Statistical Analysis for Inquiry Projects

Kevin Nowland

August 14, 2011

Contents

Acknowledgements

I would like to thank Jodi Bacon for pursuing this project on behalf of her students. I would also like to thank Worthington Schools and Mike Miller for approving funds to make this booklet possible.

Chapter 1

Introduction and How to Use This Booklet

1.1 Introduction

This booklet has been made specifically for you! That is, assuming you are going to be doing an inquiry (science fair) project. The inquiry projects are meant to help you the student make a first foray into scientific research. The business of science is finding the answers to interesting questions. Though this booklet is not designed to help you find an interesting question, though it might suggest some, it is here to help you design an experiment to find an answer more effectively.

Why does a question need to be answered *effectively* and not just *answered*? Though textbooks often give answers as absolute, this is rarely the case. For example, if you flip a coin you believe to be fair and get 8 heads and 2 tails, does this imply that the coin is actually not fair? If we claim the coin is not fair, can we put a number to our certainty? Now suppose we get 80 heads and 20 tails. Can we be more sure that the coin is not fair than when we had only 8 heads and 2 tails? And again, can we quantify our certainty?

And how can we go about answering more complicated questions? For example, how can we tell what color the sky is in a quantitative manner? Though this question can be answered easily by looking up, can we use numbers somehow to find the answer?

As it turns out, we usually can answer these questions and quantify, or at least estimate, how certain we are. This is the business of statistics, and it, along with mathematics, forms the language of science. This booklet is an introduction to the statistics used in scientific experiments, and should help you design an inquiry project which can benefit from statistical analysis.

1.2 How to Use this Booklet

This booklet should be most effective if used from the beginning of an experiment onward. In order to achieve results that can be statistically analyzed, an experiment must be designed with this goal in mind. As we shall see, there are certain things these experiments must have, such as a null hypothesis, an alternate hypothesis, and a numerical data.

Once the hypotheses are determined and data is obtained, or if it already is, this booklet can serve as a reference on how to analyze the data and draw a conclusion (even if that conclusion is “inconclusive”). Also provided are websites which can be used to find the necessary values. For those who are particularly adventurous, the details on how to use the free statistical software package R are also provided. For help with installation of R, please see Appendix A).

What this book does not provide is an extensive amount of background information on statistics. As this booklet is designed for high school students who for the most part have not studied calculus, this material is by and large too advanced, even though the basic concepts are not. References to probability and statistics books used in the creation of this booklet can be found in the bibliography, but all reference books might require up to two years of calculus as well as introductory linear algebra and probability.

Chapter 2

Designing Your Experiment or Making a Testable Hypothesis

2.1 The Importance of Numbers

A simple yet important lesson is that numbers are important. Though this may seem obvious if one wants to use statistics or mathematics, it is not always obvious in science. An example given in the introduction was that the question “What color is the sky?” has the simple answer of “blue”. However, even this is a question with possible numerical answers. Colors, for instance, correspond with certain frequencies of light. One way of quantifying the answer of “blue” is to find the average frequency of the visible light coming from the sky. If this frequency is close to the frequency of blue light, we may then conclude that the sky is in fact blue.

This is not the only way to numerically ask “What color is the sky?” The sky contains many frequencies of light, and we may instead try to find the frequency with the highest amplitude, in other words, the brightest frequency. If this frequency is close to the frequency of blue light, then maybe we conclude that the sky is blue. Unfortunately, as this example demonstrates, it is not always known which is the best way to quantify an experiment. Maybe both the mean frequency and “brightest” frequency will agree and maybe they will not. This chapter should provide some guidance on how to ask a quantifiable question, and in determining what is the best way to quantify the experiment.

Equally important is that there be a lot of data available. A large number of data points usually helps scientists be more sure about their conclusions. For example, flipping a coin twice and getting heads both times is not indicative that the coin is not fair, as there is a 25% chance of this occuring with a fair coin. Also, some tests require large numbers of data points to be accurate, such as those included in the next chapter.

