Results & Statistics: Description and Correlation

The description and presentation of results involves a number of topics. These include scales of measurement, descriptive statistics used to summarize data, and using figures and graphs to present data.

I. Scales of Measurement – A Review

There are four classes of scales of measurement: nominal, ordinal, interval and ratio. Since we have already covered this, this is a quick review.

A) Scale Types

1. A *nominal* scale is one composed of categories. Measurement consists of sorting into categories.

If we observe people on the street and sort them into categories according to hair color, then we are using a nominal scale.

2. An *ordinal* scale has the property of order. If a set of entities can be put in order, then the scale has the property of being ordinal.

If you were asked to rank order your preferences for a set of ice cream flavors from Ben & Jerry's, then you are using an ordinal scale.

3. An *interval* scale has equal intervals between adjacent points on the scale. For example, the difference between 100 degrees C and 101 degrees C is the same as the difference between 20 degrees C and 19 degrees C.

If you were asked to rate your preference for each of the ice cream flavors on a 0 (absolutely hate) to 6 (favorite, love it) scale, we would treat the differences between adjacent points on the scale as equal.

4. A *ratio* scale has equal intervals and a fixed zero point that corresponds to the absence of what is being measured.

Many physical scales have this property (e.g., length, mass, time). Psychological scales generally do not have this property (e.g. intelligence, depression).

B) Implications

The scale of measurement determines how we can summarize the data descriptively and what inferential statistics are allowed.

For example, with nominal data, we can report which category occurred most often and what percentage of observations occurred in each category. We can not report the mean or variance or other descriptive statistics that require us to add, subtract, multiply and divide. That is because the nominal scale does not have the property of equal intervals required for these arithmetic operations.

With ordinal data, we can also report percentages and results like which item was rated the highest most often. We can NOT compute average orderings because the scale must have equal intervals (to compute means) and the ordinal scale does not. We can compare orderings and compute an index of the degree to which they agree.

Our next topic is how to summarize data.
In dealing with *statistics*, we have two topics: descriptive statistics and inferential statistics. They serve different purposes and will be treated separately. Inferential statistics is the topic of the next section/chapter. *Descriptive statistics* is our topic here.

II. Descriptive Statistics

These statistics are summaries of data, intended to convey information precisely and succinctly. We will examine three types:

Central Tendency (Mean, Median, Mode)

Dispersion (Range, Variance and Standard Deviation)

Relation (Correlation or linear relation)

In addition, we will use the measures of central tendency and dispersion to create *Standardized Scores (z-scores)*.

A) Central Tendency

1. Mean - The sum of the scores divided by the number of scores.

$$\bar{X} = \sum X_i / n$$

X_i are the individual scores and n is the number of scores.

This is what is usually called the average, but average can refer to any of the measures of central tendency.

Most inferential statistics are based on the mean. Remember, data must have interval or ratio scale properties to use this descriptive statistic.

2. Median - The middle score. Place the scores in ascending (or descending) order and then take the one half way through the order. This requires ordinal, interval or ratio data.

If there is an even number of scores, then there are two in the middle. Take the value half-way between these two scores. That is the median.

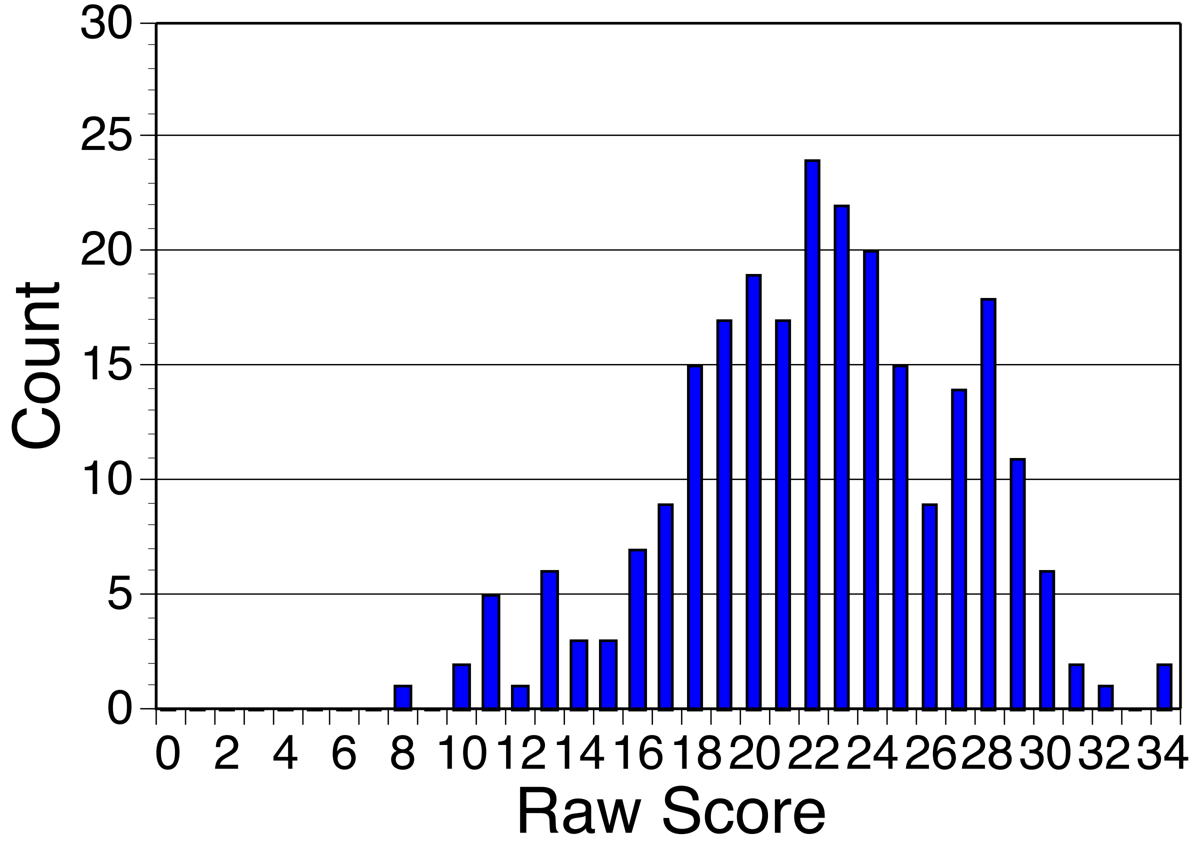
If there are an even number of scores, and the two in the middle are the same, and there are more than two scores with this value, there is a formula to determine the median.

The median is used in place of the mean because it is insensitive to extreme scores. Thus, it is sometimes more representative of the "average" score.

For example, in computing an average income for U.S. residents, the mean would be higher than the median because a relatively small number of people earn extremely large incomes. The median is more representative of what *most* people earn. The average sale price for houses has a similar issue and medians are typically used.

3. Mode - The score that occurs most often. When more than one score occurs most often, there are multiple modes. Modes can be determined for any measurement scale.

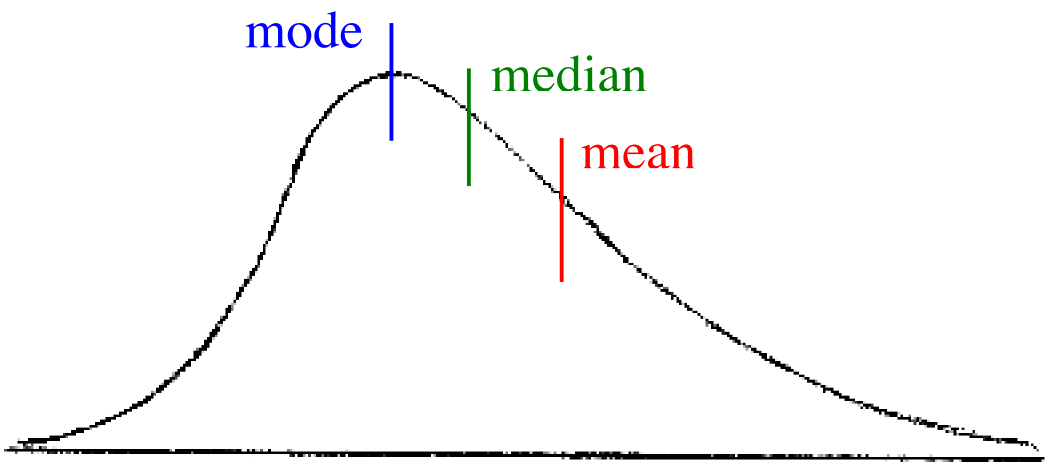
When the distribution of values or scores is symmetric, the mean, median and mode will usually be similar (assuming that there is one score that occurs most often).



In this graph, the distribution of scores on an exam is shown for a large lecture course. The distribution is roughly symmetric and the mean, median and mode are very similar.

Mean:	22.2	
Median:	22	(125^{th} score of 249)
Mode:	22	(occurs 24 times)

This next distribution is not symmetric, it is skewed (positively).



