. One slight difficulty relative to our presentation is that **Minitab** summarizes the multiple comparisons in terms of all possible 95% CIs for differences in population means. This output can be used to generate groupings by noting that the individual CIs will cover zero if and only if the corresponding 5% tests of equal means is not significant. Thus a CI for the difference in the population means that covers zero implies that the two populations are in the same group. A summary of the CIs is given below; see the earlier output. Let us see that we can recover the groups from this output.

95% CI for	Limits
$\mu_2 - \mu_1$	0.65 to 20.35
$\mu_3 - \mu_1$	-9.52 to 10.19
$\mu_3 - \mu_1$	-22.35 to -2.65
$\mu_3 - \mu_2$	-20.02 to -0.32
$\mu_4 - \mu_2$	-32.85 to -13.15
$\mu_4 - \mu_3$	-22.69 to -2.98

Discussion of the FSD Method

There are $c = .5k(k-1)$ pairs of means to compare in the second step of the FSD method. Each comparison is done at the α level, where for a generic comparison of the i^{th} and j^{th} populations

$$\alpha = \text{probability of rejecting } H_0 : \mu_i = \mu_j \text{ when } H_0 \text{ is true.}$$

This probability is called the **comparison error rate** by SAS and the **individual error rate** by **Minitab**.

The individual error rate is not the only error rate that is important in multiple comparisons. The **family error rate** (FER), or the **experimentwise error rate**, is defined to be the probability of at least one false rejection of a true hypothesis $H_0 : \mu_i = \mu_j$ over all comparisons. When many comparisons are made, you *may* have a large probability of making or or more false rejections of true null hypotheses. In particular, when all c comparisons of two population means are performed, each at the α level, then

$$\alpha < FER < c\alpha.$$

For example, in the doughnut problem where $k = 4$, there are $c = .5 * 4 * 3 = 6$ possible comparisons of pairs of fats. If each comparison is carried out at the 5% level, then $.05 < FER < .30$. At the second step of the FSD method, you could have up to a 30% chance of claiming one or more pairs of population means are different if no differences existed between population means. **Minitab** gives the actual FER for this problem as .192. SAS and most other statistical packages do not evaluate the exact FER, so the upper bound is used.

The first step of the FSD method is the ANOVA "screening" test. The multiple comparisons are carried out only if the $F-$test suggests that not all population means are equal. This screening

test tends to deflate the FER for the two-step FSD procedure. However, the FSD method is commonly criticized for being extremely liberal (too many false rejections of true null hypotheses) when some, but not many, differences exist - especially when the number of comparisons is large. This conclusion is fairly intuitive. When you do a large number of tests, each, say, at the 5% level, then sampling variation alone will suggest differences in 5% of the comparisons where the H_0 is true. The number of false rejections could be enormous with a large number of comparisons. For example, chance variation alone would account for an average of 50 significant differences in 1000 comparisons each at the 5% level.

Bonferroni Comparisons

The Bonferroni method controls the FER by reducing the individual comparison error rate. The FER is guaranteed to be no larger than a prespecified amount, say α, by setting the individual error rate for each of the c comparisons of interest to α/c. Larger differences in the sample means are needed before declaring statistical significance using the Bonferroni adjustment than when using the FSD method at the α level.

Assuming all comparisons are of interest, you can implement the Bonferroni adjustment in **Minitab** by specifying the Fisher comparisons with the appropriate **individual error rate**. **Minitab** gives the actual FER, and $100(1-\alpha/c)\%$ CI for all pairs of means $\mu_i - \mu_j$. A by-product of the Bonferroni adjustment is that we have at least $100(1-\alpha)\%$ confidence that all CI statements hold simultaneously!

If you wish to guarantee a $FER \leq .05$ on all six comparisons in the doughnut problem, then set the individual error rate to $.05/6 = .0083$. **Minitab** gives $100(1-.0083)\% = 99.17\%$ CIs for all $\mu_i - \mu_j$, and computes the actual FER. Here FER=.0382. The Bonferroni output was given earlier. Looking at the output, can you create the groups? You should get the groups given below, which implies you have sufficient evidence to conclude that the population mean absorption for Fat 2 exceeds that for Fat 4.

```
FAT 4    FAT 1    FAT 3    FAT 2
----------------------
         ------------------------
```

The Bonferroni method tends to produce "coarser" groups than the FSD method, because the individual comparisons are conducted at a lower level. Equivalently, the minimum significant difference is inflated for the Bonferroni method. For example, in the doughnut problem with $FER \leq .05$, the critical value for the individual comparisons at the .0083 level is $t_{crit} = 2.929$. You can read this off the **Minitab** output or estimate it from a $t-$table with $df = 20$. The minimum significant difference for the Bonferroni comparisons is

$$LSD = 2.929 * 8.18 * \sqrt{\frac{2}{6}} = 13.824$$

versus an LSD=9.85 for the FSD method. Referring back to our table of sample means on page 71, we see that the sole comparison where the absolute difference between sample means exceeds 13.824 involves Fats 2 and 4.

Example from Koopmans: Facial Tissue Thickness

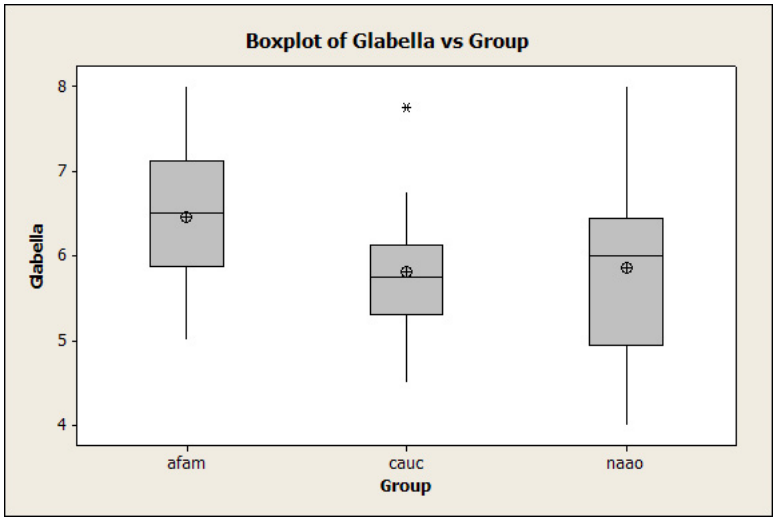
In an anthropological study of facial tissue thickness for different racial groups, data were taken during autopsy at several points on the faces of deceased individuals. The Glabella measurements taken at the bony ridge for samples of individuals from three racial groups (cauc = Caucasian, afam = African American and naao = Native American and Oriental) follow. The data values are in mm.

There are 3 groups, so there are 3 possible pairwise comparisons. If you want a Bonferroni analysis with FER of no greater than .05, you should do the individual comparisons at the $.05/3 = .0167$ level. **Minitab** output is given below. Except for the mild outlier in the Caucasian sample, the observed distributions are fairly normal, with similar spreads. I would expect the standard ANOVA to perform well here.

Let μ_c = population mean Glabella measurement for Caucasians, μ_a = population mean Glabella measurement for African Americans, and μ_n = population mean Glabella measurement for Native Americans and Orientals. At the 5% level, you would not reject the hypothesis that the population mean Glabella measurements are identical. That is, you do not have sufficient evidence to conclude that these racial groups differ with respect to their average Glabella measurement.

The Bonferroni intervals reinforce this conclusion, since each interval for a difference in population means contains zero. You can think of the Bonferroni intervals as simultaneous CI. We're (at least) 95% confident that all of the following statements hold simultaneously: $-1.62 \leq \mu_c - \mu_a \leq .32$, $-.91 \leq \mu_n - \mu_c \leq 1.00$, and $-1.54 \leq \mu_n - \mu_a \leq .33$. The individual CI have level $100(1 - .0167)\% = 98.33\%$. Any further comments?

```
 CONTENTS OF WORKSHEET: Data in Columns c1-c3, labeled

Row  cauc  afam  naao
  1  5.75  6.00  8.00
  2  5.50  6.25  7.00
  3  6.75  6.75  6.00
  4  5.75  7.00  6.25
  5  5.00  7.25  5.50
  6  5.75  6.75  4.00
  7  5.75  8.00  5.00
  8  7.75  6.50  6.00
  9  5.75  7.50  7.25
 10  5.25  6.25  6.00
 11  4.50  5.00  6.00
 12  6.25  5.75  4.25
 13        5.00  4.75
 14              6.00

 Descriptive Statistics: cauc, afam, naao

 Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
 cauc      12   0  5.813    0.241  0.833    4.500  5.313   5.750  6.125    7.750
 afam      13   0  6.462    0.248  0.895    5.000  5.875   6.500  7.125    8.000
 naao      14   0  5.857    0.298  1.117    4.000  4.938   6.000  6.438    8.000
```




```
One-way ANOVA: Glabella versus Group

Source  DF      SS     MS     F      P
Group    2   3.398  1.699  1.83  0.175
Error   36  33.461  0.929
Total   38  36.859

S = 0.9641   R-Sq = 9.22%   R-Sq(adj) = 4.18%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean   StDev  -----+---------+---------+---------+----
afam   13  6.4615  0.8947              (----------*----------)
cauc   12  5.8125  0.8334  (----------*----------)
naao   14  5.8571  1.1168    (---------*---------)
                          -----+---------+---------+---------+----
                             5.50      6.00      6.50      7.00

Pooled StDev = 0.9641


Fisher 98.33% Individual Confidence Intervals
All Pairwise Comparisons among Levels of Group

Simultaneous confidence level = 95.69%


Group = afam subtracted from:

Group    Lower   Center   Upper  ---+---------+---------+---------+------
cauc   -1.6178  -0.6490  0.3198  (-------------*-------------)
naao   -1.5365  -0.6044  0.3277   (------------*-------------)
                                 ---+---------+---------+---------+------
                                -1.40     -0.70      0.00      0.70


Group = cauc subtracted from:

Group    Lower  Center   Upper  ---+---------+---------+---------+------
naao   -0.9074  0.0446  0.9967           (-------------*------------)
                                 ---+---------+---------+---------+------
                                -1.40     -0.70      0.00      0.70
```

Further Discussion of Multiple Comparisons

The FSD and Bonferroni methods comprise the ends of the spectrum of multiple comparisons methods. Among multiple comparisons procedures, the FSD method is most likely to find differences, whether real or due to sampling variation, whereas Bonferroni is often the most conservative method. You can be reasonably sure that differences suggested by the Bonferroni method will be suggested by almost all other methods, whereas differences not significant under FSD will not be picked up using other approaches.

The Bonferroni method is conservative, but tends to work well when the number of comparisons is small, say 4 or less. A smart way to use the Bonferroni adjustment is to focus attention only on the comparisons of interest (generated independently of looking at the data!), and ignore the rest. I will return to this point later.

Two commonly used alternatives to FSD and Bonferroni are **Tukey's** honest significant difference method (HSD) and Newman-Keuls studentized range method. Tukey's method can be implemented in **Minitab** by specifying Tukey multiple comparisons (typically with FER=5%) in the one-way ANOVA dialog box. SW discuss the Newman-Keuls approach, which is not implemented in **Minitab**.

To implement Tukey's method with a FER of α, reject $H_0 : \mu_i = \mu_j$ when

$$|\overline{Y}_i - \overline{Y}_j| \geq \frac{q_{crit}}{\sqrt{2}} s_{pooled} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}},$$

where q_{crit} is the α level critical value of the studentized range distribution. For the doughnut fats, the groupings based on Tukey and Bonferroni comparisons are identical; see the **Minitab** output below.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of fat

Individual confidence level = 98.89%

fat = fat1 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat2   -2.725   10.500  23.725                  (-----*------)
fat3  -12.891    0.333  13.558            (-----*------)
fat4  -25.725  -12.500   0.725       (------*-----)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat2 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat3  -23.391  -10.167   3.058        (------*------)
fat4  -36.225  -23.000  -9.775  (------*-----)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat3 subtracted from:

fat     Lower   Center  Upper  --------+---------+---------+---------+-
fat4  -26.058  -12.833  0.391       (------*-----)
                              --------+---------+---------+---------+-
                                    -20         0        20        40
```

Checking Assumptions in ANOVA Problems

The classical ANOVA assumes that the populations have normal frequency curves and the populations have equal variances (or spreads). You can test the normality assumption using multiple normal scores tests, which we discussed earlier. An alternative approach that is useful with three or more samples is to make a single normal scores plot for the entire data set. The samples must be centered at the same location for this to be meaningful. (WHY?) This is done by subtracting the sample mean from each observation in the sample, giving the so-called **residuals**. A normal scores plot or histogram of the residuals should resemble a sample from a normal population. These two plots can be generated with the ANOVA procedure in **Minitab**, but the normal probability plot does not include a p-value for testing normality. However, the residuals can be stored in the worksheet, and then a formal test of normality is straightforward (from the path `Stat > Basic Statistics > Normality Test` — use either the Anderson Darling or the Ryan Joiner test).

Bartlett's test and Levene's test for equal population variances are obtained from `Stat > ANOVA > Test for Equal Variances`. Bartlett's test is a little sensitive to the normality assumption, while Levene's is not. I will now define **Bartlett's test**, which assumes normally distributed data. As above, let $n^* = n_1 + n_2 + \cdots + n_k$, where the n_is are the sample sizes from the k groups, and define

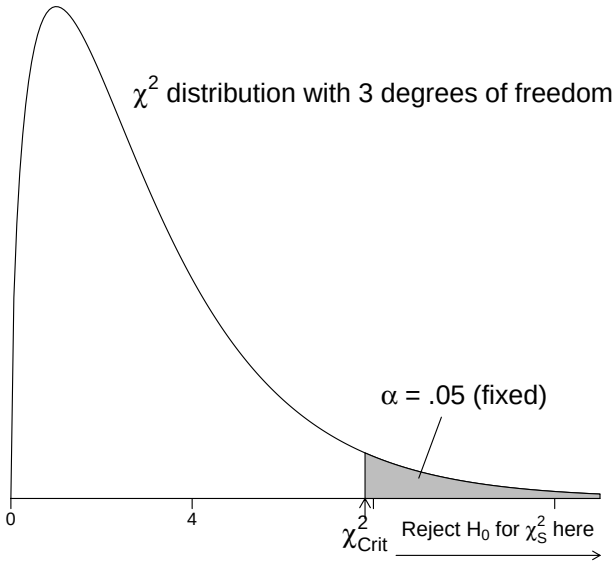
$$v = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n^* - k}\right).$$

Bartlett's statistic for testing $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ is given by

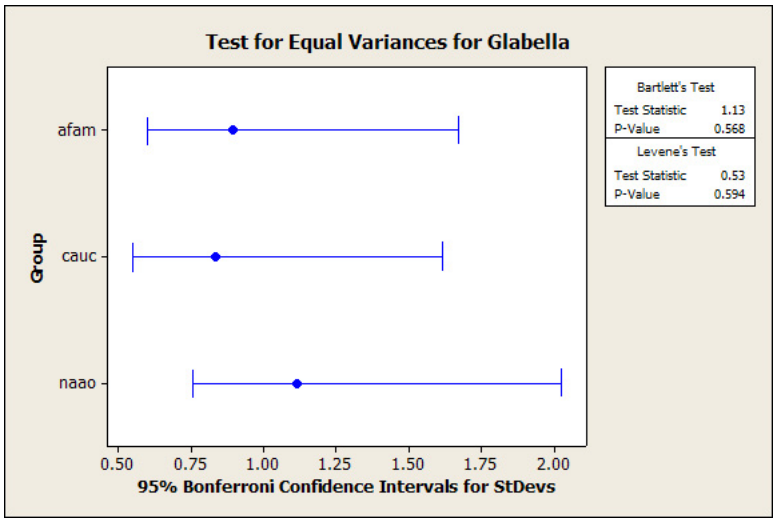
$$B_{obs} = \frac{2.303}{v}\left\{(n-k)\log s^2_{pooled} - \sum_{i=1}^{k}(n_i - 1)\log s_i^2\right\},$$

where s^2_{pooled} is the pooled estimator of variance and s_i^2 is the estimated variance based on the i^{th} sample.

Large values of B_{obs} suggest that the population variances are unequal. For a size α test, we reject H_0 if $B_{obs} \geq \chi^2_{k-1,crit}$, where $\chi^2_{k-1,crit}$ is the upper-α percentile for the χ^2_{k-1} (chi-squared) probability distribution with $k-1$ degrees of freedom. A generic plot of the χ^2 distribution is given below. SW give a chi-squared table on p. 653. A p-value for the test is given by the area under the chi-squared curve to the right of B_{obs}.



Minitab does the calculation for us, as illustrated below. Follow the menu path `Stat > ANOVA > Test for equal variances`. This result is not surprising given how close the sample variances are to each other.



Example from the Child Health and Development Study (CHDS)

We consider data from the birth records of 680 live-born white male infants. The infants were born to mothers who reported for pre-natal care to three clinics of the Kaiser hospitals in northern California. As an initial analysis, we will examine whether maternal smoking has an effect on the birth weights of these children. To answer this question, we define 3 groups based on mother's smoking history: (1) mother does not currently smoke or never smoked (2) mother smoked less than one pack of cigarettes a day during pregnancy (3) mother smoked at least one pack of cigarettes a day during pregnancy.

Let μ_i = pop mean birth weight (in lbs) for children in group i, $(i = 1, 2, 3)$. We wish to test $H_0 : \mu_1 = \mu_2 = \mu_3$ against $H_A :$ not H_0.

Several plots were generated as part of the analysis: dotplots and boxplots, normal probability plots for each sample, and a normal probability plot and histogram of the residuals from the ANOVA. These are included at the end of the notes.

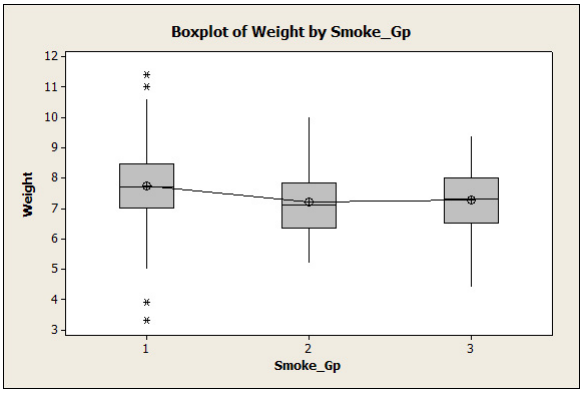
Looking at the boxplots, there is some evidence of non-normality here. Although there are outliers in the no smoking group, we need to recognize that the sample size for this group is fairly large - 381. Given that boxplots are calibrated in such a way that 7 outliers per 1000 observations are expected when sampling from a normal population, 5 outliers (you only see 4!) out of 381 seems a bit excessive. A formal test rejects the hypothesis of normality in the no and low smoker groups. The normal probability plot and the histogram of the residuals also suggests that the population distributions are heavy tailed. I also saved the residuals from the ANOVA and did a formal test of normality on the combined sample, which was significant (p-value=.029). However, I am not overly concerned about this for the following reasons - in large samples, small deviations from normality are often statistically significant and in my experience, the small deviations we are seeing here are not likely to impact our conclusions, in the sense that non-parametric methods that do not require normality will lead to the same conclusions.

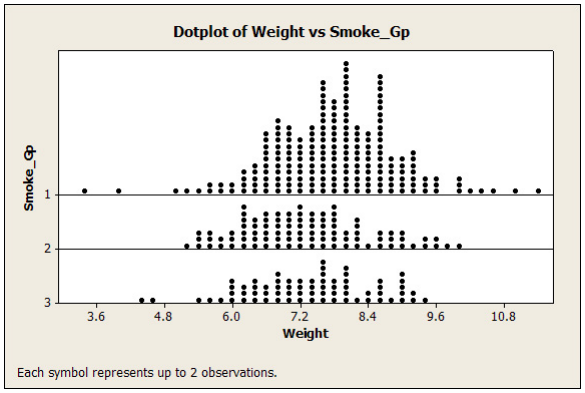
Looking at the summaries, we see that the sample standard deviations are close. Formal tests of equal population variances are far from significant. The p-values for Bartlett's test and Levene's test are greater than .4. Thus, the standard ANOVA appears to be appropriate here.

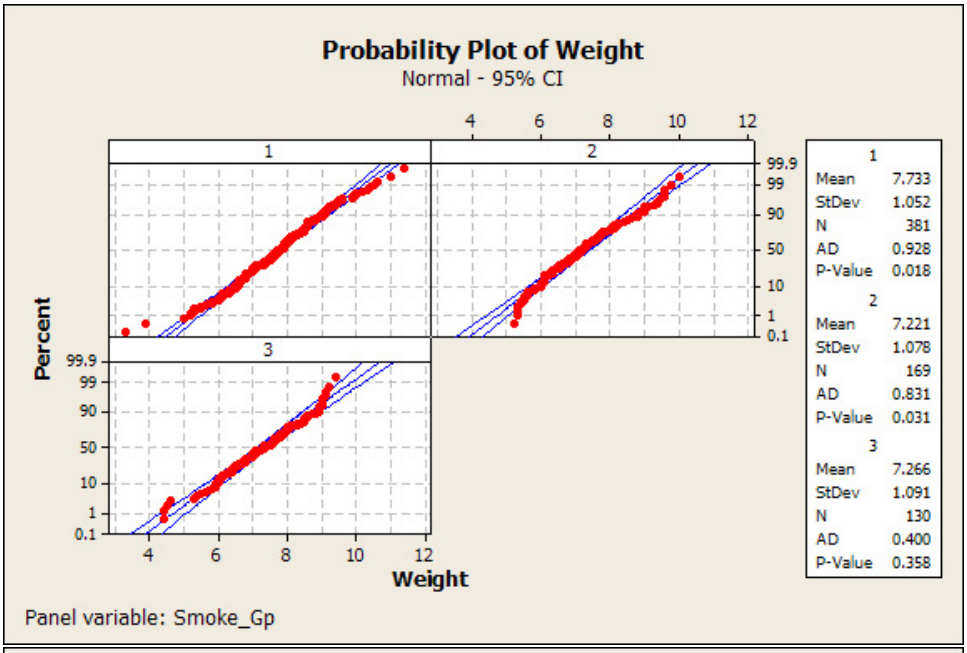
The p-value for the $F-$test is less than .0001. We would reject H_0 at any of the usual test levels (i.e. .05 or .01). The data suggest that the population mean birth weights differ across smoking status groups. The Tukey multiple comparisons suggest that the mean birth weights are higher for children born to mothers that did not smoke during pregnancy.

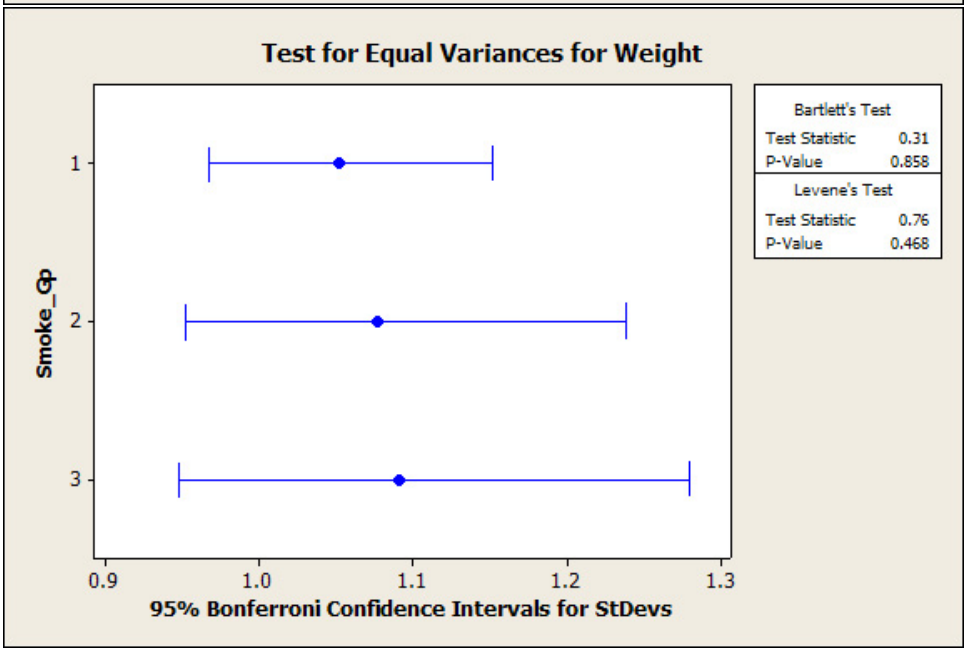
```
Descriptive Statistics: Weight