2.2 Keep It Simple

As is often noted, science does not answer questions such as ”What is the meaning of life?” or “Why does an inquiry project help me learn science?”; these questions are usually left to philosophy or sociology or other fields. This is not to say that these fields are not scientific, just that science can only be used as evidence in support of answers to these questions. Science typically asks more direct questions. For example, "What is the density of wood?" or "Do cardinals live on average longer or shorter than sagehens?" These more direct, simpler questions fall more within the domain of science.

A crucial point to remember is that the most successful experiments try to answer only one question at a time. For example, when tackling the general question of "What determines the deliciousness of a cookie more: fat content or sugar content?", a well designed experiment will hold the fat content constant while changing the sugar content. Likewise, fat content should be held constant when changing the sugar content.

An example of what not to do, would be to compare the taste of cookies which are high in both fat and sugar only with those that are low in both fat and sugar. Suppose both types of cookies are determined to be equally delicious. What would the conclusion be? Possible conclusions are that the sugar and fat contents of cookies do not affect deliciousness. However, it is also possible that both types of cookies are equally delicious because high fat content and low sugar content are both desirable and their opposites are not. The situations might have cancelled each other out. This is why holding the sugar or fat contents constant is useful.

In fact, it is much simpler to check only the effect of fat content or sugar content relative to a control of a fixed recipe. In doing so, it is easier to make a conclusion. In testing both the fat and sugar content separately, it would be hard to recommend the best level of fat and sugar, because as we have seen, we do not know if changing one affects the result of changing the other. It would be simpler to test only one, and then make a recommendation based on the control. I.e., test only the fat content and possibly conclude that "raising the fat content increased the deliciousness of the cookie as rated by the consumer".[1]

2.3 The Importance of Independence

Tackling only one question at a time is important, but there is also a question of independence. Independence is when one outcome of an experiment does not affect another outcome, and is a highly desirable property when attempting to statistically analyze data. An example of independent data comes from flipping a coin or rolling a die. If the first flip of a coin comes up heads, this does not change the chance of flipping a head on the second roll. The coin flips are independent. Similarly, rolling a 6 on a die once does not affect the chances of rolling a 6 again. The rolls are independent.

An example of a non-independent situation is drawing cards from a 52 card deck and discarding the drawn cards. If you are looking at suit color and first draw a diamond then discard the card, you no longer have the same 50% chance of drawing either red or black. The second draw depends on the first. In this case there is a relatively simple relation between the first draw and the second, but this is not always the case.

Suppose you want to study how much gamblers are likely to wager at certain points in a game of blackjack and you choose the first deal, the 25th deal, and the 50th deal. To do this, one might choose five gamblers at a table and see what they bet at each of

[1]There is a way to test the effects of both things at once. Check out Section 5.1 for ideas on how to deal with this problem.

these times. However, a player may bet more highly if they have been doing well such that the 25th deal depends somewhat on the first deal. Also, if four of the players won on the 24th deal, maybe the fifth player will bet higher on the 25th deal than he would have in isolation. To make the data more independent, the players should be at different tables. This would garner independence between the data points coming from different players. To obtain independence between data for the first, 25th, and 50th deals, different players should be used for each of these deals. (As we shall see later, more players should be used, as well in order to get more data.)

2.4 The Null Hypothesis vs. the Alternate Hypothesis

Perhaps the most basic function of a scientist is to test hypotheses. A theory is formulated, and then it is tested. Taking the coin flipping experiment as an example,

The *null hypothesis* is what is assumed to be true. This does not imply that the scientist performing the experiment believes the null hypothesis is true, it is merely assumed to be true. That which the scientist believes to be true is the *alternate hypothesis.*As an example, suppose a student has built a model volcano. With a given amount of baking soda for a constant amount of vinegar, the student knows that the volcano will erupt 5 cm above the top of the volcano. The student, desiring a more powerful explosion, decides to double the amount of baking soda. The null hypothesis in this case is that the explosion will still only reach 5 cm above the top of the volcano. The alternate hypothesis that the student wishes to test is that the explosion will reach, on average, greater than 5 cm above the top of the volcano.

