QUANTITATIVE DATA ANALYSIS

Inference on 1 & 2 - sample means
- Unpaired & paired t-tests
- ANOVA
Statistics

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting, and analyzing data as well as with drawing valid conclusions and making reasonable decisions on the basis of such analysis.
Statistics in Medicine

• The *New England Journal of Medicine* calls *Application of Statistics to Medicine* one of the most important medical developments of the past 1,000 years – ranking with such milestones as the discovery of anesthesia and the discovery of antibiotics.

Scales of Measurement

The degree of precision with which a characteristic is measured determines the statistical methods that may be used to analyze the data.

The three scales of measurement that occur most often in medicine are **nominal, ordinal, and numerical (metric)**.
Nominal Scales

Nominal variables (also called categorical variables) simply categorise sample values into an appropriate category. We cannot apply any of the rules of arithmetic to nominal data.

Example: Classification of anemias as microcytic anemias (including iron deficiency), macrocytic or megaloblastic anemia (including B\textsubscript{12} deficiency), and normocytic anemias (often associated with chronic disease).
Nominal Scales

• Data evaluated on a nominal scale are also called *qualitative* observations or categorical observations.

• Nominal or qualitative data are generally described in terms of percentages or proportions.

• Contingency tables and bar charts are most often used to display this type of information.
Ordinal Scales

If there is an inherent order among the categories, the observations are said to be measured on an ordinal scale. Observations are still classified as with nominal scales, but some observations have “more” or are “greater than” other observations.

Example: Colorectal tumors are classified as stage 1, 2, 3, or 4 (corresponding to Dukes stages A, B, C, and D), where 1 is an in situ tumor and 4 is distant or systemic disease.
Numerical Scales

Observations for which the differences between numbers have meaning on a numerical scale are sometimes called **quantitative** observations, because they measure the quantity of something. There are two types of numerical scales:

**Continuous**: height, weight, length of time of survival, laboratory values such as serum glucose, sodium, or uric acid.

**Discrete**: number of pregnancies, number of previous operations, number of risk factors
• Characteristics measured on a numerical scale are frequently displayed in a variety of tables and graphs.

• Means and standard deviations are generally used to summarize the values of numerical measures.

• Nominal and ordinal (Qualitative data) variables are inherently discrete.

• Numeric variables (Quantitative data) can be either discrete or continuous.
Quantitative & Qualitative Research - Examples

• How many parents would consult their general practitioner when their child has a mild temperature?
• What proportion of smokers have tried to give up?
• Why do parents worry so much about their children’s temperature?
• What stops people giving up smoking?
Frequency Table

Raw data – Collected data that have not been organized numerically

When summarizing large masses of raw data, it is often useful to distribute the data into classes, or categories, and to determine the number of individuals belonging to each class, called the class frequency. This tabular arrangement is called a frequency distribution, or frequency table.
Histogram

• **Grouped data** – Data organized and summarized in a frequency table

• **Histogram** is a graphic representation of a frequency distribution. Frequency curves may be symmetrical (bell shaped) or asymmetrical (skewed)
Graph Types

- Bar
- Line
- Area
- Pie
- High-Low
- Pareto
- Control
- Boxplots
- Error Bar
- Scatter
- Histogram
- Normal P-P
- Normal Q-Q
- Sequence
- Autocorrelations
- Cross-Correlations
- Spectral

Further clever and unexpected results of studying data through graphs
Are Birthdays Uniformly Distributed?

• The Birthday Problem: How may people must be in a room before the probability of two having the same birthday exceeds $\frac{1}{2}$ (the answer is 23).

• In solving the problem it is assumed that birthdays are uniformly distributed throughout the year. Do you think this is a valid assumption?
Are Birthdays Uniformly Distributed?

• There is a set of data listing the number of babies born on each day of 1978. It is available at http://www.dartmouth.edu/~chance/teaching_aids/data.html.

• Suppose we plot this data using a line graph (trend line), using days of the year on the x-axis and number of births per day on the y-axis. What would you expect to see if birthdays are uniformly distributed? Is this what you expect to see in this case?
What do you see?

• What are typical values? What are uncommon values? How different are they? What do they mean in the real world?
• Does this look like a uniform distribution?
• Do you see any patterns? That is, do births occur more or less frequently in some predictable fashion?
• What potential patterns would you like to investigate further?
A Little Closer Look

• Here is the same line graph (line graphs customarily indicate passage of time on the x-axis) with gridlines added.

