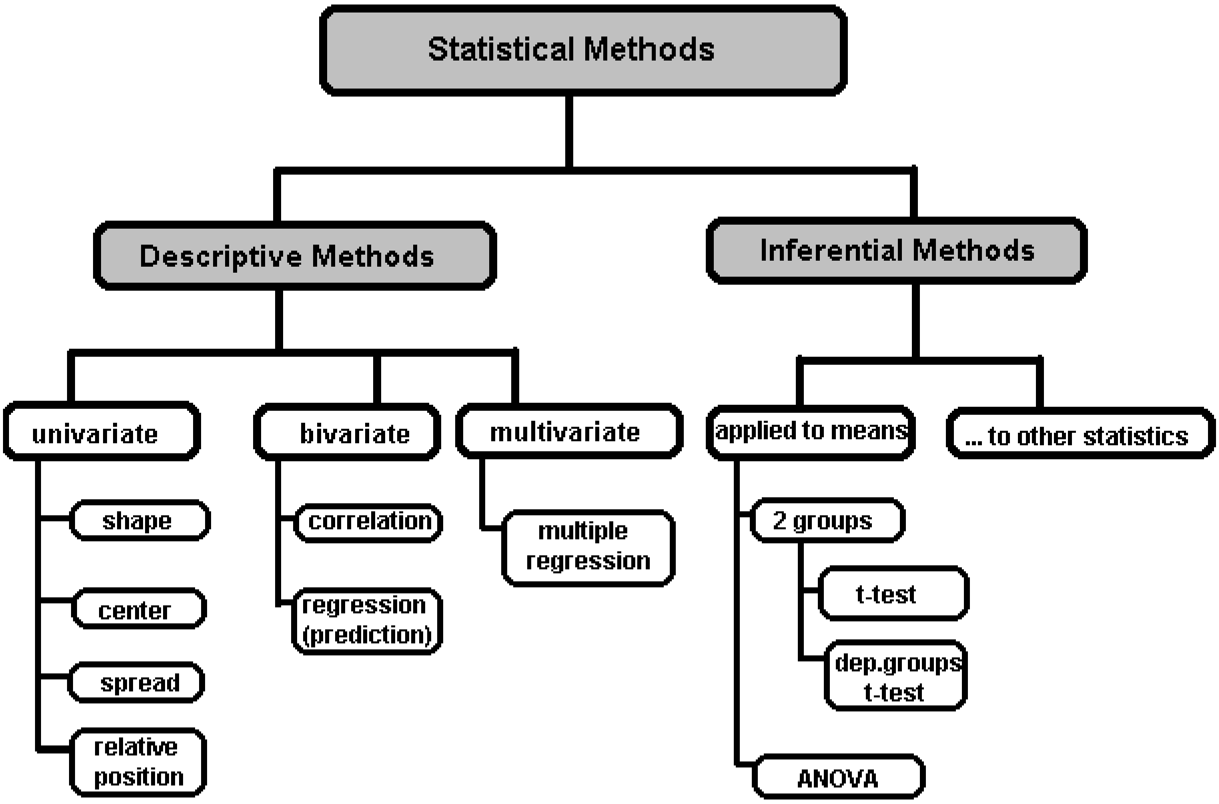
Introduction to Statistical Methods[1]



Statistical methods has two major branches: Descriptive and Inferential . The first half of this course will deal with Descriptive Statistical Methods; the second half, with Inferential Statistical Methods.

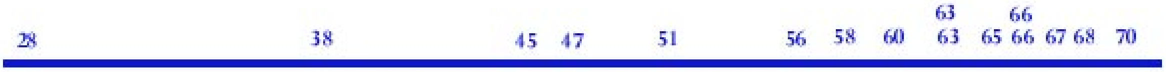
Descriptive Statistics

Example: "The average income of the 104 families in our company is $28,673."
In descriptive statistics, our objective is to describe the properties of a group of scores or data that we have "in hand," i.e., data that are accessible to us in that we can write them down on paper or type them into a spreadsheet. In descriptive statistics we are not interested in other data that were not gathered but might have been; that is the subject of inferential statistics. What properties of the set of scores are we interested in? At least three: their **center**, their **spread**, and their **shape**. Consider the following set of scores, which might be ages of persons in your bridge club:

28, 38, 45, 47, 51, 56, 58, 60, 63, 63, 65, 66, 66, 67, 68, 70

We could say of these ages that they range from 28 to 70 (**spread**), and the middle of them is somewhere around 60 (**center**). Now their shape is a property of a graph that can be drawn to depict the scores. If I marked the scores along a number line, like so

[1] http://glass.ed.asu.edu/stats/



Ages of the Bridge Club Members

then we can see that the ages tend to bunch at the older ages and trail off very gradually for the younger ages. Later we will learn that this distribution of data is said to be **negatively skewed,** because the "trailing off" is toward the negative end of the number line.

Inferential Statistics

Example: "This sample of 512 families from Maricopa county indicates with 95% confidence we can conclude that the average family income in the county is between $25,187 and $29,328."
In inferential statistics, our interest is in large collections of data that are so large that we can not have all of them "in hand." We can, however, inspect samples of these larger collections and use what we see there to make inferences to the larger collection. How **samples** relate to larger collections of data (called **populations**) from which they have been drawn is the subject of inferential statistical methods. Inferential statistics are frequently used by pollsters who ask 1000 persons whom they prefer in an election and draw conclusions about how the entire state or county will vote on election day. Scientists and researchers also employ inferential statistics to make conclusions that are more general than the conclusions they could otherwise draw on the basis of the limited number of data points they have recorded.

Tabulating and Graphing Data

The first job of any data analysis--one that is so simple that it is frequently overlooked or give short shrift--is to inspect the data for "bad" values. Bad data points are entries in a datafile that are unreasonable and represent clerical errors or misunderstandings by persons collecting or reporting the raw data, e.g., 1340 lbs. in a datafile of persons' weights. In nearly every case, bad data points are simply deleted before data analysis begins. This is known as "cleaning the datafile." So your first job is to clean the teacher salaries datafile that you have been given. Fortunately, Excel (and nearly any spreadsheet program) is a wonderful program for detecting bad data points. About all you need to do is

- Click in the lettered box at the head of a particular column in the spreadsheet;
- Select Data --> Sort from the top menu bar;
- And sort the column of data in either ascending or descending order.

The bad data points will appear at either the top or bottom (or both) of the column as "outliers," i.e., numbers that fall so far from the bulk of the remaining numbers that they raise suspicions as to their accuracy. In the case of teacher salaries, you can probably judge for yourself whether the extreme numbers are reasonable or not. When you find a bad data point or egregious "outlier," you can eliminate it in a couple of ways: highlight the cell it is in by clicking on that cell, and either delete the entry by pressing the backspace key or selecting Edit --> Cut from the top menu. **Please make a note of the data points you eliminate. Later I will ask you to send them to me.**
Once you have "cleaned the data" (be sure to clean both Desertview and Mountainview data), be sure to File --> Save a cleaned copy of the datafile to your floppy diskette.

Percentiles, 5-Number Summaries, Box-and-Whisker Plots, Frequency Distributions, Histograms

The above are all ways of describing sets of scores (or "observations of variables" to use the technical term). Box-and-Whisker plots and 5-number Summaries are constructed from percentiles; Histograms are constructed from Frequency Distributions. So let's take these two groups of things separately.

Percentiles and Other Things

A **percentile** is just a score that has a certain percent of the cases below it: the 75th Percentile is the score below which 75% of the cases fall; the 50th Percentile (sometimes written "50th %-ile") is the score below (and above) which half the scores in the datafile fall. Percentiles used to be difficult to calculate, but now thanks to computers and spreadsheets, they are real easy to find.

Find Percentiles in a Spreadsheet

1. Click in the lettered box at the head of a particular column in the spreadsheet;
2. Select Data --> Sort from the top menu bar;
3. And sort the column of data in either ascending or descending order.
4. Look at the bottom of the column to see how many cases there are for this Variable (e.g., the Variable "Teacher Salaries in Desertview"). Call this number **n**.
5. Suppose you want the 75th percentile: calculate .75(n) and count up the sorted column of scores until you rech that number (e.g., if there are 200 cases for the Variable you are working on, .75(200) = 150, so the 75th percentile will be the 150th score from the bottom (or the 50th from the top which is the same thing).
6. The Pth percentile is found by multiplying (P/100)(n) and counting up that many scores from the bottom of the column.

(There's more explanation of percentiles on pages 18-19 of the textbook.)

A **5-Number Summary** is a very concise way to describe the major features of a set of scores without getting bogged down in details. The 5 numbers in question are the 10th, 25th, 50th, 75th and 90th percentiles. In mathematical notation, these are denoted as follows: P_{10}, P_{25}, P_{50}, P_{75}, and P_{90}. P_{50} is the 50th Percentile, the score that divides the set of scores into two halves; in this sense, it is a middle score and is commonly called the **Median.** The 25th and 75th Percentiles have an obvious meaning, and noting how far they lie from the Median tells us how spread out the distribution of scores is. P_{25} and P_{75} are known by their synonyms as well: Q_1, the First Quartile, and Q_3, the Third Quartile. By what other name do you think the second quartile, Q_2, is known?

Between Q_1 and Q_3 lies half of all the scores. Knowing that much is to know quite a bit. Between P_{10} and P_{90} lies the middle 80% of all the scores, or all but the 10% highest and 10% lowest. These five numbers together, then, give a pretty informative description of the set of scores, without distracting us with too many details that may not be informative or stable. We call them the "5-Number Summary" of a distribution.

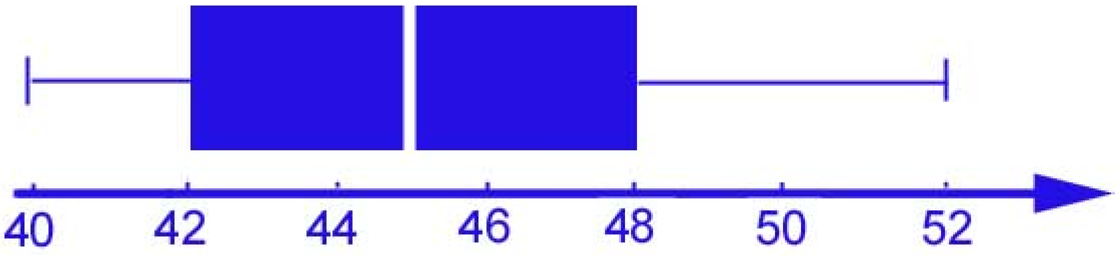
P_{10} P_{25} P_{50} P_{75} P_{90}

When the 5-Number Summary is converted to a graph, the **Box and Whisker Plot** results. We establish a ruled line horizontally across the page and mark off the full range of scores that we see in the set of scores we are analyzing. Then we draw a rectangle above the ruled scale such that the right edge is above the point on the scale corresponding to P_{75} and the left edge of the rectangle is above the 25th Percentile. We draw a line inside the rectangle at the Median. Then we draw "whiskers that extend out from the sides of the rectangle in each direction unitl they reach the 10th and 90th Percentiles. Like so:

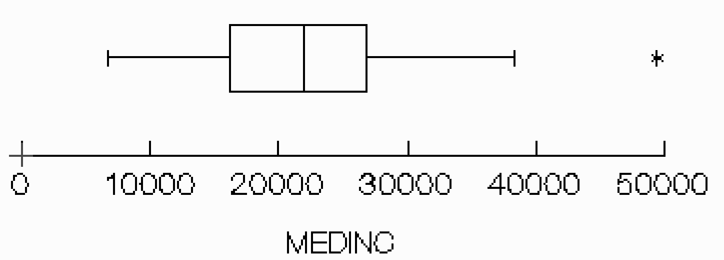
Suppose a set of 150 chidlren's heights had the following 5-Number Summary:

P_{10}= 40" **P_{25}= 42"** **P_{50}= 45"** **P_{75}= 48"** **P_{90}=52"**

Ten percent of the children are shorter than 40 inches; half the children are taller than 45". Here's what the box-and-whisker plot looks like:



Here's another box-and-whiskers plot. It describes the distribution of Median Family Incomes for 97 Elementary School Districts in Arizona (in about 1990).



What family income is exceeded by half of the Median Family Incomes for Arizona's 97 Elementary School Districts?

John Behrens offers a detailed treatment of how to construct Box-and-Whisker plots.