Variable  Smoke_Gp    N  N*    Mean  SE Mean   StDev  Minimum      Q1  Median
Weight    1         381   0  7.7328   0.0539  1.0523   3.3000  7.0000  7.7000
          2         169   0  7.2213   0.0829  1.0778   5.2000  6.3500  7.1000
          3         130   0  7.2662   0.0957  1.0909   4.4000  6.5000  7.3000

Variable  Smoke_Gp      Q3  Maximum
Weight    1         8.4500  11.4000
          2         7.8500  10.0000
          3         8.0000   9.4000

One-way ANOVA: Weight versus Smoke_Gp

Source     DF      SS     MS      F      P
Smoke_Gp    2   40.70  20.35  17.90  0.000
```

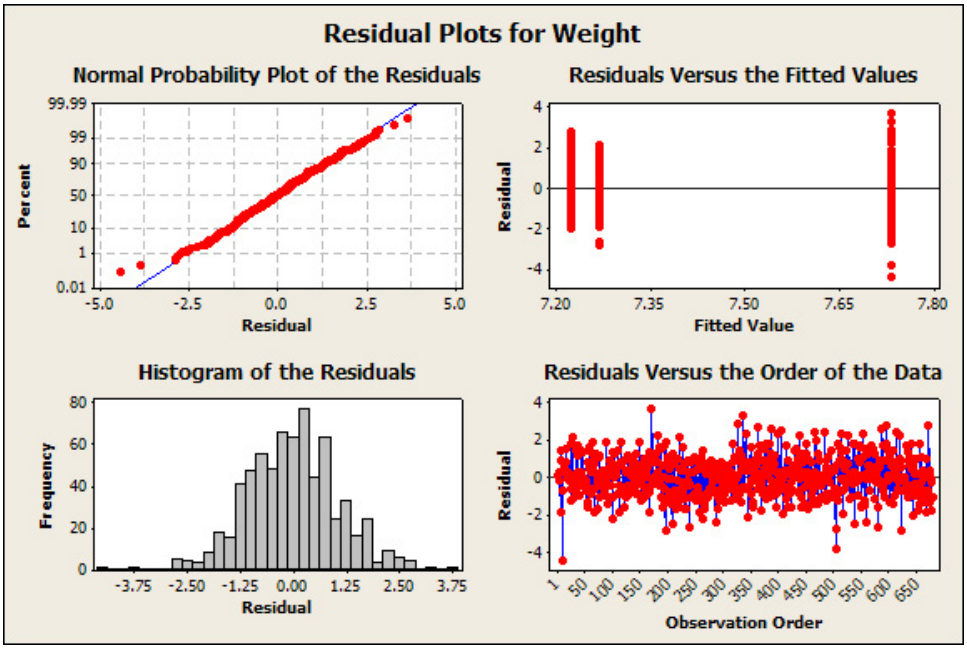
```
Error     677  769.49   1.14
Total     679  810.20

S = 1.066   R-Sq = 5.02%   R-Sq(adj) = 4.74%

                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level    N   Mean  StDev  -------+---------+---------+---------+--
1      381  7.733  1.052                              (-----*----)
2      169  7.221  1.078  (-------*-------)
3      130  7.266  1.091   (--------*--------)
                           -------+---------+---------+---------+--
                                7.20      7.40      7.60      7.80

Pooled StDev = 1.066

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Smoke_Gp

Individual confidence level = 98.05%

Smoke_Gp = 1 subtracted from:

Smoke_Gp   Lower  Center   Upper  -----+---------+---------+---------+----
2         -0.742  -0.512  -0.281  (-------*-------)
3         -0.720  -0.467  -0.213   (--------*--------)
                                 -----+---------+---------+---------+----
                                   -0.60     -0.30      0.00      0.30

Smoke_Gp = 2 subtracted from:

Smoke_Gp   Lower  Center  Upper  -----+---------+---------+---------+----
3         -0.246   0.045  0.336                  (--------*---------)
                                 -----+---------+---------+---------+----
                                   -0.60     -0.30      0.00      0.30
```






Probability Plot of Weight
Normal - 95% CI
Percent
Weight
Panel variable: Smoke_Gp
1
Mean 7.733
StDev 1.052
N 381
AD 0.928
P-Value 0.018
2
Mean 7.221
StDev 1.078
N 169
AD 0.831
P-Value 0.031
3
Mean 7.266
StDev 1.091
N 130
AD 0.400
P-Value 0.358


Test for Equal Variances for Weight
Smoke_Gp
95% Bonferroni Confidence Intervals for StDevs
Bartlett's Test
Test Statistic 0.31
P-Value 0.858
Levene's Test
Test Statistic 0.76
P-Value 0.468


Residual Plots for Weight
Normal Probability Plot of the Residuals
Percent
99.99
99
90
50
10
1
0.01
-5.0
-2.5
0.0
2.5
5.0
Residual
Residuals Versus the Fitted Values
Residual
4
2
0
-2
-4
7.20
7.35
7.50
7.65
7.80
Fitted Value
Histogram of the Residuals
Frequency
80
60
40
20
0
-3.75
-2.50
-1.25
0.00
1.25
2.50
3.75
Residual
Residuals Versus the Order of the Data
Residual
4
2
0
-2
-4
1
50
100
150
200
250
300
350
400
450
500
550
600
650
Observation Order

10 Discrete Data Analysis

SW Chapter 10

Earlier this semester we discussed inference for a single proportion problem. In this section we will generalize those methods in two directions. First we consider single sample problems involving categorical variables with multiple categories. Second, we consider problems with two or more samples.

Goodness-of-Fit Tests

Example The following data set was used as evidence in a court case. The data represent a sample of 1336 individuals from the jury pool of a large municipal court district for the years 1975-1977. The fairness of the representation of various age groups on juries was being contested. The strategy for doing this was to challenge the representativeness of the pool of individuals from which the juries are drawn. This was done by comparing the age group distribution within the jury pool against the age distribution in the district as a whole, which was available from census figures.

Age group (yrs)	Obs. Counts	Obs. Prop.	Census Prop.
18-19	23	.017	.061
20-24	96	.072	.150
25-29	134	.100	.135
30-39	293	.219	.217
40-49	297	.222	.153
50-64	380	.284	.182
65-99	113	.085	.102

A statistical question here is whether the jury pool population proportions are equal to the census proportions across the age categories. This comparison can be formulated as a **goodness-of-fit test**, which generalizes the large sample test on a single proportion to a categorical variable (here age) with $r > 2$ levels. For $r = 2$ categories, the goodness-of-fit test and large sample test on a single proportion are identical. Although this problem compares two populations, only one sample is involved because the census data is a population summary!

In general, suppose each individual in a population is categorized into one and only one of r levels or categories. Let $p_1,\ p_2, ..., p_r$ be the population proportions in the r categories, where each $p_i \geq 0$ and $p_1 + p_2 + \cdots + p_r = 1$. The hypotheses of interest in a goodness-of-fit problem are $H_0 : p_1 = p_1^0, p_2 = p_2^0, ..., p_r = p_r^0$ and H_A : not H_0, where $p_1^0,\ p_2^0, ..., p_r^0$ are given category proportions.

The plausibility of H_0 is evaluated by comparing the hypothesized category proportions to estimated (i.e. observed) category proportions $\hat{p}_1,\ \hat{p}_2,\ ..., \hat{p}_r$ from a random or representative sample of n individuals selected from the population. The discrepancy between the hypothesized and observed proportions is measured by the Pearson chi-squared statistic:

$$\chi_s^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the **observed** number in the sample that fall into the i^{th} category ($O_i = n\hat{p}_i$), and $E_i = np_i^0$ is the number of individuals **expected** to be in the i^{th} category when H_0 is true.

The Pearson statistic can also be computed as the sum of the squared residuals:

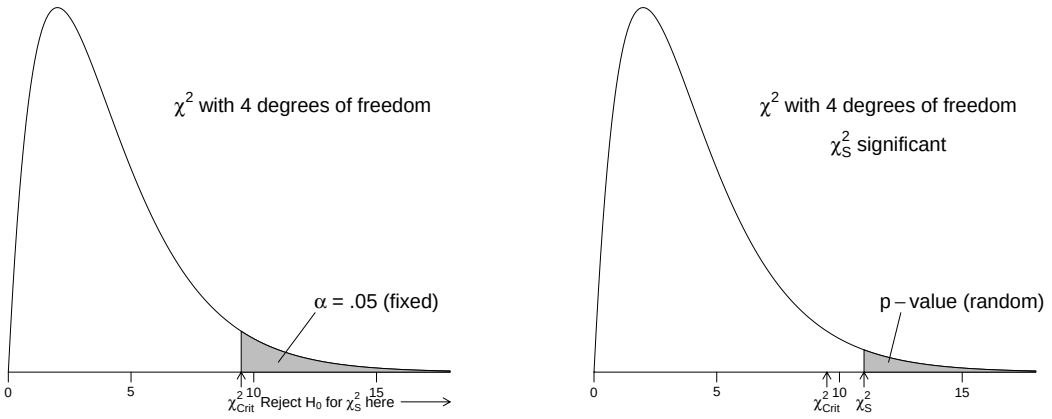
$$\chi_s^2 = \sum_{i=1}^{r} Z_i^2,$$

where $Z_i = (O_i - E_i)/\sqrt{E_i}$, or in terms of the observed and hypothesized category proportions

$$\chi_s^2 = n \sum_{i=1}^{r} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0}.$$

The Pearson statistic χ_s^2 is "small" when all of the observed counts (proportions) are close to the expected counts (proportions). The Pearson χ^2 is "large" when one or more observed counts (proportions) differs noticeably from what is expected when H_0 is true. Put another way, large values of χ_s^2 suggest that H_0 is false.

The critical value χ^2_{crit} for the test is obtained from a chi-squared probability table with $r - 1$ degrees of freedom. A chi-squared table is given on page 686 of SW. The picture below shows the form of the rejection region. For example, if $r = 5$ and $\alpha = .05$, then you reject H_0 when $\chi_s^2 \geq \chi^2_{crit} = 9.49$. The p-value for the test is the area under the chi-squared curve with $df = r - 1$ to the right of the observed χ_s^2 value.



Example (Jury pool problem) Let p_{18} be the proportion in the jury pool population between ages 18 and 19. Define p_{20}, p_{25}, p_{30}, p_{40}, p_{50} and p_{65} analogously. You are interested in testing $H_0 : p_{18} = .061$, $p_{20} = .150$, $p_{25} = .135$, $p_{30} = .217$, $p_{40} = .153$, $p_{50} = .182$ and $p_{65} = .102$ against H_A : not H_0, using the sample of 1336 from the jury pool.

The observed counts, the expected counts, and the category residuals are given in the table below. For example, $E_{18} = 1336 * (.061) = 81.5$ and $Z_{18} = (23 - 81.5)/\sqrt{81.5} = -6.48$ in the 18-19 year category.

The Pearson statistic is

$$\chi_s^2 = (-6.48)^2 + (-7.38)^2 + (-3.45)^2 + .18^2 + 6.48^2 + 8.78^2 + (-1.99)^2 = 231.26$$

on $r - 1 = 7 - 1 = 6$ degrees of freedom. Here $\chi^2_{crit} = 12.59$ at $\alpha = .05$. The p-value for the goodness-of-fit test is less than .001, which suggests that H_0 is false.

Age group (yrs)	Obs. Counts	Exp. Counts	Residual
18-19	23	81.5	-6.48
20-24	96	200.4	-7.38
25-29	134	180.4	-3.45
30-39	293	289.9	0.18
40-49	297	204.4	6.48
50-64	380	243.2	8.78
65-99	113	136.3	-1.99

Adequacy of the Goodness-of-Fit Test

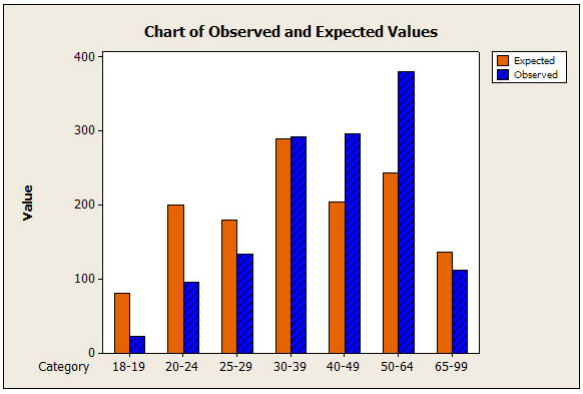
The chi-squared goodness-of-fit test is a large sample test. A conservative rule of thumb is that the test is suitable when each **expected** count is at least five. This holds in the jury pool example. There is no widely available alternative method for testing goodness-of-fit with smaller sample sizes. There is evidence, however, that the chi-squared test is **slightly conservative** (the p-values are too large, on average) when the expected counts are smaller. Some statisticians recommend that the chi-squared approximation be used when the minimum expected count is at least one, provided the expected counts are not too variable.

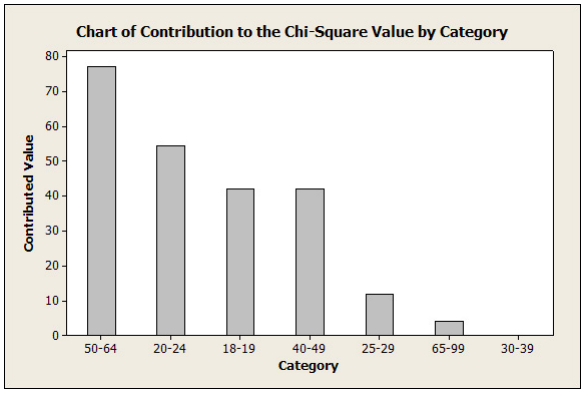
Minitab Implementation

Minitab will do a chi-squared goodness-of-fit test in the by following the menu path `Stat > Tables > Chi-Square Goodness-of-Fit Test (One Variable)`. Unlike the method we used for a single proportion of entering summarized data from a dialog box, the summarized data need to be entered into the worksheet (having counts for categories is summarized data). Following is the Minitab output for the jury pool problem:

```
Data Display

Row  Age    Count  CensusProp
  1  18-19     23       0.061
  2  20-24     96       0.150
  3  25-29    134       0.135
  4  30-39    293       0.217
  5  40-49    297       0.153
  6  50-64    380       0.182
  7  65-99    113       0.102
```

```
Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Count

Using category names in Age

                                 Test                  Contribution
Category   Observed  Proportion  Expected     to Chi-Sq
18-19            23       0.061    81.496       41.9871
20-24            96       0.150   200.400       54.3880
25-29           134       0.135   180.360       11.9164
30-39           293       0.217   289.912        0.0329
40-49           297       0.153   204.408       41.9420
50-64           380       0.182   243.152       77.0192
65-99           113       0.102   136.272        3.9743

