Data Collection, Coding and Analyses

Starting your study and collecting data can be very exciting. You are finally ready to start going!! Soon you may have some questions as to how to record the information/data you are interested in and what steps are next. Questions regarding what types of data you want to record and what variables you are interested in will be outlined in your research proposal. The topics below provide an overview of some of the statistical implications and practical considerations when collecting data but these are issues you should think about before starting to collect the data. It is advisable to consult with a statistician during the study design stages so there are no surprises once it comes time to analyze the results.

Information in this section is based on using a statistical software package SPSS (Statistical Package for Social Sciences). There are new versions of this software about once a year but the functionalities and basic commands and steps remain relatively unchanged. The library computers on each campus have this software available for your use. A student version of this software is available for purchase at a reduced rate.

Information is also provided for using Excel to perform statistics. Certainly for the step of entering your data Excel is a great and widely available tool. Information entered in Excel is easily imported in SPSS and other statistical packages. Running some basic descriptive statistics in Excel is OK. However it is not recommended for some more complicated statistical techniques for which the use of a statistical software package and/or turning to a statistician is advised.

Below is a step by step overview of the process of determining which statistical techniques to use as well as what data to collect. Some of these steps you may have already taken and it reiterates the need of thinking about statistical analyses at the very early steps of designing your research project. A step-by-step overview for many of the statistical tests and an intro to the use of SPSS can also be found in the spss online training workshop developed by faculty at Central Michigan University (http://people.cst.cmich.edu/lee1c/spss/). SPSS is available to you on several computers in the MWU library on either campus.
Steps for Data Analysis

1. Specify the question to be answered.
2. Phrase your question in terms of a biological null hypothesis and alternative hypothesis.
3. Phrase your question in terms of a statistical null hypothesis and alternative hypothesis.
4. Which variables are relevant to your question and your hypotheses?
5. What types of variables are you working with?
6. Design an experiment that controls or randomizes the confounding variables.
7. Choose the best statistical test based on the number of variables, the types of variables, whether your variables fulfill the assumptions of parametric statistics, and the hypothesis to be tested.
8. If possible, do a power analysis to determine a good sample size for the experiment.
9. Collect your data.
10. Examine the data to see if it meets the assumptions of the statistical test you chose. If it doesn't, choose a more appropriate test.
11. Apply the chosen statistical test, and interpret the result.
12. Communicate your results effectively, usually with a graph or table.
   Prepare final research report and communicate your results through a poster presentation and/or manuscript for publication.

Types of Variables

The information or data you collect is more commonly referred to as variables. There are different types of variables and they each have their own set of pros and cons and determine the type of analyses you can perform and how the results can be communicated.

Measurement variables
Measurement variables are things you can measure using a number. For example, length, weight, pH, and bone density. These are sometimes called “continuous variables”.

Nominal variables
These variables, also called "attribute variables" or "categorical variables," classify
observations into a number of categories. A good rule of thumb is that an individual observation of a nominal variable is usually a word, not a number. Examples of nominal variables include sex (the possible values are male or female), genotype (values are AA, Aa, or aa), or ankle condition (values are normal, sprained, torn ligament, or broken). Nominal variables are often used to divide individuals up into groups or classes, so that other variables may be compared among the groups.

**Ranked (Ordinal) variables**

Ranked variables, also called ordinal variables, are those for which the individual observations can be put in order from smallest to largest, even though the exact values are unknown. For example, you may want to quantify swelling without a measuring device. You can rank the degree of swelling using integers: 1, 2, 3, etc. In this case, a patient with a ranked 2 swelling will be more severe than a patient ranked 1, yet we don’t know the absolute magnitude of the difference.

**Coding, Entering data in spreadsheets**

**Naming Variables**

As a rule of thumb, variable names should be text, not numerical. This avoids the problem of confusing the variable name and the data. In SPSS, you can click the Variable View button in the lower left part of the screen and check your variable names. Each statistical software program may have restrictions on what characters can be used in the variable name. Please check as this might cause issues when importing the datafile.

**Preparing an Excel File for Use with SPSS**

Below are some guidelines on how to set up the Excel File for use in SPSS:

A. Place variable names in First Row. There are some guidelines on naming the variables; they should:

- Not be longer than 64 characters. Recommend using names not longer than 8 characters.
- Start with a letter
- Can be only letters but may have numbers or underscores
- Not contain %, $, #, @,!, +, *,",~,
- Not contain blanks
- Be unique (no duplicate variable names)
- Be on one row only

B. Include only the raw, un-summarized data. Any totals or graphs, etc need to be deleted

C. Include an identifying number (ID) for each unique case.

D. Include only one value per cell. If you are entering data for blood pressure do not enter 140/90. Instead create a variable for both systolic and diastolic blood pressure and enter that value.

E. Do not leave blank rows in the datasheet.

F. Do not use numbers and text in the same column (i.e names and ID numbers).

G. Use numeric values when feasible (see section on coding). Text (character) variables are allowed but they are not as flexible with respect to further analyses.