Frequency Distributions and Histograms

Frequency distributions and histograms are ways of portraying the complete shape of a set of scores. There is quite a bit of discussion of both these techniques in the textbook on pages 9 through 17. Suffice it to say that frequency distributions are built by dividing the range of the scores into some number of equal size classes and then counting and reporting the number (or frequency) of scores in each class. For example, if my set of scores is as follows:

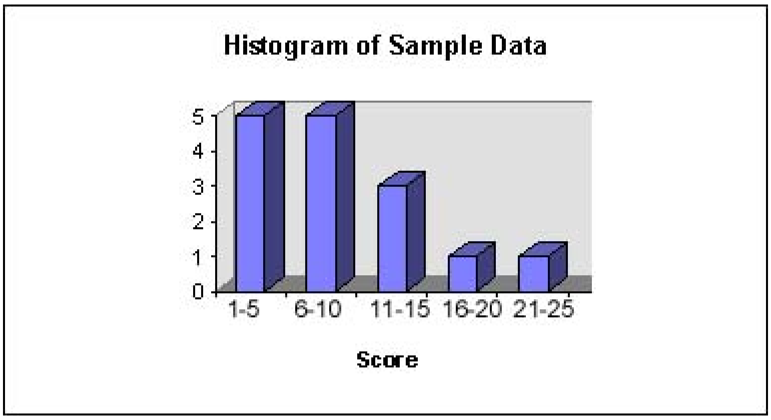
2, 3, 3, 4, 5, 6, 6, 6, 7, 10, 11, 13, 15, 16, 21

and I form the classes 1-5, 6-10, 11-15, 16-20, and 21-25, then the grouped frequency distribution looks like this:

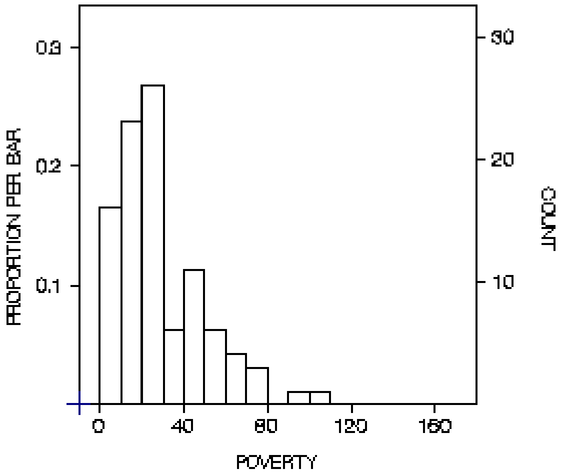
Class	Freq.
1-5	5
6-10	5
11-15	3
16-20	1
21-25	1

Question: What five scores constitute the three frequencies in the class 6-10?

A **Histogram** is simply a bar graph where the bar lengths are determined by the frequencies in each class of a grouped frequency distribution. Notice how the bar graph below (an example of a histogram) has five bars that represent the numbers of cases in each of the five classes in the above frequency distribution.



The histogram below describes the 97 elementary school districts in Arizona in terms of the proportion of poor people in the school district boundaries.



Notice that about 15 school districts are in the lowest poverty category and about 25 districts are in the third from lowest poverty category.

How to Construct a Frequency Distribution and Histogram in Excel

There are at least three methods of constructing frequency distributions (then drawing histograms) in the Excel program. One way is simple but could be tedious for very large data sets; one way is simple and powerful, if you have the right program elements in your version of Excel; and the last way is a nightmare that I hope you can avoid. **Find out about all three methods here.**

Describing Central Tendency, Variability and Skew

Describing the Central Tendency of a Set of Scores

Two properties of distributions of measures are important to describe: their **center** and their **spread**.

Set A:	**Set B:**
12,34,36,42,52 **54,68,72,81,93**	**152,154,155,155,156** **158,159,161,163,163**

- Set A has greater spread (over 80 points from 12 to 93)
- Set B has a higher center: the center of A is around 50, whereas the center of B is around 155

Measures of Center (Central Tendency)

There are three common measures of the center of a distribution:

- The **mode**

  The mode is the most frequent score in a set of measures. Some distributions have no mode, e.g., 123, 154, 167, 132. What is the mode of Set B, above?
  The mode is very unstable--minor fluctuations in the data can change it substantially; for this reason it is seldom calculated. Its principal use is colloquial:"The distribution of X is 'bimodal.'"

- The **median**

  The median is the middle score in the distribution. In the distribution 32, 35, 36, 43, 74, the median is 36. By convention, the median of the distribution 123, 154, 160, 187 is taken to be half way between the two center values: Median = (154 + 160)/2 = 157.

  Watch how a median is determined in a series of steps that make clear its definition and calculation (due to John Behrens).

- The **mean**

  The mean (average, or arithmetic mean) is the sum of the scores in the distribution divided by the number of such scores. The mean of 12, 13, 23, 43, 32 is

  Mean = (12 + 13 + 23 + 43 + 32)/5 = 24.6

  In this statistical calculation studio created by John Behrens, you can enter numbers and see how the mean is computed.

The mean (or aritmetic mean as it is called) has the property that it is the point on the number line which minimizes the sum of squared distances to all the points in the sample. That is, if the numbers 2, 4, 5, 6 and 8 have a mean of 5, then if we substitute the mean, 5, into the following formula, the sum will be as small as it can possibly be for any value of M:

$$\textbf{Sum} = (2 - M)^2 + (4 - M)^2 + (5 - M)^2 + (6 - M)^2 + (8 - M)^2$$

When M is the mean, 5, the Sum equals 9 + 1 + 0 + 1 + 9 = 20. Try putting any number but 5 in place of M and see if you can get a smaller Sum than 20; and remember, a "minus times a minus is a plus."

View a demonstration of the "minimum sum of squares" property of the mean.

Properties of Mean and Median

The most important property of the mean and median is embodied in a simple example. Observe what happens to a set of five scores when the largest one is increased by several points:

Set A: 12, 13, 23, 32, 43
Mean = 24.6
Median = 23

Set A Altered: 12, 13, 23, 32, 143
Mean= 44.6
Median = 23

The Mean changes but the Median does not.

The Mean of Combined Groups

The following situation arises not infrequently. One knows the mean of a group of some number of scores, call it Group A with n<SUBA< sub> scores in it, and the mean of another group of scores, Group B, but the original scores are not in hand and one wishes to know the mean of both groups combined. For example, Crestwood school district issues a report in which it is stated that the average salary of its 36 "probationary" teachers is $24,560, and the average salary of its 215 "tenured" teachers is $38,630. You want to know the average teacher salary in the whole district (i.e., the average of the group of probationary and tenured teaches combined).

Notice right off that the average of the combined group is NOT ($24,560 + $38,630)/2, that is, it is NOT the average of the two averages. That would only be true if the two groups being combined had equal numbers of cases in them. **As a general principle, the mean of the combined group will be closer to the mean of the larger (in terms of number of cases) group**. So we know without even making any exact calculations that the mean teacher salary in Crestwood district will be closer to $38,630 than it will be to $24,560.

But here is how the exact calculations are made even when the original scores, all 36 + 215 = 251 of them, are not available for analysis:

The mean of Groups A & B combined will be the sum of the scores in Group A plus the sum of the scores in Group B divided by the number of scores in the combined group; symbolically, it looks like this:

$$\textbf{Mean(A\&B)} = [(\textbf{Sum of A}) + (\textbf{Sum of B})] / (n_a + n_b)$$

Since Mean(A) = Sum (A) / n_a, then n_aMean(A) = Sum(A).

Consequently, Mean(A&B) = [n_aSum(A) + n_bSum(B)] / ($n_a + n_b$)

That's all there is to it. Multiply the number of cases times each group mean, add those two figures together and divide by the combined number of cases and you have the combined group mean.

Describing the Variability of a Set of Scores

There are a few common measures of variability of a distribution:

- The **Range**

  The range is the distance from the smallest to the largest score: in Set A at the top of this page, Range = 93 - 12 = 81.

- The **Hinge Spread**

  The Hinge Spread, H, is the distance from the 75th Percentile to the 25th Percentile.

- The **Semi-Interquartile Range**

  The Semi-Interquartile Range is half the Hinge Spread

  $$\mathbf{Q = H/2}$$

  In distributions that are not too terribly strange in their configuration, the Median plus and minus Q creates an interval that contains approximately the middle 50% of the distribution. If you learned that for heights of American males Median = 67" and Q = 1.5", then you could calculate that about half of all American males are between the heights of 65.5" and 68.5"

- The **Variance**

  The two most common measures of variability, namely the Variance and its close descendant the Standard Deviation, owe their popularity to the importance of the Normal Distribution, which we shall study later. Normal distributions, which play an important role in both descriptive and inferential statistics, are completely determined by two "parameters": their mean and their variance.
  The variance describes the heterogeneity of a distribution and is calculated from a formula that involves every score in the distribution. It is typically symbolized by the letter *s* with a superscript "2". The formula is

  $$\textbf{Variance} = s^2 = \textbf{Sum (Scores - Mean)}^2/(n - 1)$$

  The square root (the positive one) of the variance is known as the "standard deviation." It is symbolized by ***s*** with no superscript.

  Sq. Root of Variance = Standard Deviation, denoted by *s*

  Use the formula for the variance and standard deviation to calculate both for these scores: 8, 10, 12, 14, 16. Note that *n* = 5. Make the calculations on a piece of paper. You should get a variance of 10 and a standard deviation equal to the square root of 10 which is 3.16.

Here's an online demonstration of how the standard deviaiton is calculated, from John Behrens's stats course.

Properties of the Standard Deviation As the scores in a distribution become more heterogeneous, more "spread out" and different, the value of the standard deviation grows larger.

If I told you that the standard deviation of 6th grade students Reading grade equivalent scores was 1.68 yrs (in Grade Equivalent units) and the standard deviation of their Math scores was 0.94 yrs, you would know that the students are more varied in Reading performance than in Math performance. **Can you come up with an educationally sound explanation of why this might be so?**

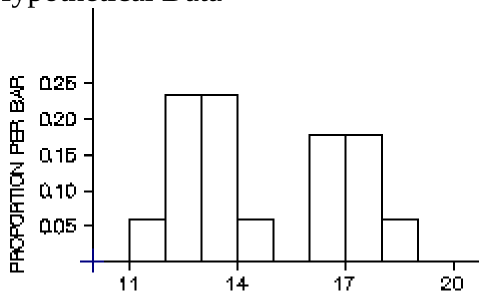
In normal distributions, roughly two-thirds of the scores lie within a distance of one standard deviation of the mean; 95% lie within two standard deviations of the mean; and 99.7% lie within 3 standard deviations.

Modality and Skewness

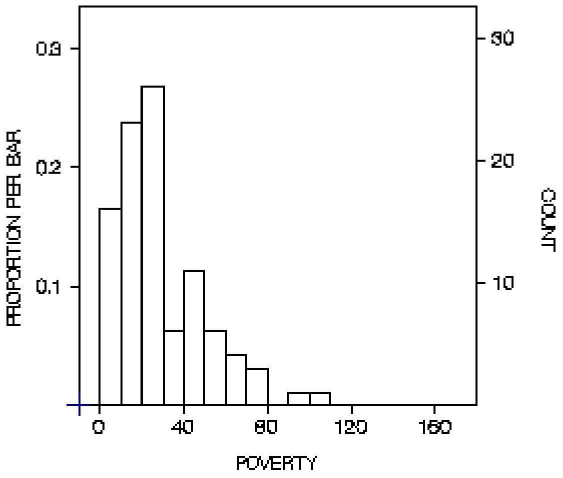
Persons talking informally about distributions of scores commonly refer to two properties: modality and skewness. **Modality** refers to the number of modes a distribution has. If the histogram of the set of scores has "one hump," it is said to be **unimodal**; two humps, and it's **bimodal**. Truly bimodal distributions are seldom encountered.

A Bimodal Distribution

Hypothetical Data



Skewness refers to the asymmetry of a histogram. If the histogram is perfectly symmetrical around its middle, the it has "no skewness." If the histogram has a hump toward the left and the right-hand tail stretches out longer than the left-hand tail, then the distribution is said to be **positively skewed.** Like this one:

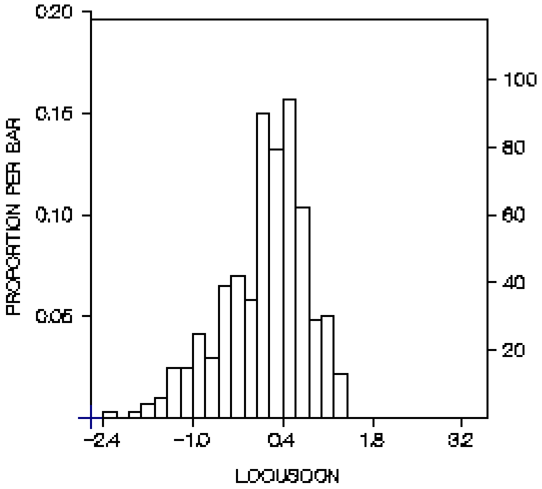


A Positively Skewed Distribution

The histogram above describes the 97 elementary school districts in Arizona in terms of the proportion of poor people in the school district boundaries.

Negative skewness is observed when the hump is to the right and the left-tail (toward the negative numbers) is elongated.

Locus of Control measures for a sample (n=600) of the High School and Beyond Survey.



A Negatively Skewed Distribution

There exists a summary statistic that measures the degree of skewness; it is not often reported, but merely inspected as to its algebraic sign to confirm an impression of either negative or positive skew from a histogram. It is roughly equivalent to an average of (standardized) third powers (cubes) of deviations of scores from the mean. Forget about it.

The Mean, the Median and Skewness

There is a relationship among the mean, the median and skewness that is important in descriptive statistics. To put it in common language, **the mean is drawn in the direction of the skew more than is the median.** That is, in a very positively skewed distribution, the mean will be higher than the median. In a very negatively skewed distribution, the mean will be lower than the median. The median is less affected by extreme scores in a distribution than is the mean. Recall the earlier example: when the largest of 5 scores is increased by several points the mean is drawn toward the elongated tail of the distribution:

Set A: 12, 13, 23, 32, 43
Mean = 24.6
Median = 23

Set A Altered: 12, 13, 23, 32, 143
Mean= 44.6
Median = 23

The Mean changes but the Median does not.