• The horizontal gridlines in blue simply mark 2000’s.

• The light blue vertical gridlines occur in 28-day intervals. The dashed brown vertical gridlines occur in 7-day intervals. Since January 1 fell on a Sunday in 1978, all the vertical gridlines fall on Sundays.
Birthdays on each Date of 1978
What do you see now?

• Is there a noticeable pattern to the dips and spikes?
• Do plausible explanations of such a pattern occur to you? If so, how might we investigate further? Would a different sort of graph help?
• Do you notice any anomalies, values that seem not to fit the pattern? (think major holidays)
Further Investigation

- It looks suspiciously as though fewer children are born on Sundays (or weekends?)
- A natural way to investigate further is to count the children born in 1978 according to the day of the week they were born on.
Children Born By Day of the Week in 1978

Day (1=Sunday)

Number Born

0 100000 200000 300000 400000 500000 600000
What do you see and conclude?

• Noticeably fewer children were born on weekends than on weekdays.
  – A data scientist may be happy to conclude there is a genuine difference based merely on this picture.
  – A data mathematician might want to verify that such a distribution is wildly improbable if children are, in fact, equally likely to be born on every day of the week (hmmm, how would we do that?)

• Notice that it makes no sense to say that the difference is significant since we have a census of the population. There is a huge difference between significant differences and important differences.
Graphical Displays of Data are Unlimited in Their Variety

• Many types of graphs and charts are so commonly useful that they have become standard and have familiar names (histograms, pie charts, line graphs, Pareto charts).

• Nevertheless clever people have often found that no conventional graph suffices to show what is important in a collection of data. This leaves room for great creativity in the invention of new graphical displays and the application of ad hoc methods tailored to unique situations.
Charles Joseph Minard’s Graphic

• In 1861 Charles Joseph Minard produced the following graphic, presenting the disastrous losses Napoleon’s army suffered in its march on Moscow in 1812 (remember the 1812 Overture?).

• This is often described as the best graphic ever produced. It displays with gripping clarity six variables: army size, two-dimensional location, direction of march, and (on the return march) date and temperature.
Historical Background

- Napoleonic started in June(?), but spring came late, depriving him of timely wheat harvests to feed his horses. Heavy rains turned the land to mud, slowing the army. Then the harsh Russian summer hit and men died of hunger, thirst, and sickness in addition to infrequent but bloody battles. Only 100,000 out of the undiverted army of 370,000 reached Moscow. When they were turned back, they faced a merciless winter in their attempt to return to France. Of the 422,000 who marched out, only 10,000 survived to return to France.
The Main Point

• Minard’s graphic is in no way standard, but it brilliantly displays what happened to Napoleon’s army. In the same way, you should see the standard collection of descriptive statistical tools as useful but not exhaustive. It takes skill to use those tools and creativity to go beyond them to produce something better on occasion if you really want to see what is in the data. Indeed sometimes no single graph will do; it may take two or three or more to tell the real story.

• Of course computers are opening up new possibilities with the ability to animate graphs and allow examination of them from many angles and in many different ways in real time.
Measures of central tendency

• The most commonly investigated characteristic of a set of data is its center, or the point about which the observations tend to cluster.
Mean

The most frequently used measure of central tendency is the arithmetic mean or average. The mean is calculated by summing all the observations in a set of data and dividing by the total number of measurements.

The sample mean is denoted by \( \bar{X} \) (read ‘X bar’) and is defined as \( \bar{X} = \frac{\sum X_i}{n} \), where
\[
\sum X_i = X_1 + X_2 + \ldots + X_n
\]

The population mean is usually denoted by \( \mu \)
Mean

Example: The mean of the numbers 8, 3, 5, 12 and 10 is $\bar{X} = (8+3+5+12+10)/5 = 38/5 = 7.6$

• It is sensitive to all of the scores. In other words, if one score in the distribution is changed, the mean will change too.

• The algebraic sum of the deviations of a set of numbers from their means is zero.