In a skewed distribution, the mean, median and mode will not be the same. Their relative positions are marked on this positively skewed distribution.

The mean is influenced by the relatively small number of extreme (high) scores and is higher than the median. The mode reflects the peak in the distribution but does not adequately reflect the larger range of scores above the median.

Skewed distributions like this often arise when there is a floor (e.g. absolute zero) on a scale but no upper limit. Examples include income and speed of response (reaction time).

B) Dispersion

Data are typically variable. Measures of dispersion show how much variability is in the data.

1. Variance (Standard Deviation) - The variance is the sum of the squared deviations from the mean divided by the number of scores. The standard deviation (s) is the square root of the variance.

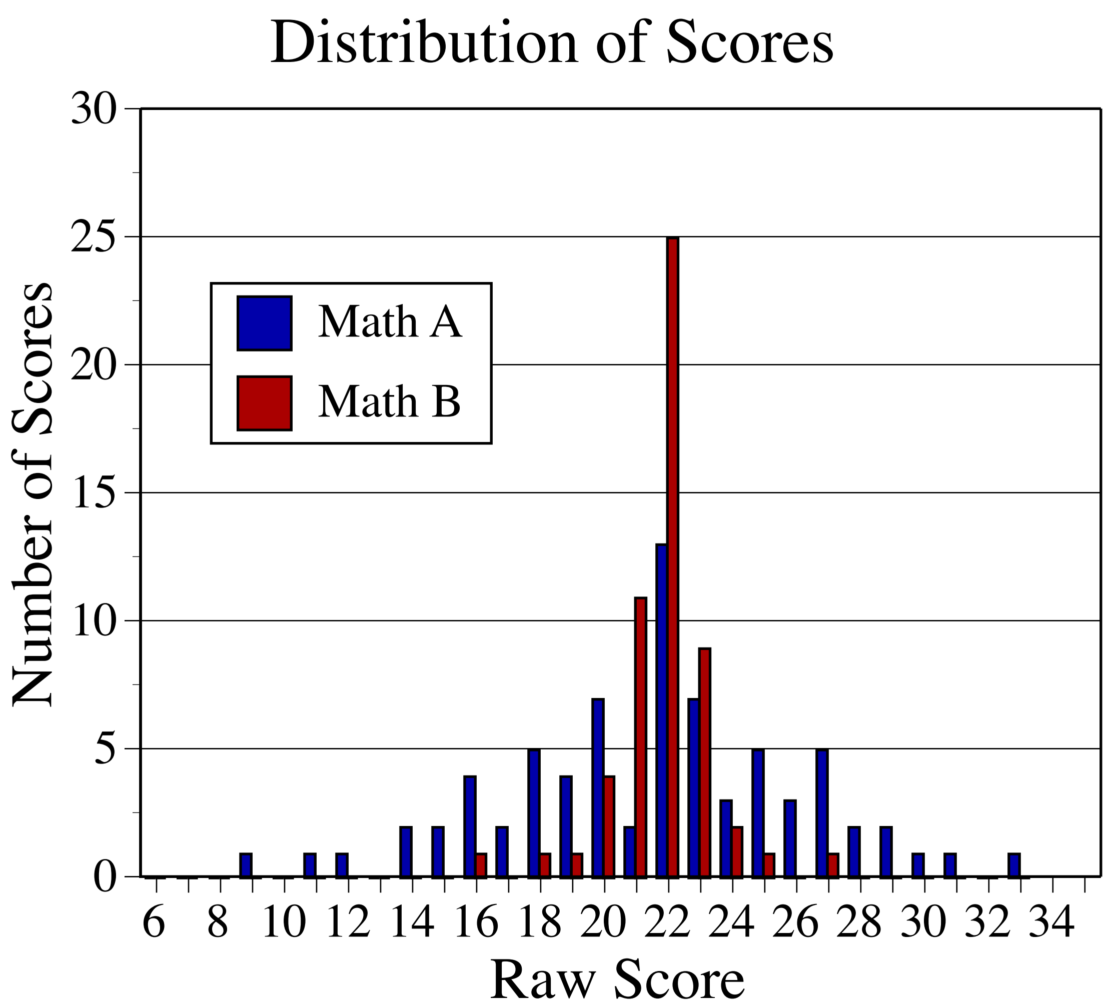
Formula for variance:

$$s^2 = (\sum (X_i - \bar{X})^2) / n$$

2. Range - The difference between the highest and lowest scores. Often, the highest and lowest scores are given rather than the difference.

3. A comparison of central tendency and dispersion.

Two distributions can have the same mean (median and mode) yet have different dispersions.



The distribution for Math B is more compact (smaller variance) than Math A.

C) A use of descriptive statistics in data analysis: Outlier removal for reaction times (RT).

In studies where speed of response is measured, extreme response times (very fast or very slow) pose a problem.

Very fast times probably do not reflect the mental processes that we hope to tap with our experiment. This is because manual responses (like pushing a button in response to the presence of a sound) require a certain minimum amount of time. If a response is faster than that, then it is probably a case where the participant was starting to respond in anticipation. We would like to avoid including these trials (response times).

Very slow times are also a problem. These probably reflect trials where the participant was not attending to the situation (daydreaming or other distraction).

These very fast and very slow response times are "outliers" when compared to the distribution of response times. How do we deal with them?

1. Using fixed thresholds for fast and slow responses.

One approach is to set limits on fast and slow responses. For example, if the process of responding manually to one of two alternatives takes a minimum of 200 ms (0.20 sec), then any response faster than this can not reflect normal processing of the stimulus and a choice response. Similarly, if the processing of the stimulus and response are relatively easy and rapid, then responses might be expected to occur within 2000 ms (2 sec) of the beginning of the stimulus. We could exclude all responses outside the range of 200 to 2000 ms and treat them as trials on which the participant did not respond (or as errors).

2. Outlier removal using the distribution of RTs.

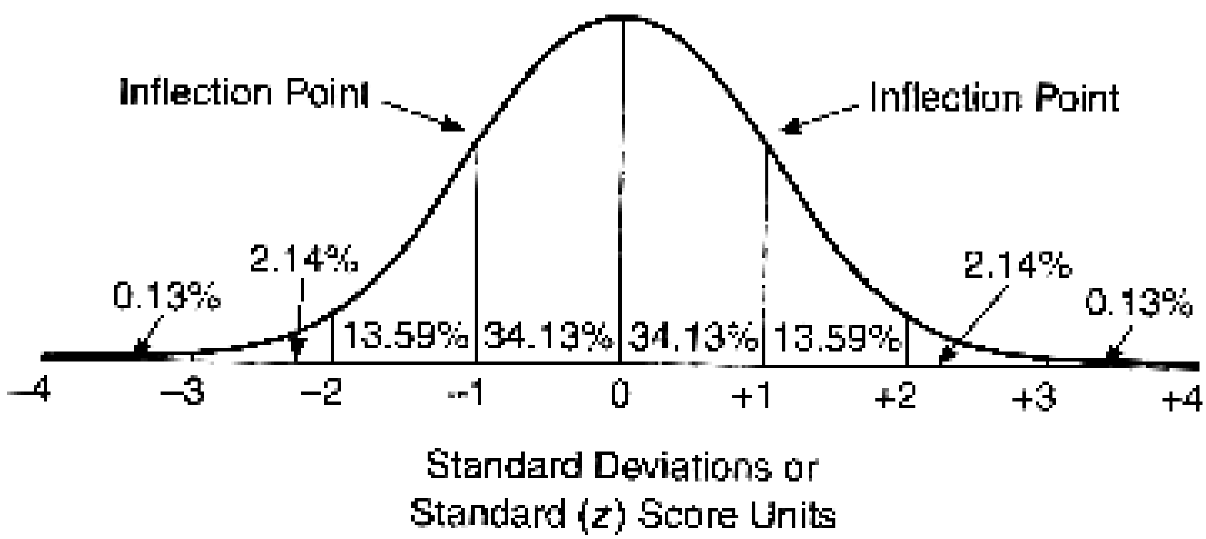
A second approach is to remove trials on which the response time is unusually fast or slow *for each participant*. That is, examine the distribution of response times for each participant and remove those that are at the extremes of the distribution for that participant.

We start by determining the mean and std. dev. for a set of trials for a participant. Then, we check each RT to see if it is within a certain distance from the mean (e.g. 2.5 standard deviations of the mean). Any trial outside these limits is treated as a trial with no response (or as an error). That is, trials far from the participant's mean are outliers.

Outlier RT removal must be done with caution since the phenomena that is being studied might produce extreme response times. This approach is typically used where there are multiple trials for each participant and the outlier removal results in the loss of no more than 2% to 5% of the trials per participant.

D) Standardized Scores and the Normal Distribution

1. The Normal Distribution



If we look at scores from various psychometric scales (e.g. depression, intelligence) and plot the distribution of scores, we generally get a "bell-shaped" curve. In this case, we assume that the data are from a population whose scores have a *Standard Normal Distribution*.