H. You can assign a numeric code to missing values such as 99 or 999, etc. Make this always one unit larger than the value of a real data value. For example; if your study includes adults 18 and over, 99 could be a real value point. The code for missing data for that specific variable could be 999. You can also leave the cell blank. You can also opt to leave missing values blank. Make certain you are consistent in your approach and if using codes for missing values in SPSS make you need to specify this value within the software (Variable View).

I. Save the spreadsheet with values only; no formulas

**Variable Coding**

Variables need to be coded so that we can run statistical analyses. Coding the variable allows the statistical software to understand our variables. It is the process of assigning numbers to qualitative descriptors. In most of the software programs
you can (and should) assign labels to the variables and their values (enter the values for the codes). If you are entering data in Excel you need to keep track of what numbers you assign to the variables and create a codebook (you can also enter this info on the bottom of the worksheet while working on the data-entry. Remember to delete this before exporting data).

**Measurement variables**
Measurement variables should contain continuous data, like weight. Therefore, the data should reflect the real value of the variable. In the case of weight, your data may be 150, 135, 180, and 175, with the values indicating pounds.

Sometimes you may also want to use categories for these types of variables. For example, instead of using exact age you may in your analyses want to describe the data in terms of age categories (i.e. 20-29; 30-40, etc). This is easily handled by the software packages you use and through a process called recoding you can have the variable age recoded into categories. Then you have the choice what to present. The recommendation is to always collect and enter the more precise data (i.e. actual weight) as this gives you the option of using either the actual values or

---

**What is Coding?**

"The process by which verbal data are converted into variables and categories of variables using numbers, so that the data can be entered into computers for analysis."

the recodes (categories of age). You do not have this option if you start by entering categories, you cannot go from less precision to more.

**Nominal variables**
Nominal variables have categorical data. Each category may be represented by a number, but the values themselves do not represent a quantity. For instance, your variable may be sex, with two categories: male or female. Each category may be represented by a number: 0 for males and 1 for females. Therefore, each sample may receive a 0 or a 1 for this variable. Typically, 0 is used as the first value for a nominal variable. It really does not matter what value is assigned to either gender as long as it is consistently applied to all cases in your datafile and you keep track of this information. Alternatively, SPSS allows you to enter and use the category’s label as data, male or female. In most statistical packages (e.g. SPSS), you should be sure that this variable is specified nominal. This will allow you to use it as a group variable in an analysis like a T-test. This will be discussed in more detail later in the chapter.

**Ranked (Ordinal) variables**
Ordinal variables will also be coded numerically, with the value having some quantitative meaning. For instance, your variable may be pain intensity, which can be ranked on a scale from 0 to 3. In this case, 0 indicates no pain, 1 indicates mild pain, 2 indicates moderate pain, and 3 indicates severe pain.

**Dates and Times**
Be careful about entering dates into statistical software. Entering dates into Excel is typically uneventful although you need to format the column accordingly and indicate it is a date you are entering and the format you would like it to be in. This is necessary as exporting the data to other software packages might not work if it was not properly formatted as a “date” column.

Working with dates can be problematic in software packages such as SPSS. SPSS does not specifically recognize dates or times as a separate type of variable. One problem with dates and times is that they are not necessarily linear measurements. For instance, January (which is often coded as month 1) is closer to December (often coded as month 12) than it is to March (often coded as month 3). SPSS and other statistical packages do not recognize the “circular” nature of these variables,
which may lead to serious data analysis errors. Therefore, special care should be used if you want to analyze date and time data.

**Missing Data**
People often use Excel to create a dataset that will be subsequently analyzed in another statistical package (e.g. SPSS). When entering data, missing values are probably best coded by *not* using a number. If you code missing values with a number, e.g. 999, then it opens the possibility for the statistical software to use this value as a real data point. One option is to code missing values in Excel with text, e.g. NA. This is fine, except that the NAs should be removed before opening the file in SPSS. If not, SPSS will assume that your variable is nominal, even if it is continuous. Some people prefer indicating a missing value by leaving the worksheet cell blank. The downside of this approach is that one might confuse this with data not yet entered. Whichever approach you take, be consistent. When using values to represent missing data this will need to be specified in the Missing column of the Variable View in SPSS.

**An example of the three types of variables in SPSS**

Data View in SPSS

![Data View in SPSS](image)

Variable View in SPSS

![Variable View in SPSS](image)
Data Collection (using surveys)

Surveys can be a good tool to collect data. There are two basic parts to a survey: the questions and the possible answers to those questions. These will shape the data you obtain and consequently, the types of statistical analyses you can perform. Typically, more statistical power can be obtained by having a variable that is measured on a continuous or ordinal scale. Also, for an ordinal variable, having a greater number of ranks is typically better than fewer, e.g. measuring something on a 1 to 10 scale is better than a 1 to 3 scale. If needed, it is always easier to collapse data, e.g. making a variable ranked 1 to 10 a variable that is ranked 1 to 3, then vice versa.