   N  DF   Chi-Sq  P-Value
1336   6  231.260    0.000
```





The term "Contribution to Chi-Square" refers to the values of $\frac{(O-E)^2}{E}$ for each category. χ_s^2 is the sum of those contributions.

Comparing Two Proportions: Independent Samples

The New Mexico state legislature is interested in how the proportion of registered voters that support Indian gaming differs between New Mexico and Colorado. Assuming neither population proportion is known, the state's statistician might recommend that the state conduct a survey of registered voters sampled independently from the two states, followed by a comparison of the sample proportions in favor of Indian gaming.

Statistical methods for comparing two proportions using independent samples can be formulated as follows. Let p_1 and p_2 be the proportion of populations 1 and 2, respectively, with the attribute of interest. Let $\hat{p}_1$ and $\hat{p}_2$ be the corresponding sample proportions, based on independent random or representative samples of size n_1 and n_2 from the two populations.

Large Sample CI and Tests for $p_1 - p_2$

A large sample CI for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2)$, where z_{crit} is the standard normal critical value for the desired confidence level, and

$$SE_{CI}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is the CI standard error.

A large sample p-value for a test of the null hypothesis $H_0 : p_1 - p_2 = 0$ against the two-sided alternative $H_A : p_1 - p_2 \neq 0$ is evaluated using tail areas of the standard normal distribution (identical to 1 sample evaluation) in conjunction with the test statistic

$$z_s = \frac{\hat{p}_1 - \hat{p}_2}{SE_{test}(\hat{p}_1 - \hat{p}_2)},$$

where

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

is the test standard error for $\hat{p}_1 - \hat{p}_2$. The **pooled proportion**

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

is the proportion of successes in the two samples combined. The test standard error has the same functional form as the CI standard error, with $\bar{p}$ replacing the individual sample proportions.

The pooled proportion is the best guess at the common population proportion when $H_0 : p_1 = p_2$ is true. The test standard error estimates the standard deviation of $\hat{p}_1 - \hat{p}_2$ assuming H_0 is true.

Example Two hundred and seventy nine French skiers were studied during two one-week periods in 1961. One group of 140 skiers receiving a placebo each day, and the other 139 receiving 1 gram of ascorbic acid (Vitamin C) per day. The study was double blind - neither the subjects nor the researchers knew who received what treatment. Let p_1 be the probability that a member of the ascorbic acid group contracts a cold during the study period, and p_2 be the corresponding probability for the placebo group. Linus Pauling and I are interested in testing whether $p_1 = p_2$. The data are summarized below as a two-by-two table of counts (a contingency table)

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

The sample sizes are $n_1 = 139$ and $n_2 = 140$. The sample proportion of skiers developing colds in the placebo and treatment groups are $\hat{p}_2 = 31/140 = .221$ and $\hat{p}_1 = 17/139 = .122$, respectively. The pooled proportion is the number of skiers that developed colds divided by the number of skiers in the study: $\bar{p} = 48/279 = .172$.

The test standard error is:

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{.172 * (1 - .172)\left(\frac{1}{139} + \frac{1}{140}\right)} = .0452.$$

The test statistic is

$$z_s = \frac{.122 - .221}{.0452} = -2.19.$$

The p-value for a two-sided test is twice the area under the standard normal curve to the right of 2.19 (or twice the area to the left of -2.19), which is $2 * (.014) = .028$ At the 5% level, we reject the

hypothesis that the probability of contracting a cold is the same whether you are given a placebo or Vitamin C.

A CI for $p_1 - p_2$ provides a measure of the size of the treatment effect. For a 95% CI

$$z_{crit}SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.221 * (1 - .221)}{140} + \frac{.122 * (1 - .122)}{139}} = 1.96 * (.04472) = .088.$$

The 95% CI for $p_1 - p_2$ is $(.122 - .221) \pm .088$, or $(-.187, -.011)$. We are 95% confident that p_2 exceeds p_1 by at least .011 but not by more than .187.

On the surface, we would conclude that a daily dose of Vitamin C decreases a French skier's chance of developing a cold by between .011 and .187 (with 95% confidence). This conclusion was somewhat controversial. Several reviews of the study felt that the experimenter's evaluations of cold symptoms were unreliable. Many other studies refute the benefit of Vitamin C as a treatment for the common cold.

Example A case-control study was designed to examine risk factors for cervical dysplasia (Becker et al. 1994). All the women in the study were patients at UNM clinics. The 175 cases were women, aged 18-40, who had cervical dysplasia. The 308 controls were women aged 18-40 who did not have cervical dysplasia. Each women was classified as positive or negative, depending on the presence of HPV (human papilloma virus).

The data are summarized below.

HPV Outcome	Cases	Controls
Positive	164	130
Negative	11	178
Sample size	175	308

Let p_1 be the probability that a case is HPV positive and let p_2 be the probability that a control is HPV positive. The sample sizes are $n_1 = 175$ and $n_2 = 308$. The sample proportions of positive cases and controls are $\hat{p}_1 = 164/175 = .937$ and $\hat{p}_2 = 130/308 = .422$.

For a 95% CI

$$z_{crit}SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.937 * (1 - .937)}{175} + \frac{.422 * (1 - .422)}{308}} = 1.96 * (.03336) = .0659.$$

A 95% CI for $p_1 - p_2$ is $(.937 - .422) \pm .066$, or $.515 \pm .066$, or $(.449, .581)$. I am 95% confident that p_1 exceeds p_2 by at least .45 but not by more than .58.

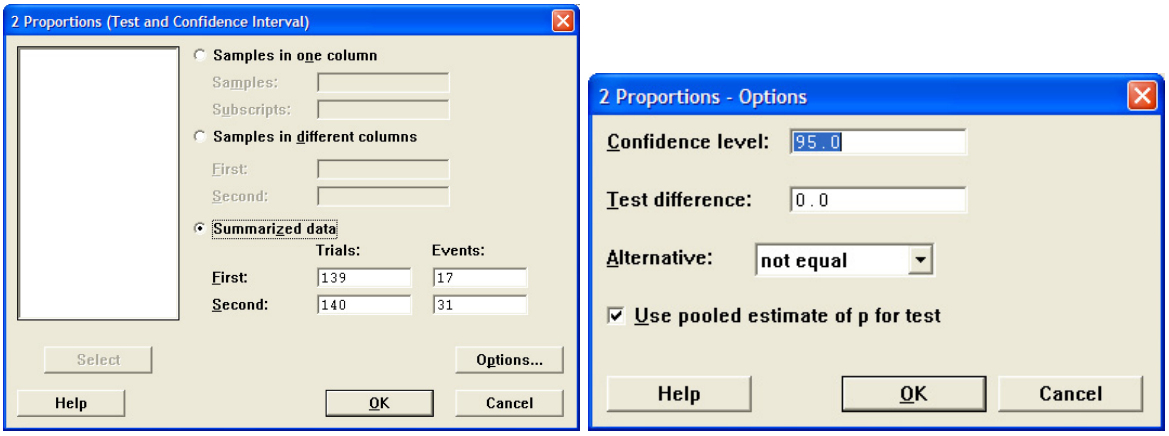
Not surprisingly, a two-sided test at the 5% level would reject $H_0 : p_1 = p_2$. In this problem one might wish to do a one-sided test, instead of a two-sided test. Let us carry out this test, as a refresher on how to conduct one-sided tests.

Appropriateness of Large Sample Test and CI

The standard two sample CI and test used above are appropriate when each sample is large. A rule of thumb suggests a minimum of at least five successes (i.e. observations with the characteristic of interest) and failures (i.e. observations without the characteristic of interest) in each sample before using these methods. This condition is satisfied in our two examples.

Minitab Implementation

For the Vitamin C example, in order to get Minitab to do all the calculations as presented, it is easiest to follow the menu path `Stat > Basic Statistics > 2 Proportions` and enter summary data as follows (you need to check the box for pooled estimate of p for test).




```
Test and CI for Two Proportions

Sample   X    N  Sample p
1       17  139  0.122302
2       31  140  0.221429


Difference = p (1) - p (2)
Estimate for difference:  -0.0991264
95% CI for difference:  (-0.186859, -0.0113937)
Test for difference = 0 (vs not = 0):  Z = -2.19  P-Value = 0.028
```

For the cervical dysplasia example, Minitab results are as follows:

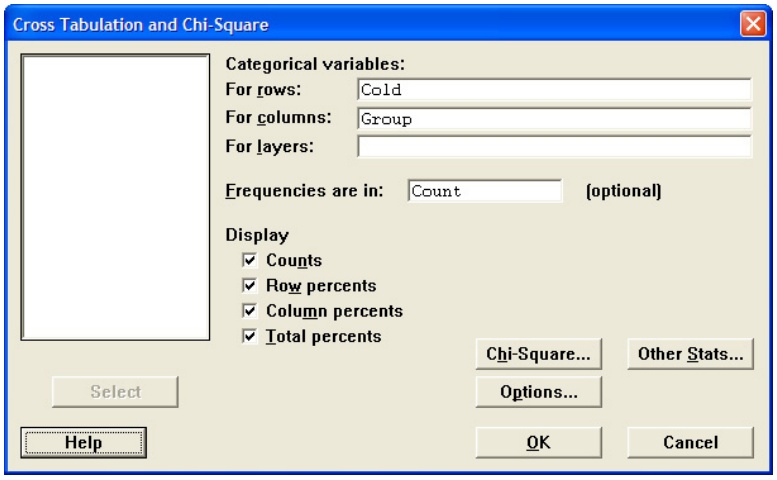
```
Test and CI for Two Proportions

Sample    X    N  Sample p
1       164  175  0.937143
2       130  308  0.422078


Difference = p (1) - p (2)
Estimate for difference:  0.515065
95% CI for difference:  (0.449221, 0.580909)
Test for difference = 0 (vs not = 0):  Z = 11.15  P-Value = 0.000
```

The above analyses are not the most common way to see data like this presented. The ability to get a confidence interval is particularly nice, and I do recommend including such an analysis. Usually, though, we present such data as a two-by-two contingency table. We need this structure in the rest of this section, so let us do that for these two examples.

The basic structure of data entry (it must be in the worksheet) is similar to our earlier use of stacked data. This is how SAS, Stata, and most other packages want it as well. For the Vitamin C example, the data are entered as follows:

```
Data Display

Row  Cold  Group      Count
  1  1Yes  1Vit C        17
  2  1Yes  2Placebo      31
  3  2No   1Vit C       122
  4  2No   2Placebo     109
```

The values for Cold could be entered as just Yes and No, but then Minitab alphabetizes in the presentation. What I have done is one way to get Minitab to present the table in the order we want it. Now we follow the menu path `Stat > Tables > Cross Tabulation and Chi-Square` and fill in the following box appropriately:



The various Display options and Other Stats are reflected in the following output. I structured this to present what I usually get out of SAS by default.

```
Tabulated statistics: Cold, Group

Using frequencies in Count


Rows: Cold   Columns: Group

         1Vit C  2Placebo     All

1Yes         17        31      48
          35.42     64.58  100.00
          12.23     22.14   17.20
           6.09     11.11   17.20
           23.9      24.1    48.0
         1.9990    1.9847       *

2No         122       109     231
          52.81     47.19  100.00
          87.77     77.86   82.80
          43.73     39.07   82.80
          115.1     115.9   231.0
         0.4154    0.4124       *

All         139       140     279
          49.82     50.18  100.00
         100.00    100.00  100.00
          49.82     50.18  100.00
          139.0     140.0   279.0
```

```
            *         *       *

 Cell Contents:      Count
                     % of Row
                     % of Column
                     % of Total
                    Expected count
                    Contribution to Chi-square

 Pearson Chi-Square = 4.811, DF = 1, P-Value = 0.028
 Likelihood Ratio Chi-Square = 4.872, DF = 1, P-Value = 0.027

 Fisher's exact test: P-Value =  0.0384925
```

The Pearson $\chi^2_s = 4.811$ is just the square of $Z_s = -2.19$, so for this case it's really an identical test (only for the two-sided hypothesis, though). The Likelihood Ratio Chi-Square is another large-sample test. Fisher's Exact test is another test that does not need large samples - I use it in practice very frequently. Minitab only performs this test for two-by-two tables — for more complicated tables, this is can be a very hard test to compute. SAS and Stata will at least try to compute it for arbitrary tables, though they do not always succeed. Let us examine the output to see what all these terms mean.

For the cervical dysplasia data, the results are:

```
Data Display

Row  HPV   Group    Count
  1  1Pos  Case       164
  2  1Pos  Control    130
  3  2Neg  Case        11
  4  2Neg  Control    178

 Tabulated statistics: HPV, Group

 Using frequencies in Count


 Rows: HPV   Columns: Group

            Case  Control     All

 1Pos        164      130     294
           55.78    44.22  100.00
           93.71    42.21   60.87
           33.95    26.92   60.87
           106.5    187.5   294.0
           31.01    17.62       *

 2Neg         11      178     189
            5.82    94.18  100.00
            6.29    57.79   39.13
            2.28    36.85   39.13
            68.5    120.5   189.0
           48.25    27.41       *

 All         175      308     483
           36.23    63.77  100.00
           100.00   100.00  100.00
           36.23    63.77  100.00
           175.0    308.0   483.0
```

```
             *        *       *

Cell Contents:       Count
                     % of Row
                     % of Column
                     % of Total
                     Expected count
                     Contribution to Chi-square


Pearson Chi-Square = 124.294, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 144.938, DF = 1, P-Value = 0.000

Fisher's exact test: P-Value =  0.0000000
```

Effect Measures in Two-by-Two Tables

Consider a study of a particular disease, where each individual is either exposed or not-exposed to a risk factor. Let p_1 be the proportion diseased among the individuals in the exposed population, and p_2 be the proportion diseased among the non-exposed population. This population information can be summarized as a two-by-two table of population proportions:

Outcome	Exposed population	Non-Exposed population
Diseased	p_1	p_2
Non-Diseased	$1 - p_1$	$1 - p_2$

A standard measure of the difference between the exposed and non-exposed populations is the **absolute difference:** $p_1 - p_2$. We have discussed statistical methods for assessing this difference.

In many epidemiological and biostatistical settings, other measures of the difference between populations are considered. For example, the relative risk

$$\text{RR} = \frac{p_1}{p_2}$$

is commonly reported when the individual risks p_1 and p_2 are small. The odds ratio

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

is another standard measure. Here $p_1/(1-p_1)$ is the odds of being diseased in the exposed group, whereas $p_2/(1-p_2)$ is the odds of being diseased in the non-exposed group.

We will discuss these measures more completely next semester. At this time I will note that each of these measures can be easily estimated from data, using the sample proportions as estimates of the unknown population proportions. For example, in the vitamin C study:

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

the proportion with colds in the placebo group is $\hat{p}_2 = 31/140 = .221$. The proportion with colds in the vitamin C group is $\hat{p}_1 = 17/139 = .122$.

The estimated absolute difference in risk is $\hat{p}_1 - \hat{p}_2 = .122 - .221 = -.099$. The estimated risk ratio and odds ratio are

$$\hat{RR} = \frac{.122}{.221} = .55$$

and

$$\hat{OR} = \frac{.122/(1-.122)}{.221/(1-.221)} = .49,$$

respectively.

Testing for Homogeneity of Proportions

Example The following two-way table of counts summarizes the location of death and age at death from a study of 2989 cancer deaths (Public Health Reports, 1983):

(Obs Counts)	Location of death			
Age	Home	Acute Care	Chronic care	Row Total
15-54	94	418	23	535
55-64	116	524	34	674
65-74	156	581	109	846
75+	138	558	238	934
Col Total	504	2081	404	2989

The researchers want to compare the age distributions across locations. A one-way ANOVA would be ideal if the actual ages were given. Because the ages are grouped, the data should be treated as categorical. Given the differences in numbers that died at the three types of facilities, a comparison of proportions or percentages in the age groups is appropriate. A comparison of counts is not.

The table below summarizes the proportion in the four age groups at each location. For example, in the acute care facility $418/2081 = .201$ and $558/2081 = .268$. The **pooled proportions** are the Row Totals divided by the total sample size of 2989. The pooled summary gives the proportions in the four age categories, ignoring location of death.

The age distributions for home and for the acute care facilities are similar, but are very different from the age distribution at chronic care facilities.

To formally compare the observed proportions, one might view the data as representative sample of ages at death from the three locations. Assuming independent samples from the three locations (populations), a chi-squared statistic is used to test whether the population proportions of ages at death are identical (homogeneous) across locations. The **chi-squared test for homogeneity** of population proportions can be defined in terms of proportions, but is traditionally defined in terms of counts.

(Proportions)	Location of death			
Age	Home	Acute Care	Chronic care	Pooled
15-54	.187	.201	.057	.179
55-64	.230	.252	.084	.226
65-74	.310	.279	.270	.283
75+	.273	.268	.589	.312
Total	1.000	1.000	1.000	1.000

In general, assume that the data are independent samples from c populations (strata, groups, sub-populations), and that each individual is placed into one of r levels of a categorical variable. The raw data will be summarized as a $r \times c$ **contingency table** of counts, where the columns correspond to the samples, and the rows are the levels of the categorical variable. In the age distribution problem, $r = 4$ and $c = 3$. (SW uses k to identify the number of columns.)

To implement the test:

1. Compute the (estimated) **expected** count for each cell in the table as follows:

$$\mathrm{E} = \frac{\text{Row Total} * \text{Column Total}}{\text{Total Sample Size}}.$$

2. Compute the Pearson test statistic

$$\chi^2_s = \sum_{\text{all cells}} \frac{(O-E)^2}{E},$$

   where O is the **observed** count.

3. For a size α test, reject the hypothesis of homogeneity if $\chi^2_s \geq \chi^2_{crit}$, where χ^2_{crit} is the upper α critical value from the chi-squared distribution with $df = (r-1)(c-1)$.

The p-value for the chi-squared test of homogeneity is equal to the area under the chi-squared curve to the right of X^2; see the picture on page 98.

For a two-by-two table of counts, the chi-squared test of homogeneity of proportions is identical to the two-sample proportion test we discussed earlier.

Minitab Analysis

Enter data as follows (just as for the two-by-two table):

```
Data Display

Row  Age    Care     Count
  1  15-54  1Home       94
  2  15-54  Acute      418
  3  15-54  Chronic     23
  4  55-64  1Home      116
  5  55-64  Acute      524
  6  55-64  Chronic     34
  7  65-74  1Home      156
  8  65-74  Acute      581
  9  65-74  Chronic    109
 10  75+    1Home      138
 11  75+    Acute      558
 12  75+    Chronic    238
```

Follow the same path `Stat > Tables > Cross Tabulation and Chi-Square` with these results:

```
Tabulated statistics: Age, Care

Using frequencies in Count

* NOTE * Fisher's exact test available only for 2 x 2 tables.

Rows: Age   Columns: Care

           1Home   Acute  Chronic     All

15-54         94     418       23     535
           17.57   78.13     4.30  100.00
           18.65   20.09     5.69   17.90
            3.14   13.98     0.77   17.90
            90.2   372.5     72.3   535.0
           0.159   5.564   33.627       *

55-64        116     524       34     674
           17.21   77.74     5.04  100.00
           23.02   25.18     8.42   22.55
            3.88   17.53     1.14   22.55
           113.6   469.3     91.1   674.0
           0.049   6.388   35.789       *

65-74        156     581      109     846
           18.44   68.68    12.88  100.00
           30.95   27.92    26.98   28.30
            5.22   19.44     3.65   28.30
           142.7   589.0    114.3   846.0
           1.249   0.109    0.250       *

75+          138     558      238     934
           14.78   59.74    25.48  100.00
           27.38   26.81    58.91   31.25
            4.62   18.67     7.96   31.25
           157.5   650.3    126.2   934.0
           2.412  13.092   98.937       *

All          504    2081      404    2989
           16.86   69.62    13.52  100.00
          100.00  100.00   100.00  100.00
           16.86   69.62    13.52  100.00
           504.0  2081.0    404.0  2989.0
               *       *        *       *

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
                    Expected count
                    Contribution to Chi-square


Pearson Chi-Square = 197.624, DF = 6, P-Value = 0.000
Likelihood Ratio Chi-Square = 200.972, DF = 6, P-Value = 0.000
```

The Pearson statistic and the likelihood ratio statistic, which is an alternative statistic for testing homogeneity, both report a p-value of 0 to three places. The data strongly suggest that there are differences in the age distributions among locations. The likelihood ratio statistic leads to the same conclusion. The various summaries help us to explain what is the nature of the differences.

Testing for Homogeneity in Cross-Sectional and Stratified Studies

Two-way tables of counts are often collected either by **stratified sampling** or by **cross-sectional** sampling.

In a stratified design, distinct groups, strata, or sub-populations are identified. Independent samples are selected from each group, and the sampled individuals are classified into categories. The HPV study is an illustration of a stratified design (and a case-control study). Stratified designs provide estimates for the strata (population) proportion in each of the categories. A test for **homogeneity of proportions** is used to compare the strata.

In a **cross-sectional design**, individuals are randomly selected from a population and classified by the levels of **two** categorical variables. With cross-sectional samples you can test homogeneity of proportions by comparing either the row proportions or by comparing the column proportions.

Example The following data (*The Journal of Advertising,* 1983, p. 34-42) are from a cross-sectional study that involved soliciting opinions on anti-smoking advertisements. Each subject was asked whether they smoked and their reaction (on a five-point ordinal scale) to the ad. The data are summarized as a two-way table of counts, given below:

	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	8	14	35	21	19	97
Non-smoker	31	42	78	61	69	281
Col Total	39	56	113	82	88	378

The row proportions are

(Row Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	.082	.144	.361	.216	.196	1.000
Non-smoker	.110	.149	.278	.217	.245	1.000

For example, the entry for the (Smoker, Str. Dislike) cell is: $8/97 = .082$.

Similarly, the column proportions are

(Col Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like
Smoker	.205	.250	.310	.256	.216
Non-smoker	.795	.750	.690	.744	.784
Total	1.000	1.000	1.000	1.000	1.000

Although it may be more natural to compare the smoker and non-smoker row proportions, the column proportions can be compared across ad responses. There is no advantage to comparing "rows" instead of "columns" in a formal test of homogeneity of proportions with cross-sectional data. The Pearson chi-squared test (and the likelihood ratio test) treats the rows and columns interchangeably, so you get the same result regardless of how you view the comparison. However, one of the two comparisons may be more natural to interpret.

Note that checking for homogeneity of proportions is meaningful in stratified studies only when the comparison is across strata! Further, if the strata correspond to columns of the table, then the column proportions or percentages are meaningful whereas the row proportions are not.
Question: How do these ideas apply to the age distribution problem?

Testing for Independence in a Two-Way Contingency Table

The row and column classifications for a population where each individual is cross-classified by two categorical variables are said to be independent if each **population** cell proportion in the two-way table is the product of the proportion in a given row and the proportion in a given column. One can show that independence is equivalent to homogeneity of proportions. In particular, the two-way table of population cell proportions satisfies independence if and only if the population column proportions are homogeneous. If the population column proportions are homogeneous then so are the population row proportions.

This suggests that a test for independence or **no association** between two variables based on a cross-sectional study can be implemented using the chi-squared test for homogeneity of proportions. This suggestion is correct. If independence is not plausible, I tend to interpret the dependence as a deviation from homogeneity, using the classification for which the interpretation is most natural.

Example

```
Data Display

Row  Smoker  Opinion         Count
  1  1Yes    1 Str. Dislike      8
  2  1Yes    2 Dislike          14
  3  1Yes    3 Neutral          35
  4  1Yes    4 Like             21
  5  1Yes    5 Str. Like        19
  6  No      1 Str. Dislike     31
  7  No      2 Dislike          42
  8  No      3 Neutral          78
  9  No      4 Like             61
 10  No      5 Str. Like        69

Tabulated statistics: Smoker, Opinion

Using frequencies in Count

* NOTE * Fisher's exact test available only for 2 x 2 tables.