Because of this sensitivity of the mean to extreme scores, it is sometimes not favored for describing central tendency of very skewed dstributions. Often, distributions of financial statistics (income, poverty rates, expenditures and the like) are very skewed. One will find the median preferred for describing the centers of skewed distributions.

Exercises

You can use the online stats calculator to make some very quick calculations of means, medians, variances, standard deviations, and skewness so that you get a feel for what these summary statistics mean.

Online Stats Calculator

Exercises

1. Enter the following numbers into the online calculator and observe the mean, median, variance and standard deviation:

   12.3 21.4 34.5 32.8 42.3 18.6 25.2 28.3 27.1 24.3 31.7

2. Here's the same set of scores as in #1 above except, the largest score, 42.3, has been increased to 68.2. What will happen to the mean and median of thisgroup of scores compared to the group of scores in #1?

   12.3 21.4 34.5 32.8 68.2 18.6 25.2 28.3 27.1 24.3 31.7

Using Excel to Calculate Summary Statistics

Fortunately, it is a whole lot easier to calculate things like means, medians, variances and standard deviations in Excel than it was to construct a frequency distribution.

Calculating the Mean in Excel

1. Suppose that the numbers whose mean you want are in Rows a3 through a156 of the spreadsheet.
2. First, find an empty spot in the spreadsheet where you want the answer to appear and click on it, e.g., cell e5.
3. Click on the Function icon (it looks like this, remember: $\boldsymbol{f}_x$).
4. In the left box that appears, click on "Statistical." In the right hand box, click on "AVERAGE." Then click on Next at the bottom of the dialogue box.
5. Then in the dialogue box that appears next, type this in the first window labeled **number 1** $\boldsymbol{f}_x$: a3:a156 . (Obviously, if your data are in some other rows, enter the proper symbols, e.g., b1:b100). Finally, click on "Finish" at the bottom of the dialogue box.
6. That's all; by now you should be seeing the mean of the numbers in Rows a3 through a156 in the cell at e5.

Calculating the Standard Deviation in Excel

- Do exactly as you did to calculate the Mean, only pick "STDEV" in the "Statistical" dialogue box instead of "AVERAGE."

And now, experiment with these other **Statistical Functions** that you'll find in your **Excel** spreadsheet:

- **COUNT**
- **MAX**
- **MEDIAN**
- **MIN**
- **MODE**
- **PERCENTILE**

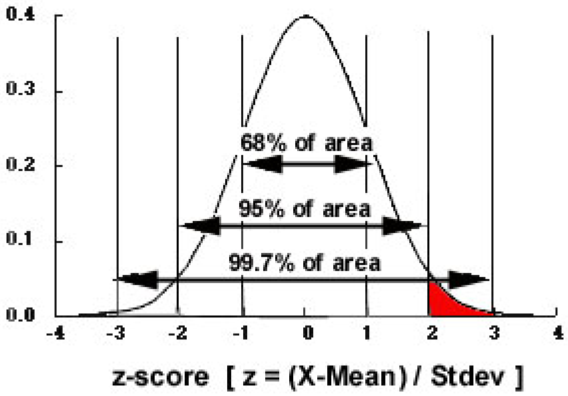
- **SKEW**
- **VAR**

If the Excel program you are working on has the **Tools ---> Data Analysis** package installed on it, then you can get all these statistics in one fel swoop. Just select **Tools ---> Data Analysis ---> Descriptive Statistics** and when you reach the Descriptive Statistics dialgoue box, fill in the "Input Range" with the location of your scores (e.g., a3:a391) and be sure to check Summary Statistics at the very bottom of the box. That's all there is to it.

Normal Distribution & Standard Scores

Normal Distribution

The normal (or Gaussian) distribution is the familiar unimodal, symmetric curve with thin tails that every introductory psychology textbook calls the "bell curve." (Many years ago when I was teaching at the University of Illinois, which was a leader in accommodating students with disabilities, I lectured to a class that included a blind student. I scribbled a replica of the normal curve on the chalk board and described it as looking like a bell. After class, the student politely explained to me that there are many kinds of bells--door bells, sleigh bells and the like--and would I please be a bit more specific. Touche!) The normal curve looks like a vertical cross section of the Liberty bell--with all the top attachments removed--oh, forget it.
Here is a picture of a normal distribution showing the important facts about areas under the normal curve within various standard devistion units of the center.



Certain things follow from the facts about areas in the graph for the normal curve:

- 50% of the area (and hence, half the cases in a set of data that is normally distributed) lies below the middle or mean.
- 34% of the area lies between the mean and a point one standard deviation above. Likewise, there is 34% between the mean and a standard deviation below the mean.
- It follows, then, that 16% of the data in a normal distribution lies below a point one standard deviation below the mean.

Answer each of the following questions--write down your answers. At the end, you can click on the Answers.

1. What percent of the normal distribution lies between one and two standard deviations above the mean?
2. What percent of the normal distribution lies above three standard deviations above the mean?
3. If there were 100,000 persons arrayed in a normal distribution of heights, how many would be expected to lie more than three standard deviations above the mean?

Answers to the first set of Normal Curve Questions

It is not a simple matter to calculate the area under the normal curve between to arbitrary points like 1.25 and 2.38 standard deviations above the mean.
You have four options:

1. look up values of areas under the normal curve in a printed table in a statistics textbook,
2. hope that these good people in the Netherlands have their server functioning when you need a quick reading of normal curve areas. **Please note: when you enter a z-value of 1.5, say, into the calculator in the Netherlands, the area returned is the probability of being greater than 1.5 or less than -1.5; i.e., it is a two-tailed probability and must be divided by 2 to give a single tail area.**
3. **your best bet,** if your browser will handle the Java, is to use Gary McClelland's nice Java program from the University of Colorado.
4. and, finally, just in case none of these utilities is available when you need them on the internet, you can always resort to the old-fashion way of finding normal curve areas by looking them up in a table like this one.

Please exercise either option now in answering the following questions:

1. What percent of the normal distribution lies below a point .675 standard deviations above the mean?
2. What percent of the normal distribution lies above a point that is 1.96 standard deviations above the mean?

Answers to the second set of Normal Curve Questions

Unit Normal Scores: the z-Score

All normal distributions have the same "shape" but they can have different means and different standard deviations. Once one specifies the mean and standard deviation of a normal distribution, everything alse about it is fixed (e.g., the percent of area between any two points).
For this reason, all the various normal distributions (of people's heights and weights and IQ scores) can be referred to a single table of the normal curve by standardizing a variable to a common mean and standard deviation. The simplest standardization measures the position of any point in a normal distribution in terms of its distance above or below the mean in units of the standard deviation. Thus, a *standard unit normal variable* has the formula

$$\mathbf{z = (X\text{-}m)/s}\ ,$$

where **m** is the mean, and **s** is the standard deviation of the distribution of scores. Consequently, a person with a **z** score of +1.5 lies one and one-half standard deviations above the mean.

> You are given that the distribution of adolescents IQ scores is normal in shape with a mean of 100 and a standard deviation of 15 points.

1. What is the percentile rank of a child whose IQ is 120?
2. What percent of the population of adolescents have IQ scores below 90?

Answers to the third set of Normal Curve Questions

Standard Scores

Unit normal z-scores are useful, but their properties are sometimes viewed as a disadvantage for particular applications. In these cases, one transforms them to scales that have more convenient means and standard deviations. For example, if one would multiply each z-score by 200 and then add 1000 to the product, the resulting new standard scores would have a mean of 1000 and a standard deviation of 200.
There are several particular standard score scales in such common use that it is useful to look more closely at them. In general, if a *z-score* is transformed via the following formula:

$$\mathbf{Z = Bz + A\ ,}$$

then the *Z-score* has a mean of **A** and a standard deviation of **B**.

Some Popular Standard Scores

A **Mean**	**B** **St Dev**	**Scale Name**
500	100	SAT; GRE; LSAT; GMAT
100	15	Wechsler IQ
100	16	Stanford Binet IQ
20	5	ACT (Amer College Testing Co.)
50	10	T-scale (MMPI)

Standard scores vs. percentiles

If all one does with standard scores is convert them to percentiles, then why have both? Percentiles and standard scores have slightly different information in them. Another way to put this is that the transformation from standard scores to their normal curve percentile equivalents is a "non-linear transformation." Very large differences between extremely large of extremely small standard scores correspond to small differences in percentiles; likewise, very small differences in standard scores near the mean correspond to large differences in percentiles.
Consider two groups of three persons each whose heights are measured both in inches and in percentiles among adult males:

```
                 Heights-inches      Heights-percentiles
                 ______________      ___________________
         Group A: 70", 72", 84"       80, 92, 99.999
         Group B: 70", 74", 76"       80, 95, 99.9


                         Group A     Group B
                         _______     _______
     Mean in inches      75.333       73.333
     Mean Percentiles     90.67       91.63
```

Notice that Group A is taller than Group B when heights are expressed in inches, but Group B is "taller" when heights are expressed in percentiles. Is this possible? Or did I make a calculation error?

Collateral Reading

Here's more on the Normal Curve, courtesy of John Behrens.

Correlation & Regression

Data Layout

The table below is an example of how data are arranged for purposes of correlation. The question that correlation analysis answers is this: "When scores on one variable (X) increase, do the scores on a second variable (Y) also, increase, or do they decrease, or is there no systematic increase or decrease?" For example, are higher cholesterol measures associated with high body weight? Are increases in numbers of hours a student spends on homework (X) associated with increases in the student's GPA (Y)?
In a correlation analysis, there are two variables (X and Y) that are mesured on one group of "units." Units can be persons, families, schools, school districts or other thigns that can be measured in two or more ways. An example using units other than persons might be this: For schools (the units), is school size (X) measured by number of pupils associated with school violence (Y) as measured by number of reported assualts among students?
When data are collected for several units on at least two variables, they can be tabulated like this:

	IQ	Spelling
Sue	115	34
Alma	87	18
Carlos	104	28
Anita	121	26
Hans	96	19
Linda	99	20
Louis	136	26

A Note About Cleaning Datafiles for Correlation Analysis

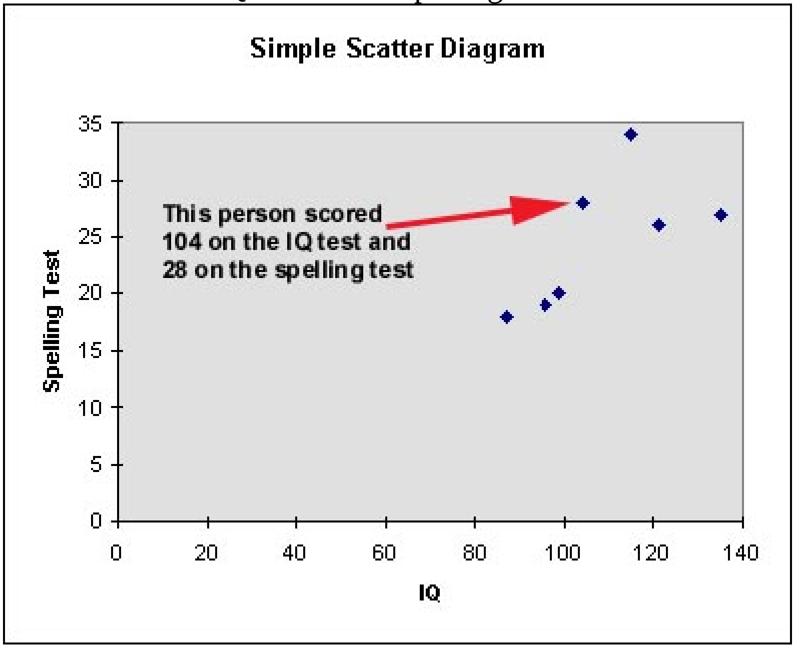
Recall that when we first opened a datafile that we either acquired from someone else or built ourselves, that we needed to inspect it for "bad data" points, i.e., outliers or mistypes or other errors that distort the data set. The easiest way to clean data in Excel is to Sort a column and onspect the top and bottom for extreme scores that make no sense. When you are performing correlation analyses, there is a slight complication you must heed. Unlike with the first datafile we used (the teacher salary data for tow school districts), data arranged for correlation analysis have rows that are not arbitrary. In our example of teacher salaries, Row 5, for instance, had no particular meaning; it was simply the fifth person whose salary was recorded. And the person in Row 5 for Desertview bore no special relationship to the person in Row 5 for Mountainview. **But in a correlation spreadsheet, the situation is different.** The rows are tied together, in a sense. The data for each unit occupies its own row in the spreadsheet. For example, in the above table, Sue's IQ score is 115 and her Spelling test score is 34; we can't just arbitrarily put Anita's Spelling score up on Sue's row and move Sue's score down to Anita's row and make any sense of the correlation analysis.

So what does this have to do with sorting the datafile? Well, the Sort function in Excel breaks up the rows. If you Sort the above table in Excel, Anita's IQ score will come out in Row 1 of the IQ variable, but Sue's Spelling score will occupy Row 1 of the Spelling variable. The solution to this problem is simple. Just Select and Copy the entire worksheet into a new Excel worksheet before sorting. Sort and find the bad data points in the temporary worksheet, and make any corrections on the origincal worksheet.

In a correlation analysis, we ask and answer three questions about the relationship between X and Y: a) What is the **shape** of the relationship between X and Y; b) what is the **direction** of the relationship between X and Y; and c) what is the strength of the relationship?
Each of these questions can be answered by inspecting a graph of the relationship called a **scatter diagram** or **scatter plot**; and questions b) and c) can be given an even more precise answer by calculation of the **correlation coefficient**.