For instance, for the question, What is your level of pain?, one set of answers may be: 1) low, 2) moderate, or 3) extreme. Another set of answers may be: rate your pain on a scale from 1 to 10, with 10 being the highest. An additional question with answers that can be ranked is: What is your perceived need for selecting the proper statistical test to analyze data? A possible set of answers is: 1) need further basic instruction, 2) able to perform with close supervision, 3) able to perform with minimal supervision, 4) able to perform independently. Even though the answers are descriptions (i.e. not numerical), these answers can be ranked from 1 to 4, in terms of how well the participant knows statistics.

Note that one can and should identify the values of the missing data. If not, SPSS will assume that those are valid numbers which leads to all sorts of errors.

Note that the type of variable can be specified in the Measure column.
Probability and Statistical Analyses
The goal of a statistical test is to identify a null hypothesis, collect some data, then estimate the probability of getting the observed data if the null hypothesis were true. If the probability of getting a result like the observed one is low under the null hypothesis, then you would conclude that the null hypothesis is probably not true.

Hypothesis Testing

Null and alternative hypotheses
In general, the null hypothesis is that things are the same as each other, or the same as a theoretical expectation. For example, you measure the tail length of individuals in two mouse populations, the null hypothesis could be that there is no difference in tail length between these populations. The alternative hypothesis is that things are different from each other, or different from a theoretical expectation. For example, tail length of individuals differs between two mouse populations because one population lives in a high food resource area and the other population does not.

Testing the null hypothesis
The primary goal of a statistical test is to determine whether an observed dataset is so different from what you would expect under the null hypothesis that you should reject the null hypothesis.
Probability values
A probability value (i.e. p value) can be defined as the probability of getting the observed result, or a more extreme result, if the null hypothesis is true. Therefore, a small p value indicates that the null hypothesis is likely false.

A p value is the probability of an observed (or more extreme) result arising only from chance.

Probability of getting different numbers of males when sampling 48 individuals, when the known proportion of males in the population is 0.5. Note that the outcome with the highest probability is 24 males, which is 50% of 48. The p value for obtaining 17 or fewer males is 0.030. Similarly, the p value for obtaining nearly all males or nearly no males from a sample of 48 individuals is extremely small.

Significance levels
How small does the p value need to be to know that the null hypothesis should be
rejected? This value varies a bit across disciplines, but a p value of 0.05 is the most commonly accepted value. Therefore, if the probability value (p) is less than 0.05, you reject the null hypothesis. If p is greater than or equal to 0.05, then you don't reject the null hypothesis. There is nothing statistically or biologically important about 0.05. The convention could have been a significance level of 0.04, or 0.025, or 0.071.

The significance level is also related to the probability of rejecting the null hypothesis, even when it is true. With a significance level of 0.05, you have a 5 percent chance of rejecting the null hypothesis, even if it is true. For example, 100 studies may not have a real biological effect, yet 5 percent of them may yield a statistically significant p value by chance alone.

**Statistical vs. biological significance**

It is important to remember that statistical significance does not necessarily equal biological significance. If you conduct a study with a very large sample size (e.g. several hundred), then a statistical test can detect a very small effect size. Essentially, the effect size is the magnitude of the effect in your study. For instance, you may be interested in testing whether a new diet increases the weight of mice. You have one group of mice that only eat the new diet and another group that consumes a control diet. You weigh the mice in each group and determine that the mice eating the new diet weigh 0.1 gram more than the mice in the control group. If your study contains several hundred mice in each group, then a T-test may yield a statistically significant p value, indicating a difference between the two groups. Yet, a 0.1 gram difference between the two groups may not be biologically meaningful, depending on the goals of the research.
**Power Analysis**

When you are designing an experiment, it is a good idea to estimate the sample size you'll need. This is especially true if you're proposing to do a project that uses live subjects, is very time-consuming and/or is expensive. In order to do a power analysis, you need to specify an effect size. This is the size of the difference between your null hypothesis and the alternative hypothesis that you hope to detect. Though, when doing biological research, you often don't know how big a difference you're looking for.

You need several pieces of information to conduct a power analysis. The effect size is the deviation from the null hypothesis that you hope to detect. For example, if you are treating hens with something that you hope will change the sex ratio of their chicks, you might decide that the minimum change in the proportion of sexes that you're looking for is 10 percent. You might have a good reason for choosing the effect size; if not, you might want to see what kind of effects other people have found in similar experiments. If you don't have a particular effect size in mind, you might want to try different effect sizes and produce a graph of effect size vs. sample size.

Alpha is the significance level of the test (the p-value), the probability of rejecting the null hypothesis even though it is true (a false positive). The usual value is alpha=0.05. Some power calculators allow the user to set a one-tailed or two-tailed alpha. Be sure you know which you are using. One-tailed tests will provide more power, but should only be used if you have a directional prediction, e.g. the treatment group will exhibit lower pain levels compared to the control group, or the amount of time spent exercising per week is positively correlated with longevity. In contrast, a two-tailed test can be used without making a directional prediction, e.g. pain levels will differ between treatment and control groups, or the amount of time spent exercising per week is related to longevity.