In the normal curve, the mean, median and mode are the same. The point at which the curve changes from concave down to concave up is called the inflection point and is one standard deviation from the mean.

68% of all scores are within one standard deviation of the mean. 95% are within two standard deviations and 99.74% are within three standard deviations.

2. Based on the mean and standard deviation of a set of scores, we can convert an individual score to a standard score (*z-score*). The standard score for an individual is their score, minus the mean, divided by the standard deviation:

$$(X - \bar{X}) / s$$

Computing standardized scores allows us to compare the scores from different samples of individuals or for the same individuals on different measures.

For example, if a course has four exams that differ in their overall difficulty (mean), converting each individual's score to a standardized score (for each exam) and then averaging (mean) the standardized scores will eliminate the differences in difficulty between the exams.

Put another way, the standard score tells us how an individual performed relative to others in the same sample. When we convert raw scores to z-scores (standard scores), we are removing factors such as differences in the difficulty of the task from the scores. If the measurements of each individual are taken using different devices (e.g. exams), then this process removes any differences in average difficulty (was each exam hard or easy overall) and scales the individual scores so that each score reflects performance relative to the group.

III. Relations Between Data - Correlation

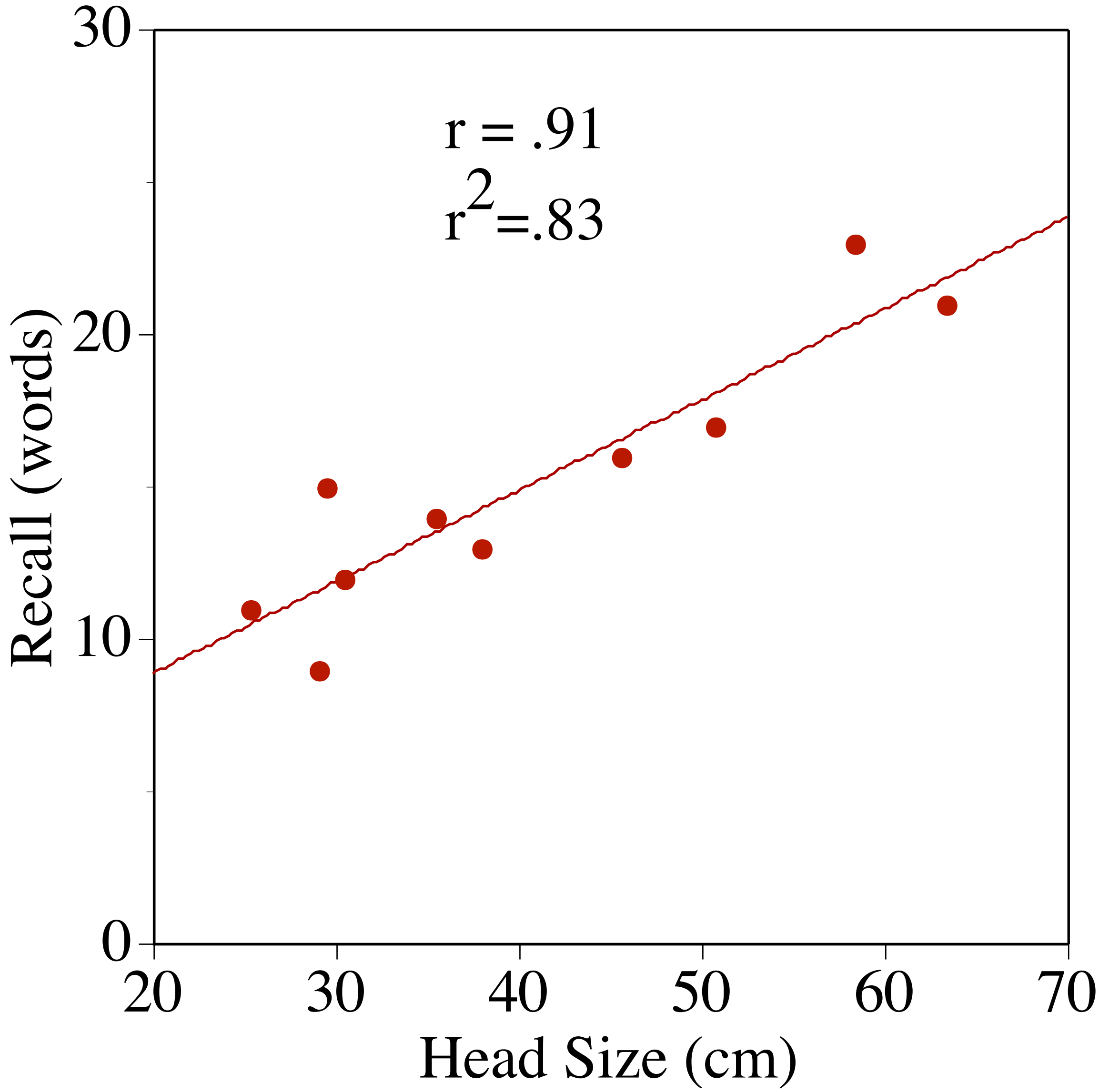
In some cases, we would like to know if the scores on two measures go together. That is, can we predict the value on one score by knowing the other?

One measure of this relation is the *correlation coefficient*. The correlation coefficient reflects the degree to which the relationship between two variables is linear. If we were to plot the data and determine the straight line that best fits the plot, the correlation coefficient would tell us the degree to which the straight line fits the data. There are different coefficients (formulas) for different types of data (different scales of measurement).

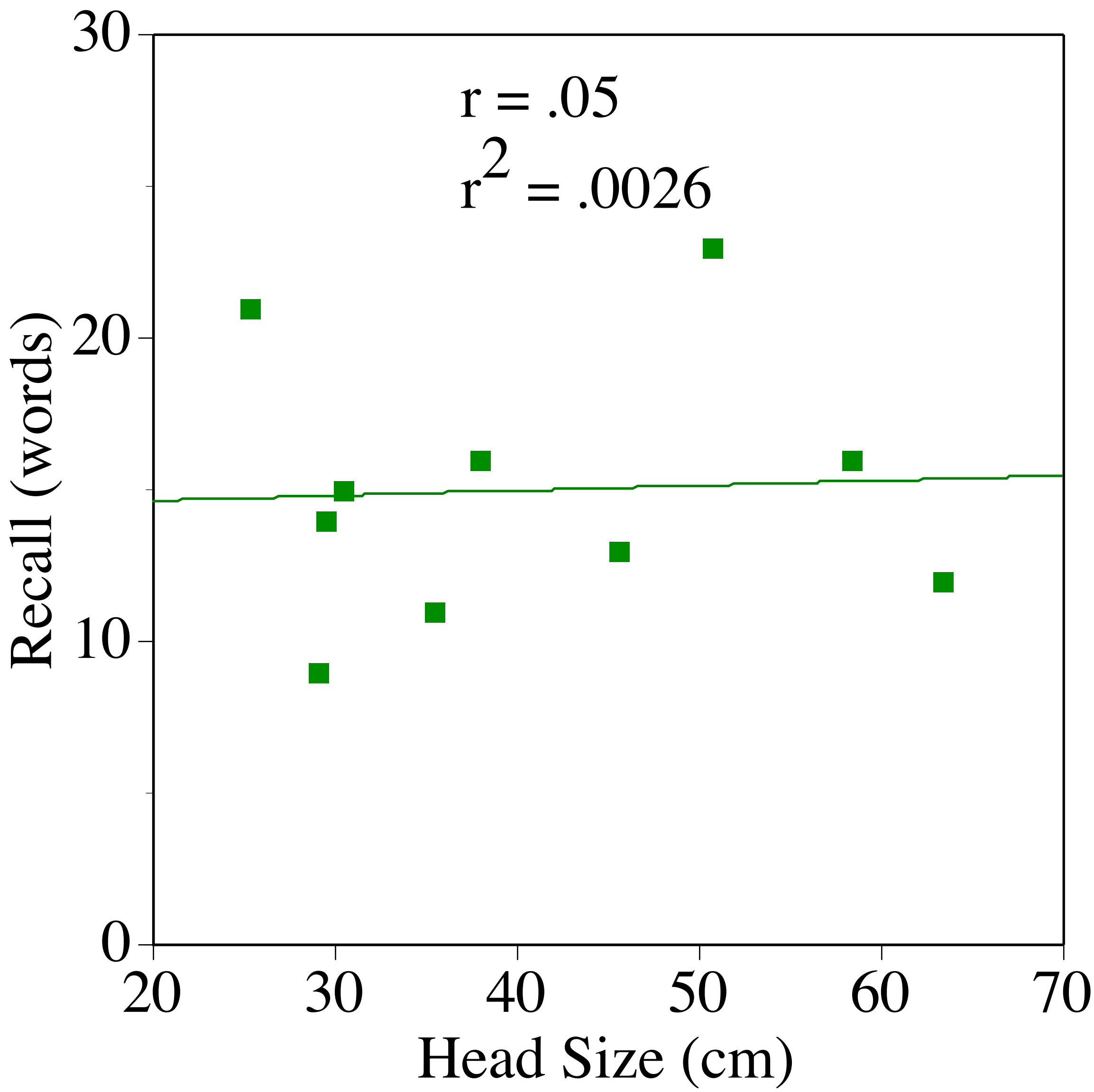
A) Examples

In the next four figures, correlations are shown for four sets of data. These illustrate correlations with data that are (or are not) well described by a straight line.