Note that the null hypothesis specifies an amount, in this case 5 cm. However, the alternate hypothesis only says that the amount will be greater than 5 cm and does not specify an amount. We will mostly consider in this booklet the three most common alternate hypotheses: that the new amount is greater than the original, is less than the original, or is simply not equal to the original. Never will an actual amount be specified. There is one other type of alternate hypothesis that is considered in Section 5.1. If the experimental results are very unlikely (how unlikely is to be determined later) if the null hypothesis were true, then the alternate hypothesis could be accepted with some degree of certainty.

The alternate hypothesis should usually be decided ahead of time, but can be determined after. If the new baking soda amount led to the volcano not erupting beyond the top of the volcano in 100 trials, the student might have used the alternate hypothesis that the new solution actually decreased the height of the eruption.

2.5 Are the Results Significant?

For all of the tests included in this booklet, the null hypothesis can only be rejected with a certain level of confidence. For example, it is possible that a fair coin will come up heads four flips in a row. In fact, this has a .0625 probability (or a 6.25% chance) of occurring. If the null hypothesis is that the coin is fair and the alternate hypothesis is that the coin is not fair, we could not reject the null hypothesis at the 5% significance (probability) level. However, if we only required 10% significance, we would reject the null hypothesis and accept that the coin is not fair.

The key idea, though, is that we can only reject the null hypothesis at a certain *significance*. Rejection of the null hypothesis, or even acceptance of it, does not imply certainty, just a certain amount of "confidence". This confidence is what we call significance. As a note, it is very common to use a 5% significance level, and the significance level should be decided before beginning hypothesis testing begins. The higher the number chosen, the less rigorous the test. The lower the number, the harder it is to reject the null hypothesis in favor of the alternate hypothesis, but the more certain it is that the alternate hypothesis is true. Significance is denoted by α, alpha.

2.6 Percent Likelihood, Mean, and Variance

This booklet will cover testing for three different things. The first, is the percent chance of something happening. A scientist wishes to see what the chance is of some specific event occuring. He observes data, and sees how often the event occurred. This is like the coin flip experiment, where a coin, believed to be fair, is flipped 200 times. If 160 of the flips result in heads, we would like to conclude that the coin was not fair.

The second, and perhaps most common, value to test for is the mean, or average. For example, a scientist might want to see if high school students, either male or female, are as tall as the average adult of the same sex. Given this average height, the scientist might measure the height of 50 students from the high school and try to conclude based on their average height whether the average high schooler is taller or shorter than the average adult.

A third test is for the variance. The variance is a measure of how much values differ from the mean. High variance implies that values are more frequently very far from the average value, while variance close to zero implies that most values are close to the average. Variance is always positive.

Variance can be more difficult to test for than mean, since it is often less obvious what the variance is expected to be. (Do you know the variance for flipping a fair coin, for example? This is often hard to figure out without knowledge of probability distributions.) It is more common to test the variance of two independent group and see if they are different or the same, as this does not require attempting to determine the variance ahead of time.

There are also several other types of testing which are covered in the "Other Types of Data" chapter. Most commonly used from this chapter is the calculation of the correlation between two variables.

2.7 General Hypothesis Testing Process

Examples for each test will be given, but a general procedure is outlined here.

- Decide which test is most appropriate based on your experiment and determine how much data to collect. (Subsection 2.7.1.)

- Form a null hypothesis. This usually has a specific value.

- Form an alternate hypothesis. This does not involve specific values.
- Determine the significance level of the test, α.
- Record data.
- Calculate appropriate statistical value, such as the mean and variance of a group of data.
- Calculate a *test statistic* which relates the statistical values.
- Calculate the percentage likelihood associated with the test statistic. This is called a *p-value*, and is the chance of getting the value of the test statistic or a value that is even less likely if the null hypothesis is true.
- If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. Otherwise, continue to accept the null hypothesis.