Beta, in a power analysis, is the probability of accepting the null hypothesis, even though it is false (a false negative), when the real difference is equal to the minimum effect size. The power of a test is the probability of rejecting the null hypothesis when the real difference is equal to the minimum effect size, or 1−beta. There is no clear consensus on the value to use; a power of 80% (equivalent to a beta of 20%) is probably the most common, while powers of 50% and 90% are...
also sometimes used. The cost to you of a false negative should influence your choice of power; if you really, really want to be sure that you detect your effect size, you'll want to use a higher value for power (lower beta), which will result in a bigger sample size. Some power calculators ask you to enter beta, while others ask for power (1−beta). Be sure you know which you are using.

Conducting a power analysis for some statistical tests also requires that you estimate the standard deviation for each measurement variable. This can come from pilot experiments or from similar experiments in the published literature.

The actual calculations needed to conduct a power analysis can be quite complicated. Fortunately, there are online calculators for doing power analyses for many statistical tests. The first link below is to a java based website and the second link is to a stand-alone free software package, G*Power.

http://www.stat.uiowa.edu/~rlenth/Power/index.html

http://wwwpsycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
**Entering Data in Excel**

Below is an example of a data sheet that can be created in Excel. This project examined the weight of mice with a low calorie vs. high calorie diet. A t-test may be an appropriate statistical test to examine the potential weight difference in these two groups. In this example, the ID number for each mouse is in the first column (named SampleID), the treatment group that each mouse is a member of is in the second column (named Group). It is often beneficial to assign each group name a numerical code, in this case, 0 indicates the low calorie diet group and 1 indicates the high calorie diet group. The fourth column contains the weight for each mouse. Please note that not all columns (variables) in the dataset may be used in the actual data analysis. For instance, the SampleID is not important for conducting the actual t-test or ANOVA, yet this information may be useful if an outlier is detected in the analysis. Knowing the SampleID of the outlier might give us the opportunity to more easily check for data entry errors.
Descriptive Statistics

Summarizing data for a measurement variable requires a number that represents the "middle" of a set of numbers (known as a "statistic of central tendency" or "statistic of location"), along with a measure of the "spread" of the numbers (known as a "statistic of dispersion"). The arithmetic mean is the most common statistic of central tendency, while the variance or standard deviation is usually used to describe the dispersion.

Mean

The arithmetic mean is the most common type of mean calculated for a measurement variable. It is the sum of the observations divided by the number of observations. It is the most common statistic of central tendency, and when someone says simply "the mean" or "the average," this is what they mean. It is sensitive to extreme values, which makes it not work well for data that are highly skewed. You can calculate the arithmetic mean in Excel by typing the following in a cell

=AVERAGE(A1:A10)

This code will calculate the arithmetic mean for data in cells A1 through A10. You can replace A1 and A10 with other cells that contain your data. The colon designates that all consecutive cells between A1 and A10 also will be included in the calculation. If you want to calculate a mean for non-consecutive data points then use a comma. For instance, =AVERAGE(A1, A3, A5, A10)

This will calculate the arithmetic mean of the data in cells A1, A3, A5 and A10. Using the colon and comma in this way can be applied to any formula in Excel.

Median

When the data are sorted from lowest to highest, this is the value of the data set that is in the middle. For an odd number of samples, the median is the single middle value in the sorted data; for an even number, it is the arithmetic mean of the two values in the middle. The median is useful when dealing with highly skewed distributions. You can calculate the median value in Excel by using

=MEDIAN(A1:A10)

*You should replace A1 and A10 with the cells that contain your data.
Mode
This is the most common value in a data set. In Excel you can use the function

=MODE(A1:10)
*You should replace A1 and A10 with the cells that contain your data.

Sum of Squares
This is not really a statistic of dispersion by itself, but it forms the basis of the variance and standard deviation. Subtract the sample mean from an observation and square this "deviate". Squaring the deviates makes all of the squared deviates positive and has other statistical advantages. Do this for each observation, then sum these squared deviates. This sum of the squared deviates from the mean is known as the sum of squares. You can calculate this in Excel using the function:

=DEVSQ(A1:A10)
*You should replace A1 and A10 with the cells that contain your data.

Sample Variance
The sample variance is a measure of variation in your data. To calculate your sample variance, divide the sum of squares by n−1, where n is the number of samples. It is important to remember that the value of your sample variance is in squared units. You can use the standard deviation (next section) if you want to calculate the variation in your data using the original units of measurement. In Excel, there are two functions that calculate the sample variance. Older versions of Excel use =VARA, while newer versions of Excel can use =VARA as well as =VAR.S

Standard Deviation
Taking the square root of the variance gives a measure of dispersion that is in the original units. The sample standard deviation requires a rather complicated correction factor and can be calculated in Excel using the function:

=STDEV(A1:A10)
*You should replace A1 and A10 with the cells that contain your data.
In addition to being more understandable than the variance as a measure of the amount of variation in the data, the standard deviation summarizes how close observations are to the mean in a very nice way. Many variables in biology fit the normal probability distribution fairly well. If a variable fits the normal distribution, 68.3 percent (or roughly two-thirds) of the values are within one standard deviation of the mean, 95.4 percent are within two standard deviations of the mean, and 99.7 (or almost all) are within 3 standard deviations of the mean. Here's a histogram that illustrates this:

The proportions of the data that are within 1, 2, or 3 standard deviations of the mean are different if the data do not fit the normal distribution, as shown for these two very non-normal data sets:
Left: Frequencies of 5,000 numbers randomly generated to fit a distribution skewed to the right. Right: Frequencies of 5,000 numbers randomly generated to fit a bimodal distribution.
Conducting Descriptive Statistics in SPSS

A wide variety of descriptive (and related) statistics are available in SPSS under the Analyze → Descriptive Statistics → Descriptives or Frequencies.
**Student's t-test**

**What is it**
The t-test allows you to examine whether two groups exhibit a significantly different mean value for a particular variable.

**When to use it**
When you have one nominal variable and one measurement variable, and you want to compare the mean values of the measurement variable. Multiple observations of the measurement variable are made for each value of the nominal variable. The nominal variable must have only two values, such as "male" and "female" or "treated" and "untreated."

**How the test works**
The test statistic, which in this case is the t value, is calculated using a formula that has the difference between the means in the numerator; this makes t larger as the means get further apart. The denominator is the standard error of the difference in the means, which gets smaller as the sample variances decrease or the sample sizes increase. Thus t gets larger as the means get farther apart, the variances get smaller, or the sample sizes increase. In other words, the t value can also be thought of as the magnitude of the difference between the two groups.

**Assumptions**
The t-test assumes that the observations within each group are normally distributed and the variances are equal in the two groups. The t-test is robust to moderate violations of these assumptions. Though, if severe violations occur, then the data should be transformed or a non-parametric version of the t-test should be conducted.

**Example**
You are interested in examining the height difference in two groups of men. One group contains men that were malnourished as a child, and the other group contains men with proper childhood diets. You collect height data on 25 men that were malnourished and 25 men that had proper childhood nutrition. There are two variables: one measurement variable, height, and one nominal variable, nutrition group. It is possible to use a third variable, an identification code for each individual. Yet, this latter variable will not be included in the analysis.
Height is measured in centimeters and you can code individuals in the malnourished group as a 0, and individuals that had proper nutrition as a 1.

The data setup would look like this example:

<table>
<thead>
<tr>
<th>Height (cm)</th>
<th>Nutrition Group Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>152</td>
<td>0</td>
</tr>
<tr>
<td>154</td>
<td>0</td>
</tr>
<tr>
<td>157</td>
<td>0</td>
</tr>
<tr>
<td>160</td>
<td>0</td>
</tr>
<tr>
<td>166</td>
<td>0</td>
</tr>
<tr>
<td>195</td>
<td>1</td>
</tr>
<tr>
<td>180</td>
<td>1</td>
</tr>
<tr>
<td>177</td>
<td>1</td>
</tr>
<tr>
<td>170</td>
<td>1</td>
</tr>
<tr>
<td>171</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: 152.4cm = 5 feet; 167.6cm = 5 feet, 6 inches; 183cm = 6 feet; 198cm = 6 feet, 6 inches

You can download the example dataset in Excel format, [here](#).
Conducting a t-test in SPSS

The t-test instructions and screen shots are based on SPSS version 17. Newer versions are similar but may be slightly different. The instructions and screen shots for the ANOVA example are based on SPSS version 19.

If your data were entered in Excel, then you need to import the data into SPSS. You can do this by starting SPSS, then clicking the top toolbar menu:

File → Open → Data → a pop-up window will appear → In the pop-up window, select Excel in the Files of Type dropdown menu → choose the location of the data file → click OK in the Opening Excel Data Source pop-up window

In the t-test dataset, you will see the two columns of data.

The above image is the dataset, in the Data View. The Data View and Variable View tabs are located in the lower left portion of the window. The next screenshot is the same dataset in the Variable View.
The above screenshot is the same dataset in the Variable View. The Variable View is useful for several actions, including specifying the type of variable, the number of decimal places for a continuous variable, or labels for a variable code (e.g. 0 = healthy, 1=ill).

We can now start the t-test analysis.

At the top menu bar, click: Analyze → Compare Means → Independent Samples T Test → you will then have a pop-up window appear.

The window (below) will have a list of your variables in the box on the left.

Click on the variable you want to test, which is Height, and click the upper purple right arrow to move the designate Height as your Test Variable. Next, click on the variable that represents your groups, in this case, Nutrition Group Code, and then click on the lower purple right arrow to designate it as your Grouping Variable.
Click on the Define Groups button and specify the names of each group, in this case 0 and 1. Click Continue.

You will then return to the Independent-Samples T Test window. Click OK.

SPSS will then open a new Output window.

The Output window consists of two panes. The left pane lists shortcut links to the detailed results in the right pane.
The right pane contains the details of the results. Here are the results tables:

The first table provides descriptive statistics for each group: the number of samples, the mean value, and the standard deviation.