Graphing Correlations (Scatter Diagrams)

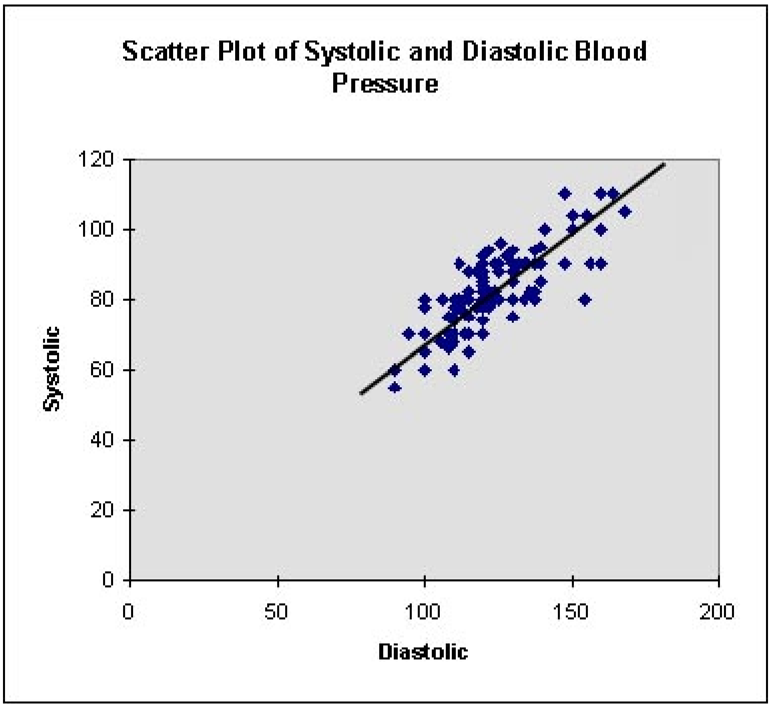
The shape and direction of a correlation can be seen in the scatter diagram. Scatter diagrams can be constructed in Excel, or they can be drawn by hand. Two axes (one for X the other for Y) are drawn at right angles to each other and labeled with the scale of measurement of the two variables X and Y. A unit (person, for example) is then represented as a point in the diagram at the intersection of that unit's X and Y scores. The following picture shows the scatter diagram for the data in the above table on IQ scores and Spelling test scores.



Linear vs. Curvilinear

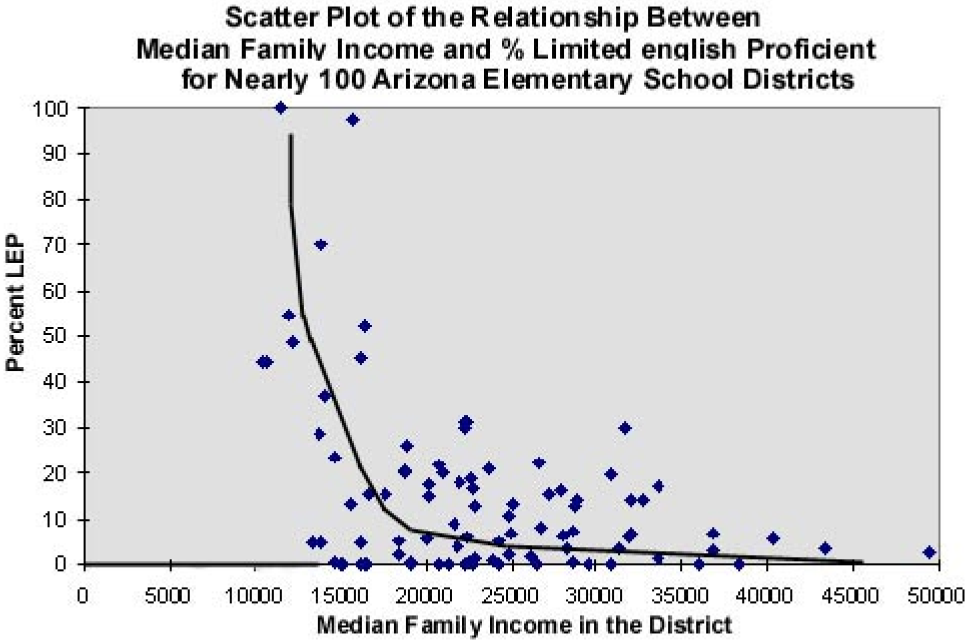
The first thing to notice about a scatter diagram depicting the correlation of X and Y is whether the relationship is linear (straight line) or curvilinear (curved). **The correlation coefficient we will use—the Pearson correlation—only describes linear relationships. If the relationship between X and Y is curved, the Pearson will either misrepresent the relationship or simply make it look weaker than it really is.** (Note: coefficients that describe curvilinear relationships are beyond the scope of this course; a reference to how one deals with them appears in footnote #2 on page 93 of your textbook.)
Here is the scatter plot of a linear relationship:

A Linear Scatter Plot



The scatter plot below shows a curvilinear relationship:

A Curvilinear Scatter Plot



The **direction** of a two-variable relationship can be either direct (positive) or inverse (negative).

- When higher values of X tend to be paired with higher values of Y, then X and Y are said to be **positively** correlated.
- When higher values of X tend to be paired with **lower** values of Y, then X and Y are said to be **negatively** correlated.

In the above two scatter plots, the correlation is direct, or positive, in the first example, of a linear relationship. And the relationship is inverse, or negative, in the second example of a curvilinear relationship.

Later on we'll see how to use Excel to draw scatter plots; but if you want to try out some examples with an online plotter, **try this one from John Behrens.**

r, the Measure of Correlation

```
                    r = Sum[z(X)z(Y)]/(n-1),

           where n is the number of units or pairs of data points,
          z(X) is the z-score for each unit on the X variable, and
            z(Y) is the z-score for each unit on the Y variable.
```

You'll recall from Lesson 3 that a z-score is the difference between a raw score and its group mean divided by the group's standard deviaiton. But, don't spend too much time contemplating the above formula; computer programs like Excel can calculate the value of the Pearson product-moment correlation coefficient (which **r** is, for you.

Pearson's **r** has the following properties:

1. **r** can assume values from -1 to +1, inclusive;
2. **r** is positive for direct relationships;
3. **r** is negative for indirect relationships;
4. **r** is zero (0) when there is no linear relationship between the two variables or when there exists particular kinds of curvilinear relationships between X and Y.

Try this online correlation calculator. Use the data in the table at the beginning of this lesson and verify that the correlation coefficient **r** equals **.615942** . That's rather more decimal accuracy than anyone needs in such circumstance, so we would likely report the result as **r = .62** , and say of it that "IQ and Spelling score were positively related with a correlation of .62." People often ask, "How big does a correlation have to be to be significant, or important, or large, or strong?" There is no simple answer to such questions. Essentially the person is asking, "How do I translate a numerical value of **r** into words?" The answer to that is , "Don't." No words (of the type "large," "strong," "weak," etc.) will adequately substitute for the picture of the scatter diagram and the reported value of **r**. Correlation coefficients of .20 can be extremely important, e.g., the correlation between smoking and ling cancer, and other correlations of .95 can be useless, e.g., the correlation between one's height at 8 a.m. and one's height at 10 p.m. The skill one develops, however, is the skill of inspecting a scatter diagram and calculating the associated value of **r** so that one can visualize a typical scatter plot when one hears of or reads a value of **r**--because many values of **r** are reported without their scatter plots.
Study the following figures in your textbook and try to develop some skill at associating a value of **r** with the strength of association pictured in the scatter diagram: Figure 6.1 on page 91; Figure 6.3 on page 97; Figure 6.4 on page 98; the five figures on page 101; the figure at the bottom of page 105; Figure 7.1 on page 109; Figure 8.2 on page 126; Figure 8.5 on page 139. Now you are ready to test yourself. Here is an instructional device, courtesy of the Univ. of Illinois, that helps you learn to relate scatter diagrams and values of **r**. You may need to run Microsoft Internet Explorer for this one. Have fun with it.
Here's an example of a research report that makes extensive use of correlational analysis and has linked reported values of *r* to the corresponding scatter plots in an innovative way.

Scale Invariance: An Important Property of r

The Pearson correlation coefficient has an important property: no mater how you would change the scale of measuring either X or Y, so long as you change them only by multiplying or adding, X and Y will still have the same value of **r**. So, for example, if I tell you that weight in kilograms and height in centimeters have a correlation of .74, you know immediately that weight in pounds and height in inches also correlate .74—because kilograms can be converted to pounds merely by a "linear transformation" (multiplying and/or adding) and so can centimeters be converted to inches (one inch = 2.54 centimeters).

Test out this property of **r** by entering the numbers in Table A into an online calculator, noting the value of **r**, then entering the numbers in Table B into the calculator. Note that the numbers in B are just multiples of the numbers in A.

Table A

	IQ	Spelling
Sue	115	34
Alma	87	18
Carlos	104	28
Anita	121	26
Hans	96	19
Linda	99	20
Louis	136	26

Table B

	IQ	Spelling
Sue	11.5	68
Alma	8.7	36
Carlos	10.4	56
Anita	12.1	52
Hans	9.6	38
Linda	9.9	40
Louis	13.6	52

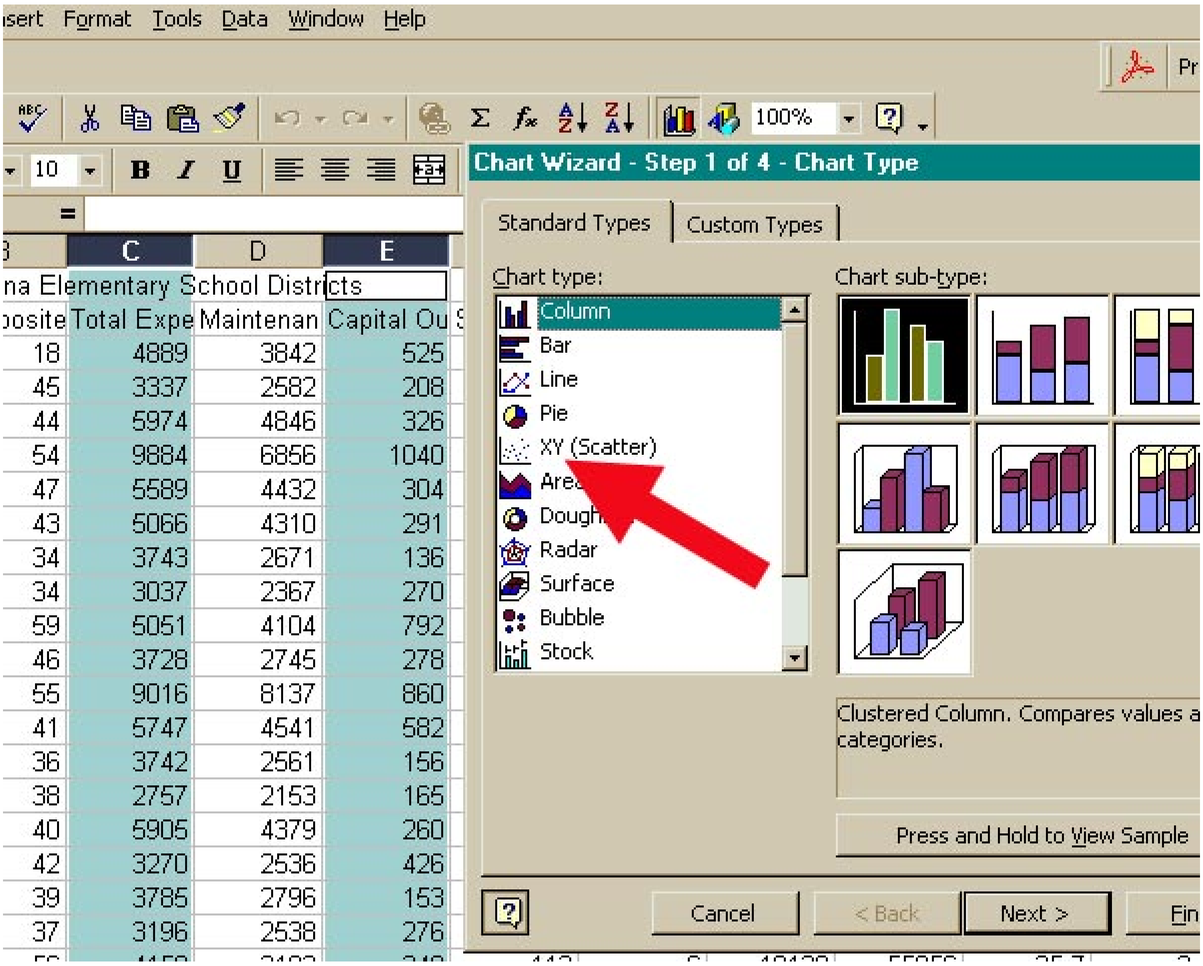
Drawing Scatter Plots and Calculating Correlations with Excel

Excel is a fairly convenient tool for doing both of these.

To draw a scatter plot

- With the data showing in the spreadsheet, think of the two variables you want to plot.
- Select one of these variables by clicking at the head of the column in which it resides.
- **Now, while holding down the Ctrl key,** click at the head of the column for the second variable.
- Now click on the Chart Wizard button on the menu bar at the top of the spreadsheet. (It's next to the two Sort buttons.)