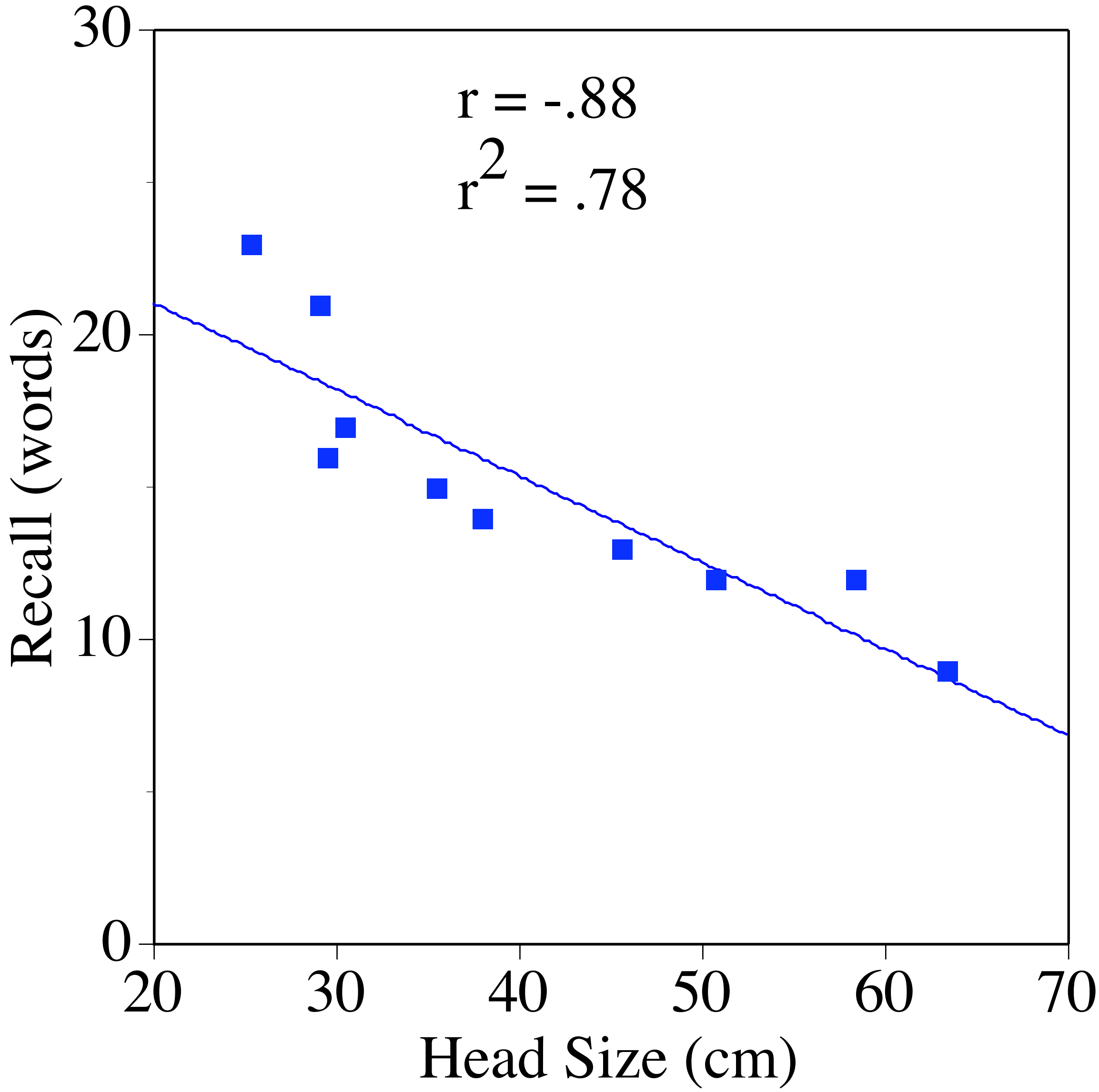
Data showing a strong positive correlation and the best fitting straight line.



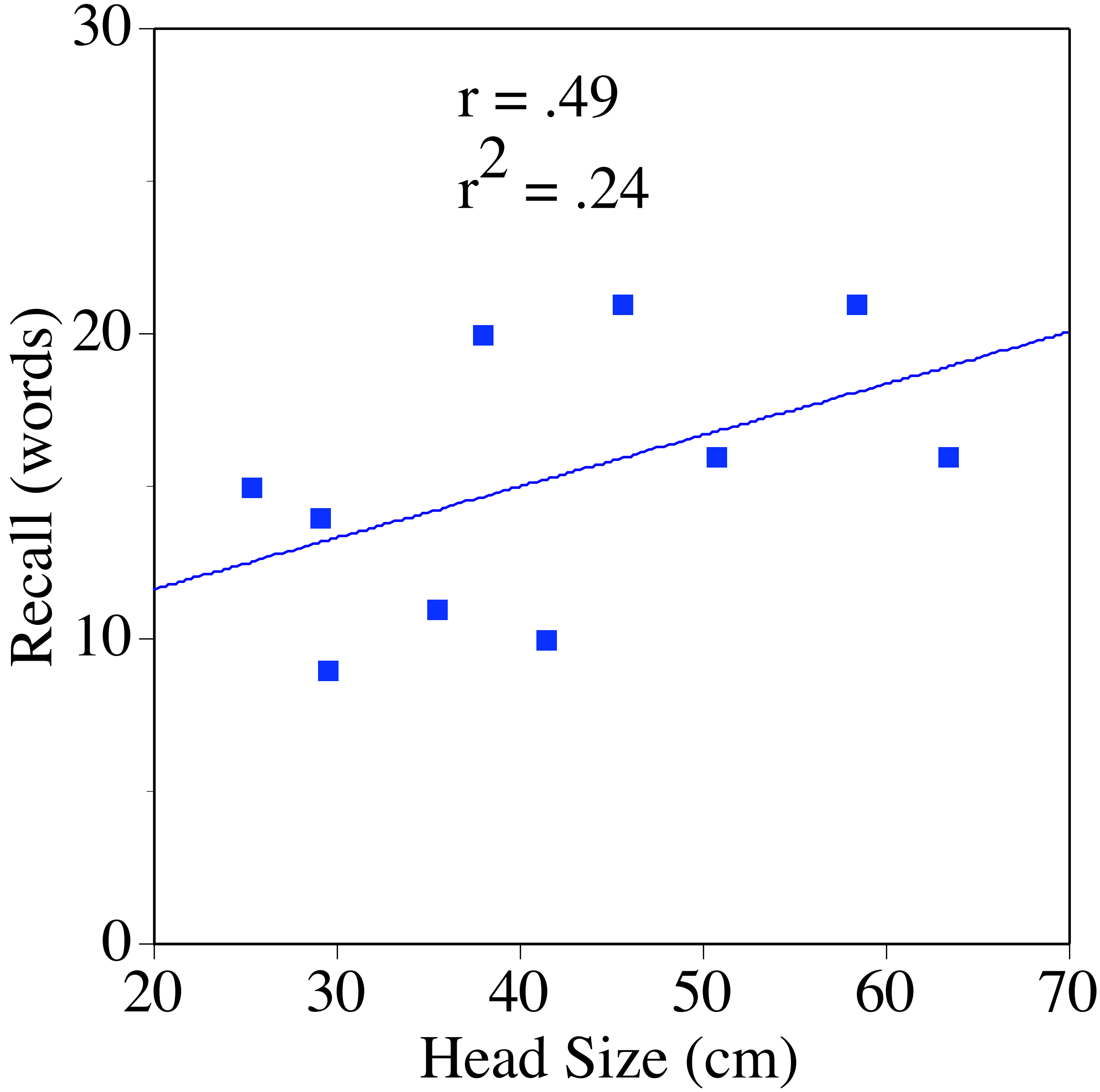
Data showing a near zero correlation and the best fitting straight line.



Data showing a strong negative correlation and the best fitting straight line.



Data showing a moderate positive correlation and the best fitting straight line.



Positive correlations mean that low scores on one scale go with low scores on the other and high with high. In other words, the best fitting line has a positive slope.

Negative correlations mean that high scores on one scale go with low scores on the other and vice-versa (low with high). The best fitting line has a negative slope.