• The sum of the squared deviations about the mean is less than the sum of the squared deviations about any other value.
Median

The median of a set of numbers arranged in order of magnitude is either the middle value (if ‘n’ is odd) or the mean of the two middle values (if ‘n’ is even).
**Median**

Example:

The median of the numbers 6,4,3,8,5,4,8,10,8 is 6
- Arranged in order: 3,4,4,5,6,8,8,8,10 (n is odd)

The set of numbers 12,5,18,7,5,11,9,12 has median
\[ \frac{1}{2}(9+11) = 10 \]
- Arranged in order: 5,5,7,9,11,12,12,18 (n is even)
Mode

• The mode of a set of numbers is that value which occurs with the greatest frequency; that is, it is the most common value. The mode may not exist, and even if it does it may not be unique.
Mode

The set 2, 12, 5, 10, 2, 10, 9, 11, 7, 9, 18, and 9 has mode 9.

The set 3, 5, 8, 10, 12, 15, and 16 has no mode.

The set 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, and 9 has two modes, 4 and 7, and is called \textit{bimodal}.

A distribution having only one mode is called unimodal.
Empirical relation

• For *unimodal* frequency curves that are moderately skewed (asymmetrical), we have the empirical relation:

\[
\text{Mean-Mode}=3(\text{Mean-Median})
\]
Measures of dispersion

The degree to which numerical data tend to spread about an average value is called the dispersion, or variation, of the data.

**Range:** The range of a set of numbers is the difference between the largest and smallest numbers in the set.

Example: The range of the set 10,2,6,5,3,5,8,4,7, and 12 is 12-2=10
Variance

The *variance* quantifies the amount of variability or spread about the mean of a sample. It is calculated by subtracting the mean of a set of data values from each of the observations, squaring these deviations, adding them up, and dividing by the number of observations in the data set. Representing the variance by $s^2$,

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

The *standard deviation* of a set of data is the square root of the variance and is denoted by $s$. The population variance and standard deviation are usually denoted by $\sigma^2$ and $\sigma$ respectively.

In practice, the standard deviation is used more frequently than the variance. In a comparison of two groups of data, the group with the smaller standard deviation has the more homogeneous observations; the group with the larger standard deviation exhibits a greater amount of variability.
Coefficient of Variation

The coefficient of variation is a useful measure of relative spread in data and is used frequently in biologic sciences. It is defined as the standard deviation divided by the mean times 100%.

The formula for the coefficient of variation is:
\[ CV = \left( \frac{s}{\bar{x}} \right) \times 100\% \]
Percentiles

Values of a quantitative variable that divide the ordered data into groups so that a certain percentage is above and another percentage is below. For example, the median is the 50% percentile, the value below which 50% of the cases fall.

The 25th and 75th percentiles are known as the first and third quartiles. The median is the second quartile.

The three quartiles divide the frequency distribution into four equal parts.
Normal (Gaussian) distribution

The normal distribution is a continuous symmetric distribution that follows the familiar bell-shaped curve. The distribution is uniquely determined by its mean and variance. It has been noted empirically that many measurement variables have distributions that are at least approximately normal.

Many frequently used statistical tests make the assumption that the data come from a normal distribution.

Skewness and kurtosis are statistics that describe the shape and symmetry of the distribution.
Normal Distribution

Mean, Std. dev. 0,1
Example - Distribution of blood pressure

Distribution of blood pressure can be approximated as a normal distribution with mean 85 mm. and standard deviation 20 mm. A histogram of 1,000 observations and the normal curve is shown below.
Standard Deviation

• A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one SD of the mean and 95% of cases fall within 2 SD.

• For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
Area properties of the Normal Distribution
Skewness

A measure of the asymmetry of a distribution. The normal distribution is symmetric, and has a skewness value of zero. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. A skewness value greater than 1 generally indicates a distribution that differs significantly from a normal distribution.
Left or negatively skewed
Right or positively skewed
Skewness

- If the frequency distribution is left (negatively) skewed, then the mean is less than the median which in turn is less than the mode.

- If the frequency distribution is right (positively) skewed, the mean is greater than the median which in turn is greater than the mode.

- If the frequency distribution is symmetric, then all three measures will be identical.
Kurtosis

A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is 0. Positive kurtosis indicates that the observations cluster more and have longer tails than those in the normal distribution and negative kurtosis indicates the observations cluster less and have shorter tails.
A Population can be finite or infinite
Statistical Inference

If a sample is representative of a population, important conclusions about the population can often be inferred from analysis of the sample. The phase of statistics dealing with conditions under which such inference is valid is called inductive statistics, or statistical inference. Because such inference cannot be absolutely certain, the language of probability is often used in stating conclusions.