- You'll now see a menu of chart types; pick "XY Scatter" (see the illustration below.) Click on Next.
- If prompted to choose a scatter plot type, pick only the type that does not involve connecting dots with lines.
- Follow your nose at this point, and everything should be OK.



3	C	D	E
na Elementary School Districts			
posite	Total Expe	Maintenan	Capital Ou
18	4889	3842	525
45	3337	2582	208
44	5974	4846	326
54	9884	6856	1040
47	5589	4432	304
43	5066	4310	291
34	3743	2671	136
34	3037	2367	270
59	5051	4104	792
46	3728	2745	278
55	9016	8137	860
41	5747	4541	582
36	3742	2561	156
38	2757	2153	165
40	5905	4379	260
42	3270	2536	426
39	3785	2796	153
37	3196	2538	276

Clustered Column. Compares values a categories.

To calculate the value of r

- Click on an empty cell in the spreadsheet where you would like the answer to appear.
- Click on the Function button on the menu bar; it looks like this: ***f*** *x*.
- Select "Statistical" in the left side of the dialog box and CORREL in the right side.
- In the next dialog box enter as **array1** the location of the X scores, e.g., e2:e150, and as **array2** the location of the Y scores, e.g., g2:g150. (Note: the range of these entries will have to match; in other words, you can't have f1:f15 and h1:h25, for example).
- When you click on Finish, you should see the value of **r** appear in the cell you chose in the first step above.

Regression

Reading Assignment for Regression

Chapter 8. Sections 8.1 through 8.12. This is one big wad of statistical methods to try and swallow; don't worry if you gag on it the first time. If it happens that you will need regession analysis in life after this course, you'll have occasion to study it again (and again). It is not a simple topic.

The statistical technique known as *regression* is a very widely used technique at the more advanced levels of statistical analysis. It is applied both a) to predict the future values of variables and, in more advanced applications, b) to analyze (decompose) the structure of functional relationships (for example, how students' aspirations, parental support and abilities combine and interact to mutually influence their success in school; or to measure how students' efforts influence their achievement when the influence of their intellectual ability is "removed," i.e., held constant).

The techniques of regression and correlation are closely connected. In fact, the very symbol for correlation, **r**, stems from this association. (When Karl Pearson devised the measure of correlation between two variables now known as the Pearson product-moment correlation coefficient in the 19th Century, he was working with Francis Galton on the phenomenon of cross-generational regression.)

Suppose that I can calculate the GPA of a high school senior and wish to use it to predict the student's first-year college GPA. We'll call the High School GPA by the name **X** and the Freshman GPA by the name **Y**. Suppose further that both **X** and **Y** are transformed to standard score form, i.e.,

$$\mathbf{z(X) = (X - M_x)/s(x)}$$
$$\mathbf{z(Y) = (Y - M_y)/s(y)}$$

where $\mathbf{M_x}$ is the mean of X.

Please note that z(X) denotes the z-score on X and not the product of z and X; likewise, s(X) is the standard deviation of X and not the product of s and X. Similarly for z(Y) and s(Y).

Assume that once we have information on **z(X)** we will use the following simple equation to predict what **z(Y)** will be:

Predicted z(Y) = c z(X) **Equation 1**

In the above equation, the coefficient **c** is a multiplier that changes **z(X)** into the best estimate of what **z(Y)** will someday be. It remains to specify exactly how **c** is chosen, i.e., by what criterion.

The universally accepted criterion for determining **c** in Equation 1 is known as the *Least-squares criterion*. The Least-squares criterion states that the best **c** would be that **c** which makes the sum of the squared differences between **c z(X)** and a bunch of the actual associated **z(Y)'s** as small as possible (minimal, as the mathematician says). So, you see that finding a value for **c** will require actually gathering a sample of **X** and **Y** scores (actually measuring High School GPA and Freshman GPA for a sample of students in a *derivation sample*). From this experience, we can find the value of **c** that satisfies the Least-squares criterion in the derivation sample and then hope that it will hold up in future applications to samples where **X** is known but **Y** has yet to occur.

In a derivation sample, the value of **c** that satifies the Least-squares criterion is given by Equation 2 below:

c = Sum[z(X)z(Y)]/(n-1)

The interesting thing about c is that it is precisely the same formula as that for Pearson's product-moment correlation coefficient.

$$\mathbf{c = r_{xy}}$$

So if one wishes to make a good prediction of a person's standard score on some future variable **Y**, one should multiply the person's standard score on **X** by the correlation between **X** and **Y** from a derivation sample.

$$\textbf{Predicted } z(Y) = r[z(X)]$$

Consider a simple example. Adults males in the U.S. have an average height of about 69" with a standard deviaiton of about 2.5". Furthermore, it is well known that the correlation between a father's and his son's height is about .60. How tall will we predict a son to be at adulthood if his father is two standard deviations above average in height?

$$\textbf{Son's Predicted } z(Y) = r[z(X)] = .60(2.0) = 1.20$$

So a father is who is two standard deviaitons above average in height is predicted to have a son who will grow to 1.2 standard deviaitons above average. Tall fathers are predicted to have shorter sons. Likewse, shorter than average fathers are predicted to have sons who will be taller than their fathers, though the prediction will not exceed the national avaerage. This phenomenon is known as "regression to the mean," and it has a fascinating and checkered history.

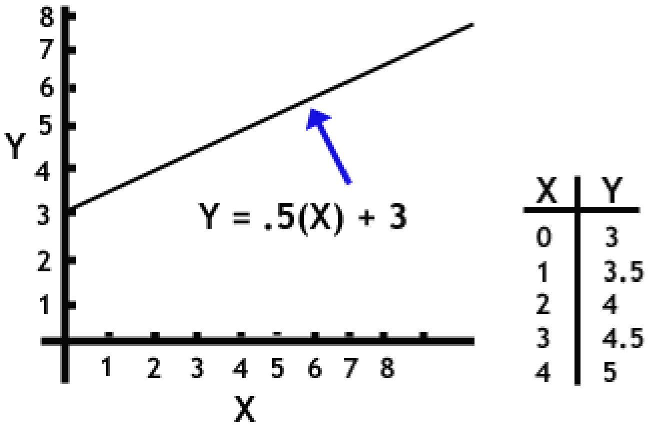
The Prediction Equation in Raw-Score Form

By algebraically expanding the above standard score form of the prediction equation, we can obtain the Least-squares linear prediction equation in raw-score form, i.e., in a form that can be applied to the data as originally gathered:

$$\textbf{Predicted-}Y = (r)(s_y/s_x)X + M_y - (r)(s_y/s_x)M_x$$

where s_x and s_y are the standard deviations of **X** and **Y**, respectively,
and M_x and M_y are the means of **X** and **Y**.
The above prediction or "regression equation" is the algebraic equation of a straight line when graphed in a two-dimensional coordinate system. Every straight line equation has two numbers that determine where it will lie in the graph: the **slope** and the **intercept**. Slope is the number that multiplies **X** and it got that name because as **X** increases by one unit, the line will rise (or fall, if slope is negative) by the slope. The intercept is where the line cuts the Y-axis (the ordinate of the graph) above an X-value of zero. So, the equation

$Y = .5(X) + 3$ has a slope of .5 and an intercept of +3. So it looks like this:



X	Y
0	3
1	3.5
2	4
3	4.5
4	5

Graph of the straight line Y = .5(X) + 3,
illustrating a slope of .5 and an intercept of 3.

Consequently, for the least-squares regression line, the

$$\textbf{Slope} = (r)(s_y/s_x), \textbf{ and the}$$
$$\textbf{Intercept} = M_y - (r)(s_y/s_x)M_x$$

Let's go back to our example of fathers' and sons' heights. Fathers have average height of 69" with a standard deviaiton of 2.5", and so does the generation of their sons. The correlation of heights across the generation is .60. A father who is 6'2" (74") tall will have a son whose height is predicted to be what?

Predicted Son's height = .60(2.5/2.5)X + 69 – .60(2.5/2.5)69 =

.60(2.5/2.5)74 + 69 – .60(2.5/2.5)69 =

72.0"

So a father who is 74" tall is predicted to have a son who is 72.0" tall.

The Standard Error of Estimate

The least-squares predicted value of **Y** is a best guess, in a sense, of what **Y** will be when we have **X** and the prediction equation to do the guessing. That guess, however, may not be very good. The measure of hwo good the predictions can be expected to be is given by an expression called the **standard error of estimate, S_e.**

$$s_e = s_y[\text{Sqroot}(1 - r^2)].$$

Here is the use that is made of the standard error of estimate. For all those sons predicted to have a height of 70.8", 68% of them will have actual heights between 70.8" plus s_e and 70.8" minus s_e. Furthermore, 95% of the sons predicted to be 70.8" tall will have actual adult heights between 70.8" plus $2s_e$ and 70.8" minus $2s_e$. (If "one standard ... 68%" and "two standard ... 95%" ring a bell, they should. This interpretation is based on the assumption that for fathers of any given height, say, 74", their sons heights will be normally distributed.)

In our example with heights, the standard error of estimate is

$$s_e = 2.5[\text{Sqroot}(1 - .60^2)]$$

= 2.5[Sqroot(.64)]

= 2.5(.80)

= 2.0"

Our prediction of sons' heights then takes this form: For fathers who are 74" tall, we predict that their sons will be 70.8" tall and two-thirds (68%) of them will have heights between 68.8" and 72.8" (plus and minus one standard error of estimate).

Using Regression Analysis to Remove the Influence of a Variable

Consider a situation with which many of us are faced these days. The state, or the U.S. Dept. of Education, or the school district you work for, says that they are going to hold teachers accountable for their students' learning—by which they mean, for their students' scores on standardized tests. It doesn't take much reflection to realize that it is unfair to the teachers to compare them in terms of their class's avaerage test score at the end of the school year, because the classes started out at different points, e.g., Jones's Grade 3 tested 3.1 GE in September and 4.4 GE in June, whereas Smith's Grade 3 started at 3.9 GE and finished at 4.5 GE. No one truly thinks that Smith taught more than Jones. [Dear reader: it is, perhaps, more difficult for me to write about these matters in such gross oversimplifications than for you to read them. I have many misgivings about this entire approach to "accountability," but they lie far afield of this lesson in statistics. Nonetheless, this illustration has been chosen because it is distressingly realistic; people are actually advocating such things.]

It might be fairer to take the September to June "gain" for the class as the measure of the teacher's contribution: Jones scores 4.4 – 3.1 = 1.3 GE; Smith scores 4.5 – 3.9 = 0.6 GE; Jones wins.

The simple "gain" approach has disadvantages that are overcome to some extent by the approach that "predicts" June from September and then subtracts the prediction from the actual score and the result is the regression corrected residual gain score for the teachers. In this sense, then, the September class averages are used to "explain" some of the variation in the June class averages; by subtracting the prediction of June from the actual June average, we will have "removed the influence" of September differences from the June differences. The situation we are examining here is not really a "prediction" situation, since both September and June averages are available; there's no need to make any decision about a teacher in advance of obtaining the June average. Nonetheless, we use the regression method remove from the June scores differences in the "inputs" to the class in September."

This logic—that of removing the influence of Variable *X* from Variable *Y*—has a wide range of applications. In fact, it is applied over and over in economics, the social sciences and in all manner of statistical analyses. (How well it achieves it purpose is hotly debated.) For example, one might wish to study the variation in per pupil expenditures across school ditricts after removing the influence of "percent of students classified as 'Special Education,' since everyone knows that per pupil special educaiton costs are much greater than others. So, if we "regressed per pupil expenditures" onto "% spec. ed."—this is how we speak of "predicting" per pupil expend. from "% spec. ed."—and then subtracted the predicted per pupil expenditure from the actual per pupil expenditure, the resulting *residual* would be a comparison of per pupil expenditures of the districts *as if* they served equal percentages of special education students.

Let's go back to the teacher accountability example to see how all this works out with actual numbers. Just suppose that in a relatively small school district the elementary school teachers' classes are tested on the SAT9 on September 1st and again on the following June 1st. (Ignore the fact that students come and students go during the school year, and the fact that some are absent at one time and not the other, and that one of the rural schools only has 4 students in the fifth grade...well, I needn't go on in this vein.) And we obtained the following scores in Reading (Grade Equivalent units) for the teachers:

September-to-June Class Average Test Scores (SAT9 Reading)
for Elementary School Teachers (Grades 3 - 6) in the
Brookside Unified School District

Teacher	Sept(X)	June(Y)
a	3.2	3.5
b	2.8	4.1
c	2.5	4.2
d	3.4	4.3
e	3.5	3.7
f	3.1	4.3
g	2.8	3.1
h	3.9	5.2
i	2.5	2.9
j	2.3	3.0
k	3.2	3.9
l	3.4	3.8
m	4.2	5.7
n	3.9	4.3
o	3.8	5.2
p	3.6	5.4
q	4.6	6.8
r	4.1	5.5
s	3.9	3.9
t	4.2	4.8
u	4.1	4.6
v	4.5	5.9
w	3.8	4.5
x	3.7	5.6
y	5.1	6.7
z	4.8	5.5
aa	4.7	5.3
bb	5.2	5.7
cc	5.5	6.3
dd	5.1	6.2
ee	4.2	6.3
ff	4.3	5.0
gg	5.3	5.9
hh	5.5	6.9
ii	4.8	5.4
jj	4.6	5.9
kk	5.9	5.9
ll	5.6	6.4
mm	6.1	7.2
nn	6.2	7.8
oo	5.6	6.3
pp	5.9	6.2
qq	6.1	6.5
rr	5.8	7.1
ss	5.7	6.3
tt	6.6	7.3
uu	5.8	6.2

vv	5.2	6.8

Now, the summary staistics needed to calculate the regression equation are as follows:

```
n = 48
X: Sept.  Mean-X = 4.471     st-dev X = 1.126
Y: June   Mean-Y = 5.402     st dev Y = 1.235
r= .892
```

So the regression equation for estimating June average scores from September average scores is as follows:

Predicted Y = .978(X) + 1.03

Let's use Teacher *a* to illustrate how the regression equation removes the influence of September variation from June variation. Teacher *a* (a Grade 3 teacher) had a class average score on SAT9 Reading of 3.2 GE yrs. Consequently, we would predict that Teacher *a*'s class will score the following score in June:

Predicted Y = .978(3.2) + 1.03 = 4.16 GE yrs.