Correlations range from -1.0 (perfect negative) through 0.0 (no *linear* relation) to +1.0 (perfect positive).

As an index of how well we can predict scores on one scale from knowing the scores on the other scale, we square the correlation coefficient (r^2). This is the proportion of variance (variability in the data) that can be attributed to (explained by) the linear relationship.

The significance of a correlation is based on the value of r and the number of data points (sample size).

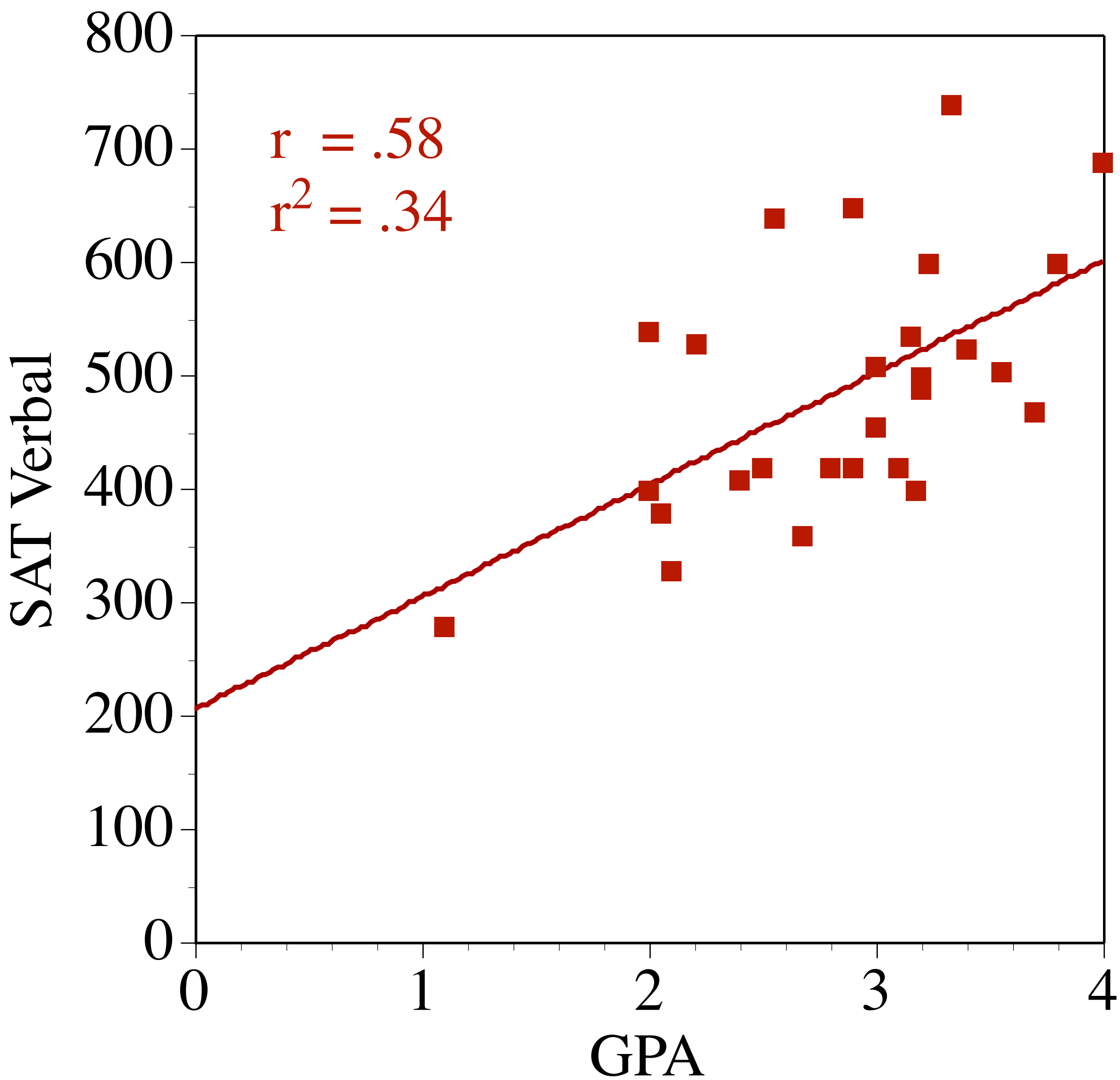
Some cautionary notes:

1. A correlation can be significant because of a large sample size yet account for very little of the variability in the data. For example, with a sample of 100, an r of .23 is significant, but accounts for only 5% of the variability in the data.

2. I*f the range of a variable is constricted or truncated, the correlation will be reduced*. Limited, homogeneous sampling usually results in a truncated range. As an example, GREs of students admitted to graduate school do not predict performance in graduate school very accurately. (Only a limited, upper GRE range is admitted to graduate school.)

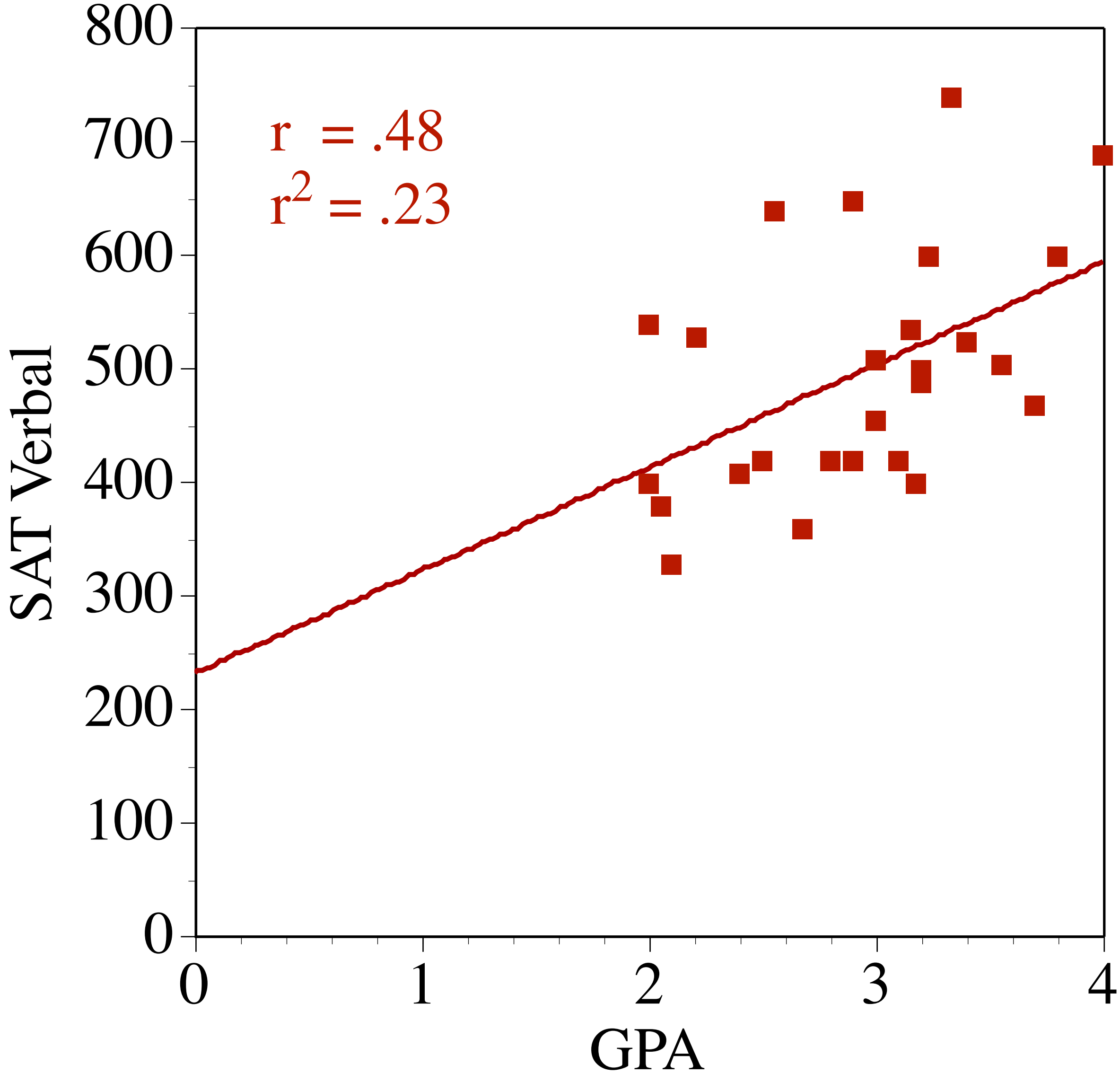
3. *If outlier data are included, the correlation will generally be inflated*. Outliers, in particular, are likely to represent situations where some other factor is influencing the relation.

Example Correlation between SAT Verbal and GPA



Low grade and SAT score included.