Inference from data: Quantifying evidence concerning hypothesis
Steps in hypothesis testing

1. The first step in hypothesis testing is to specify the null hypothesis (H_0) and the alternative hypothesis (H_1). If the research concerns whether one method of presenting pictorial stimuli leads to better recognition than another, the null hypothesis would most likely be that there is no difference between methods (H_0: µ1 - µ2 = 0). The alternative hypothesis would be H_1: µ1 ≠ µ2. If the research concerned the correlation between grades and SAT scores, the null hypothesis would most likely be that there is no correlation (H_0: ρ = 0). The alternative hypothesis would be H_1: ρ ≠ 0.

2. The next step is to select a significance level. Typically the .05 or the .01 level is used.

3. The third step is to calculate a statistic analogous to the parameter specified by the null hypothesis. If the null hypothesis were defined by the parameter µ1 - µ2, then the statistic M1 - M2 would be computed.
Type I & Type II errors

There are two kinds of errors that can be made in significance testing: (1) a true null hypothesis can be incorrectly rejected and (2) a false null hypothesis can fail to be rejected.

The former error is called a Type I error and the latter error is called a Type II error. These two types of errors are defined in the table.

<table>
<thead>
<tr>
<th>Statistical decision</th>
<th>True state of null hypothesis</th>
<th>Ho True</th>
<th>Ho False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject Ho</td>
<td>Type I error</td>
<td>Correct</td>
<td></td>
</tr>
<tr>
<td>Do not reject Ho</td>
<td>Correct</td>
<td></td>
<td>Type II error</td>
</tr>
</tbody>
</table>
Power of the test

Pr(type I error) = \(\alpha\)
Pr(type II error) = \(\beta\)

We also speak of the complements of \(\alpha\) and \(\beta\). The probability of not making a type I error \((1-\alpha)\) is called confidence. The probability of not making a type II error \((1-\beta)\) is called power. Thus:

\[
\begin{align*}
\Pr(\text{avoiding a type I error}) &= 1 - \alpha = \text{confidence} \\
\Pr(\text{avoiding a type II error}) &= 1 - \beta = \text{power}
\end{align*}
\]

Conventional levels for confidence are .90, .95 and .99. Conventional levels for power are .80, .90, and .95.

- Greater sample size increases power of a test.
- The significance level \((\alpha)\). Power decreases with decreasing \(\alpha\). For example, reducing \(\alpha\) from 0.05 to 0.01 to reduce the probability of making a Type I error but increases the probability of making a Type II error.
- Higher variation reduces power.
- The difference between means, \(\mu_1 - \mu_2\). The larger the difference between the population means, the greater the probability of rejecting Ho.
Comparing Treatment Groups

• Differences in outcomes between the treatment groups may be due to:
  - treatment effect
  - bias
  - chance

• Groups must be alike in all important aspects except for the treatments under evaluation (avoid bias)

• Large enough sample (limit chance imbalances)

• In any comparison in a medical context, differences are almost bound to occur. The problem is separating real effects from random variation

• It is the job of the analysts to decide how much variation should be ascribed to chance, so that any remaining variation can be assumed to be a real effect – This is the art of Statistics
Hypothesis testing

• In a randomised trial the treatment groups differ either because of the effect of the treatment or by chance.

• If the treatment has no effect then chance is the only explanation.

• We can work out what sort of differences are likely to be produced by chance.

• If the difference is much bigger than this then pure chance is not a plausible explanation.
• Many research problems are questions about differences between mean values of a quantitative outcome. The three common types of questions are:

1. What is the best estimate of true mean value for this group and how different is it from that of an asserted value or a "standard" value?

2. What is the best estimate of the true mean difference in this group before and after treatment?

3. What is the best estimate of the true mean difference between two treatment groups?
Procedure

• Set up the null and alternative hypotheses

• Decide on the significance level $\alpha$ (usually 0.05 or 0.01)

• Get the relevant sample value (this might be a sample mean, or difference between two sample means, and so on)

• Determine the p-value, i.e. the probability of getting this sample result (or one even worse) by chance when the null hypothesis is true

• If this p-value is less than 0.05 or 0.01, whichever significance level has been set, reject the null hypothesis, otherwise do not reject the null hypothesis
Confidence Intervals (CI)

An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called a confidence interval.