The second table provides the results of the t-test. The first columns are related to testing an assumption of t-tests. SPSS automatically tests for the equality of group variances using Levene’s test. If your groups have similar variances, which is an assumption of the t-test, then the p value for this test (noted as Sig.) will be >0.05. In this case, the value is 0.908, indicating that the assumption is satisfied. If the p value for this test is <0.05, then it may indicate that your groups have very
different variances. Yet, Levene’s test is overly sensitive and often produces low p values when groups do not differ greatly. In this case, it is best to create box plots of your data to visually inspect the variation (see below).

The next columns are related to the t-test itself. Note that there are two rows. The top row of results when group variances are equal and the bottom row of results when group variances are not equal. We will focus on the top row, though in this case, both rows are nearly identical. The test statistic, t, is -12.837, with an associated p value of 0.000 (indicated here by Sig. 2-tailed). This indicates that there is a statistically significant difference between the two groups. Please note that a p value can never be zero. The results table in SPSS only provides three decimal places. In this case, when presenting the results, it is best to state that the p value is <0.001. The degrees of freedom (df) in the test is 48. This is based on the number of samples in the test (50) minus the number of groups in the test (2). The table also contains the mean difference between the two groups, -18.600, and 95% confidence intervals of this difference, which has a lower bound of -21.153 and an upper bound of -15.687. These values indicate the expected difference between the two groups if the analysis was re-run with a different sample from the same population.

You can also visually represent your data/results by creating a box plot. This is also useful for inspecting the amount of variance in each group.

**Creating a boxplot of your results**

A boxplot shows five statistics (minimum, first quartile, median, third quartile, and maximum) about the data in each group. It can also alert you to outliers.

In the top menu bar, click Graphs ➔ Chartbuilder ➔ a pop-up window appears ➔ Click Boxplot under the Gallery tab (lower left part of the pop-up window) ➔ Click the Simple Boxplot icon in the lower pane and drag it to the upper pane ➔
Click your nominal variable (Nutrition Group) from the upper left pane and drag it to the X-axis in the graph preview pane → Click your measurement variable (Height) from the upper left pane and drag it to the Y-axis in the graph view pane → Click OK → The graph will be produced in your Output window
Interpreting the boxplot:

The boxplot visually summarizes several pieces of information for each group. It displays the maximum and minimum values within each group, the median value, and the first and third quartiles. For this example, group 0 (malnourished) has a minimum value of 151 and maximum value of 170, and a median value of 161. The first quartile is 155 (25% of the data is less than 155), and the third quartile is 163 (75% of the data is less than 163).

An outlier is not present in this data. If it was, it would appear as an asterisk in the boxplot with a number next to it. The number indicates the case number of the outlier in your dataset.

You can edit your graph if you double click it in the Output window.
**Writing-up the results of the t-test**

Typically, the most important values to present in a results section of a research report are the t value, p value, and df. In the above example, you may write:

We found a statistically significant difference between the height of men that were malnourished as a child compared to those with proper childhood diets (t = -12.837, p <0.001, df = 48).

You may also want to include the mean value for each group and perhaps some measure of variance, e.g. standard error.

The text can also be accompanied with a box plot of the results.
One-way ANOVA

What is it
This test allows you to examine whether three or more groups exhibit a significantly different mean value for a particular variable.

When to use it
In a one-way analysis of variance (ANOVA), there is one measurement variable and one nominal variable, with the nominal variable having three or more categories. Multiple observations of the measurement variable are made for each value of the nominal variable. For example, you could measure the amount of transcript of a particular gene for multiple samples taken from arm muscle, heart muscle, brain, liver, and lung. The transcript amount would be the measurement variable, and the tissue type (arm muscle, brain, etc.) would be the nominal variable.

How the test works
The basic idea is to calculate the mean of the observations within each group, then compare the variance among these means to the average variance within each group. Under the null hypothesis that the observations in the different groups all have the same mean, the among-group variance will be the same as the within-group variance. As the means get further apart, the variance among the means increases. The test statistic, which is the F statistic, is the ratio of the variance among means divided by the average variance within groups.

Assumptions
ANOVA assumes that the data within each group are normally distributed and the each group has the same variance. The ANOVA is robust to moderate violations of these assumptions. Though, if severe violations occur, then the data should be transformed or a non-parametric version of the ANOVA should be conducted.
Example
You are interested in examining whether the average age of patients differ among three hospitals. The measurement variable is the age of each patient and the nominal variable is the hospital name. Each hospital can be given a code for analysis.

The data setup would look like this example:

<table>
<thead>
<tr>
<th>Age of Patient</th>
<th>Group</th>
<th>Group Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Diamondback Hospital</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>Diamondback Hospital</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>Diamondback Hospital</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>Diamondback Hospital</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>Diamondback Hospital</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>Sun Hospital</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>Sun Hospital</td>
<td>2</td>
</tr>
<tr>
<td>47</td>
<td>Sun Hospital</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>Sun Hospital</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>Sun Hospital</td>
<td>2</td>
</tr>
<tr>
<td>31</td>
<td>Cardinal Hospital</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>Cardinal Hospital</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>Cardinal Hospital</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>Cardinal Hospital</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>Cardinal Hospital</td>
<td>3</td>
</tr>
</tbody>
</table>

You can download the example dataset in Excel format, [here](#).
Conducting an ANOVA in SPSS

The following instructions and screen shots for the ANOVA example are based on SPSS version 19. In contrast, the instructions and screen shots for the t-test example are based on SPSS version 17. These two versions are very similar.