2.7.1 Determining Which Test is Appropriate

Follow this chart to help figure out which test is appropritae for your experiment. Then, read the appropriate sections. It is almost always helpful to have 30 data points, so if possible, get this much data or more. If fewer data points are recorded, usually more tests are required before proceeding.

1. Testing how often something occurs (prcentage likelihood): See assumptions for number of data points needed (usually 30 are sufficient). This is the coin flip experiment, or attempting to find what percentage of students from two elementary schools will graduate from a common high school.
   (a) One group of data: Subsection 3.3.1
   (b) Two groups of data: Subsection 3.3.2
2. Testing the average or most likely value of something (mean or median). For example, the average height of a plant or the average number of points scored by a football team over the course of 3 years versus those scored by the same team 20 years ago.
   (a) One group of data (no control)
      i. 30 data points or more: Subsection 4.3.1
      ii. Less than 30 data points: check if data are approximately normal
         A. Approximately normal: Subsection 4.4.1
         B. Not normal (median): Subsection 6.3.2
   (b) Two groups of data (usually control and exsperimental groups)
      i. Both groups of data have 30 or more points: Subsection 4.3.2
      ii. At least one group of data has fewer than 30 data points: check if data are approximatey normal

A. Approximately normal: Subsection 4.4.2
B. Not normal (median): Subsection 6.3.3

3. Testing how much things vary or how often things are different from what is expected (variance): need 30 data points for every group of data. For example, do the chapters of a book vary greatly in number of typos (some chapters have none, others have 20) or just a little (every page chapter has exactly 20 typos).
   (a) One group of data: Subsection 4.5.1
   (b) Two groups of data: Subsection 4.5.2
4. Testing if two variables are related (correlation): Subsection 5.2.2. For example, as price goes up, do lightbulbs tend to last longer?
5. Testing survery results with many categories. For example, political or psychology polls.
   (a) Testing whether data are in categories as expected (goodness of fit): 5.1.1. (Do eye colors in the school follow what would be predicted by genetics?)
   (b) Testing whether two big categories (like gender) affects the chance of belonging to other, smaller categories (independence): Subsection 5.1.2. (Is gender independent of the number of AP classes taken?)

2.7.2 Examples

Here are 10 example experiments. Determine which test is appropriate. For more examples, each test is accompanied by a possible experiment and its accompanying data.

1. Ariel and Kristin want to test whether students who play a musical instrument are better at math than non-musicians. They plan to survey band students anonymously to ask what math class they are in and what grade they currently have in that class, and compare this data to a random sample survey of students who do not play any instrument.
2. Ellen wants to test whether the color of the paper on which a quiz is given affects the students' grades on the quiz. She is working with a 1st grade teacher. They will administer the weekly spelling quiz over 8 weeks on different colors of paper. Ellen has read that blue is a color associated with higher scores.
3. Marcus and Joe want to test whether increasing the CO2 in an environment makes plants grow better as defined by final dry mass. They will set up two environments and grow bean plants in each environment. One will have normal air and the other will have CO2 added regularly.
4. Tyler and Emily read that students who eat breakfast maintain a healthier weight than those who skip breakfast. They plan to define "breakfast" as any food or drink containing at least 250 calories and 5 grams of protein and eaten within 1 hour of

waking up. They define healthy weight as a BMI in the normal range (18.5 to 24.9). They will survey high school students anonymously to obtain their data.

5. Aaron and Julia want to test whether increased light affects the reproduction of fish. They have two large tanks each containing an equal number of male and female guppies. They keep all conditions the same except for light. Over one tank, a 100 Watt CFL bulb is left on continuously. The other tank is placed near a window and receives normal daylight and is covered at night. They plan to count the baby guppies that are born over a two-month period.