The important point is that we can use a CI not only to inform us about the likely value of a population parameter, but also to test a belief or hypothesis we might hold about the value of that population parameter.

If the CI includes the believed or hypothesized value we can conclude, with a 95% or 99% certainty anyway, that our hypothesis is true. If the CI does not include the value we previously believed the population parameter to be equal to, then we must reject that belief or hypothesis. So, not only does a CI estimate it also tests!
One-Sample T Test

The One-Sample T Test procedure tests whether the mean of a single variable differs from a specified constant.

Example:
A researcher might want to test whether the average IQ score for a group of students differs from 100. Or, a cereal manufacturer can take a sample of boxes from the production line and check whether the mean weight of the samples differs from 1.3 pounds at 95% confidence interval.
One-Sample T Test

Data:
To test the values of a quantitative variable against a hypothesized test value, choose a quantitative variable and enter a hypothesized test value

Assumptions:
This test assumes that the data are normally distributed; however this test is fairly robust to departures from normality
Problem 1:
Imagine that you walk down to casualty one day and notice that a sign on the wall says “Average waiting time is now 30 minutes”. You jot down the time spent waiting by the last 20 patients and since the mean waiting time of all arriving patients is not 30 minutes, you would like to test the hypothesis that the population mean equals to 30 minutes.

Data:
Time spent waiting by 20 consecutive arrivals in a casualty department (minutes)

20, 30, 45, 40, 15, 50, 30, 25, 30, 50, 80, 20, 10, 50, 60, 25, 20, 20, 35, 60
To Obtain a One-Sample T Test

From the menus choose:
Statistics
    Compare Means
        One-Sample T Test...

Select one or more continuous, numeric variables to be tested against the same value.

Enter a numeric test value against which each sample mean is compared.
Paired Samples

Pairing involves matching up individuals in two samples so as to minimize their dissimilarity except in the factor under study. For example, in pre-test/post-test studies, each subject is paired (matched) with himself, so the difference between the pre-test and post-test responses can be attributed to the change caused by taking the test, and not to differences between the individuals taking the test. Such data are analyzed by examining the paired differences.
Paired-Samples T Test

The Paired-Samples T Test procedure compares the means of two variables for a single group. It computes the differences between values of the two variables for each case and tests whether the average differs from 0.

Example:

In a study on high blood pressure, all patients are measured at the beginning of the study, given a treatment and measured again. Thus, each subject has two measures, often called before and after measures.
Problem 2:

In chronic hypercapneic respiratory failure patients a low carbohydrate diet may improve the oxygen carrying capacity of the blood. You measure the arterial oxygen tension of 8 patients before and after the introduction of a low-carbohydrate diet plan. What is the estimated true mean difference in arterial oxygen before and after introduction of the diet? Is the diet clinically effective in improving blood oxygen?

Arterial Oxygen Tension (mm Hg)

Before: 70 59 53 54 44 58 64 43
After: 82 66 65 62 74 77 68 59
To Obtain a Paired-Samples T Test

From the menus choose:

Statistics
  Compare Means
    Paired-Samples T Test...

Select a pair of variables, as follows:
Click each of two variables. The first variable appears in the Current Selections group as VARIABLE 1, and the second appears as VARIABLE 2.

After you have selected a pair of variables, click the arrow button to move the pair into the Paired Variables list. You may select more pairs of variables. To remove a pair of variables from the analysis, select a pair in the Paired Variables list and click the arrow button.
Independent-Samples T Test

This procedure compares means for two groups of cases. Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors.
T-test

• The two sample t test for unpaired data is defined as: $H_0: \mu_1 = \mu_2 \quad H_a: \mu_1 \neq \mu_2$

• Test Statistic: $T = \frac{Y_1 - Y_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}$

where $N_1$ and $N_2$ are the sample sizes, $Y_1$ and $Y_2$ are the sample means, $s_1^2$ and $s_2^2$ and are the sample variances.