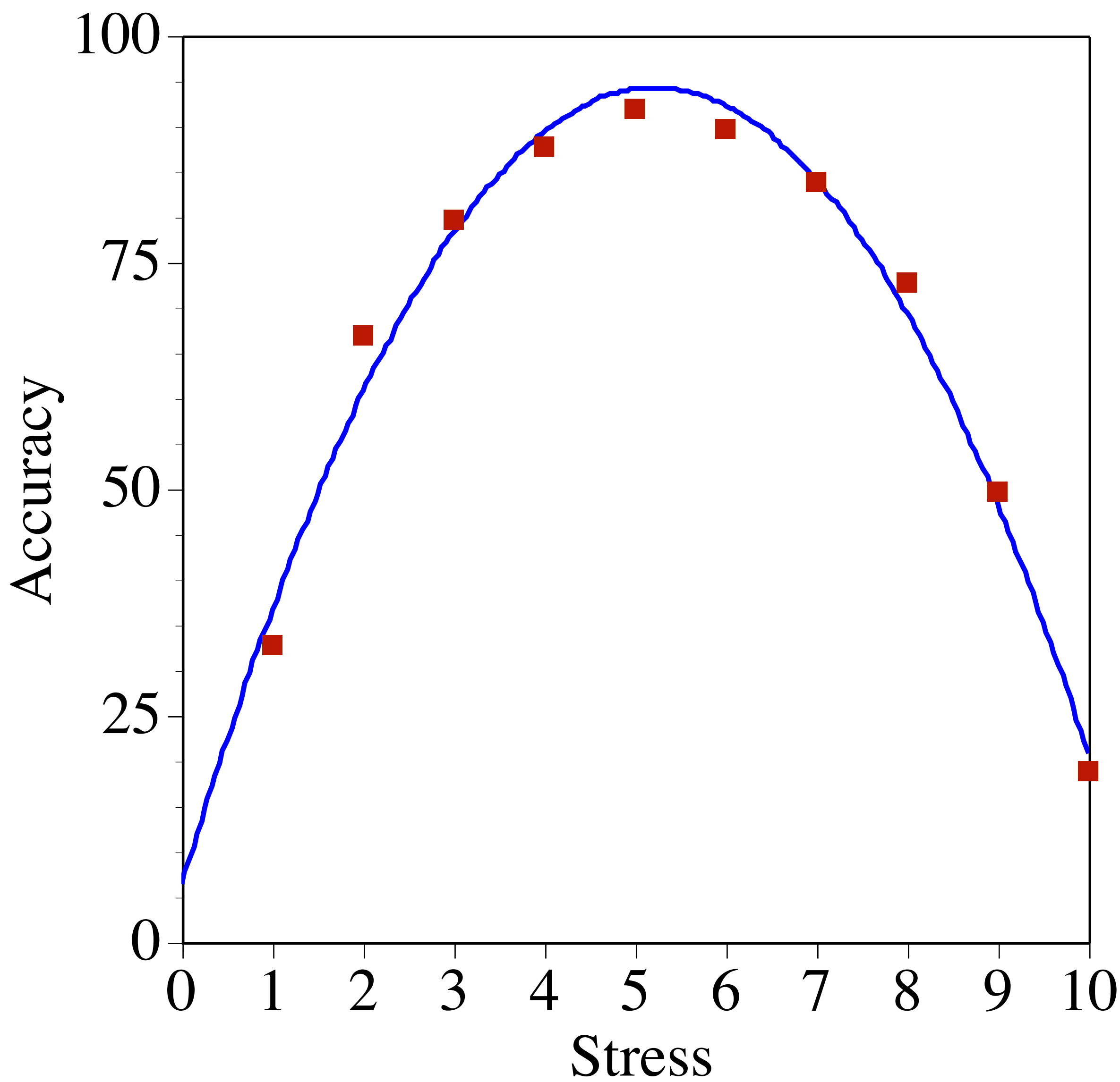
Example Correlation between SAT Verbal and GPA



Low grade and SAT score removed.

4. If the relation between two variables is *not* linear, then the correlation coefficient will underestimate the strength of the relation.

Yerkes-Dodson Law



The relation between accuracy in this task and stress is quite orderly, but clearly not linear.

B) Interpreting Correlation Coefficients

Correlation does not indicate causation (or direction of relation).

Even a sizable correlation (i.e. 0.89 or -0.89) does not imply the existence of a causal relationship between variables. A third variable may be responsible for the strong correlation (negative or positive). Also, a causal relation, if present, could go in either direction.

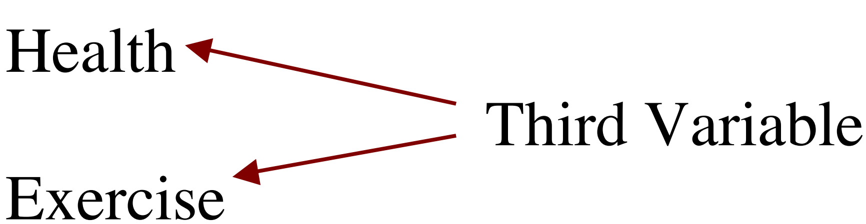
Causation can only be determined from experimental designs.

For example: The relation between physical exercise and overall health.

We find that over a three year period those who exercise maintain better overall health than those who don't. Because of possible confounding variables we can *not* conclude that better health is *caused* by exercise.

1. Direction: Do overall healthy people choose to exercise more or does exercise lead to better health?

2. Third variables: It may be that people who exercise are generally more motivated to control their diets, or sleep more regularly, or avoid high levels of stress, avoid excess alcohol intake, etc.



C) Regression Equations

In the graphs of data of head size versus memory recall, the line that "best fits" the data is also shown. This line is from the *regression equation* that relates head size to recall.

The equation has the form:

$$Y = a + bX$$

Where Y is the variable of memory recall and X in the variable of head size.

For the graph showing a positive correlation of 0.91 between head size and recall, the equation is:

$$\text{Recall} = 2.6 + 0.3 \text{ x Head Size}$$

This regression equation allows us to predict recall (the criterion variable) for an individual if we know their head size (the predictor variable).

Regression equations are useful for predicting an individual's unknown score on some measure from knowledge of their score on another measure (the predictor variable).

Before we do this, we have to establish the relationship between the criterion variable and the predictor variable. Furthermore, the relationship should be reasonably strong (a moderate to high correlation).

IV. Advanced Correlational Techniques

Our discussion so far has been about simple correlations between two variables.

However, there is often more than one causative factor in behavior. We would like techniques that allow us to use more than one predictor variable for predictions.

We would also like techniques for dealing with potential third variables in correlations.

A) Multiple Correlation

The multiple correlation (R) is the predicted performance for a measure based on two or more predictor variables. In this case, we are predicting performance on the criterion variable based on a linear combination of scores on the predictor variables.

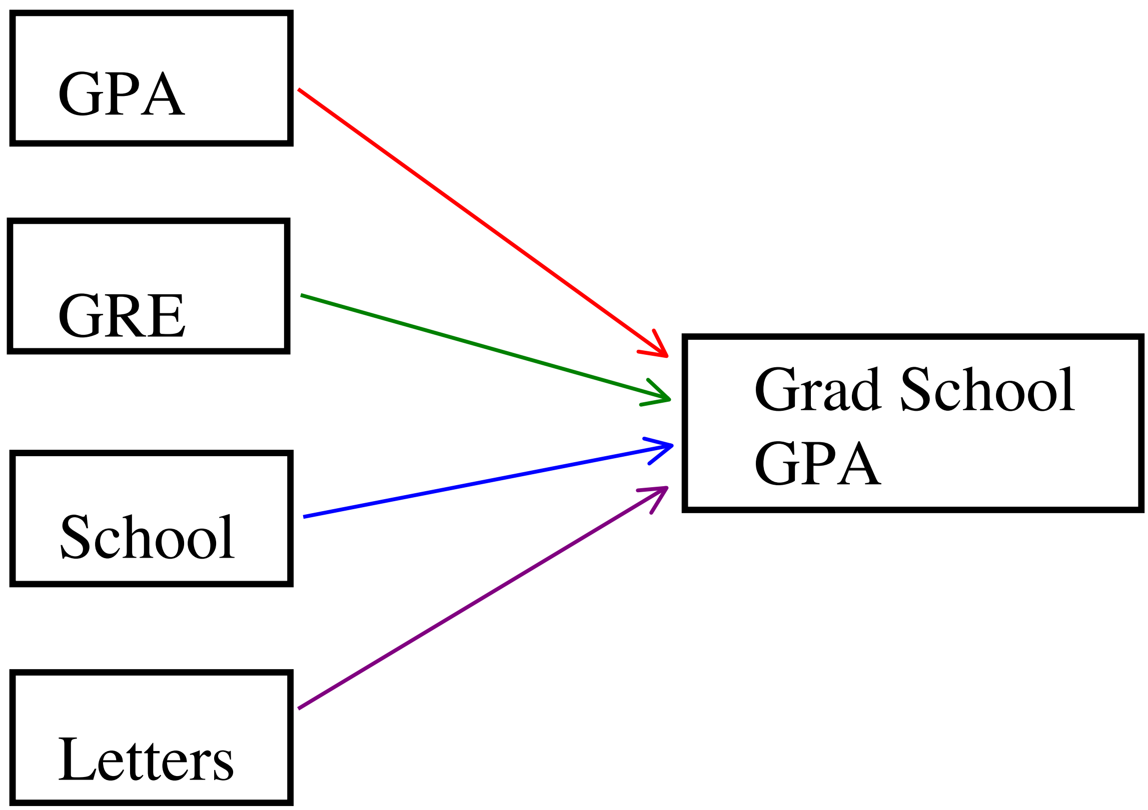
For example, to predict performance in graduate school, we might use 4 predictor variables:

1. Undergraduate GPA
2. GRE total score
3. A rating of difficulty of undergraduate school/program
4. A rating on the favorability of the letters of recommendation

The multiple regression equation would combine these predictor variables to yield a predicted score on the criterion variable.