Rows: Smoker   Columns: Opinion

         1 Str.
        Dislike  2 Dislike  3 Neutral   4 Like  5 Str. Like     All

1Yes          8         14         35       21           19      97
           8.25      14.43      36.08    21.65        19.59  100.00
          20.51      25.00      30.97    25.61        21.59   25.66
           2.12       3.70       9.26     5.56         5.03   25.66
          10.01      14.37      29.00    21.04        22.58   97.00
        0.40286    0.00955    1.24259  0.00009      0.56819       *

No           31         42         78       61           69     281
          11.03      14.95      27.76    21.71        24.56  100.00
          79.49      75.00      69.03    74.39        78.41   74.34
           8.20      11.11      20.63    16.14        18.25   74.34
```

```
          28.99      41.63      84.00      60.96      65.42  281.00
        0.13907    0.00330    0.42894    0.00003    0.19614       *

All          39         56        113         82         88     378
          10.32      14.81      29.89      21.69      23.28  100.00
         100.00     100.00     100.00     100.00     100.00  100.00
          10.32      14.81      29.89      21.69      23.28  100.00
          39.00      56.00     113.00      82.00      88.00  378.00
              *          *          *          *          *       *

Cell Contents:       Count
                     % of Row
                     % of Column
                     % of Total
                     Expected count
                     Contribution to Chi-square