You can start your analysis after you open your data in SPSS (see the section “Conducting a t-test in SPSS” for instructions).

If you open the example dataset in SPSS, you will see:

To start the ANOVA, click the top menu bar at: Analyze → Compare Means → One-way ANOVA → a pop-up window will appear

The Dependent List panel contains the variables you want to test. In this case, move the Age of Patient variable to the Dependent List box. The GroupCode
variable should be moved to the Factor box, as this is the variable that contains the groups.

There are three additional buttons in the upper right corner of the One-Way ANOVA pop-up box that provide more detailed results for the ANOVA, all of which are optional: Contrasts, Post Hoc, and Options. Clicking the Post Hoc button allows you to test for differences between specific groups in your study. Remember, an ANOVA tests for a difference among all groups in your dataset, yet it does not specify which particular groups differ from each other. The additional options in the Post Hoc menu allow you to find which specific groups differ from one another (if any). Clicking the Options button allows you to produce various basic statistics related to your data.

Click the OK button in the One-Way ANOVA pop-up box when you are finished choosing your options (if any). Your results will appear in the Output window.

The main results table produced by the analysis is:

<table>
<thead>
<tr>
<th>Age of Patient</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>6970.300</td>
<td>2</td>
<td>3485.150</td>
<td>183.887</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td>1080.300</td>
<td>57</td>
<td>18.953</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8050.600</td>
<td>59</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The most pertinent values in this table are the df (degrees of freedom), F statistic, and Sig. (the p value). The df is based on the number of samples and number of groups in the analysis. In this case, we have 60 samples and three groups. The Between Groups df is the number of groups minus one. The Within Groups df is the total sample size minus the number of groups. The F statistic is a measure of the difference among groups calculated by dividing the Between Group variance (3485.15) by the mean Within Group variance (18.953); a higher value indicates that the groups differ more. This ANOVA produced a p value of <0.05 so we accept that the groups are significantly different from each other. Also, remember that a p value of 0 is not possible. The results tables in SPSS only produce three decimal places. Therefore, we can say that the p value is <0.001.
Writing-up the results of an ANOVA

Typically, the most important values to present in a results section of a research report are the $F$ value, $p$ value, and df. In the above example, you may write:

We found a statistically significant difference among the age of patients at the three hospitals we examined ($F = 183.89$, $p < 0.001$, df = 2, 57).

You may also want to include the mean value for each group and perhaps some measure of variance, e.g. standard error.

The text can also be accompanied with a box plot of the results. Here is a boxplot from the above example.
**Testing the Assumptions of Parametric Statistics**

The t-test and ANOVA are types of parametric statistics. These methods have a set of assumptions that should be satisfied before conducting the analysis. That being said, parametric statistics are robust to moderate violations of their assumptions. More severe deviations should be addressed by either transforming the dataset (e.g. using the $\log_{10}$ of each value) or by using a non-parametric equivalent.

There are statistical tests to help you examine whether your data fulfill the assumptions of parametric statistics. These methods include [Levene’s test](https://en.wikipedia.org/wiki/Levene%27s_test) for examining whether groups have similar variances. Yet, most biostatisticians consider these tests overly sensitive. Therefore, the best way to examine your data is by visually inspecting plots of your data.

A [histogram](https://en.wikipedia.org/wiki/Histogram) can be created that will allow you to examine whether your data follow a normal distribution or not. A histogram displays the value of the data (e.g. weight, length, height, etc.) on the X axis and the frequency of those data on the Y axis.

Here are examples of data that are considered normally distributed:
And examples of data that strongly deviate from normality (below):
You can create a histogram of your data using SPSS. If your datasheet contains data for more than one group you will have to create separate histograms for each group. If the data for all groups are in a single column, then you need to select the cases belonging to a single group. You can do this by clicking the top menu bar: Data → Select Cases → a pop-up window will appear.

Highlight the variable you want to use as your selection criterion, in this case Nutrition Group. Then, click the “If condition is satisfied” button, then If . . .

Another pop-up window will appear.
This window allows you to specify the criteria for selecting cases. In this instance, we want to select all cases that have NutritionCode 0. To do this, we highlight NutritionCode then click the right arrow button. This moves the variable from the left pane to the upper pane. Then will click the “=” button and then type 0. Once this is done, click Continue, then click OK in the next window. If you go back to your SPSS data window, you will see that a new column was produced called “filter_$”. The column contains 1s indicating the samples that are selected and 0s for the samples not selected. This corresponds to selecting only samples with NutritionCode 0. Now we are ready to create a histogram for the data in NutritionCode 0.