As with basic correlational data, we would first have to measure all of these variables and establish that there is a reasonably strong relationship of each predictor to the criterion variable.

Conceptually, our example can be visualized as:



Each one of our predictor variables should contribute to the accuracy of the prediction (account for a significant increase in the overall variance accounted for).

B) Partial Correlation

How do you deal with the possible influence of a third variable on understanding the correlation between two variables?

Using our exercise and health example, how do we control for the effects of socioeconomic status (income), stress, diet, and other factors on the relationship between exercise and health?

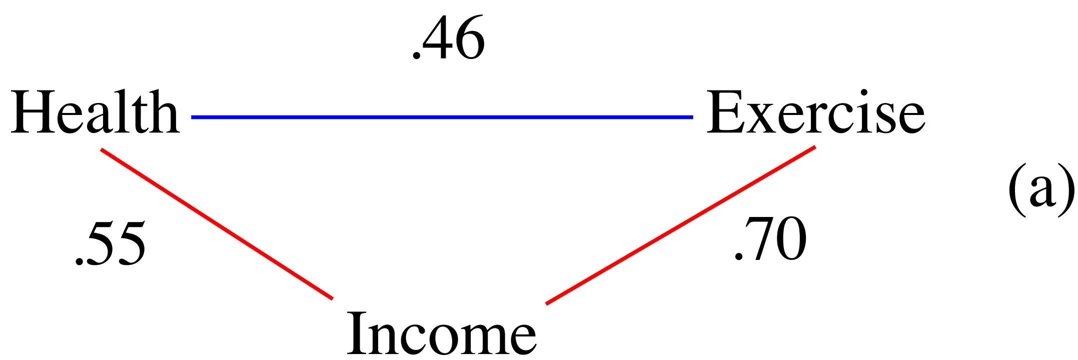
Using experimental techniques, we can control third variables. *Partial Correlation* is a statistical technique for doing this.

In Partial Correlation, we control third variables statistically and remove their influence from the data. We examine the correlation between our variables before and after each third variable is removed.

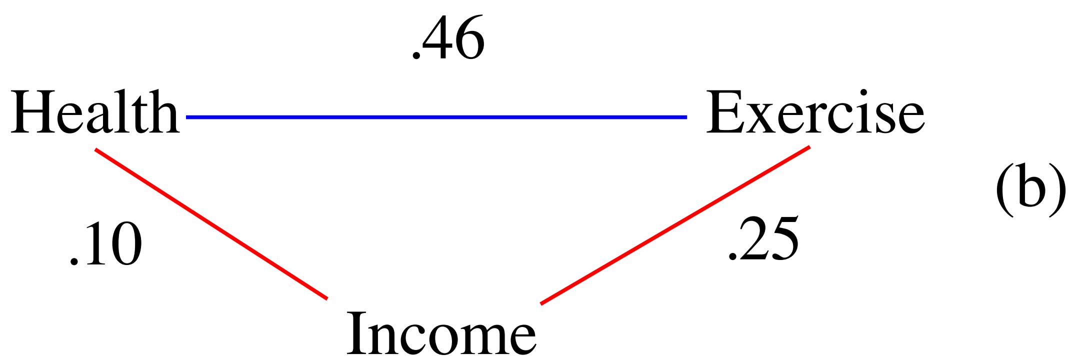
In order to do this, we must measure our participants on the third variables that we are concerned with. *We can not remove the effect of a possible third variable if we have not measured it*. This limits the usefulness of this technique. The researcher must plan ahead on possible third variables and measure them as a part of the study.

In our health and exercise example, suppose that the correlation between health and exercise is .46 which is a moderate relationship. We also measure family income for our participants.

In panel (a), below, income correlates strongly with both health and exercise. When the effect of income is removed using Partial Correlation, the correlation between health and exercise that remains is very low (near zero).



In panel (b), income does not correlate strongly with health (it does correlate with exercise). Here, partial Correlation reveals that the correlation between health and exercise remains (largely unchanged) after removing the effect of income.



So, we can control for the effect of income in our example by measuring it and removing its influence statistically.

Remember, though, that you can not measure every third variable and you can only remove the effects of those that you do measure. Also, partial correlation does not address the issue of the *direction* of any relationship.

C) Path Analysis and Structural Models

Another approach to dealing with correlational data is to propose a model of the pattern of interrelations among variables. The data are then collected on behavioral measures of the various components and the model is evaluated for how well it predicts the relations among the variables. This is *Structural Equation Modeling*.

Path Analysis is similar and used to evaluate the possible causal paths for the relationships among variables.

Both techniques are primarily used in Social and Clinical Psychology were differences between individuals (subject variables) are explored in terms of their relationship to feelings, personality and behavior.

V. Graphs for Data Summary and Presentation

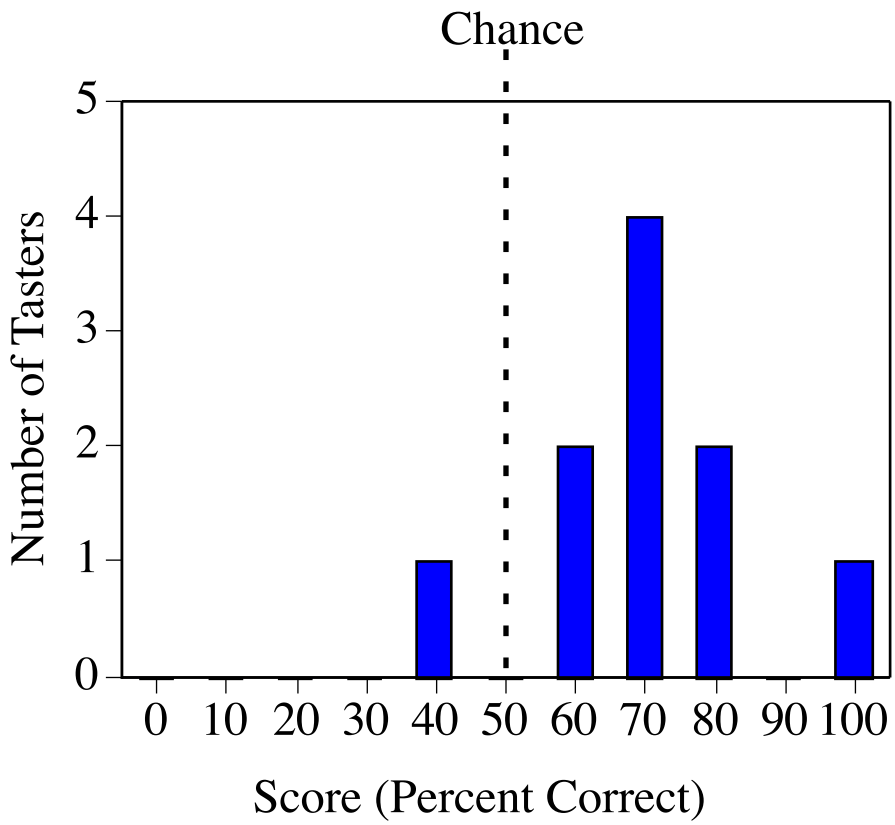
A) Graph Types

The choice of the type of graph (pie chart, bar graph or line graph) depends on the scale on which the data are measured and the interpretation that the researcher wishes to convey.

For nominal scale data a bar graph (of frequency or percentages) or a pie chart (of percentages) is appropriate. A line graph is *not* appropriate because the scale has no order. Without order, it is inappropriate to connect the points.