• If equal variances are assumed, then the formula reduces to: $T = \frac{Y_1 - Y_2}{s_p \sqrt{1/N_1 + 1/N_2}}$ where $s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$
Decision criterion

- Significance Level: $\alpha$

- Critical Region: Reject the null hypothesis that the two means are equal if $T < -t(\alpha/2, \nu)$ or $T > t(\alpha/2, \nu)$

where $t(\alpha/2, \nu)$ is the critical value of the t-distribution with $\nu$ degrees of freedom where

$$\nu = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)}$$

- If equal variances are assumed, then $\nu = N_1 + N_2 - 2$
Example:

Patients with high blood pressure are randomly assigned to a placebo group and a treatment group. The placebo subjects receive an inactive pill and the treatment subjects receive a new drug that is expected to lower blood pressure. After treating the subjects for two months, the two-sample t test is used to compare the average blood pressure for the placebo group and the treatment group. Each patient is measured once and belongs to one group.
Data:
The values of the quantitative variable of interest are in a single column in the data file. The procedure uses a grouping variable, with two values to separate the cases into two groups. The grouping variable can be numeric (values such as 1 and 2) or short string (such as YES and NO). As an alternative, you can use a quantitative variable, such as AGE, to split the cases into two groups by specifying a cut point.
Assumptions:

For the equal-variance t test, the observations should be independent, random samples from normal distributions with the same population variance. For the unequal-variance t test, the observations should be independent, random samples from normal distributions. The two-sample t test is fairly robust to departures from normality. When checking distributions graphically, look to see that they are symmetric and have no outliers.
Problem 3:

You’ve read that the diagnostic hallmark of carcinoid syndrome is an increased urinary secretion of 5-hydroxyindoleacetic acid (5-HIAA). You measure the 5-HIAA levels in 16 patients with carcinoid heart disease and 12 controls. You want to know whether there is a true difference between the mean urine 5-HIAA levels in these two groups of patients.

**Null hypothesis**: The mean urine 5-HIAA levels in the two groups of patients is equal

**Alternative hypothesis**: The mean urine 5-HIAA levels in the two groups of patients is not equal
# Data

<table>
<thead>
<tr>
<th>MAX5HIAA</th>
<th>GROUP</th>
<th>60.00</th>
<th>2.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>263.00</td>
<td>1.00</td>
<td>119.00</td>
<td>2.00</td>
</tr>
<tr>
<td>288.00</td>
<td>1.00</td>
<td>153.00</td>
<td>2.00</td>
</tr>
<tr>
<td>432.00</td>
<td>1.00</td>
<td>588.00</td>
<td>2.00</td>
</tr>
<tr>
<td>890.00</td>
<td>1.00</td>
<td>124.00</td>
<td>2.00</td>
</tr>
<tr>
<td>450.00</td>
<td>1.00</td>
<td>196.00</td>
<td>2.00</td>
</tr>
<tr>
<td>1270.00</td>
<td>1.00</td>
<td>14.00</td>
<td>2.00</td>
</tr>
<tr>
<td>220.00</td>
<td>1.00</td>
<td>23.00</td>
<td>2.00</td>
</tr>
<tr>
<td>350.00</td>
<td>1.00</td>
<td>43.00</td>
<td>2.00</td>
</tr>
<tr>
<td>283.00</td>
<td>1.00</td>
<td>854.00</td>
<td>2.00</td>
</tr>
<tr>
<td>274.00</td>
<td>1.00</td>
<td>400.00</td>
<td>2.00</td>
</tr>
<tr>
<td>580.00</td>
<td>1.00</td>
<td>73.00</td>
<td>2.00</td>
</tr>
<tr>
<td>285.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>524.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>135.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>120.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To Obtain an Independent-Samples T Test

From the menus choose:

Statistics
   Compare Means
      Independent-Samples T Test...

Select one or more continuous, numeric test variables. A separate t test is computed for each variable.

Select a dichotomous grouping variable (a categorical variable that divides cases into two groups)

Click Define Groups and specify the values of the grouping variable that define the two groups.

The grouping variable can be a string (alphanumeric) variable or a numeric variable that uses numeric codes to represent categories (e.g., 0=Male, 1=Female).
Don’t just use statistics as the drunken man uses the lamp post - for support rather than illumination

USE STATISTICS FOR SUPPORT AND ILLUMINATION