Teacher *a*'s score in June with the influence of the September score removed is

3.5 - 4.16 = -.67

This corrected score is called a *residual* around the regression line and what it means, in this instance, is that Teacher *a*'s class scored .67 GE yrs (or almost 7 grade-equivalent months) below what would be expected. In other words, taking out the influence of September status, Teacher *a* is performing 6.7 GE months below expectation.

The complete table of residual scores for all 48 teachers follows:

Teacher	Sept	June	Pre-Y	Residual
a	3.2	3.5	4.16	-0.66
b	2.8	4.1	3.77	0.33
c	2.5	4.2	3.48	0.72
d	3.4	4.3	4.36	-0.06
e	3.5	3.7	4.45	-0.75
f	3.1	4.3	4.06	0.24
g	2.8	3.1	3.77	-0.67
h	3.9	5.2	4.84	0.36
i	2.5	2.9	3.48	-0.58
j	2.3	3	3.28	-0.28
k	3.2	3.9	4.16	-0.26
l	3.4	3.8	4.36	-0.56
m	4.2	5.7	5.14	0.56
n	3.9	4.3	4.84	-0.54
o	3.8	5.2	4.75	0.45
p	3.6	5.4	4.55	0.85
q	4.6	6.8	5.53	1.27
r	4.1	5.5	5.04	0.46
s	3.9	3.9	4.84	-0.94
t	4.2	4.8	5.14	-0.34
u	4.1	4.6	5.04	-0.44
v	4.5	5.9	5.43	0.47
w	3.8	4.5	4.75	-0.25

x	3.7	5.6	4.65	0.95
y	5.1	6.7	6.02	0.68
z	4.8	5.5	5.72	-0.22
aa	4.7	5.3	5.63	-0.33
bb	5.2	5.7	6.11	-0.41
cc	5.5	6.3	6.41	-0.11
dd	5.1	6.2	6.02	0.18
ee	4.2	6.3	5.14	1.16
ff	4.3	5	5.24	-0.24
gg	5.3	5.9	6.21	-0.31
hh	5.5	6.9	6.41	0.49
ii	4.8	5.4	5.72	-0.32
jj	4.6	5.9	5.53	0.37
kk	5.9	5.9	6.80	-0.90
ll	5.6	6.4	6.51	-0.11
mm	6.1	7.2	6.99	0.21
nn	6.2	7.8	7.09	0.71
oo	5.6	6.3	6.51	-0.21
pp	5.9	6.2	6.80	-0.60
qq	6.1	6.5	6.99	-0.49
rr	5.8	7.1	6.70	0.40
ss	5.7	6.3	6.60	-0.30
tt	6.6	7.3	7.48	-0.18
uu	5.8	6.2	6.70	-0.50
vv	5.2	6.8	6.11	0.69

From this point you can download a Windows Excell Spreadsheet that contains the data above and the results of the regression analysis that Excell performs. Take this Macintosh version if that is your flavor. (You really want to have the data analysis toolpak available in Excell to perform these regression analyses, otherwise you have to make separate calculations of slope and intercept using the $\mathbf{f}_x$ functions, and then form your own equation to calculate residuals.

Putting It All Together

Here is a Java instructional program created by John Behrens and his students that allows you to enter data, see it plotted in a scatter diagram and observe the calculated value of the correlation coefficient. If your browser will handle Java, you can learn a lot about correlation by playing with this program for a few minutes. (The program works well in Microsoft's Internet Explorer 4.0 or better, and may not work so well in Netscape. In fact, it doesn't work at all in my Netscape Communicator 4.51 for Windows. And I notice that even Netscape 6.2 requires the installation of a special "plug-in," that is a bit of a hassle.)

Proportions & Contingencies

Proportions

Statisticians use propotions to convey information about the occurrence of "nominal" characteristics.

The Kinds of Questions Answered by Proportions:

- What proportion of Arizona teachers belong to the AEA?
- What is the proportion of high school graduates who immediately enroll in a community college?

(If you have never encountered the term "nominal measurement" before, take this mini-lesson on Scales of Measurement then come back.)

Proportions are really just like percents except that they range from 0 to 1.00 instead of 0 to 100. If the proportion of high school athletes who report using anabolic steroids is .053, the 5.3% of high school athletes say they use steroids. (Multiply a proportion by 100 to get a percent.)

The statistical definition of a proportion is simple: count the number of cases in the group of **n** who have the characteristic you are interested in--let's say that number turns out to be **f** (for "frequency")--and then divide **f** by **n** and you have the **proportion** sybolized by **p**:

$$p = f/n$$

Question: A sample of 200 eigth-grade students revealed 140 students who answered "Yes" to the questionnaire item "Is too much homework assigned to you?" and 60 who answered "No" to this question. **What is the proportion of eigth-graders who responded that too much homework was being assigned?**

$$n = 200 \text{ and } f = 140 \text{ , so } p = f/n = 140/200 = .70$$

Contingencies

Contingency table analysis is, perhaps, the most often used statistical technique in the social sciences and in researching the professions. When we studied the association between two variables that were measured on a numeric scale with ordinal or better properties, we spoke of the "correlation" between X and Y. But when we examine the association between nominally measured characteristics (e.g., gender and political affiliation), we speak of the "contingency' between the two characteristics. Contingencies are displayed in tables like the following, called, naturally enough "contingency tables":

	Political Affiliation	
	Republican	**Democrat**
Gender		
Male	55	41
Female	52	70

In this table, we see that in the sample studied, 55 persons are male and Republican, while 70 are Female and Democrat. It is generally more revealing to transform these frequencies into proportions (or percents), as in the following table:

Adults Classified by Gender & Political Affiliation with Proportions by Rows

	Political Affiliation	
	Republican	**Democrat**

Gender		
Male	.57	.43
Female	.43	.57

Now we see that .57, or 57%, of Males are Republican, and .43, or 42%, of Females are Republican. (The .57 comes from the division of 55 by 96; the number of Male Republicans divided by the total number of Males.) Thus, a woman is more likely to be a Democrat than a man is likely to be a Democrat. Or, to put it slightly differently, "Males prefer Republican as a political affiliation more often than females do." Or, in the jargon of statistics, "There is a slight association between Gender and Political Affiliation with men more likely to be Republican than are women." (Notice how the proportions add to 1.0 across the rows in the above table.) Suppose we had chosen to calculate the proportions in the above example by columns instead of by rows. The following table would have resulted.

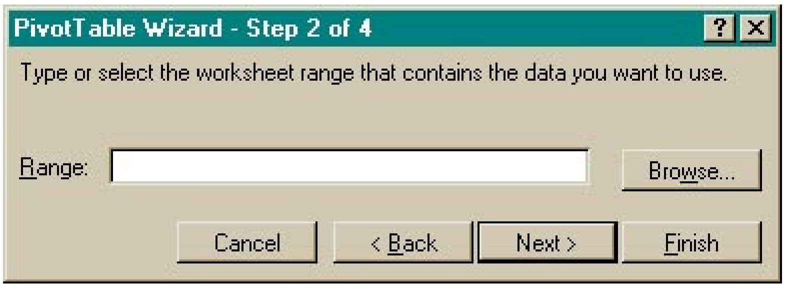
Adults Classified by Gender & Political Affiliation with Proportions by Columns

	Political Affiliation	
Gender	Republican	Democrat
Male	.51	.37
Female	.49	.63

Even though the two above tables contain different numbers, they are both true in what they report; they are simply reporting on different things. The table containing proportions by rows shows how males distribute themselves between the two major political parties, and then, in the second row, how females distribute themselves. The second table, showing proportions by columns, reports how Republicans are divided between males and females, then, in the second column, how Democrats are so divided. It is generally informative to report both sets of proportions (or percents).

Calculating Contingency Tables in Excel

Excel has a weird name for contingency tables; it calls them "pivot tables." To begin the construction of a pivot table, first click in an empty cell somewhere to the right of the data in your worksheet. Next, you'll find the Pivot Table option about half way down the list of menu items under the Data option on the top menu bar in Excel. When you take the Pivot Table option, you'll see Step 1 of $ of the Pivot Table Wizard; just click on Next at the bottom. You next see Step 2 that looks like this:



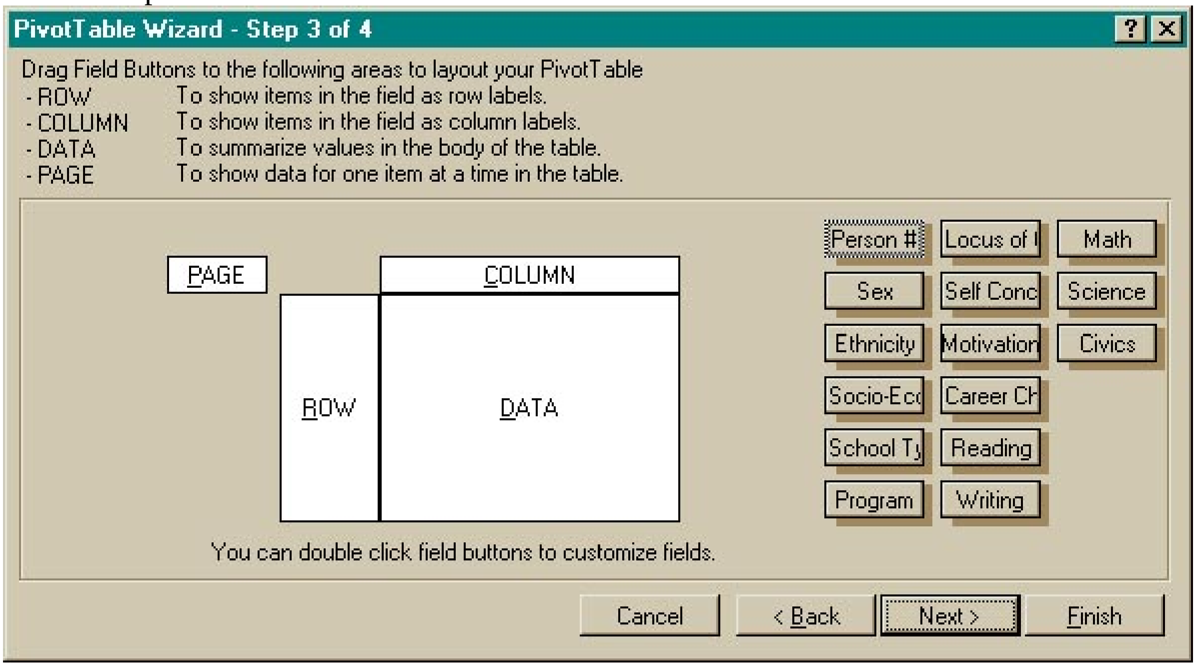
PivotTable Wizard - Step 2 of 4

Type or select the worksheet range that contains the data you want to use.

Range:

The above is the dialogue box for entering the location of the data that you will want to tabulate in a contingency table. The easiest thing to do is to enter a description of the location of all the data in your worksheet, e.g., for the High School & Beyond worksheet that you downloaded, the data occupy a rectangular area from Row 4, Column A to Row 603, Column O. **Please Note: Excel uses your variable names as well in creating a contingency table; notice that the HSB variables names are in Row 3. So you must enter the following code into the dialogue box: a3:o603**. Now click on the Next button.

Now you should see the dialog box below, called Step 3 of 4. This is the box where most of the action takes place.