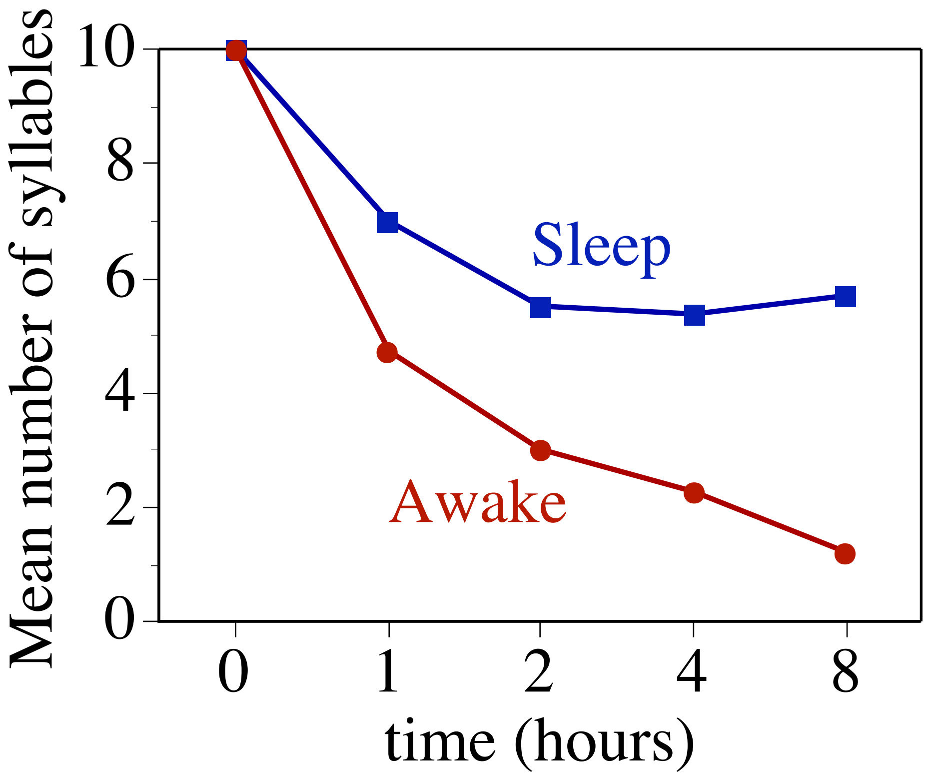
When a scale has the property of order, then data on percentages or frequencies can be displayed using a line graph.

Distribution of Cola Labeling Scores



In this example, a bar graph is used to show the distribution of performance in a cola tasting test. Every taster was asked to sample a set of colas (either Coke or Pepsi) and label each one as Coke or Pepsi. The percent correct was computed for each taster and the number of tasters at each level of accuracy is shown here.

Line Graph Example



Effect of activity (sleep versus awake) on recall. Data from Jenkins & Dallenbach (1924)

In this example of a line graph, the x-axis represents time. This scale has the property of order, so the points can be connected using a line graph.

B) Lies, Damned Lies, and Statistics (M. Twain)

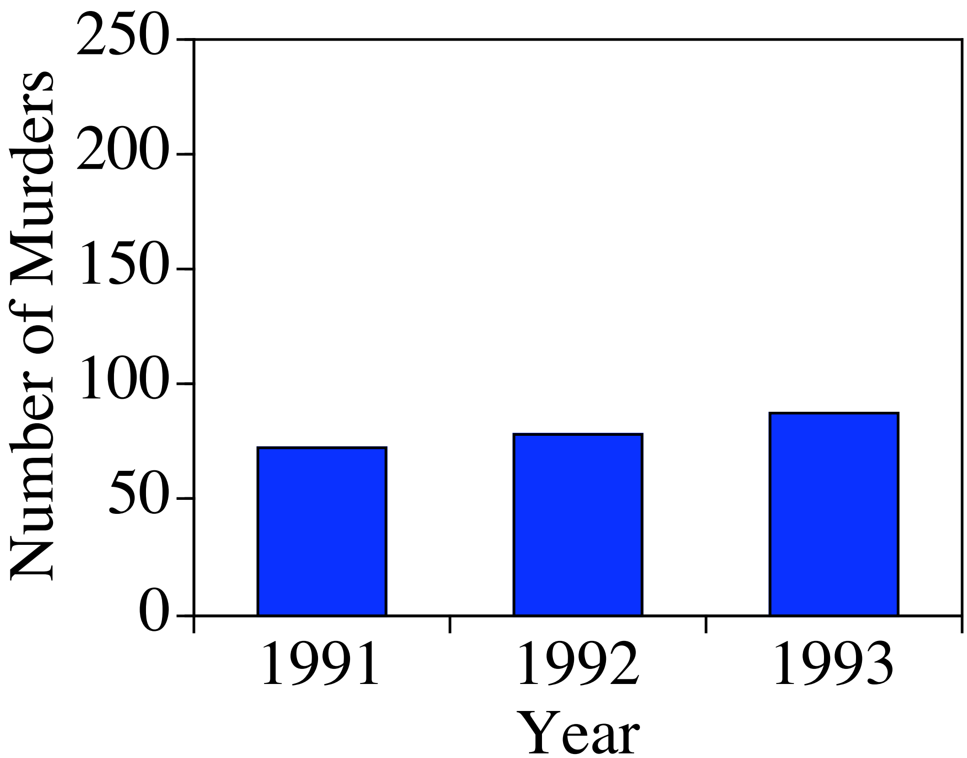
The vertical axis scale can be used to exaggerate or understate the size of any trend in the data. The next two figures show this with the same data plotted on two different scales.

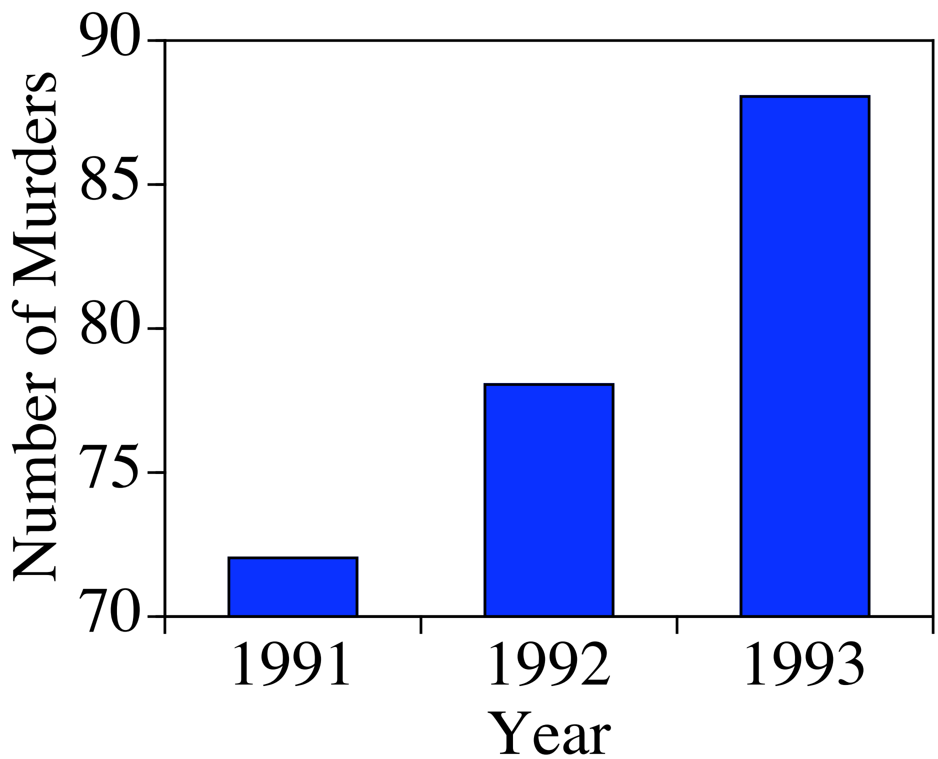
In a sense, both are lying with the data. The compressed scale in the top panel goes up to a frequency count that does not occur. This minimizes the difference between the two years.

The bottom graph expands the vertical (frequency) scale. By doing this, the difference between the two years is made to look large.

The scale should be chosen based on the possible range of scores and to reflect the variation that can (does) happen. Different graphs should use the same scales. Otherwise, it is difficult or misleading when the reader tries to compare the results across the graphs.

Effect of scale on graph of data.





Answers to Chapter 11 Sample Questions

1) – d; 2) – d; 3) – b; 4) – c

Chapter 12 Sample Questions

1) If we found that self-esteem and grade point average were correlated, this would imply: a) self-esteem causes grade point average b) intelligence causes both self-esteem and grade point average c) self-esteem and grade point average are not causally related d) none of the above

2) The standard deviation: a) is the average deviation around the mode b) is a measure of the dispersion or variability in the data c) can be computed for any scale of measurement d) b & c above

3) If the data from a study contained a small number of scores that differ *highly* from all other scores, which measure of central tendency should be used to prevent sensitivity to these extreme scores? a) the standard score b) the median c) the mean d) a & c above

4) In general, we can not infer causal relationships from correlational data. Why? a) The direction of any causal relation between the variables is not usually known. b) We can not know for certain if one variable caused the other, or if there is an underlying third factor that caused the relation between the variables studied. c) We can not establish the size of the relationship between the variables. d) a & b above