Pearson Chi-Square = 2.991, DF = 4, P-Value = 0.559
Likelihood Ratio Chi-Square = 2.980, DF = 4, P-Value = 0.561
```

The Pearson chi-squared test is not significant (p-value = .561). The observed association between smoking status and the ad reaction is not significant. This suggests, for example, that the smoker's reactions to the ad were not statistically significantly different from the non-smoker's reactions, which is consistent with the smokers and non-smokers attitudes being fairly similar.

11 Correlation and Regression

SW, Chapter 12.

Suppose we select $n = 10$ persons from the population of college seniors who plan to take the MCAT exam. Each takes the test, is coached, and then retakes the exam. Let X_i be the pre-coaching score and let Y_i be the post-coaching score for the i^{th} individual, $i = 1, 2, \cdots, n$. There are several questions of potential interest here, for example: Are Y and X related (associated), and how? Does coaching improve your MCAT score? Can we use the data to develop a mathematical model (formula) for predicting post-coaching scores from the pre-coaching scores? These questions can be addressed using **correlation** and **regression** models.

The **correlation coefficient** is a standard measure of **association** or relationship between two features Y and X. Most scientists equate Y and X being correlated to mean that Y and X are associated, related, or **dependent** upon each other. However, correlation is only a measure of the strength of a **linear relationship.** For later reference, let ρ be the correlation between Y and X in the population and let r be the sample correlation. I define r below. The population correlation is defined analogously from population data.

Suppose each of n sampled individuals is measured on two quantitative characteristics called Y and X. The data are pairs of observations $(X_1,\ Y_1)$, $(X_2,\ Y_2)$, $\cdots (X_n,\ Y_n)$, where $(X_i,\ Y_i)$ is the $(X,\ Y)$ pair for the i^{th} individual in the sample. The sample correlation between Y and X, also called the **Pearson product moment correlation coefficient**, is

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2 \sum_i (Y_i - \overline{Y})^2}},$$

where

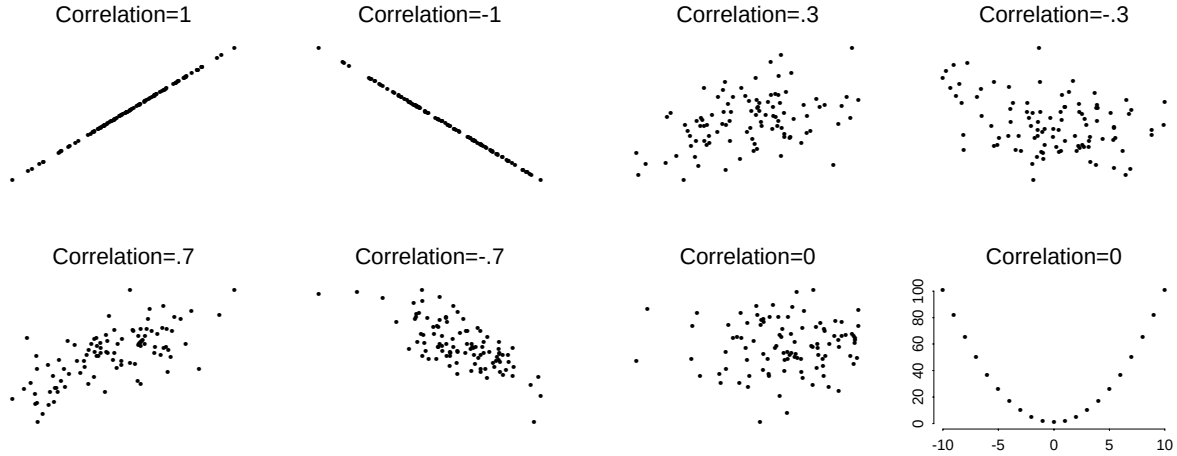
$$S_{XY} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

is the **sample covariance** between Y and X, and $S_Y = \sqrt{\sum_i (Y_i - \overline{Y})^2/(n-1)}$ and $S_X = \sqrt{\sum_i (X_i - \overline{X})^2/(n-1)}$ are the standard deviations for the Y and X samples. Here are eight important properties of r:

1. $-1 \leq r \leq 1$.
2. If Y_i tends to increase linearly with X_i then $r > 0$.
3. If Y_i tends to decrease linearly with X_i then $r < 0$.
4. If there is a perfect linear relationship between Y_i and X_i with a positive slope then $r = +1$.
5. If there is a perfect linear relationship between Y_i and X_i with a negative slope then $r = -1$.
6. The closer the points (X_i, Y_i) come to forming a straight line, the closer r is to ± 1.
7. The magnitude of r is unchanged if either the X or Y sample is transformed linearly (i.e. feet to inches, pounds to kilograms, Celsius to Fahrenheit).
8. The correlation does not depend on which variable is called Y and which is called X.

If r is near ± 1, then there is a strong linear relationship between Y and X in the sample. This suggests we might be able to accurately predict Y from X with a linear equation (i.e. linear regression). If r is near 0, there is a weak linear relationship between Y and X, which suggests that a linear equation provides little help for predicting Y from X. The pictures below should help you develop a sense about the size of r.

Note that $r = 0$ does not imply that Y and X are not related in the sample. It only implies they are not linearly related. For example, in the last plot $r = 0$ yet $Y_i = X_i^2$.



Testing that $\rho = 0$

Suppose you want to test $H_0 : \rho = 0$ against $H_A : \rho \neq 0$, where ρ is the population correlation between Y and X. This test is usually interpreted as a test of no association, or relationship, between Y and X in the population. Keep in mind, however, that ρ measures the strength of a linear relationship.

The standard test of $H_0 : \rho = 0$ is based on the magnitude of r. If we let

$$t_s = r\sqrt{\frac{n-2}{1-r^2}},$$

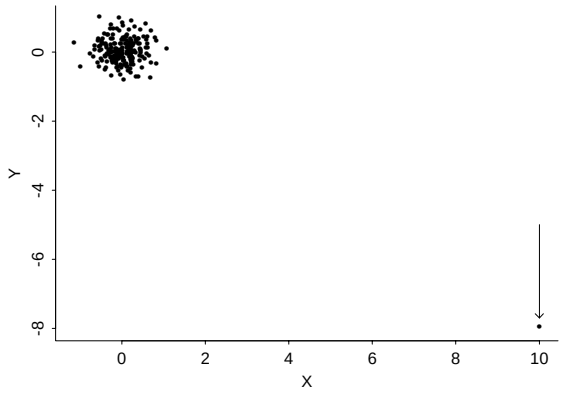
then the test rejects H_0 in favor of H_A if $|t_s| \geq t_{crit}$, where t_{crit} is the two-sided test critical value from a t-distribution with $df = n - 2$. The p-value for the test is the area under the t-curve outside $\pm t_s$ (i.e. two-tailed test p-value).

This test assumes that the data are a random sample from a **bivariate normal population** for $(X,\ Y)$. This assumption implies that all linear combinations of X and Y, say $aX + bY$, are normal. In particular, the (marginal) population frequency curves for X and Y are normal. At a minimum, you should make boxplots of the X and Y samples to check marginal normality. For large-sized samples, a plot of Y against X should be roughly an elliptical cloud, with the density of the points decreasing as the points move away from the center of the cloud.

The Spearman Correlation Coefficient

The Pearson correlation r can be highly influenced by outliers in one or both samples. For example, $r \approx -1$ in the plot below. If you delete the one extreme case with the largest X and smallest Y

value then $r \approx 0$. The two analyses are contradictory. The first analysis (ignoring the plot) suggests a strong linear relationship, whereas the second suggests the lack of a linear relationship. I will not strongly argue that you should (must?) delete the extreme case, but I am concerned about any conclusion that depends heavily on the presence of a single observation in the data set.



Spearman's rank correlation coefficient r_S is a sensible alternative to r when normality is unreasonable or outliers are present. Most books give a computational formula for r_S. I will verbally describe how to compute r_S. First, order the X_is and assign them ranks. Then do the same for the Y_is and replace the original data pairs by the pairs of ranked values. The Spearman rank correlation is the Pearson correlation computed from the pairs of ranks.

The Spearman correlation r_S estimates the **population rank correlation coefficient**, which is a measure of the strength of linear relationship between population ranks. The Spearman correlation, as with other rank based methods, is not sensitive to the presence of outliers in the data. In the plot above, $r_S \approx 0$ whether the unusual point is included or excluded from the analysis. In samples without unusual observations and a linear trend, you often find that $r_S \approx r$.

An important point to note is that the magnitude of the Spearman correlation does not change if either X or Y or both are transformed (monotonically). Thus, if r_S is noticeably greater than r, a transformation of the data might provide a stronger linear relationship.

Example

Eight patients underwent a thyroid operation. Three variables were measured on each patient: weight in kg, time of operation in minutes, and blood loss in ml. The scientists were interested in the factors that influence blood loss. Minitab output for this data set is a separate document.

weight	time	blood loss
44.3	105	503
40.6	80	490
69.0	86	471
43.7	112	505
50.3	109	482
50.2	100	490
35.4	96	513
52.2	120	464

Comments:

1. (Pearson correlations). Blood loss tends to decrease linearly as weight increases, so r should be negative. The output gives $r = -.77$. There is not much of a linear relationship between blood loss and time, so r should be close to 0. The output gives $r = -.11$. Similarly, weight and time have a weak negative correlation, $r = -.07$.

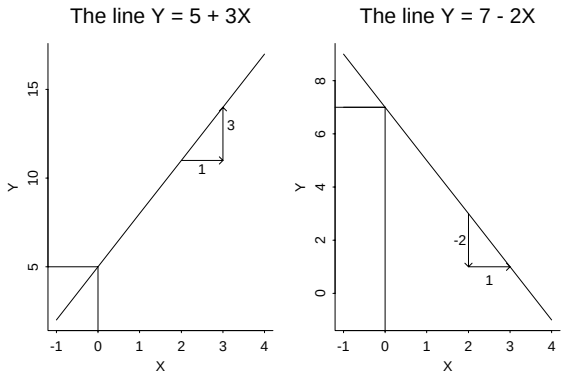
2. The Pearson and Spearman correlations are fairly consistent here. Only the correlation between blood loss and weight is significant at the $\alpha = 0.05$ level (the p-values are given below the correlations).

Simple Linear Regression

In linear regression, we are interested in developing a linear equation that best summarizes the relationship in a sample between the **response variable** Y and the **predictor variable** (or **independent variable**) X. The equation is also used to predict Y from X. The variables are not treated symmetrically in regression, but the appropriate choice for the response and predictor is usually apparent.

Linear Equation

If there is a perfect linear relationship between Y and X then $Y = \beta_0 + \beta_1 X$ for some β_0 and β_1, where β_0 is the Y-intercept and β_1 is the slope of the line. Two plots of linear relationships are given below. The left plot has $\beta_0 = 5$ and $\beta_1 = 3$. The slope is positive, which indicates that Y increases linearly when X increases. The right plot has $\beta_0 = 7$ and $\beta_1 = -2$. The slope is negative, which indicates that Y decreases linearly when X increases.



Least Squares

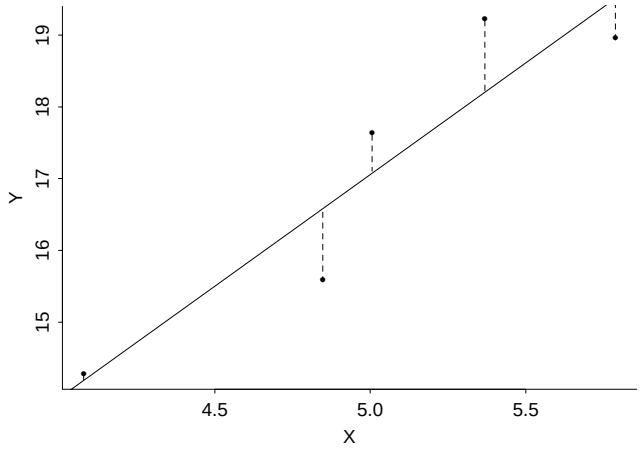
Data rarely, if ever, fall on a straight line. However, a straight line will often describe the **trend** for a set of data. Given a data set (X_i, Y_i), $i = 1, ..., n$ with a **linear trend**, what linear equation "best" summarizes the observed relationship between Y and X? There is no universally accepted

definition of "best", but many researchers accept the **Least Squares** line (LS line) as a reasonable summary.

Mathematically, the LS line chooses the values of β_0 and β_1 that minimize

$$\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

over all possible choices of β_0 and β_1. These values can be obtained using calculus. Rather than worry about this calculation, note that the LS line makes the sum of squared deviations between the responses Y_i and the line as small as possible, over all possible lines. The LS line typically goes through "the heart" of the data, and is often closely approximated by an eye-ball fit to the data.



The equation of the LS line is

$$\hat{y} = b_0 + b_1 X$$

where the intercept b_0 satisfies

$$b_0 = \overline{Y} - b_1\overline{X}$$

and the slope is

$$b_1 = \frac{\sum_i (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_i (X_i - \overline{X})^2} = r\frac{S_Y}{S_X}.$$

As before, r is the Pearson correlation between Y and X, whereas S_Y and S_X are the sample standard deviations for the Y and X samples, respectively. The **sign of the slope** and the **sign of the correlation** are **identical** (i.e. + correlation implies + slope).

Special symbols b_0 and b_1 identify the LS intercept and slope to distinguish the LS line from the generic line $Y = \beta_0 + \beta_1 X$. You should think of $\hat{Y}$ as the **fitted value** at X, or the value of the LS line at X.

Minitab Implementation

The separate document shows Minitab output from a least squares fit.

For the **thyroid operation data** with Y = Blood loss in ml and X = Weight in kg, the LS line is $\hat{Y} = 552.44 - 1.30X$, or Predicted Blood Loss = 552.44 − 1.30 Weight. For an $86kg$ individual, the Predicted Blood Loss = $552.44 - 1.30 * 86 = 440.64ml$.

The LS regression coefficients for this model are interpreted as follows. The intercept b_0 is the predicted blood loss for a 0 kg individual. The intercept has no meaning here. The slope b_1 is the predicted increase in blood loss for each additional kg of weight. The slope is -1.30, so the predicted *decrease* in blood loss is 1.30 ml for each increase of 1 kg in weight.

Any fitted linear relationship holds only approximately and does not necessarily extend outside the range of the data. In particular, nonsensical predicted blood losses of less than zero are obtained at very large weights outside the range of data.

ANOVA Table for Regression

The LS line minimizes

$$\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

over all choices for β_0 and β_1. Inserting the LS estimates b_0 and b_1 into this expression gives

$$\text{Residual Sums of Squares} = \sum_{i=1}^{n}\{Y_i - (b_0 + b_1 X_i)\}^2.$$

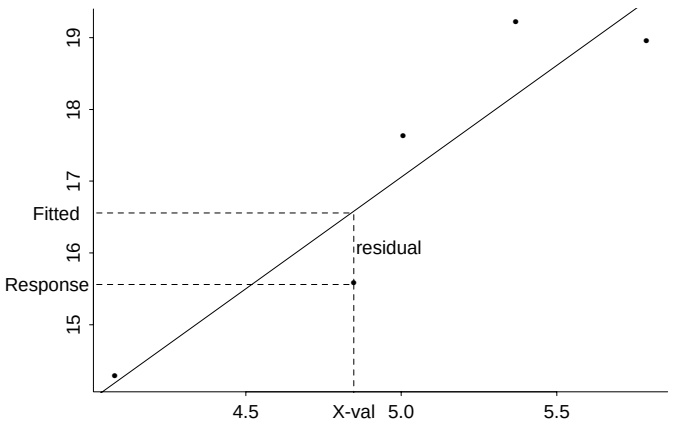
Several bits of notation are needed. Let

$$\hat{Y}_i = b_0 + b_1 X_i$$

be the **predicted** or fitted $Y-$value for an $X-$value of X_i and let $e_i = Y_i - \hat{Y}_i$. The fitted value $\hat{Y}_i$ is the value of the LS line at X_i whereas the **residual** e_i is the distance that the observed response Y_i is from the LS line. Given this notation,

$$\text{Residual Sums of Squares} = \text{Res SS } = \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

Here is a picture to clarify matters:



The Residual SS, or sum of squared residuals, is *small* if each $\hat{Y}_i$ is *close to* Y_i (i.e. the line closely fits the data). It can be shown that

$$\text{Total SS in Y} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \geq \text{Res SS} \geq 0.$$

Also define

$$\text{Regression SS} = \text{Reg SS} = \text{Total SS} \ - \ \text{Res SS} = b_1 \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}).$$

The Total SS measures the variability in the $Y-$sample. Note that

$$0 \leq \text{Regression SS} \leq \text{Total SS}.$$

The percentage of the variability in the $Y-$ sample that is **explained by the linear relationship** between Y and X is

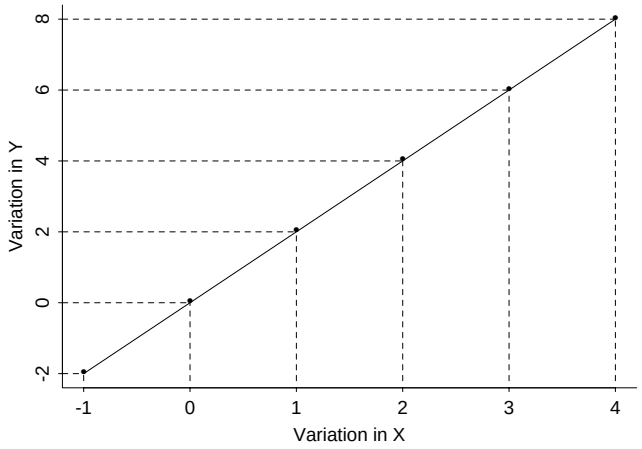
$$R^2 = \text{coefficient of determination} = \frac{\text{Reg SS}}{\text{Total SS}}.$$

Given the definitions of the Sums of Squares, we can show $0 \leq R^2 \leq 1$ and

$$R^2 = \text{square of Pearson correlation coefficient } = r^2.$$

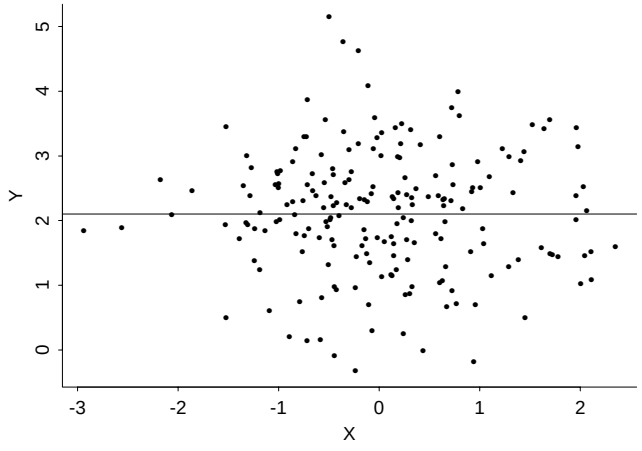
To understand the interpretation of R^2, at least in two extreme cases, note that

$$\begin{aligned}
\text{Reg SS} = \text{Total SS} \quad &\Leftrightarrow \quad \text{Res SS} = 0 \\
&\Leftrightarrow \quad \text{all the data points fall on a straight line} \\
&\Leftrightarrow \quad \text{all the variability in } Y \text{ is explained by the linear relationship with } X \\
&\qquad\quad \text{(which has variation)} \\
&\Leftrightarrow \quad R^2 = 1. \quad \text{(see the picture below)}
\end{aligned}$$



Furthermore,

$$\begin{aligned}
\text{Reg SS} = 0 &\Leftrightarrow \text{Total SS} = \text{Res SS} \\
&\Leftrightarrow b_1 = 0 \\
&\Leftrightarrow \text{LS line is } \hat{Y} = \overline{Y} \\
&\Leftrightarrow \text{none of the variability in } Y \text{ is explained by a linear relationship} \\
&\Leftrightarrow R^2 = 0.
\end{aligned}$$



Each Sum of Squares has a corresponding *df* (degrees of freedom). The Sums of Squares and *df* are arranged in an analysis of variance (ANOVA) table:

Source	*df*	SS	MS
Regression	1		
Residual	$n-2$		
Total	$n-1$		

The Total *df* is $n-1$. The Residual *df* is n minus the number of parameters (2) estimated by the LS line. The Regression *df* is the number of predictor variables (1) in the model. A Mean Square is always equal to the Sum of Squares divided by the *df*. SW use the following notation for the Residual MS: $s^2_{Y|X} = Resid(SS)/(n-2)$.

Brief Discussion of Minitab Output for Blood Loss Problem

1. Identify fitted line: Blood Loss = 552.44 - 1.30 Weight (i.e. $b_0 = 552.44$ and $b_1 = -1.30$).
2. Locate Analysis of Variance Table. More on this later.
3. Locate Parameter Estimates Table. More on this later.
4. Note that $R^2 = .5967 = (-.77247)^2 = r^2$.

The regression model

The following statistical model is assumed as a means to provide error estimates for the LS line, regression coefficients, and predictions. Assume that the data (X_i, Y_i), $i = 1, ..., n$ are a sample of (X, Y) values from the population of interest, and

1. The mean in the population of all responses Y at a given X value (called $\mu_{Y|X}$ by SW) falls on a straight line, $\beta_0 + \beta_1 X$, called the population regression line.

2. The variation among responses Y at a given X value is the same for each X, and is denoted by $\sigma^2_{Y|X}$.

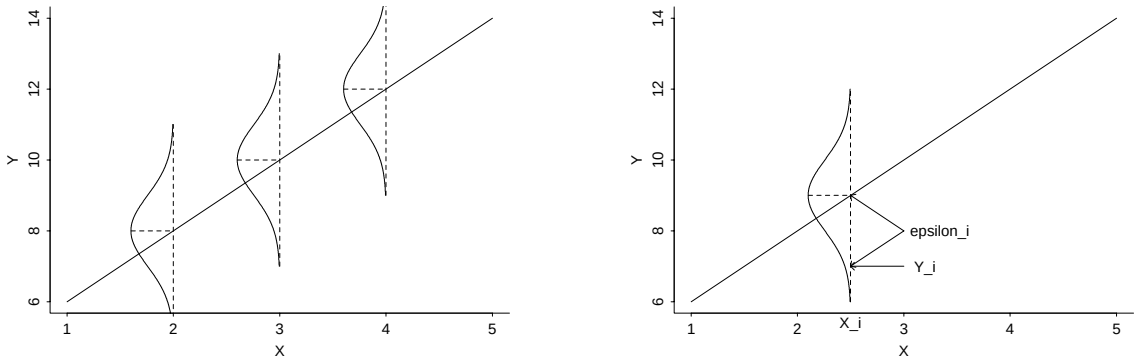
3. The population of responses Y at a given X is normally distributed.

4. The pairs (X_i, Y_i) are a random sample from the population. Alternatively, we can think that the X_is were fixed by the experimenter, and that the Y_i are random responses at the selected predictor values.

The model is usually written in the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

(i.e. Response = Mean Response + Residual), where the ϵ_is are, by virtue of assumptions 2, 3 and 4, independent normal random variables with mean 0 and variance $\sigma^2_{Y|X}$. The following picture might help see this. Note that the population regression line is unknown, and is estimated from the data using the LS line.



Back to the Data

There are three unknown population parameters in the model: β_0, β_1 and $\sigma^2_{Y|X}$. Given the data, the LS line

$$\hat{Y} = b_0 + b_1 X$$

estimates the population regression line $\beta_0 + \beta_1 X$. The LS line is our best guess about the unknown population regression line. Here b_0 estimates the intercept β_0 of the population regression line and b_1 estimates the slope β_1 of the population regression line.

The i^{th} **observed residual** $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = b_0 + b_1 X_i$ is the i^{th} **fitted value**, estimates the **unobservable residual** ϵ_i. (ϵ_i is unobservable because β_0 and β_1 are unknown.) See the picture on page 10 to refresh your memory on the notation. The Residual MS from the ANOVA table is used to estimate $\sigma^2_{Y|X}$:

$$s^2_{Y|X} = \text{Res MS} = \frac{\text{Res SS}}{\text{Res df}} = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{n-2}.$$

CI and tests for β_1

A CI for β_1 is given $b_1 \pm t_{crit} SE_{b_1}$, where the standard error of b_1 under the model is

$$SE_{b_1} = \frac{s_{Y|X}}{\sqrt{\sum_i (X_i - \overline{X})^2}},$$

and where t_{crit} is the appropriate critical value for the desired CI level from a $t-$distribution with df =Res df.

To test $H_0 : \beta_1 = \beta_{1,0}$ (a given value) against $H_A : \beta_1 \neq \beta_{1,0}$, reject H_0 if $|t_s| \geq t_{crit}$, where

$$t_s = \frac{b_1 - \beta_{1,0}}{SE_{b_1}},$$

and t_{crit} is the $t-$critical value for a two-sided test, with the desired size and df =Res df. Alternatively, you can evaluate a p-value in the usual manner to make a decision about H_0.

The parameter estimates table in Minitab gives the standard error, $t-$statistic, and p-value for testing $H_0 : \beta_1 = 0$. Analogous summaries are given for the intercept, but these are typically of less interest.

Testing $\beta_1 = 0$

Assuming the mean relationship is linear, consider testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$. This test can be conducted using a t-statistic, as outlined above, or with an ANOVA $F-$test, as outlined below.

For the analysis of variance (ANOVA) F-test, compute

$$F_s = \frac{\text{Reg MS}}{\text{Res MS}}$$

and reject H_0 when F_s exceeds the critical value (for the desired size test) from an $F-$table with numerator $df = 1$ and denominator $df = n - 2$; see SW, page 654. The hypothesis of zero slope (or no relationship) is rejected when F_s is large, which happens when a significant portion of the variation in Y is explained by the linear relationship with X. Minitab gives the $F-$statistic and p-value with the ANOVA table output.

The p-values from the $t-$test and the $F-$test are always equal. Furthermore this p-value is equal to the p-value for testing no correlation between Y and X, using the $t-$test described earlier. Is this important, obvious, or disconcerting?

A CI for the population regression line

I can not overemphasize the **power** of the regression model. The model allows you to estimate the mean response at any X value in the range for which the model is reasonable, even if little or no data is observed at that location.

We estimate the mean population response among individuals with $X = X_p$

$$\mu_p = \beta_0 + \beta_1 X_p,$$

with the fitted value, or the value of the least squares line at X_p:

$$\hat{Y}_p = b_0 + b_1 X_p.$$

X_p is not necessarily one of the observed X_is in the data. To get a CI for μ_p, use $\hat{Y}_p \pm t_{crit} SE(\hat{Y}_p)$, where the standard error of $\hat{Y}_p$ is

$$SE(\hat{Y}_p) = s_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}}.$$

The $t-$critical value is identical to that used in the subsection on CI for β_1.

CI for predictions

Suppose a future individual (i.e. someone not used to compute the LS line) has $X = X_p$. The best prediction for the response Y of this individual is the value of the least squares line at X_p:

$$\hat{Y}_p = b_0 + b_1 X_p.$$

To get a CI (prediction interval) for an individual response, use $\hat{Y}_p \pm t_{crit} SE_{pred}(\hat{Y}_p)$, where

$$SE_{pred}(\hat{Y}_p) = s_{Y|X}\sqrt{1 + \frac{1}{n} + \frac{(X_p - \overline{X})^2}{\sum_i (X_i - \overline{X})^2}},$$

and t_{crit} is identical to the critical value used for a CI on β_1.

For example, in the blood loss problem you may want to estimates the blood loss for an $50kg$ individual, and to get a CI for this prediction. This problem is different from computing a CI for the mean blood loss of all $50kg$ individuals!

Comments

1. The prediction interval is wider than the CI for the mean response. This is reasonable because you are less confident in predicting an individual response than the mean response for all individuals.