PivotTable Wizard - Step 3 of 4

Drag Field Buttons to the following areas to layout your PivotTable

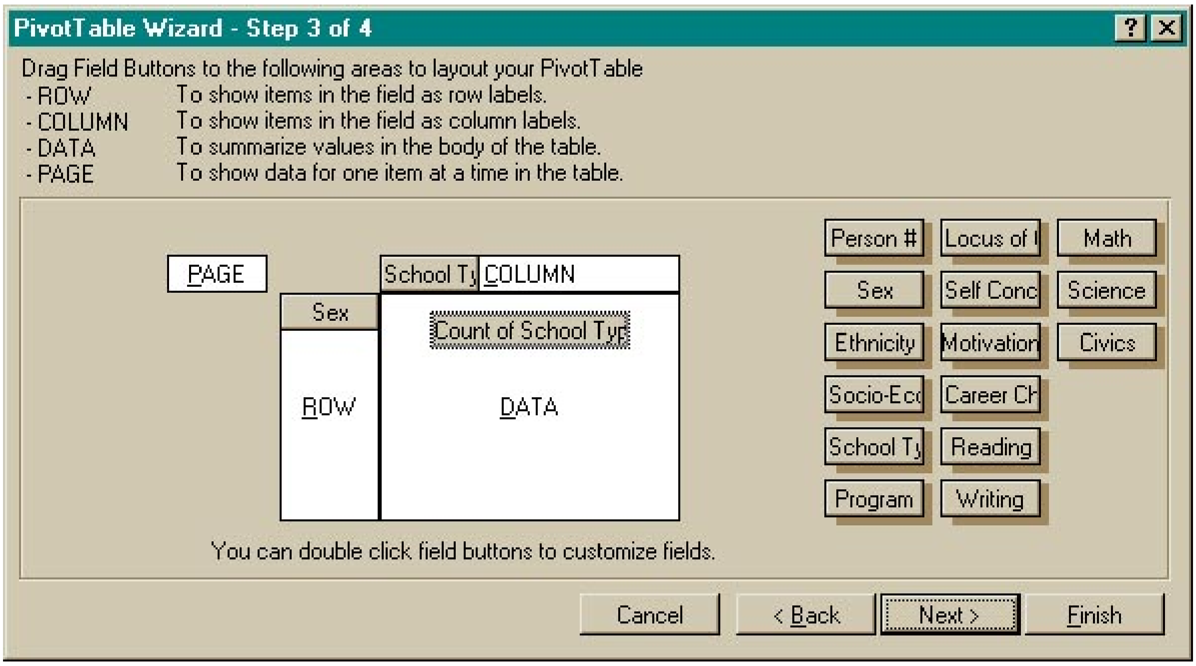
- ROW	To show items in the field as row labels.
- COLUMN	To show items in the field as column labels.
- DATA	To summarize values in the body of the table.
- PAGE	To show data for one item at a time in the table.

You can double click field buttons to customize fields.

The Step 3 dialogue box is where you choose which variable will constitute the rows of your contingency table and which variable will be the columns. Observe the labels for the variables off to the right side.

- Click and drag a variable for rows--e.g., Sex. Drag the label to the left and drop it in the tall rectangle named ROW.
- Now, click, drag and drop the label for School Ty(pe) into the horizontal rectangle named COLUMN.
- Finally, click and drag either the Sex or the School Type label **from the list of variables on the right** and drop it into the center box named DATA.
- If the label you have just dropped into the DATA box suddenly changes its name to "Count of School Type" or "Count of Sex," you are in good shape. If, however, its name becomes something like "Sum of School Type" or "Sum of Sex," you will need to double click that label and pick the Count option.

When the Step 3 dialogue box looks like the example below, you are ready to click on the Finish button:

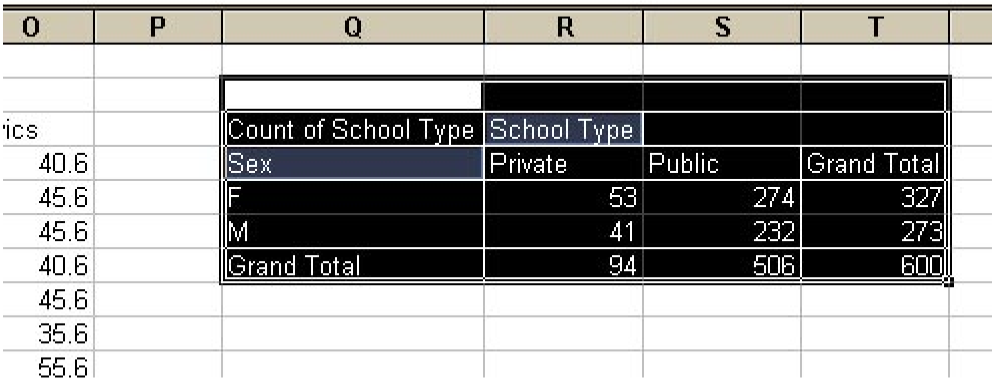


PivotTable Wizard - Step 3 of 4

Drag Field Buttons to the following areas to layout your PivotTable
- ROW To show items in the field as row labels.
- COLUMN To show items in the field as column labels.
- DATA To summarize values in the body of the table.
- PAGE To show data for one item at a time in the table.

You can double click field buttons to customize fields.

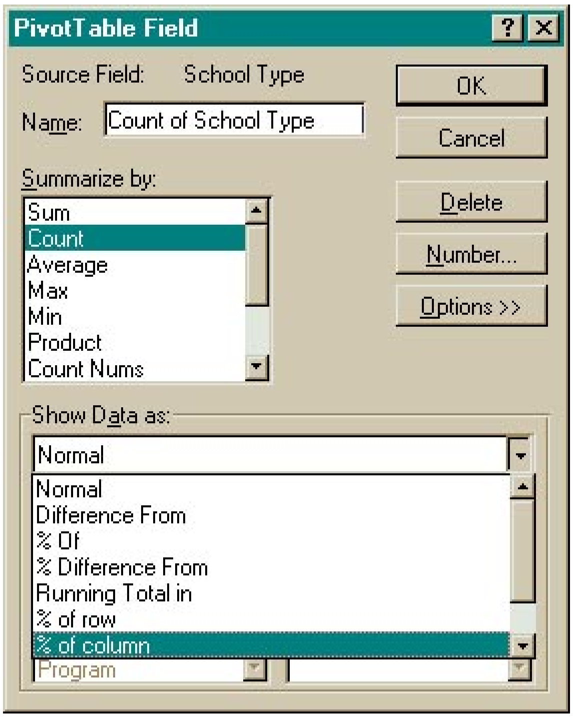
Clicking on the Finish button in the Step 3 dialogue box will produce the following pivot table (contingency table) results:



O	P	Q	R	S	T
ics		Count of School Type	School Type		
40.6		Sex	Private	Public	Grand Total
45.6		F	53	274	327
45.6		M	41	232	273
40.6		Grand Total	94	506	600
45.6					
35.6					
55.6					

Observe what we learn from this: 53 out of 327 females in the High School & Beyond sample attended private schools, whereas 41 out of 273 males went to private schools. This is very informative and would have taken a long time to tally by hand, but there is much more we can learn about this issue when we look at proportions or percents.

Let's click on another empty cell in the worksheet and select the Pivot Table option again from the Data menu. Now, after arranging the Step 3 dialogue box as above and before clicking on the Finish button, let's double click the small variabe label that reads "Count of School Type" (or "Count of Sex"). A dialogue like the one below will appear.



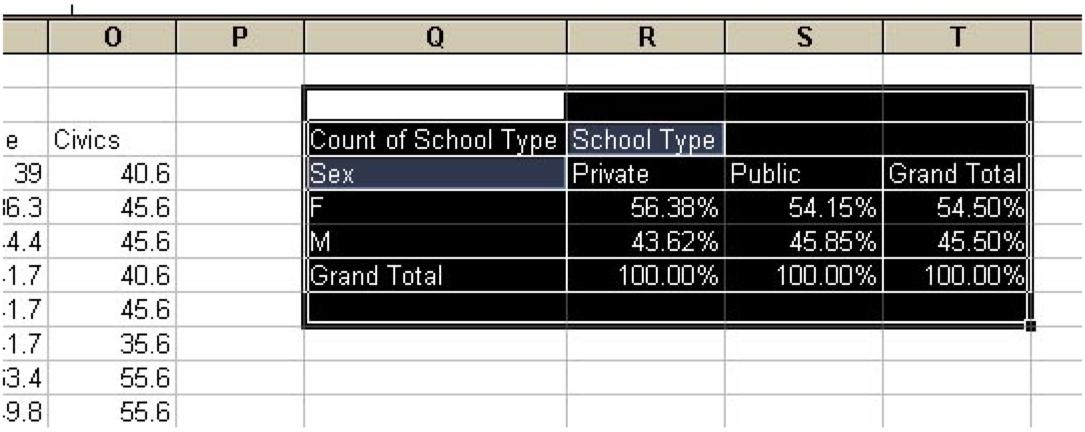
Source Field: School Type

Name: Count of School Type

Summarize by: Count

Show Data as: % of column

From among the five buttons on the right of the box below, click on the Options >> button. When you then click on the tiny arrow beside the box that reads "Show Data as:" you will see a list of choices. Pick the item that reads "% of column." Then following contingency table will result:



	O	P	Q	R	S	T
e	Civics		Count of School Type	School Type		
39	40.6		Sex	Private	Public	Grand Total
6.3	45.6		F	56.38%	54.15%	54.50%
4.4	45.6		M	43.62%	45.85%	45.50%
1.7	40.6		Grand Total	100.00%	100.00%	100.00%
1.7	45.6					
1.7	35.6					
3.4	55.6					
9.8	55.6					

This contingency table is even more easily interpreted. Among all Private school students who finished high school, 56% are female; among all Public school students who finished high school, 54% are female. Thus, the percents of Private and Public school graduates who are female are nearly the same. In statistical jargon, we would say, "There is no association between Sex and School Type." (If the results had proven to be something like 70% of Private school grads are female and only 40% of Public school grads are female, we would have concluded that Sex and School Type are associated.)

One final word about contingency tables: we don't construct contingency tables from data that are measured like height and weight or age or the test score data in the High School & Beyond worksheet. Contingency tables describe the relationships between nominally measured variables.

Introduction to Statistical Inference

Statistical Inference

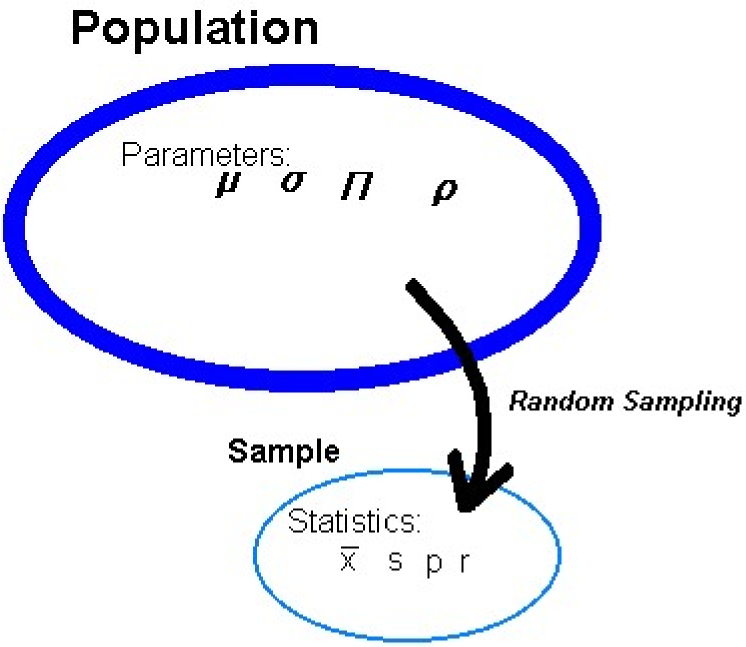
The last portion of this course—and the later stage of most introductory statistics courses—deals with **statistical inference**.
Statistical inference combines the methods of descriptive statistics with the theory of probability for the purpose of learning what **samples** of data tell about the characteristics of **populations** from which they were drawn.

Probability

It is unnecessary to understand the theory of probability to acquire a good working knowledge of inferential statistics. So little more will be said about probability theory per se other than that your intuitive notion will suffice. The probability of an event is the proportion of times it will occur in a long string of independent opportunities. Flip a fair coin 10,000 times. Toss a pair of dice 1,000 times. Guess at the answers to the 500 true-false questions on your final exam.

- The probability of a coin coming up heads is .50
- The probability of a thrown die coming up with a 6 is 1/6 = .17
- The probability that a baby born is male is about .51

Enough said.



Populations & Samples; Parameters & Statistics

The statistical characteristics of populations are called **parameters**. The mean, variance and correlation of x and y in a population are examples of parameters. It is conventional to represent parameters with **Greek letters** (such as mu, μ, pi, π, and rho, ρ, for mean, proportion and correlation, respectively). Note the parameters in the above figure.
Samples are taken from populations to learn something about the parameters. Naturally, one would want the sample to be **representative** of the population if it is to reflect on the population's characteristics. But if the population is so unknown as to have to be sampled to learn about it, then how could it be known well enough to determine whether a sample was representative of it? One could, perhaps, reason that the population of U.S. voters is roughly half male and half female so that a sample of voters shouldl\ likewise be split 50-50 between the sexes. But if sex is unrelated to candidate preference, it counts not at all in favor of the sample's representativeness that it too is half male and half female.
So, the inferential problem confronts a dilemma immediately. If the population is unknown, how can one have confidence that a particular sample from it is representative of it? The solution to the dilemma uses the theory of probability. **If the sample is drawn *according to***

***the laws of probability*, then the degree to which the sample is representative of the population can be calculated in probabilistic terms**. Hence, one will be able to say that the probability is 95% that the sample is representative of the population to a certain degree. Hence, the concept of a "representative sample" is replaced by the concept of a **randomly representative** sample.

Random Samples

A **random sample** is a sample drawn in such a way that each element of the population has an equal and independent chance of being included in the sample—or so the statistician says. So, how do you draw a random sample?