In the top menu, click Graphs → Chart Builder. This will bring up the same window that we used to create the box plots. Choose Histogram in the lower left pane. Then click and drag the leftmost histogram icon to the large pane at the top of the window.
Then click and drag Height from the left pane to the X axis on the graph. Do not move anything to the Y axis.

Then click OK. The histogram will appear in your SPSS output window.

To create a histogram for NutrionGroup 1, repeat the above steps except use NutrionGroupCode=1 in the select cases window.
Both the t-test and ANOVA assume that each group has a similar variance (also known as exhibiting homoscedasticity). If groups have different variances, then you are more likely to obtain a spurious p value. This problem is even more likely when the groups have very different sample sizes. You can also create a box plot of your data, which allows you to compare the amount of variation within each group. The box plot consists of a measure of central tendency, either the median or mean value for each group, as well as the variation within each group, defined by statistics such as the quartiles, standard deviations, or standard errors. Outliers within each group can also be displayed.

Here are examples of boxplots. In these cases, the groups are on the X axis and the values of the data are on the Y axis. These boxplots illustrate groups with similar variances:

These boxplots illustrate groups with different variances:

Creating boxplots is discussed in the t-test section.
Data Transformations

Log Transformation
If your data fail to meet the assumptions of the t-test or ANOVA, then one option is to transform your data and re-check the assumptions. To transform data, you perform a mathematical operation on each observation, then use these transformed numbers in your statistical test. There are several transformations you can do, with the most common being log transformation. Log transforming your data (using either log base 10 or the natural log), minimizes the effects of outliers and reduces the absolute magnitude difference between datapoints. In Excel, you can log transform your data by typing =LOG(A1), for log base 10, or =LN(A1), for the natural log. A1 refers to the location of your datapoint in the spreadsheet; in this case, Column A, Row 1. For example:

Note: You cannot log transform zero. If your dataset includes zeros, then you can add a constant to each datapoint, e.g. add 0.5 to each datapoint.

Square-Root Transformation
This consists of taking the square root of each observation. This transformation is
commonly used for count data, e.g. the number of patients waiting in the Emergency Room at different times during the day. Square-root transformation can be accomplished in Excel by typing the formula: =SQRT(A1). A1 refers to the location of your datapoint in the spreadsheet; in this case, Column A, Row 1. For example:

Note: If you have negative numbers, you can't take the square root; you should add a constant to each number to make them all positive.
Nonparametric Statistics

These methods are commonly used when your data do not satisfy the assumptions of parametric tests such as the t-test and ANOVA. This is especially true for data that are not normally distributed, which is a common occurrence with small datasets. Nonparametric statistics are performed on ranked data, so your continuous variable is converted to a ranked variable, with data sorted (automatically by the software) by their rank in the overall data set: the smallest value gets a rank of 1, the next smallest gets a rank of 2, and so on. The loss of information involved in substituting ranks for the original values can make this a less powerful test than parametric methods, so techniques such as the t-test and ANOVA should be used if the data meet the assumptions.

Two common nonparametric methods are the **Kruskal–Wallis test** and the **Mann–Whitney U-test**. The Kruskal–Wallis test is most commonly used when there is one nominal variable and one measurement variable, and the measurement variable does not meet the normality assumption of an ANOVA. It is the non-parametric analogue of a one-way ANOVA. The Mann–Whitney U-test (also known as the Mann–Whitney–Wilcoxon test, the Wilcoxon rank-sum test, or the Wilcoxon two-sample test) is limited to nominal variables with only two values; it is the non-parametric analogue to Student's t-test. It uses a different test statistic (U instead of the H of the Kruskal–Wallis test), but the p-value is mathematically identical to that of a Kruskal–Wallis test.

Although both tests do not assume that the data are normally distributed, they do assume that the observations in each group come from populations with the same shape of distribution. Therefore, if different groups (e.g. treatment vs. control) have different shapes (one is skewed to the right and another is skewed to the left, for example, or they have different variances), then these tests may give inaccurate results.
Conducting Nonparametric Statistics in SPSS

Analyze → Nonparametric Tests → Independent Samples

A pop-up window will appear.
Click the Objective tab at the top of the window → Click the button next to “Automatically compare . . . “ → click the Run button

Another pop-up window will appear.
Move your categorical variable to the “Groups” box and your continuous variable to the Test Fields box.

If you click the Run button, the SPSS will automatically choose the nonparametric test that fits your data. Alternatively, you can customize your test by clicking on the Settings tab.

After you click Run, the results will appear in the output window.
The results of the test is presented as a statement related to the null hypothesis, the name of the test run, the significance level (i.e. p value) for the test, and statement regarding to reject or accept the null hypothesis.

In this case, a Mann-Whitney U test was conducted because the categorical variable contained two groups. If the categorical variables contained more than two groups, then a Kruskal-Wallis test would be conducted.

*Parts of this chapter were adapted from the web version of: McDonald, J.H. 2009. Handbook of Biological Statistics (2nd ed.). Sparky House Publishing, Baltimore, Maryland.  
http://udel.edu/~mcdonald/statintro.html