2. The CI for the mean response and the prediction interval for an individual response become wider as X_p moves away from $\overline{X}$. That is, you get a more sensitive CI and prediction interval for X_ps near the center of the data.

3. In `Stat > Regression > Fitted Line Plot` Minitab will plot a band of 95% confidence intervals and a band of 95% prediction intervals on the data plot, along with the fitted LS line.

A further look at the blood loss data (Minitab Output)

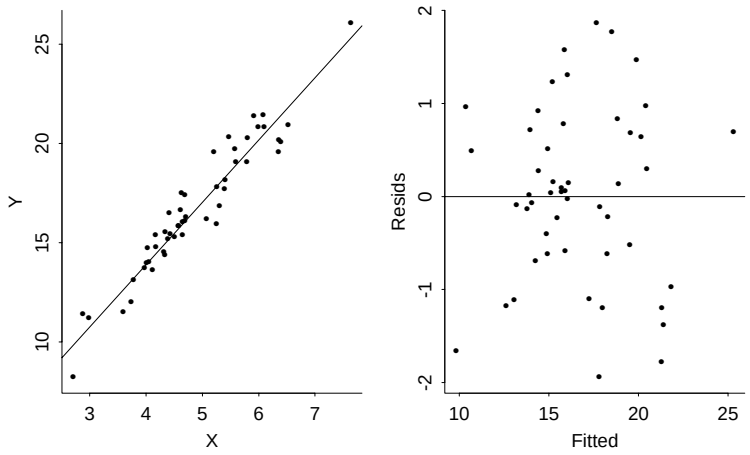
- The LS line is: Predicted Blood Loss = 552.442 - 1.30 Weight.
- The R^2 is .597 (i.e. 59.7%).
- The $F-$statistic for testing $H_0 : \beta_1 = 0$ is $F_{obs} = 8.88$ with a $p - value = .025$. The Error MS is $s^2_{Y|X} = 136.0$; see ANOVA table.
- The Parameter Estimates table gives b_0 and b_1, their standard errors, and $t-$statistics and p-values for testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. The $t-$test and $F-$ test p-values for testing that the slope is zero are identical. We could calculate a 95% CI for β_0 and β_1. If we did so (using the t critical value) we find we are 95% confident that the slope of the population regression line is between -2.37 and -.23.
- Suppose we are interested in estimating the average blood loss among all 50kg individuals. The estimated mean blood loss is $552.442 - 1.30033 * 50 = 487.43$. Reading off the plot, we are 95% confident that the mean blood loss of all 50kg individuals is between (approximately) 477 and 498 ml. A 95% prediction interval for the blood loss of a single 50 kg person is less precise (about 457 to 518 ml).

As a summary we might say that weight is important for explaining the variation in blood loss. In particular, the estimated slope of the least squares line (Predicted Blood loss = 552.442 - 1.30 Weight) is significantly different from zero (p-value = .0247), with weight explaining approximately 60% (59.7%) of the variation in blood loss for this sample of 8 thyroid operation patients.

Checking the regression model

A regression analysis is never complete until the assumptions of the model have been checked. In addition, you need to evaluate whether individual observations, or groups of observations, are unduly influencing the analysis. A first step in any analysis is to plot the data. The plot provides information on the linearity and constant variance assumption. For example, the data plot below shows a linear relationship with roughly constant variance.

In addition to plotting the data, a variety of methods for assessing model adequacy are based on plots of the residuals, $e_i = Y_i - \hat{Y}_i$ (i.e. Observed $-$ Fitted values). For example, an option in Minitab is to plot the e_i against the fitted values $\hat{Y}_i$, as given below. This residual plot should exhibit no systematic dependence of the sign or the magnitude of the residuals on the fitted values.

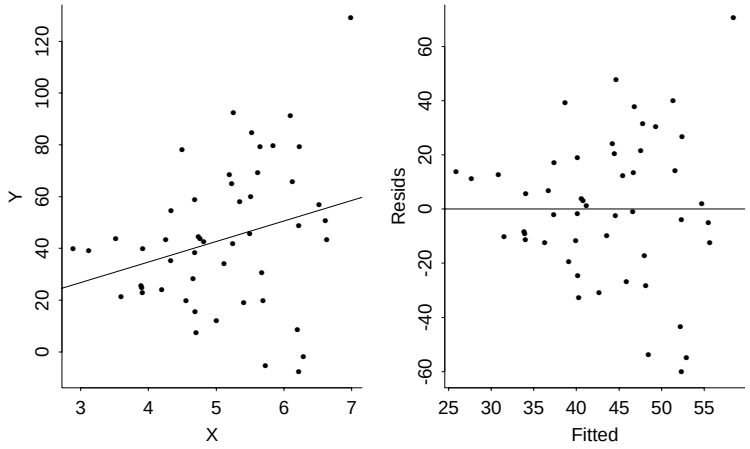


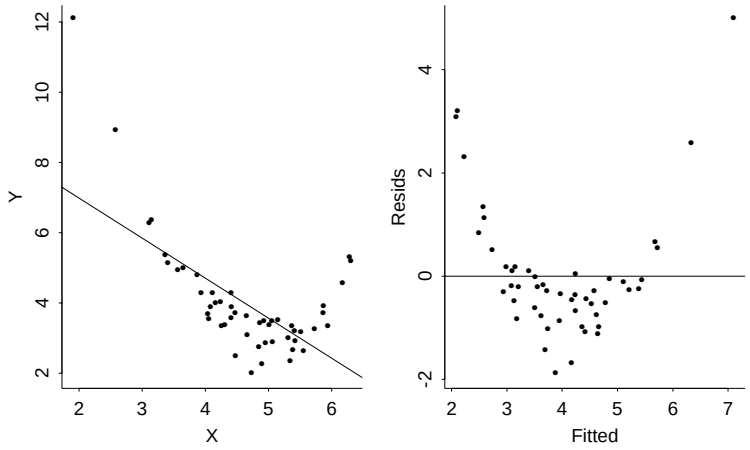
The real power of this plot is with multiple predictor problems (multiple regression). For simple linear regression, the information in this plot is similar to the information in the original data plot, except that the residual plot eliminates the effect of the trend on your perceptions of model adequacy.

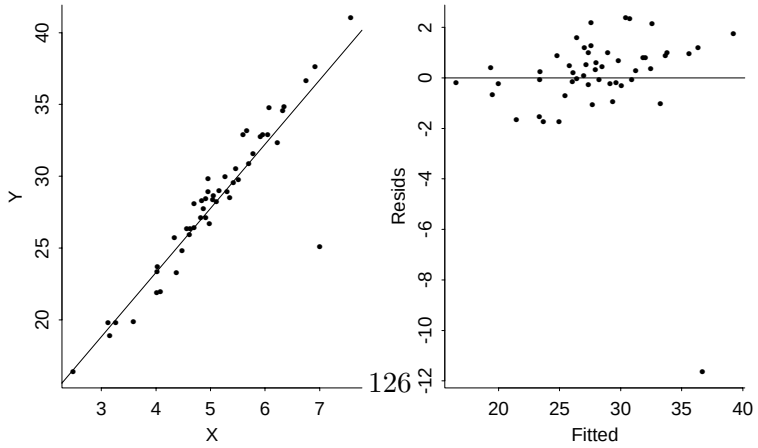
The following plots show how inadequacies in the data plot appear in a residual plot. The first plot shows a roughly linear relationship between Y and X with non-constant variance. The residual plot shows a megaphone shape rather than the ideal horizontal band. A possible remedy is a **weighted least squares** analysis to handle the non-constant variance, or to transform Y to stabilize the variance. Transforming the data may destroy the linearity.

The second plot shows a nonlinear relationship between Y and X. The residual plot shows a systematic dependence of the sign of the residual on the fitted value. A possible remedy is to transform the data.

The last plot shows an outlier. This point has a large residual. A sensible approach is to refit the model after deleting the case and see if any conclusions change.


Y
0
20
40
60
80
100
120
3
4
5
6
7
X
Resids
-60
-40
-20
0
20
40
60
25
30
35
40
45
50
55
Fitted


Y
2
4
6
8
10
12
2
3
4
5
6
X
Resids
-2
0
2
4
2
3
4
5
6
7
Fitted


Y
20
25
30
35
40
3
4
5
6
7
X
Resids
-12
-10
-8
-6
-4
-2
0
2
20
25
30
35
40
Fitted

Checking normality

The normality assumption can be evaluated with a boxplot or a normal quantile plot of the residuals. A formal test of normality using the residuals can be computed as discussed earlier this semester.

Checking independence

Diagnosing dependence among observations usually requires some understanding of the mechanism that generated the data. There are a variety of graphical and inferential tools for checking independence for data collected over time (called a time series). The easiest thing to do is plot the r_i against time index and look for any suggestive patterns.

Outliers

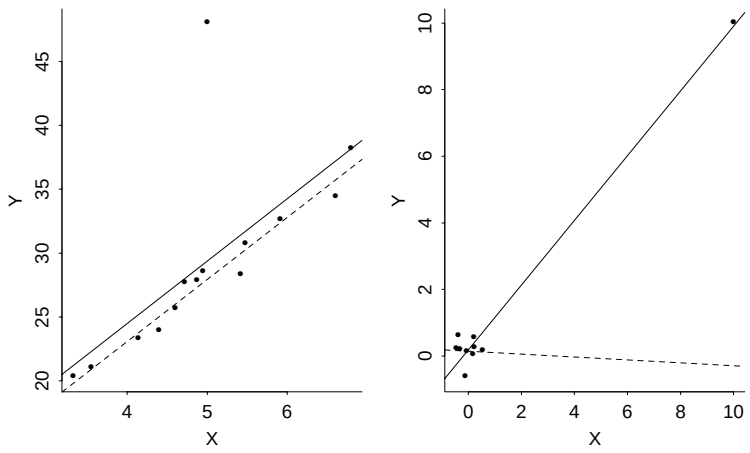
Outliers are observations that are poorly fitted by the regression model. The response for an outlier is far from the fitted line, so outliers have large positive or negative values of the residual e_i.

What do you do with outliers? Outliers may be due to incorrect recordings of the data or failure of the measuring device, or indications or a change in the mean or variance structure for one or more cases. Incorrect recordings should be fixed if possible, but otherwise deleted from the analysis.

Routine deletion of outliers from the analysis is not recommended. This practice can have a dramatic effect on the fit of the model and the perceived precision of parameter estimates and predictions. Analysts who routinely omit outliers without cause tend to overstate the significance of their findings and get a false sense of precision in their estimates and predictions. At the very least, a data analyst should repeat the analysis with and without the outliers to see whether any substantive conclusions are changed.

Influential observations

Certain data points can play a very important role in determining the position of the LS line. These data points may or may not be outliers. For example, the observation with $Y > 45$ in the first plot below is an outlier relative to the LS fit. The extreme observation in the second plot has a very small e_i. Both points are highly **influential observations** - the LS line changes dramatically when these observations are deleted. The influential observation in the second plot is not an outlier because its presence in the analysis determines that the LS line will essentially pass through it! In these plots the solid line is the LS line from the full data set, whereas the dashed line is the LS line after omitting the unusual point.



There are well defined measures of the influence that individual cases have on the LS line, and they are available in Minitab. On the separate output I calculated Cook's D (labelled COOK1) – large values indicate influential values. Which observations are most influential according to this measure? For simple linear regression most influential cases can be easily spotted by carefully looking at the data plot. If you identify cases that you suspect might be influential, you should hold them out (individually) and see if any important conclusions change. If so, you need to think hard about whether the cases should be included or excluded from the analysis.

12 Introduction to Multiple Linear Regression

In **multiple linear regression**, a linear combination of two or more predictor variables is used to explain the variation in a response. In essence, the additional predictors are used to explain the variation in the response not explained by a simple linear regression fit.

It can be a lot more interesting than that sounds, however, since predictors can operate much differently together than alone. Fitting multiple predictors *adjusts* the estimated effects of a predictor for the other predictors. An apparently important predictor can have little effect if adjusted for other variables, or an apparently insignificant predictor can appear very important after adjusting for other variables. The upshot is that the simple linear regression models we worked with last week are inadequate in many circumstances (though they are the basis for what we will see in this section).

As an illustration, I will consider the following problem. Anthropologists conducted a study to determine the long-term effects of an environmental change on systolic blood pressure. They measured the blood pressure and several other characteristics (weight, age, years since migration, pulse rate, skin fold measures) of 39 Indians who migrated from a very primitive environment high in the Andes into the mainstream of Peruvian society at a lower altitude. All of the Indians were males at least 21 years of age, and were born at a high altitude.

A question we consider concerns the long term effects of an environmental change on the systolic blood pressure. In particular, is there a relationship between the systolic blood pressure and how long the Indians lived in their new environment as measured by the fraction of their life spent in the new environment? (fraction = years since migration/age)

A plot of systolic blood pressure against fraction (see the scatterplot in the separate Minitab output) suggests at best a weak linear relationship. Nonetheless, consider fitting the regression model

$$\text{sys bp} = \beta_0 + \beta_1 \text{ fraction} + \epsilon.$$

The least squares line is given by

$$\widehat{\text{sys bp}} = 133 - 15.8 \text{ fraction},$$

and suggests that average systolic blood pressure decreases as the fraction of life spent in modern society increases (if half of life is spent in modern society then that should account for almost an 8 point drop in systolic blood pressure). However, the $t-$test of $H_0 : \beta_1 = 0$ is not significant at the 5% level (p-value=.089). That is, the weak linear relationship observed in the data is not atypical of a population where there is no linear relationship between systolic blood pressure and the fraction of life spent in a modern society (and if we constructed a 95% CI for β_1 it would contain 0).

Even if this test were significant, the small value of $R^2 = 7.6\%$ suggests that fraction does not explain a **substantial** amount of the variation in the systolic blood pressures. If we omit the individual with the highest blood pressure (see the plot) then the relationship would be weaker still. Minitab identifies two unusual observations, one of which is that individual. We should identify the other one too, and see if the fit changes much with those two observations removed. I will show you how to do that later in this section, and Erik will help you with it in lab.

Taking Weight into Consideration

At best, there is a weak relationship between systolic blood pressure and fraction alone. However, it is usually accepted that systolic blood pressure and weight are related; see the scatterplot matrix for confirmation. A natural way to take weight into consideration is to include weight and fraction both as predictors of systolic blood pressure in the multiple regression model:

$$\text{sys bp} = \beta_0 + \beta_1 \text{ fraction} + \beta_2 \text{ weight} + \epsilon.$$

As in simple linear regression, the model is written in the form:

$$\text{Response} = \text{Mean of Response} + \text{Residual},$$

so the model implies that that average systolic blood pressure is a linear combination of fraction and weight. As in simple linear regression, the standard multiple regression analysis assumes that the responses are normally distributed with a constant variance $\sigma^2_{Y|X}$. The parameters of the regression model $\beta_0,\ \beta_1,\ \beta_2$ and $\sigma^2_{Y|X}$ are estimated by LS.

Minitab output from fitting the multiple regression model is given at the end of the handout.

Important Points to Notice About the Regression Output

1. (Parameter Estimates) The LS estimates of the intercept and the regression coefficient for fraction, and their standard errors, change from the simple linear model to the multiple regression model. For the simple linear regression

   $$\widehat{\text{sys bp}} = 133 - 15.8 \text{ fraction}.$$

   For the multiple regression model

   $$\widehat{\text{sys bp}} = 60.9 - 26.8 \text{ fraction} + 1.22 \text{ weight}.$$

2. (ANOVA Table) Comparing the simple linear regression and the multiple regression models we see that the Regression (model) *df* has increased to 2 from 1 (2=number of predictor variables) and the Residual (error) *df* has decreased from 37 to 36 ($=n-1$- number of predictors). Adding a predictor increases the Regression *df* by 1 and decreases the Residual *df* by 1.

3. (ANOVA Table) The Residual SS decreases by 6033.4 - 3441.4 = 2592.0 upon adding the weight term. The Regression SS increased by 2592.0 upon adding the weight term term to the model. The Total SS does not depend on the number of predictors so it stays the same. The Residual SS, or the part of the variation in the response unexplained by the regression model never increases when new predictors are added.

4. The proportion of variation in the response explained by the regression model:

   $$R^2 = \text{Regression SS / Total SS}$$

never decreases when new predictors are added to a model. The R^2 for the simple linear regression with fraction as the only predictor was 7.6%, whereas $R^2 = 47.3\%$ for the multiple regression model. Adding the weight variable to the model increases R^2 by 40%. That is, weight explains an additional 40% of the variation in systolic blood pressure not already explained by fraction alone.

This is not as simple as it sounds. If you use weight as the only predictor, then $R^2 = 27.2\%$ (confirm this), so adding fraction to that model gives $R^2 = 47.3\%$ and fraction can be seen as explaining an additional 20% of variation in systolic blood pressure not already explained by weight alone. This is far more than explained by fraction alone! The two variables together act in a much more interesting manner than a simple sum of individual behavior.

5. (ANOVA table) The estimated variability about the regression line

$$\sqrt{\text{Residual MS}} = s_{Y|X}$$

decreased dramatically after adding the weight effect. For the simple linear regression model (fraction alone) $s_{Y|X} = 12.77$, whereas $s_{Y|X} = 9.78$ for the multiple regression model. This suggests that an important predictor has been added to model.

6. (ANOVA table) The F-statistic for the multiple regression model

$$F_{obs} = \text{Regression MS / Residual MS} = 16.16$$

(which is compared to a F-table with 2 and 36 *df*) tests $H_0 : \beta_1 = \beta_2 = 0$ against $H_A :$ not H_0. This is a test of no relationship between the average systolic blood pressure and fraction and weight, assuming the relationship is linear. If this test is significant than either fraction or weight, or both, are important for explaining the variation in systolic blood pressure. The p-value for this test is 0.

7. (Parameter Estimates Table) Given the model

$$\text{sys bp} = \beta_0 + \beta_1 \text{ fraction} + \beta_2 \text{ weight} + \epsilon,$$

one interest is testing $H_0 : \beta_2 = 0$ against $H_A : \beta_2 \neq 0$. The t-statistic for this test

$$t_{obs} = \frac{b_2 - 0}{SE(b_2)} = \frac{1.2169}{.2337} = 5.21$$

is compared to a $t-$critical value with Residual $df = 36$. Minitab gives a p-value of .000, which suggests $\beta_2 \neq 0$. The t-test of $H_0 : \beta_2 = 0$ in the multiple regression model tests whether adding weight to the simple linear regression model explains a significant part of the variation in systolic blood pressure not explained by fraction. In some sense, the $t-$test of $H_0 : \beta_2 = 0$ will be significant if the increase in R^2 (or decrease in Residual SS) obtained by adding weight to this simple linear regression model is substantial. We saw a big increase in R^2, which is deemed significant by the $t-$test. A similar interpretation is given to the $t-$test for $H_0 : \beta_1 = 0$.

8. The $t-$tests for $\beta_0 = 0$ and $\beta_1 = 0$ are conducted, assessed, and interpreted in the same manner. The p-value for testing $H_0 : \beta_0 = 0$ is .000, whereas the p-value for testing $H_0 : \beta_1 = 0$ is .001. This implies that fraction is important in explaining the variation in systolic blood pressure **after** weight is taken into consideration (by including weight in the model as a predictor).

9. (Seq SS Table) I wish Minitab gave you p-values here. The t-tests above do not depend upon what order terms are entered into the model. This table shows how much of the total SS is accounted for in a sequential manner. If fraction alone is entered into the model, the regression SS is 498.1 (also seen in the simple linear regression model). If weight then is added, an additional 2592.0 is accounted for (as calculated above). These also are known as SAS Type I SS. An F-test is constructed by dividing the sequential SS by residual MS (MSE = 95.96), and the critical value for each test is an F with numerator degrees of freedom 1 and denominator degrees of freedom 36 (residual df). Minitab gives you a value of $F_{crit} = 4.11317$. Those test statistics are $F_s = \frac{498.1}{95.6} = 5.033$ and $F_s = \frac{2592.0}{95.6} = 27.11$ for fraction and weight, respectively. Both are significant. It seems contradictory that fraction is significant here but not in the simple linear regression. The difference is that MSE is much smaller in the 2-predictor model.

10. We could compute CIs for the regression parameters in the usual way: $b_i + t_{crit} SE(b_i)$, where t_{crit} is the $t-$critical value for the corresponding CI level with df = Residual df.

11. Residual plots show no striking problems.

12. Minitab does flag 3 unusual observations (outliers in some sense).

Understanding the Model

The $t-$test for $H_0 : \beta_1 = 0$ is highly significant (p-value=.0007), which implies that fraction is important in explaining the variation in systolic blood pressure **after weight is taken into consideration** (by including weight in the model as a predictor). Weight is called a **suppressor variable**. Ignoring weight suppresses the relationship between systolic blood pressure and fraction.

The implications of this analysis are enormous! Essentially, the correlation between a predictor and a response says very little about the importance of the predictor in a regression model with one or more additional predictors. This conclusion also holds in situations where the correlation is high, in the sense that a predictor that is highly correlated with the response may be unimportant in a multiple regression model once other predictors are included in the model.

Another issue that I wish to address concerns the interpretation of the regression coefficients in a multiple regression model. For our problem, let us first focus on the fraction coefficient in the fitted model

$$\widehat{\text{sys bp}} = 60.89 - 26.76 \text{ fraction} + 1.21 \text{ weight}.$$

The negative coefficient indicates that the predicted systolic blood pressure decreases as fraction increases **holding weight constant**. In particular, the predicted systolic blood pressure decreases by 26.76 for each unit increase in fraction, holding weight constant at any value. Similarly, the predicted systolic blood pressure increases by 1.21 for each unit increase in weight, holding fraction constant at any level.

Considering the effect of unusual observations

Minitab has warned us about a large residual and two influential values, observations 1, 8, and 39. We should make certain that these few values are not driving our entire analysis. I created a new variable named flag, where I assigned a value of 0 to those three observations and a value of 1 to the others. Next I constructed a matrix plot with groups where flag is the group variable. This plots the three special points separately. You can see from the plots that these correspond to the most extreme values of weight and systol.

Under regression options you can enter a variable of weights. I constructed flag with this in mind. A weight of 0 throws the observation out of the analysis, while a weight of 1 keeps it in normally. There other possibilities we will not consider this semester. Using flag as the weight variable throws those three unusual observations out of the analysis, so we can see how much effect they really have.

The values of coefficients are a little different with these values out of the analysis, with an indication that both fraction and weight have smaller effects (though still significant) than before. Most striking is the drop in R^2 from 47.3% to 27.3%. The effects of both fraction and weight are still pronounced, but the three observations are making a big difference in total variability.

Another Multiple Regression Example

The data below are selected from a larger collection of data referring to candidates for the General Certificate of Education (GCE) who were being considered for a special award. Here, Total denotes the candidate's Total mark, out of 1000, in the GCE exam, while Comp is the candidate's score in the compulsory part of the exam, which has a maximum score of 200 of the 1000 points on the exam. SCEL denotes the candidate's score, out of 100, in a School Certificate English Language paper taken on a previous occasion.