To draw a simple random sample you must

- Give each element of the population an ID number
- Use a table of random digits to select those elements that enter the sample

For example, the 589 school districts in your state are given ID numbers from 001 through 589. You go to a table of random digits in the back of a statistics book and you close your eyes and pick a place to start other than the upper-left of the page. If the first three digits you come to are 412, then school district #412 on your list goes in the sample. Continue in this way until you have drawn all the elements in your sample. (How many? No easy answer. As many as you can afford, unless that gets to be too many. How many is too many? Again, no easy answer. There are lots of phony answers in statistics books about the "right" number of cases to pick for a sample but these answers are generally highly arbitrary and made to look less so. Don't trust them.)

Randomness

Almost nothing other than this tedious process of assigning ID numbers and resorting to tables of random digits will do (though sometimes drawing slips of paper out of a hat isn't a bad approximation to a random sample). You must trust a mechanical process to pick the elements that will be sampled—because, **humans are incapable of random behavior.** Even when you think you are achieving randomness, you aren't.

Do this exercise before reading further:

Drawing from the digits 0 through 9 with repeats, write a string of 20 digits in random order, i.e., 2,6,1,8 etc. Just write them on a scrap of paper right now. Then read on.

Humans are incapable of behaving randomly. Evan when what they do feels haphazard, further analysis reveals unconscious patterns and regularities. For example, in the above exercise, you undoubtedly produced digits with certain nonrandom patterns. Here is a section of a truly random digits from a table in a textbook:

```
6036 5946 4653 3507 5339
4942 6142 9297 0191 8283
1683 7994 2402 5662 3344
4234 9944 1374 7007 1147
3632 9600 7405 3640 9832
3299 3854 1600 1113 3075
```

Now take a look at the string of 20 digits you tried to write down at random and see if there is even a single instance of a digit appearing beside itself, e.g., 2,5,5,1,7 . You probably didn't repeat any digit beside itself because something in you that was producing these digits felt that that wouldn't be random enough. In fact, digits repeat beside themselves quite often in truly random sequences. Look at the above set of six strings of 20 random digits. In only one of the six is a digit not repeated beside itself (in the second string), and in the other strings repeated digits occur more than once. Look at the last string! 99, 00, 111? Does that look random? Not to me; and not to you , I'd guess, but it is. **Moral: Don't try to select randomly by yourself; use a mechanical procedure like a table of random digits**.

The Form of Inferential Statistical Results

Inferential statistics come in two varieties: **interval estimation** and **hypothesis testing**.

Interval Estimation

This form of statistical inference produces an interval of values (e.g., -.12 to +.35) by a process that has a known probability of including the true but unknown parameter value on the interval (e.g., the value of a correlation coefficient in a population). The interval is known as a **confidence interval** and any confidence interval has associated with it a **confidence coefficient** that gives the probability that the interval will capture the parameter. The confidence coefficient is under the control of the data analyst and typically assumes values near 1.0 like .90, .95 and .99. A typical result of interval estimation might take this form:

A random sample of 185 males is drawn from the U.S. population and their heights are calculated. The average is 68.34 inches. The 95% confidence interval for the population mean extends from 66.04 inches to 70.64 inches.

In the above example, the details of how the confidence interval on the population mean is calculated are omitted.

Confidence Intervals for Proportions, Correlations & Means

Correlations

Suppose that you draw a random sample of *n* cases from a population and you wish to learn something about the correlation between two variables, *X* and *Y*, in that population. From the sample you can calculate the estimate of the population correlation coefficient; call that estimate, as usual, *r*. We wish to calculate the 95% confidence interval around *r* and estimating the population correlation, ρ. There are several ways to obtain this confidence interval.

- Use the "nomograph" on page 263 of our textbook. See if you can reproduce this example using the graph on page 263. A sample of *n* equal 50 produces a value of *r* equal to .30. The 95% confidence interval for the population correlation coefficient extends from .02 to .54. The interval you construct from the graph could differ slightly (not more than .02 or .03) from my answer depending on your eyesight. The interpretation of this confidence interval goes like this: The interval from .02 to .54 was generated by a process that has .95 probability of capturing the population value of the correlation coefficient between its bounds. It is somewhat unclear and confusing to say "The probability is 95% that the population correlation is between .02 and .54," though it is not altogether inaccurate—better to say, "This interval was constructed so as to have 95% probability of capturing the true correlation," or more simply, the shorthand version understood by people in the business, "The 95% CI on rho, ρ, is .02 to .54."
- Or, if you have access to the WWW, then you can use an interactive WWW page that will calculate confidence intervals around Pearson product-moment correlation coefficients. Give it a try with *r*= .30 and *n*= 50. Compare the results to the answer you obtained with the graph on page 263 in the textbook.

Proportions

Suppose you draw a random sample of *n* cases from a population in which the proportion of elements possessing some characteristic ("Are left-handed" for example) is equal to some unknown number, call it π. In the sample, the proportion of elements (persons, say) with the characteristic is *p*. You wish to construct an interval by a procedure that has a known probability of including π. If that probability is 95%, for example, then you wish to find the "95% confidence interval on π around *p*.

There are at least three ways to get the job done. I present all three here, but you may find the third method below the easiest and most convenient.

- If the sample you have drawn is fairly large, let's say 100 cases or more, and you are fairly certain that you are not trying to estimate a population proportion π that is extremely small (less than .05, say) or extremely large (greater than .95, say), then a simple approximate 95% confidence interval may do quite well.

The 95% confidence interval on π is given approximately by the following formulas:

Upper-limit = p + 1.96 sqrt[p(1-p)/n],

Lower-limit = p - 1.96 sqrt[p(1-p)/n],

where sqrt stands for square root and *p* and *n* are the sample proportion and sample size, respectively.

Here's an example:

A random sample of 900 persons above the age of 60 revealed that .25 proportion of them had enrolled in college at sometime during their life. What is the 95% confidence interval on the proportion of college "attenders" in the population of persons sample?

UL = .25 + 1.96 sqrt[.25(.75)/900]
= .25 + .0289
= .279

LL = .25 - .0289
= .221

So the 95% confidence interval on the proportion of college attenders in the population of persons over age 60 extends from .22 on the low end to .28 on the high end. Anyone who claimed that a third (.33) of persons over 60 had attended college would be claiming something not in accord with the results of this analysis. They could be correct in their claim, but the sample drawn and analyzed here would have to be one of the 5% of samples that fail to capture the true value of π for the claim that π equals .33 to be true.

- Another method of calculating confidence intervals on sample proportions is to use nomographs similar to that used above with correlation coefficients. On pages 231 of your textbook, there appears a graph that can be used to obtain approximate 95% confidence intervals on the population proportion π.
- A final method of confidence interval for proportions uses the WWW to submit and analyze the data. Enter the sample value of *p* and the sample size to calculate a confidence interval for the population proportion.

Confidence Intervals on Means

Suppose a random sample of *n* cases is drawn from a population that has a mean of μ. The 95% confidence interval around the sample mean is constructed by calculating a formula like the following:

Lower-limit of 95% CI on μ = Mean - t(st.dev.)/sqrt(n), and

Upper-limit of 95% CI on μ = Mean + t(st.dev.)/sqrt(n),

where "sqrt" stands for "square root" and *t* is a number roughly equal to 2, which depends on the size of the sample (for an *n* of 100, *t* equals 1.96, and for an *n* of 15, *t* equals 2.13.
Consider this example. A random sample of *n* = 50 cases is drawn from the population of all beginning 5th grade students in the Mesa School District. The research office is interested in checking on whether beginning 5th-graders in their district score at the national norm level (5.0) in spelling, a subtest of the Language Arts standardized test. In the sample of 50 pupils, the mean equals 4.72 and the standard deviation is 1.15. Using the above formulas, the lower-limit of the 95% confidence interval for the mean of all 5th-graders in the Mesa School District is as follows:
Lower-limit of 95% CI on μ= 4.72 - 2 (1.15/7.07) = 4.72 - .33 = 4.40.

Upper-limit of 95% CI on μ= 4.72 + 2 (1.15/7.07) = 4.72 + .33 = 5.04.

The 95% confidence interval on the population mean, μ, extends from 4.39 to 5.05, and since it includes the national norm value of 5.0, the researchers conclude that there is no reason to suspect that the Mesa 5th-graders are below norm in spelling.
You can use the form supplied here to submit data and receive in return the 95% Confidence Interval on a Mean.

Chi-square Test of Association for Contingency Tables

Statistical inference with contingency tables presents a slightly different set of problems from those addressed with confidence interval estimation. In particular, the inferential question that we ask about a contingency table is a complex relational question concerning several population proportions. So, rather than calculate confidence intervals, we calculate what is called a "test statistic" (in this case a chi-square test statistic); this chi-square statistic is used to reference a table of probabilities that will tell us how probable are the different possible answers to our question.
But first, what is the question we ask of the population contingency table? It is called the **"hypothesis of independence"** or the **"hypothesis of no association."** The reference to independence or association is to the factors used to classify the cases in the table. For example, Sex and Political Affiliation, or Opinion on Abortion (Favor v. Disfavor) and Church Membership (Y v. N).
What does it mean for the two factors of classification in a contingency table to be **independent** or **not associated**? Suppose that 45% of all women are registered Republicans and 45% of all men are registered Republicans. Then, whether a person is male or female, we can say, they are 45% likely to be a registered Republican. This last statement is not conditional on whether a person is a male or female; regardless of Sex, a person is 45% to be a registered Republican. We would say, then, that Sex and Registered Political Affiliation (Republican or Democrat) are "independent" or "not associated." Suppose, however, that 40% of all women—in the population—are registered Republicans, but 50% of all men are registered Republicans. Then, in this case, Sex and Registered Political Affiliation are "not independent," rather they are "associated." We say this because the likelihood of being a registered Republican depends on one's sex; it's 40% if you are female and it's 50% if you are male.
In which of the following **populations** is the hypothesis of "No association" or "independence" between the row and column classifications true?

- **Example 1**

	Male	Female
Golfer	30%	25%

Swimmer	25%	30%
Tennis Player	45%	45%

- **Example 2**

	Republican	Democrat
Favors "School Choice"	50%	50%
Opposes "School Choice"	50%	50%

- **Example 3**

	Boy	Girl
Left-handed	11%	11%
Right-handed	89%	89%

Consult the correct answers here and see how you did.

Note that the above three tables show percents in a population of persons. In reality, we have only samples of cases to tabulate in a contingency table. So, we need a way of deciding whether the sample data are consistent or inconsistent with a population in which the two factors of classification are independent. The chi-square calculation and statistic provides that way.

For a sample contingency table, the **chi-square** statistic is calculated and if it is very large in value (it must always be positive or at least zero), we reject the hypothesis of independence in the population from which the sample was randomly drawn. If the chi-square statistic is small, then we accept the hypothesis of independence of the factors of classification in the population. Whether the calculated value of the chi-square statistic is small or large is decided by the probability of obtaining at least the value of chi-square that was observed when the hypothesis of independence in the population is true. This probability routinely accompanies the reported value of the chi-square statistic.

An example may help:

Suppose that we draw a random sample of 400 school administrators in Arizona and classify them as follows:

	Male	Female
Principal	80	39
Asst. Super.	46	8
Superintendent	24	3

As you can see, in the sample of 200 persons, there are 80 male school principals, 46 male assistant superintendents, etc.

Now, for the above data, the calculated value of chi-square (using either the method in Table 12.3 on page 243 of your textbook or the online statistical analysis programs linked to below) is

Chi-square = 9.60 Prob. = .0082

What this calculation shows is that the probability of obtaining a chi-square statistic this large when there is independence of Sex and Administrative Role in the population samples is 8 in 1000; hence, it is very improbable that Sex and Administrative Role are independent in the population of school administrators in Arizona. By studying the table you can see that a woman is relatively more likely to be a school principal and less likely to be an assistant superintendent or superintendent than a man is.

When is the chi-square test significant? In this example, the probability of obtaining the sample chi-square value was so small (8 in 1000) when there is independence in the population that there is hardly any question that independence of Sex and Administrative Role does not exist in the population. But what if the probability of the chi-square value had been 1 in 100 (.01) or 5 in 100 (.05) or .10, .20 or even .30? At which point does one say, "Yes, this chi-square value is sufficiently probable to be seen when the population has independence that I'll conclude that the two factors are indeed independent"? Well, there is no single probability that separates "significant sample association" from "non-significant sample association." Different circumstances will call for different degrees of evidence. It is conventional to conclude that the two factors of classification are associated (non-independent) in the population when the probability of the chi-square value obtained is .05 or smaller. But it is a convention honored by history only and not by good reasons. Perhaps the best one can do is to report the probability of the chi-square statistic and let all who wish to make their arguments do so.

Online Calculation of the Chi-square Test of Association

Unfortunately, Excel does not have a chi-square contingency table significance test built into it—or at least, I can't find one. Fortunately, there are several places online where you can enter the counts from the cells of your contingency table and have the chi-square significance test calculations made for you.

- Java from South Carolina

Click on the Stat option on the top menu once you are into the spreadsheet portion of the site. Then pick the "Contingency table" option. All you have to do is enter the frequencies (Counts) into the cells of the table you specify (it asks you for numbers of Rows and Columns) and the program calculates the value of chi-square and tells you the probability (pValue) of obtaining such a value if the sample is a random sample form a population in which the row and column classifications are "independent," (not associated).

- A handy online calculator with some explanations from a physics department somewhere in California, I think. Use the general contingency table analyzer and not the special one provided for 2X2 tables.
- A nice little 2X2 contingency table test from Norway that uses Java to calculate Fisher's exact test of association. Don't use this one unless you understand Fisher's test.