```
Data Display
Row  Total  Comp  SCEL
  1    476   111    68
  2    457    92    46
  3    540    90    50
  4    551   107    59
  5    575    98    50
  6    698   150    66
  7    545   118    54
  8    574   110    51
  9    645   117    59
 10    690   114    80
 11    634   130    57
 12    637   118    51
 13    390    91    44
 14    562   118    61
 15    560   109    66
```

A goal here is to compute a multiple regression of the Total score on Comp and SCEL, and make the necessary tests to enable you to comment intelligently on the extent to which current performance in the compulsory test (Comp) may be used to predict aggregate Total performance on the GCE exam. You also want to know whether previous performance in the School Certificate English Language (SCEL) has any predictive value independently of what has already emerged from the current performance in the compulsory papers.

I will lead you through a number of steps to help you answer this question. Let us answer the following straightforward questions based on the Minitab output.

1. Plot Total against Comp and SCEL individually, and comment on the form (i.e. linear, non-linear, logarithmic, etc.), strength, and direction of the relationships.

2. Plot Comp against SCEL and comment on the form, strength, and direction of the relationship.

3. Compute the correlation between all pairs of variables. Do the correlation values appear reasonable, given the plots?

In parts 4 through 9, ignore the possibility that Total, Comp or SCEL might ideally need to be transformed.

4. Which of Comp and SCEL explains a larger proportion of the variation in Total? Which would appear to be a better predictor of Total? (Explain).

5. Consider 2 simple linear regression models for predicting Total one with Comp as a predictor, and the other with SCEL as the predictor. Do Comp and SCEL individually appear to be important for explaining the variation in Total (i.e. test that the slopes of the regression lines are zero). Which, if any, of the output, support, or contradicts, your answer to the previous question?

6. Fit the multiple regression model

$$\text{Total} = \beta_0 + \beta_1 \text{ Comp} + \beta_2 \text{ SCEL} + \epsilon.$$

Test $H_0 : \beta_1 = \beta_2 = 0$ at the 5% level. Describe in words what this test is doing, and what the results mean here.

7. In the multiple regression model, test $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$ individually. Describe in words what these tests are doing, and what the results mean here.

8. How does the R^2 from the multiple regression model compare to the R^2 from the individual simple linear regressions? Is what you are seeing here appear reasonable, given the tests on the individual coefficients?

9. Do your best to answer the question posed above, in the paragraph on page 117 that begins "A goal .... ". Provide an equation (LS) for predicting Total.

Comments on the GCE Analysis

I will give you my thoughts on these data, and how I would attack this problem, keeping the ultimate goal in mind. As a first step, I plot the data and check whether transformations are needed. The plot of Total against COMP is fairly linear, but the trend in the plot of Total against SCEL is less clear. You might see a non-linear trend here, but the relationship is not very strong. When I assess plots I try to not allow a few observations affect my perception of trend, and with this in mind, I do not see any strong evidence at this point to transform any of the variables.

One difficulty that we must face when building a multiple regression model is that these two-dimensional (2D) plots of a response against individual predictors may have little information about the appropriate scales for a multiple regression analysis. In particular, the 2D plots only tell us

whether we need to transform the data in a simple linear regression analysis. If a 2D plot shows a strong non-linear trend, I would do an analysis using the suggested transformations, including any other effects that are important. However, it might be that no variables need to be transformed in the multiple regression model.

Although SCEL appears to be useful as a predictor of Total on its own, the multiple regression output indicates that SCEL does not explain a significant amount of the variation in Total, once the effect of Comp has been taken into account. In particular, the SCEL effect in the multiple regression model is far from significant (p-value=.30). Hence, previous performance in the SCEL exam has little predictive value independently of what has already emerged from the current performance in the compulsory papers (Comp).

What are my conclusions? Given that SCEL is not a useful predictor in the multiple regression model, I would propose a simple linear regression model to predict Total from Comp:

$$\text{Predicted Total} = 128.5 + 3.95\text{Comp}.$$

Output from the fitted model was given earlier. A residual analysis of the model showed no serious deficiencies.

13 Logistic Regression

The data below are from a study conducted by Milicer and Szczotka on pre-teen and teenage girls in Warsaw. The subjects were classified into 25 age categories. The number of girls in each group (sample size) and the number that reached menarche (# RM) at the time of the study were recorded. The age for a group corresponds to the midpoint for the age interval.

Sample size	# RM	Age	Sample size	# RM	Age
376	0	9.21	200	0	10.21
93	0	10.58	106	67	13.33
120	2	10.83	105	81	13.58
90	2	11.08	117	88	13.83
88	5	11.33	98	79	14.08
105	10	11.58	97	90	14.33
111	17	11.83	120	113	14.58
100	16	12.08	102	95	14.83
93	29	12.33	122	117	15.08
100	39	12.58	111	107	15.33
108	51	12.83	94	92	15.58
99	47	13.08	114	112	15.83
			1049	1049	17.58

The researchers were interested in whether the proportion of girls that reached menarche (# RM/ sample size) varied with age. One could perform a test of homogeneity by arranging the data as a 2 by 25 contingency table with columns indexed by age and two rows: ROW1 = # RM and ROW2 = # that have not RM = sample size − # RM. A more powerful approach treats these as regression data, using the proportion of girls reaching menarche as the "response" and age as a predictor.

A plot of the observed proportion of girls that have reached menarche (labelled Proportion on page 1 of the Minitab output) shows that the proportion increases as age increases, but that the relationship is nonlinear. This is reinforced by the **Lowess smoother** superimposed on the data plot. The plot and smoother are described in the output.

The observed proportions, which are bounded between zero and one, have a lazy S-shape (a **sigmoidal function**) when plotted against age. The change in the observed proportions for a given change in age is much smaller when the proportion is near 0 or 1 than when the proportion is near 1/2. This phenomenon is common with regression data where the response is a proportion.

The trend is nonlinear so linear regression is inappropriate. A sensible alternative might be to transform the response or the predictor to achieve near linearity. A better approach is to use a non-linear model for the proportions. A common choice is the **logistic regression model**.

The Simple Logistic Regression Model

The simple logistic regression model expresses the population proportion p of individuals with a given attribute (called a success) as a function of a single predictor variable X. The model assumes that p is related to X through

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X \tag{1}$$

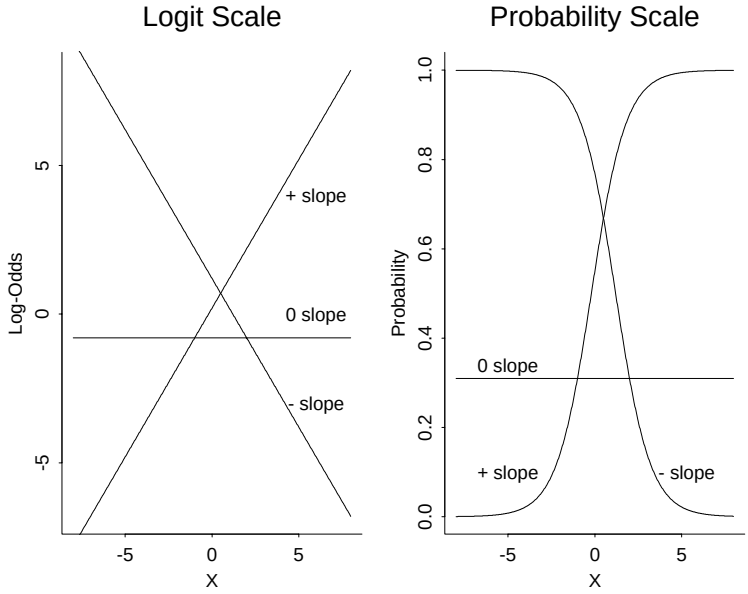
or, equivalently, as

$$p = \frac{exp(\alpha + \beta X)}{1 + exp(\alpha + \beta X)}.$$

The logistic regression model is a **binary response model**, where the response for each case falls into one of 2 exclusive and exhaustive categories, often called success (cases with the attribute of interest) and failure (cases without the attribute of interest). In many biostatistical applications, the success category is presence of a disease, or death from a disease.

I will often write p as $p(X)$ to emphasize that p is the proportion of all individuals with score X that have the attribute of interest. In the menarche data, $p = p(X)$ is the population proportion of girls at age X that have reached menarche.

The odds of success are $p/(1-p)$. For example, the odds of success are 1 (or 1 to 1) when $p = 1/2$. The odds of success are 9 (or 9 to 1) when $p = .9$. The logistic model assumes that the log-odds of success is linearly related to X. Graphs of the logistic model relating p to X are given below. The sign of the slope refers to the sign of β. A corresponding plot for the menarche data appears in the Minitab output.



There are a variety of other binary response models that are used in practice. The **probit** regression model or the **complementary log-log** regression model might be appropriate when the logistic model does not fit the data.

Data for Simple Logistic Regression

For the formulas below, I assume that the data is given in summarized or **aggregate** form:

X	n	D
X_1	n_1	d_1
X_2	n_2	d_2
.	.	.
.	.	.
X_m	n_m	d_m

where d_i is the number of individuals with the attribute of interest (number of diseased) among n_i randomly selected or representative individuals with predictor variable value X_i. The subscripts identify the group of cases in the data set. In many situations, the sample size is 1 in each group, and for this situation d_i is 0 or 1. There are four different forms in which Minitab accepts this type of data - I discuss this in the separate Minitab output. The preceding format is the one used in the analysis.

Estimating Regression Coefficients

The principle of maximum likelihood is commonly used to estimate the two unknown parameters in the logistic model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X.$$

The **maximum likelihood estimates** (MLE) of the regression coefficients are estimated iteratively by maximizing the so-called Binomial likelihood function for the responses, or equivalently, by minimizing the **deviance** function (also called the likelihood ratio LR chi-squared statistic)

$$\text{LR} = 2\sum_{i=1}^{m}\left\{d_i\log\left(\frac{d_i}{n_i p_i}\right) + (n_i - d_i)\log\left(\frac{n_i - d_i}{n_i - n_i p_i}\right)\right\}$$

over all possible values of α and β, where the p_is satisfy

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i.$$

The ML method also gives standard errors and significance tests for the regression estimates.

The deviance is an analog of the residual sums of squares in linear regression. The choices for α and β that minimize the deviance are the parameter values that make the observed and fitted proportions as close together as possible in a "likelihood sense".

Suppose that $\hat{\alpha}$ and $\hat{\beta}$ are the MLEs of α and β. The deviance evaluated at the MLEs:

$$\text{LR} = 2\sum_{i=1}^{m}\left\{d_i\log\left(\frac{d_i}{n_i\tilde{p}_i}\right) + (n_i - d_i)\log\left(\frac{n_i - d_i}{n_i - n_i\tilde{p}_i}\right)\right\},$$

where the fitted probabilities $\tilde{p}_i$ satisfy

$$\log\left(\frac{\tilde{p}_i}{1-\tilde{p}_i}\right) = \hat{\alpha} + \hat{\beta}X_i,$$

is used to test the adequacy of the model. The deviance is small when the data fits the model, that is, when the observed and fitted proportions are close together. Large values of LR occur when

one or more of the observed and fitted proportions are far apart, which suggests that the model is inappropriate.

If the logistic model holds, then LR has a chi-squared distribution with $m-r$ degrees of freedom, where m is the number of groups and r (here 2) is the number of estimated regression parameters. A p-value for the deviance is given by the area under the chi-squared curve to the right of LR. A small p-value indicates that the data does not fit the model.

Age at Menarche Data: Minitab Implementation

A logistic model for these data implies that the probability p of reaching menarche is related to age through

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \text{AGE}.$$

If the model holds, then a slope of $\beta = 0$ implies that p does not depend on AGE, i.e. the proportion of girls that have reached menarche is identical across age groups. However, the power of the logistic regression model is that if the model holds, and if the proportions change with age, then you have a way to quantify the effect of age on the proportion reaching menarche. This is more appealing and useful than just testing homogeneity across age groups.

A logistic regression model is fit by following the path `Stat > Regression > Binary Logistic Regression`. I discuss the various options for entering the data on the separate Minitab output. Minitab is a lot more flexible about structuring the data for this procedure than are most packages. There also are available ordinal and nominal logistic regression to handle cases with more than two response categories.

The **Logistic Regression Table** gives the MLEs of the parameters: $\hat{\alpha} = -21.23$ and $\hat{\beta} = 1.63$. Thus, the fitted or predicted probabilities satisfy:

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = -21.23 + 1.63\text{AGE}$$

or

$$\tilde{p}(AGE) = \frac{\exp(-21.23 + 1.63\text{AGE})}{1 + \exp(-21.23 + 1.63\text{AGE})}.$$

The p-value for testing $H_0 : \beta = 0$ (i.e. the slope for the regression model is zero) based upon the Z-test in the **Logistic Regression Table** is 0 (the area outside ± 27.68 in a standard normal distribution is 0), which leads to rejecting H_0 at any of the usual test levels. Thus, the proportion of girls that have reached menarche is not constant across age groups.

The **Goodness-of-Fit Tests** table gives the deviance chi-square statistic as 26.70 on 23 df, with a p-value of .269. The large p-value suggests no gross deficiencies with the logistic model. The Pearson and Hosmer-Lemeshow tests are also checks on model fit.

The **Test that all slopes are zero** gives the logistic regression analog of the F-test for the model in multiple regression. In general, the chi-squared statistic provided here is used to test the hypothesis that the regression coefficients are zero for each predictor in the model. There is a single predictor here, AGE, so this test and the test for the AGE effect in the **Logistic Regression Table** are both testing $H_0 : \beta = 0$. This test is not just the square of the Z-test for Age, however.

Probably the most commonly reported part of the output is the odds ratio in the Logistic Regression Table. In order to understand that we need to review some properties of logs and exponentials.

1. If $y = \log x$ (natural log) then $e^y = x$, i.e. $e^{\log(x)} = x$ where $e = 2.71828\ldots$.
2. $\log e^y = y$.
3. $e^a e^b = e^{a+b}$.
4. $\frac{e^a}{e^b} = e^{a-b}$.
5. $\log(ab) = \log(a) + \log(b)$ and $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

Now consider the odds of reaching menarche for a given value of Age (any given value) vs. one year older (Age + 1). Minitab's estimated log odds of reaching menarche at the given value of Age is $\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = -21.23 + 1.63\text{AGE}$. The estimated log odds at Age + 1 then is $-21.23 + 1.63(\text{AGE} + 1)$. Now the estimated log of the odds ratio at Age + 1 vs. Age is $\log\left(\frac{\text{Odds at Age+1}}{\text{Odds at Age}}\right) = \log(\text{Odds at Age+1}) - \log(\text{Odds at Age}) = \{-21.23 + 1.63(\text{AGE}+1)\} - \{-21.23 + 1.63(\text{AGE})\} = 1.63$. If the log of the odds ratio is 1.63, then the odds ratio is $e^{1.63} = 5.11$, which is the value reported in **Logistic Regression Table**. The estimated odds of RM for a 15-year old is 5.11 that of a 14-year old, that for a 16-year old 5.11 that of a 15-year old, etc. If $\hat{\beta}$ is the estimate of a coefficient in the logit scale, then the odds ratio for a one unit change in the associated predictor variable is $e^{\hat{\beta}}$. The 95% CI reported by Minitab is obtained by first computing $\hat{\beta} \pm 1.96\ SE$ and exponentiating the endpoints. The odds for a 2 unit change is 5.11^2, by an identical derivation.

Logistic Regression with Two Effects: Leukemia Data

Feigl and Zelen reported the survival time in weeks and the white cell blood count (WBC) at time of diagnosis for 33 patients who eventually died of acute leukemia. Each person was classified as AG+ or AG- (coded as IAG = 1 and 0, respectively), indicating the presence or absence of a certain morphological characteristic in the white cells. The researchers are interested in modelling the probability p of surviving at least one year as a function of WBC and IAG. They believe that WBC should be transformed to a log scale, given the skewness in the WBC values.

As an initial step in the analysis, consider the following model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 \text{IAG},$$

where LWBC = log WBC. This is a logistic regression model with 2 effects both of which must be entered in the model portion of the dialog box. The parameters α, β_1 and β_2 are estimated by maximum likelihood.

The model is best understood by separating the AG+ and AG- cases. For AG- individuals, IAG=0 so the model reduces to

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 * 0 = \alpha + \beta_1 \text{LWBC}.$$

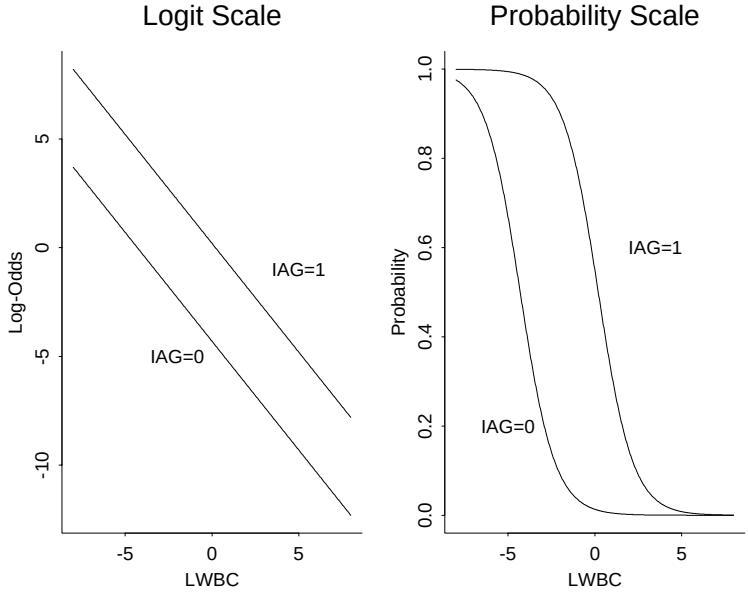
For AG+ individuals, IAG=1 and the model implies

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 * 1 = (\alpha + \beta_2) + \beta_1 \text{LWBC}.$$

The model without IAG (i.e. $\beta_2 = 0$) is a simple logistic model where the log-odds of surviving one year is linearly related to LWBC, and is independent of AG. The reduced model with $\beta_2 = 0$ implies that there is no effect of the AG level on the survival probability once LWBC has been taken into account.

Including the **binary predictor** IAG in the model implies that there is a linear relationship between the log-odds of surviving one year and LWBC, with a constant slope for the two AG levels. This model includes an effect for the AG morphological factor, but more general models are possible. Thinking of IAG as a **factor**, the proposed model is a logistic regression analog of ANCOVA.

The parameters are easily interpreted: α and $\alpha + \beta_2$ are intercepts for the population logistic regression lines for AG- and AG+, respectively. The lines have a common slope, β_1. The β_2 coefficient for the IAG indicator is the difference between intercepts for the AG+ and AG- regression lines. A picture of the assumed relationship is given below for $\beta_1 < 0$. The population regression lines are parallel on the logit (i.e. log odds) scale only, but the order between IAG groups is preserved on the probability scale.



The Minitab worksheet contains **raw data** for individual cases. There are four columns: the binary or **indicator variable** IAG (with value 1 for AG+, 0 for AG-), WBC (continuous), LIVE (with value 1 if the patient lived at least 1 year and 0 if not), and Log WBC (natural log of WBC). Note that a frequency column is not needed with raw data and that the success category corresponds to surviving at least 1 year.

Before looking at output for the equal slopes model, note that the data set has 30 distinct IAG and LWBC combinations, or 30 "groups" or samples that could be constructed from the 33 individual cases. Only two samples have more than 1 observation. The majority of the observed proportions surviving at least one year (number surviving $\geq$ 1 year/ group sample size) are 0 (i.e. 0/1) or 1 (i.e. 1/1). This sparseness of the data makes it difficult to graphically assess the suitability of the logistic model (Why?). Although significance tests on the regression coefficients

do not require large group sizes, the chi-squared approximation to the deviance is suspect in sparse data settings. With small group sizes as we have here, most researchers would not interpret the p-values for the deviance literally. Instead, they would use the p-values to informally check the fit of the model. Diagnostics would be used to highlight problems with the model.

The large p-value (.684) for the lack-of-fit chi-square (i.e. the deviance) indicates that there are no gross deficiencies with the model. Given that the model fits reasonably well, a test of $H_0 : \beta_2 = 0$ might be a primary interest here. This checks whether the regression lines are identical for the two AG levels, which is a test for whether AG affects the survival probability, after taking LWBC into account. The p-value for this test is .021. The test is rejected at any of the usual significance levels, suggesting that the AG level affects the survival probability (assuming a very specific model).

The estimated survival probabilities satisfy

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52\text{IAG}.$$

For AG- individuals with IAG=0, this reduces to

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC},$$

or equivalently,

$$\tilde{p} = \frac{\exp(5.54 - 1.11\text{LWBC})}{1 + \exp(5.54 - 1.11\text{LWBC})}.$$

For AG+ individuals with IAG=1,

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52 * (1) = 8.06 - 1.11\text{LWBC},$$

or

$$\tilde{p} = \frac{\exp(8.06 - 1.11\text{LWBC})}{1 + \exp(8.06 - 1.11\text{LWBC})}.$$

Using the **logit scale**, the difference between AG+ and AG- individuals in the estimated log-odds of surviving at least one year, at a fixed but arbitrary LWBC, is the estimated IAG regression coefficient:

$$(8.06 - 1.11\text{LWBC}) - (5.54 - 1.11\text{LWBC}) = 2.52.$$

Using properties of exponential functions, the odds that an AG+ patient lives at least one year is $\exp(2.52) = 12.42$ times larger than the odds that an AG- patient lives at least one year, regardless of LWBC.

Although the equal slopes model appears to fit well, a more general model might fit better. A natural generalization here would be to add an **interaction**, or product term, IAG $*$ LWBC to the model. The logistic model with an IAG effect and the IAG $*$ LWBC interaction is equivalent to fitting separate logistic regression lines to the two AG groups. This interaction model provides an easy way to test whether the slopes are equal across AG levels. I will note that the interaction term is not needed here.

14 Introduction to Survival Analysis

In many biomedical studies, the outcome variable is a survival time, or more generally a time to an event. We will describe some of the standard tools for analyzing survival data.

Most studies of survival last a few years, and at completion many subjects may still be alive. For those individuals, the actual survival time is not known – all we know is how long they survived from their entry in the study. Similarly, certain individuals may drop out from the study or be lost to follow-up. Each of these cases is said to be *censored*, and the recorded time for such individuals is their time until the censoring event.

Example: HPA staining for breast cancer survival

We consider data from a retrospective study of 45 women who had surgery for breast cancer. Tumor cells, surgically removed from each woman, were classified according to the results of staining on a marker taken from the Roman snail, the *Helix pomatia agglutinin* (HPA). The marker binds to cancer cells associated with metastasis to nearby lymph nodes. Upon microscopic examination, the cancer cells stained with HPA are classified as positive, corresponding to a tumor with the potential for metastasis, or negative. It is of interest to determine the relationship of HPA staining and the survival of women with breast cancer.

The survival times in months T_i and staining results ($x_i = 0$ for negative and $x_i = 1$ for positive) for the 45 women are presented in the following table. Also included is a *censoring indicator* d_i. Contrary to the normal definition of an indicator variable, the censoring indicator is zero if the observation is right-censored, and one if the observation is uncensored. So it's really a *non-censoring* indicator! A woman's survival time was right censored if the woman was alive at the end of the study or if the woman died of causes unrelated to breast cancer.

T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d	T	x	d
23	0	1	47	0	1	69	0	1	70	0	0	71	0	0	100	0	0	101	0	0	148	0	1	181	0	1	198	0	0
208	0	0	212	0	0	224	0	0	5	1	1	8	1	1	10	1	1	13	1	1	18	1	1	24	1	1	26	1	1
26	1	1	31	1	1	35	1	1	40	1	1	41	1	1	48	1	1	50	1	1	59	1	1	61	1	1	68	1	1
71	1	1	76	1	0	105	1	0	107	1	0	109	1	0	113	1	1	116	1	0	118	1	1	143	1	1			
154	1	0	162	1	0	188	1	0	212	1	0	217	1	0	225	1	0												

This is the general format the data should be in to work with it in packages like **Minitab** and **Stata**, though Minitab is flexible about the actual censoring indicator you use. Succinctly, the *sorted* survival times for the negative stained women are

$$23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*,$$

where $*$ denotes a right-censored observation. The survival times for the positive stained group are

$$5, 8, 10, 13, 18, 24, 26, 26, 31, 35, 40, 41, 48, 50, 59, 61, 68, 71, 76^*, 105^*,$$

$$107^*, 109^*, 113, 116^*, 118, 143, 154^*, 162^*, 188^*, 212^*, 217^*, 225^*.$$

In the breast cancer study, 8 individuals in the negative stained group, and 11 in the positive stained group are censored. Although it is common for studies to have *right-censored* cases, such as we have here, left-censoring and interval-censoring are found in other clinical studies.

Survival Curves

A first step in survival analysis is often to estimate the *survival curve*, or *survival time distribution*. Suppose we are considering a single (homogeneous) population. Let T be the survival time (from some reference point) for a randomly selected individual from the population. Where t is any arbitrary positive value, the survival time distribution is defined to be

$$\begin{aligned} S(t) &= Pr(T \geq t) \\ &= \text{probability randomly selected individual survives at least until time } t \\ &= \text{proportion of population that survives at least until time } t. \end{aligned}$$

The function might look like Figure 1.

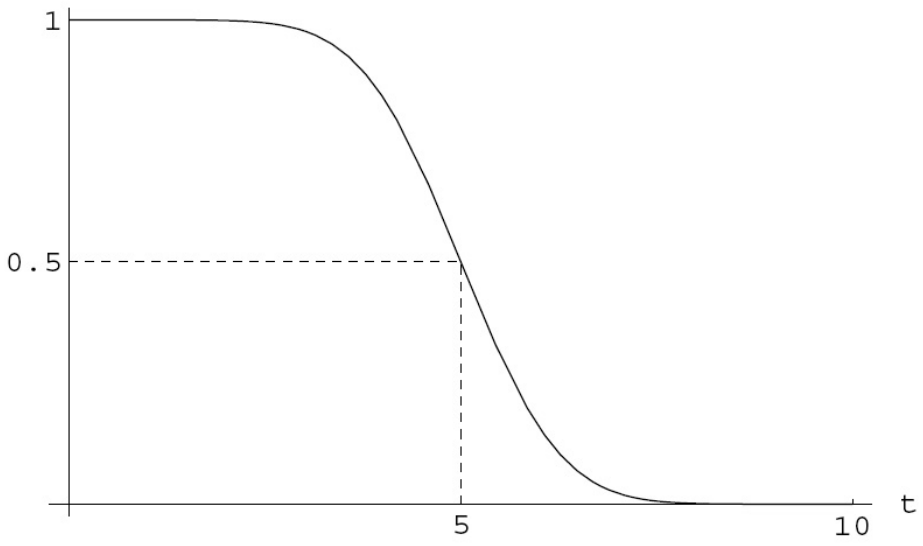


Figure 1: $S(t)$ versus t; median survival time for population is 5.

Estimating the Survival Curve

Case I: No censoring

If we have a random sample from the population, we use the *empirical survival function*:

$$\hat{S}(t) = \text{sample proportion that survive at least until time t}$$

to estimate $S(t)$. This is easy to compute and plot as a function of t.

Suppose we have a sample of 5 survival times (in days): 5, 8, 20, 30, and 33. $\hat{S}(t)$ has "jumps" of size 1/5 (i.e. 1 divided by the sample size) at each survival time; see Figure 2.

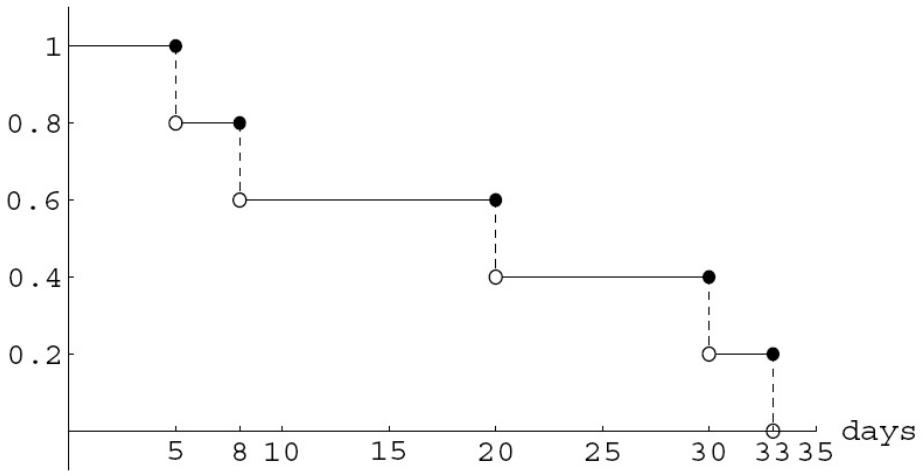


Figure 2: Empirical survival function $\hat{S}(t)$ for the data 5, 8, 20, 30, and 33.

Case II: Right censoring

Recall the data on the survival of women with breast cancer whose cells were negatively stained with HPA:

$$23, 47, 69, 70^*, 71^*, 100^*, 101^*, 148, 181, 198^*, 208^*, 212^*, 224^*,$$

where the $*$ superscript identifies a right-censored observation.

The following algorithm describes the Kaplan-Meier (KM) method for estimating the survival curve (Kaplan-Meier product-limit estimate).

1. Identify times for non-censored cases $0 = t_0 < t_1 < t_2 < \cdots < t_r$. That is, t_1 is the smallest non-censored survival time, t_2 is the second smallest, et cetera. For the example $r = 5$ and $t_0 = 0$, $t_1 = 23$, $t_2 = 47$, $t_3 = 69$, $t_4 = 148$, and $t_5 = 181$.

2. For the j^{th} interval, where $t_{j-1} \leq t < t_j$, evaluate

$$\begin{aligned} n_j &= \text{number at risk (of dying) at beginning of interval,} \\ d_j &= \text{number of deaths in interval,} \\ \frac{n_j - d_j}{n_j} &= \text{estimated probability of surviving past } t_{j-1}, \\ &\quad \text{given you are at risk at time } t_{j-1} \\ &= \hat{P}(T \geq t_{j-1} | T \geq t_{j-2}). \end{aligned}$$

3. For $t_{j-1} \leq t < t_j$,

$$\begin{aligned}\hat{S}(t) &= \hat{P}(T \geq t) \\ &= \hat{P}(T \geq t_{j-1} | T \geq t_{j-2}) \times \\ &\quad \hat{P}(T \geq t_{j-2} | T \geq t_{j-3}) \times \cdots \times \\ &\quad \hat{P}(T \geq t_1 | T \geq t_0) \\ &= \frac{n_j - d_j}{n_j} \times \frac{n_{j-1} - d_{j-1}}{n_{j-1}} \times \cdots \times \frac{n_1 - d_1}{n_1}.\end{aligned}$$

Remark: Censored observations are taken into account by being treated as cases at risk at the beginning of the interval in which they fail.

To illustrate the calculation for our data, consider the table:

j	Interval	n_j	d_j	$\frac{n_j - d_j}{n_j}$	$\hat{S}(t)$
1	$0 \leq t < 23$	13	0	$\frac{13-0}{13} = 1$	1.0
2	$23 \leq t < 47$	13	1	$\frac{13-1}{13} = \frac{12}{13} \doteq 0.923$	$1.0 \times 0.923 = 0.923$
3	$47 \leq t < 69$	12	1	$\frac{12-1}{12} = \frac{11}{12} \doteq 0.917$	$0.923 \times 0.917 = 0.846$
4	$69 \leq t < 148$	11	1	$\frac{10}{11} \doteq 0.909$	$0.846 \times 0.909 = 0.769$
5	$148 \leq t < 181$	6	1	$\frac{5}{6} \doteq 0.833$	$0.769 \times 0.833 = 0.641$
6	$181 \leq t$	5	1	$\frac{4}{5} = 0.8$	$0.641 \times 0.8 = 0.513$

To obtain the KM estimate in **Minitab**, follow the path `Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Nonparametric Distribution Analysis`. Enter the failure time variable in `Variables`, check the `By variable` and enter group, click on `Censor` and enter the censoring variable and value, on `Estimate` enter Kaplan-Meier, on `Graphs` check Survival Plot, and ask for full results. In Figure 3 we have a picture of $\hat{S}(t)$ from the negatively stained group as well as the estimate from the positively stained group. Note that the negatively stained group tends to live longer, as we would expect. Output follows. We will discuss this in class.

```
Distribution Analysis: time by group

Variable: time
group = 0

Censoring Information  Count
Uncensored value           5
Right censored value       8

Censoring value: cens = 0


Nonparametric Estimates
```

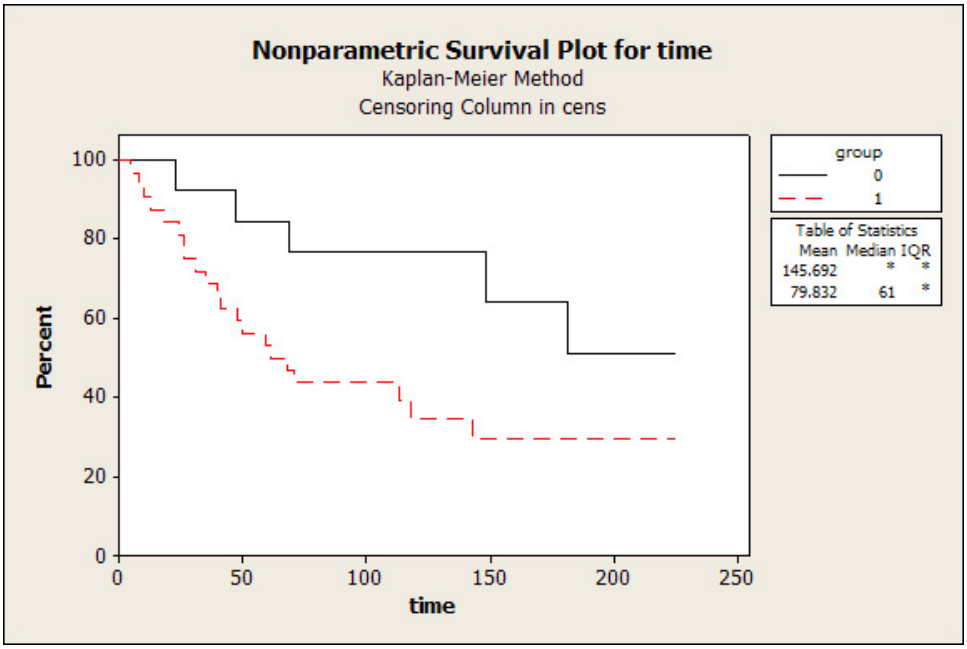


Figure 3: KM survival curves for positively and negatively stained groups.

```
Characteristics of Variable

           Standard   95.0% Normal CI
Mean(MTTF)    Error    Lower    Upper
  145.692  17.6423  111.114  180.271

Median = * IQR = *  Q1 = 148  Q3 = *


Kaplan-Meier Estimates

      Number
          at  Number     Survival  Standard   95.0% Normal CI
Time    Risk  Failed  Probability     Error     Lower    Upper
  23      13       1     0.923077  0.073905  0.778225  1.00000
  47      12       1     0.846154  0.100068  0.650024  1.00000
  69      11       1     0.769231  0.116855  0.540200  0.99826
 148       6       1     0.641026  0.152249  0.342623  0.93943
 181       5       1     0.512821  0.167285  0.184948  0.84069


Empirical Hazard Function

        Hazard
Time  Estimates
  23   0.076923
  47   0.083333
  69   0.090909
 148   0.166667
 181   0.200000


Distribution Analysis: time by group
```

Variable: time group = 1

Censoring Information	Count
Uncensored value	21
Right censored value	11

Censoring value: cens = 0

Nonparametric Estimates

Characteristics of Variable

Mean(MTTF)	Standard Error	95.0% Normal CI Lower	Upper
79.8320	9.70863	60.8035	98.8606

Median = 61 IQR = * Q1 = 26 Q3 = *

Kaplan-Meier Estimates

Time	Number at Risk	Number Failed	Survival Probability	Standard Error	95.0% Normal CI Lower	Upper
5	32	1	0.968750	0.0307578	0.908466	1.00000
8	31	1	0.937500	0.0427908	0.853632	1.00000
10	30	1	0.906250	0.0515270	0.805259	1.00000
13	29	1	0.875000	0.0584634	0.760414	0.98959
18	28	1	0.843750	0.0641862	0.717947	0.96955
24	27	1	0.812500	0.0689981	0.677266	0.94773
26	26	2	0.750000	0.0765466	0.599972	0.90003
31	24	1	0.718750	0.0794804	0.562971	0.87453
35	23	1	0.687500	0.0819382	0.526904	0.84810
40	22	1	0.656250	0.0839617	0.491688	0.82081
41	21	1	0.625000	0.0855816	0.457263	0.79274
48	20	1	0.593750	0.0868207	0.423584	0.76392
50	19	1	0.562500	0.0876951	0.390621	0.73438
59	18	1	0.531250	0.0882155	0.358351	0.70415
61	17	1	0.500000	0.0883883	0.326762	0.67324
68	16	1	0.468750	0.0882155	0.295851	0.64165
71	15	1	0.437500	0.0876951	0.265621	0.60938
113	10	1	0.393750	0.0891735	0.218973	0.56853
118	8	1	0.344531	0.0905972	0.166964	0.52210
143	7	1	0.295313	0.0900371	0.118843	0.47178

Empirical Hazard Function

Time	Hazard Estimates
5	0.031250
8	0.032258
10	0.033333
13	0.034483
18	0.035714
24	0.037037
26	0.040000
31	0.041667
35	0.043478
40	0.045455
41	0.047619
48	0.050000
50	0.052632
59	0.055556
61	0.058824
68	0.062500
71	0.066667

```
 113   0.100000
 118   0.125000
 143   0.142857

Distribution Analysis: time by group

Comparison of Survival Curves


Log-Rank Statistic

Variable         1         2
          -4.56513   4.56513


Variance/Covariance of Log-Rank Statistic

Variable         1         2
       1   5.92900  -5.92900
       2  -5.92900   5.92900


Wilcoxon Statistic

Variable     1    2
          -159  159


Variance/Covariance of Wilcoxon Statistic

Variable         1         2
       1   6048.14  -6048.14
       2  -6048.14   6048.14


Test Statistics

Method    Chi-Square  DF  P-Value
Log-Rank     3.51499   1    0.061
Wilcoxon     4.17997   1    0.041
```

Some remarks:

- The estimated survival curve "drops to zero" only if the last case is not censored.
- The KM curve allows us to estimate percentiles of the survival distribution, with a primary interest being the median survival time (50^{th} percentile). In the example above, the 90^{th} percentile is approximately 47 months (i.e. we estimate that 90% of the population will survive at least 47 months). The median cannot be estimated here – all we can say is that we estimate the median to be at least 181 months.
- The KM estimate is the usual empirical estimate if no cases are censored.
- Statistical methods are available to
  - Estimate the mean survival time.
  - Get a C.I. for the survival curve.
  - Compare survival curves across groups – you can think of this as the censored data analogue of (non-parametric) ANOVA.

The Cox Proportional Hazards Model

Note: Minitab does not do this. I have included old notes from Stata. This is a way to use regression methods similar to those used in logistic regression. Minitab includes methods more common in engineering; proportional hazards methods are more common in biostatistics.

The risk of failing at time t is defined to be the probability of an individual dying in the "next instant" (e.g. in a time frame of length Δ) given this individual has survived at least until time t:

$$P(t \leq T < t + \Delta | t \leq T).$$

We define the *hazard function* $h(t)$ such that for small enough Δ,

$$P(t \leq T < t + \Delta | t \leq T) = h(t)\Delta.$$

The hazard function is proportional to the instantaneous "risk of failing" at any time t, given that an individual has lived at least to time t.

Now consider two individuals, 1 and 2, each with their own hazard functions $h_1(t)$ and $h_2(t)$. If we assume that one individual's instantaneous rate of failing is a constant multiple of the other's, i.e. $h_2(t) = ah_1(t)$ for some constant a, then these two individuals have *proportional hazard functions*. Figure 4 shows an example of this phenomenon where the hazard ratio is 1/2.

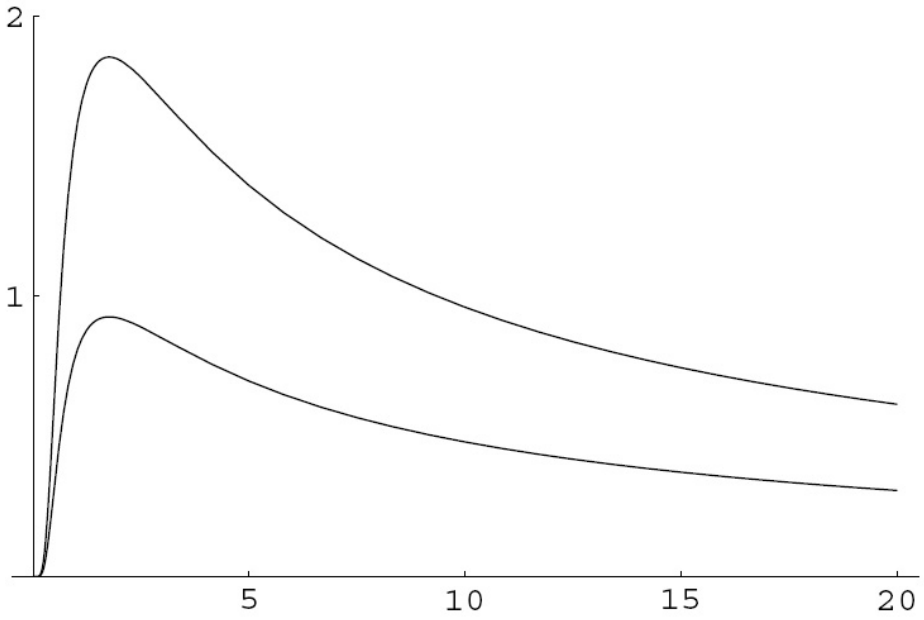


Figure 4: An example of proportional hazard functions; here the constant of proportionality is 0.5.

Proportional hazards may or may not be a reasonable assumption to make. For example, consider two people, roughly the same age and demographic except that at the age of 20, person 2 takes up smoking while person 1 does not. You will hopefully agree with me that initially, the smoker and the non-smoker will most likely have *identical* hazards. As the years roll by, and

smoking takes its toll, we would think that the smoker's instantaneous rate of failing, which is proportional to the probability of dying in the next minute, say, will increase relative to the hazard for the non-smoker. In this example proportional hazards probably is an unreasonable assumption.

The proportional hazards *model* generalizes the above concept for n individuals, each with their own covariate value x_i or set of p covariate values $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. In the case where the n individuals only have one covariate, the model stipulates for individuals i and j, with a hazard functions $h_i(t)$ and $h_j(t)$ respectively, that

$$h_i(t)e^{-\beta x_i} = h_j(t)e^{-\beta x_j}.$$

Note that this implies

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta x_i}}{e^{\beta x_j}} = e^{\beta(x_i - x_j)}.$$

Here, $e^{\beta(x_i-x_j)}$ is the relative risk of instantaneous failure at *any time* t for individuals i and j. That is the power of the proportional hazards assumption: the relative risk of dying for two individuals is a simple function of the model parameters and holds for all t, independent of the value of t. If individual i has covariate value $x+1$ and individual j has covariate value x, i.e. their covariate values only differ by 1 unit on the covariate measurement scale, then

$$\frac{h_i(t)}{h_j(t)} = \frac{e^{\beta(x+1)}}{e^{\beta x}} = e^{\beta}.$$

Thus, e^{β} is the relative risk of failing in the next instant when we increase the covariate by one unit. Note that if x_i is a simple zero/one variable denoting which group individual i falls into, then e^{β} is the relative risk of failing in the next instant for the group denoted by $x_i = 1$ versus $x_i = 0$.

The breast cancer data are loaded with the commands `infile time group cens using c:/breast.txt` and the Cox PH model is fit via `cox time group, dead(cens)`. The survival time, followed by the predictor variable(s) is specified. The non-censoring indicator is included in the subcommand `dead`. We obtain the following output:

```
------------------------------------------------------------------------------
        time |
        cens |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .9080157   .5009228     1.81   0.070    -.0737749    1.889806
------------------------------------------------------------------------------
```

We have an estimate of $\hat{\beta} = 0.908$ and the estimated relative risk is $e^{\hat{\beta}} = e^{0.908} \stackrel{\bullet}{=} 2.5$. That is, those with positive staining are estimated to have a risk of dying in the next instant about 2.5 times as great as those with negative staining. Note that the p-value for $H_0 : \beta = 0$ is small but not significant at the 5% level. There is definitely *some* indication that staining affects survival, with positive staining decreasing survival. A 95% C.I. for the risk may be obtained by exponentiating the endpoints for the C.I. for β. Here, we estimate the relative risk of expiring (for positive compared to negative staining) to be within $(e^{-0.073}, e^{1.89}) = (0.93, 6.62)$ with 95% confidence.

Remark: The hazard function for individual i can be defined to be a scale multiple $e^{x_i\beta}$ of a *baseline* hazard function denoted $h_0(t)$. The model may be recast as $h_i(t) = h(t|x_i) = e^{x_i\beta}h_0(t)$. This baseline hazard function $h_0(t)$ and β thus completely determine the model. The baseline hazard $h_0(t)$ may be estimated from the data as well as survival curves, median and mean survival, et cetera, for any covariate value x. These sorts of inferences are quite easy to get out of **Stata** but a bit beyond what is comfortable to cover in this class.

A final example

We examine a data set consisting of the time spent running on a treadmill for 14 people aged 15 and older. Each subject's gender and age were recorded. It is of interest to the experimenter how age and gender affects ones endurance.

We define a numeric indicator variable for the gender variable by taking g to be 0 for a male subject and 1 for a female subject. When fitting the PH model with gender and age as main effects,

$$h(t|age, g) = e^{age\beta_1 + g\beta_2} h_0(t),$$

the baseline group (i.e. those with covariates $age = 0$ and $g = 0$, and thus a hazard function of $e^{0\beta_1 + 0\beta_2} = e^0 h_0(t) = h_0(t)$) consists of males of age zero, which is not interpretable in this context. Observations were censored due to a subject having to leave the treadmill for reasons other than being tired. The data follow:

Obs	gender	age	minutes	cens	weight	g
1	male	34	16	1	215	0
2	male	15	35	0	135	0
3	female	22	55	0	145	1
4	female	18	95	1	97	1
5	male	18	55	0	225	0
6	female	32	55	1	185	1
7	female	37	25	1	155	1
8	female	67	15	1	142	1
9	female	55	22	1	132	1
10	male	55	13	1	183	0
11	male	62	13	1	168	0
12	female	33	57	0	132	1
13	female	17	52	0	112	1
14	male	24	54	1	175	0

The fit of the model with only gender $h(t|g) = e^{g\beta_1} h_0(t)$:

minutes	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
g	-.6786811	.7161483	-0.95	0.343	-2.082306	.7249439

The test for a gender effect yields a p-value of 0.343. We would accept at any reasonable significance level that there is not a gender effect. The estimate of β_1 is $\hat{\beta}_1 = -0.679$ so the fitted model is $h(t|g) = e^{-0.678g} h_0(t)$ implying that $h(t|g = 1) = 0.507 h(t|g = 0)$ and finally $h(t|g = 1)/h(t|g = 0) = 0.507$ for all t. That is, the probability of a randomly picked woman failing (stepping off the treadmill) in the next second is estimated be half the probability of a randomly picked male.

Rephrased, we see that, assuming proportional hazards is reasonable, females are about half (the hazard ratio is $e^{-0.679} = 0.507$) as likely to step off the treadmill at any instant versus males. We obtain an approximate 95% C.I. for this ratio by first considering the 95% C.I. for the regression effect: (-2.08, 0.72). Exponentiate both endpoints to obtain a 95% C.I. for the hazard ratio: (0.12, 2.07). The hazard ratio interval includes one (no difference in the hazard functions for males and females) because the regression effect interval includes zero.

Let's look at the model fit with only age $h(t|age) = e^{age\beta_1} h_0(t)$:

minutes	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.1116606	.0385688	2.90	0.004	.0360672	.187254

A year from now, a randomly selected individual will be $e^{0.1117} = 1.118$ times as likely to step off the treadmill after 15 minutes (or any amount of time) than now. In ten years it will be $1.118^{10} = 3.05$ times as likely. When we fit the model with both of these predictors $h(t|age, g) = e^{age\beta_1 + g\beta_2} h_0(t) = e^{age\beta_1} e^{g\beta_2} h_0(t)$ we see that estimated regression effects, and therefore model interpretation, change somewhat:

```
------------------------------------------------------------------------------
     minutes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |  -3.551859    1.57856    -2.25   0.024    -6.645779   -.4579388
         age |   .2186267   .0855601     2.56   0.011      .050932    .3863214
------------------------------------------------------------------------------
```

At a given age, a random male running alongside a random female is about $1/e^{-3.55} = 1/0.029 = 35$ times as likely to step off the treadmill at any time. A woman 20 years older than another woman is about $e^{0.218 \times 20} = 80$ times as likely to step off compared to the younger woman. Note that in the presence of age, gender is now significant, although marginally, gender is not a significant factor. In this case age is said to be a *suppressor* variable. The **Stata** commands for this analysis are:

```
infile age minutes cens weight g1 using c:/running.txt
cox minutes g1, dead(cens)
cox minutes age, dead(cens)
cox minutes g1 age, dead(cens)
```

In the model fit that included an interaction between age and gender, the interaction term was not significant.