6. Chris and Brent want to test whether more expensive skateboard ball bearings provide less friction than cheaper bearings. They plan to put different types of bearings into a skateboard and roll it down an inclined plane and measure the distance the board travels across a smooth surface at the base of the plane. They do this 30 times for the cheap bearings and 30 times for the more expensive bearings.

7. Craig read that garlic is a natural inhibitor of bacteria in food. He plans to place a filter paper disc soaked in mashed garlic cloves at the center of a Petri dish inoculated with bacteria from yogurt. He will measure the zone of inhibition around the filter paper after 2 days of incubation. He does the same thing but without soaking the paper in garlic cloves. He uses 50 filters for each group.

8. Annie thinks that students who play action video games regularly are better at multi-tasking than students who never play action video games. She wants to test this by giving students several tasks to complete at once (such as talking to a friend on the phone while washing dishes and studying for a short memory quiz). She will rate their multi-tasking ability by measuring the time it takes them to complete all the tasks along with their grade on the memory quiz. Each participant will also estimate how many hours per week he or she spends playing action video games (will be defined as certain games).

9. James thinks that swimmers will have a faster time if wearing the new super-streamlined Speedo swimsuits than if wearing a traditional suit. He will measure the times of his teammates during practice. Swimmers will wear different suits on different days.

10. Greg thinks that students will contribute more money to a charitable cause if they are with a friend than if they are alone. He plans to set up a collection table at a sporting event and give away cookies for a donation. He will note how much each person donates and whether he or she was alone or accompanied by friends.

Chapter 3

Testing Hypotheses: Percentage Likelihood

3.1 Binomial Data: Coin Flip Experiments

The simplest thing to test for is the percentage likelihood of something occurring. The classic example, as seen earlier in Section 2.5 is the coin flip. If you expect 50% of coin flips to come up heads and find that 80% of them do out of 200 flips, we want to determine whether the coin is fair or not. Though the coin flipping experiment is the classic example, it can be used for any percent chance of something happening, such as the chance of a light bulb failing after lit for 10 hours straight.

3.2 Assumptions

In order for the test to be reasonably accurate, we need a large number of data points. Let p_0 be the assumed chance of an event occurring (this is the value from the null hypothesis). A sufficiently large number of data points for the test is that the number of data points times p_0 and times $(1 - p_0)$ must both be above 5. For the coin flip experiment where $p_0 = .5$, this would mean that no fewer than 10 flips of the coin should be used. For the two-sample test, this condition must be met for both p_x and p_y, the percantage likelihood for both the first and second populations.

For example, in a one sample test, the student expects that 50% of coin flips will come up heads. Then $p_0 = 0.5$. In order to use the test, $0.5n$ must be larger than five. The student therefore needs at least ten data points. If the student expects 70% to come up heads, then both $0.7n$ and $0.3n$ would have to be larger than 5. Therefore the student would require 17 coing flips.

3.3 Tests for Percentage Likelihood

3.3.1 One-Sample Test

1. Form a null hypothesis, i.e., choose the probability p_0 (percent divided by 100) that you think the given event has of occurring.

2. Form an alternate hypothesis, this is that the true probability is greater than p_0, less than p_0, or simply that it is different than p_0.

3. Choose a significance level, α. Usually 5%.

4. After collecting data, let $\hat{p}$ be the percentage of the n trials where the desired event occurred.

5. Calculate $S_p = \sqrt{\frac{p_0(1-p_0)}{n}}$.

6. Calculate the test statistic
$$z = \frac{\hat{p} - p_o}{S_p}.$$

7. To calculate the p-value for this particular value of z, go to http://www.changbioscience.com/stat/ztest.html. If your test is that the true percentage is simply different than p_0, choose "Two-tailed p-value" from the drop-down menu above the "calculate p" button. Otherwise, leave it at "One-tailed p-value". Enter the value of z you calculated in the previous step in the field to the right of "z=(y-mu)/(sigma/sqrt(n))". Press the "Calculate p" button. Multiply the resultant number by 100%. To use R, see Subsection A.2.1.

8. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

It is important to note that if the alternate hypothesis is that the true percentage likelihood is greater than p_0 but the experimentally found percentage $\hat{p}$ is less that p_0, then you must add 50% to the p-value found in step 7. In other words, it is highly unlikely that you will reject the null hypothesis in this case. Similarly if the hypothesis is that the true percentage likelihood is less than p_0 and the found value is greater than p_0 50% must be added to the p-value.

Example 1. *Gladys wants to test what percentage of rats can successfully memorize the way out of a simple maze that she has designed. She decides to put a piece of food at the end of the maze and let each rat find the food 5 times before she begins testing. She then removes the food, and lets each of 20 different rats try to get to the end in 3 minutes. Her null hypothesis is that 50% ($p_0 = .5$) of the rats will complete the maze, and her alternate hypothesis is that more than 50% of the rats will find the end of the maze before the time limit. She tests at the $\alpha = 5\%$ significance level.*

She finds that 14 of the 20 rats successfully reach the end, i.e., 70% and $\hat{p} = .7$, and calculates $S_p = 0.112$. The test statistic is then $z = 1.79$. The p-value associated to this *value of z is 3.67%, below the 5% significance level that she chose to test at. Therefore with 95% confidence Gladys concludes that more than 50% of the rats had learned to complete the maze.*

3.3.2 Two-Sample Test

Let p_x be the percentage likelihood of the event for the first sample of data and let p_y be the percentage likelihood of the second sample. Similarly, n_x and n_y are the number of data points for the first and second samples, respectively.

1. Form a null hypothesis, i.e., choose $D = p_x - p_y$ the assumed difference between the percentage likelihoods of the two groups.

2. Form an alternate hypothesis, this is that the difference $p_x - p_y$ is less than, greater than, or simply different than D. Note that the difference is still the percentage likelihood of the second sample subtracted from that of the first sample.

3. Choose a significance level, α. Usually 5%.

4. After collecting data, let $\hat{p}_x$ be the percentage of the n_x trials where the desired event occurred in the first sample, and let $\hat{p}_y$ be the percentage for the n_y trials of the second sample.

5. Calculate $S_p = \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$.

6. Calculate the test statistic
$$z = \frac{\hat{\hat{p}}_x - \hat{p}_y - D}{S_p}.$$

7. To calculate the p-value for this particular value of z, go to http://www.changbioscience.com/stat/ztest.html. If your test is that the true percentage difference is simply different than D, choose "Two-tailed p-value" from the drop-down menu above the "calculate p" button. Otherwise, leave it at "One-tailed p-value". Enter the value of z you calculated in the previous step in the field to the right of "z=(y-mu)/(sigma/sqrt(n))". Press the "Calculate p" button. Multiply the resultant number by 100%. To use R, see Subsection A.2.1.

8. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 2. *Daniel goes to his middle school and high school libraries, and wants to determine the percentage of books have been written in by students who have checked them out. His null hypothesis is that 10% fewer books, or $D = .1$, will have been marked in the high school library than in the middle school library (he designates the books from the high school library to be the first sample). He randomly selects 50 books from each library and leafs through them to look for pencil and pen markings. His alternate hypothesis is that the* *difference is less than 10%. He finds that 14% of the books in the high school library have been marked while 22% of the middle school books are marked.*

Daniel calculates that $S_p = 0.0649$ *and that* $z = -0.924$*. The p-value corresponding to this value of* z *is 17.8%. Therefore Daniel continues to accept the null hypothesis that 10% fewer of the books in the high school library are marked than those in the middle school library.*

Chapter 4

Testing Hypotheses: Mean and Variance

This chapter contains tests for hypotheses about the mean and variance of one or two different sets of data.

4.1 Assumptions: Large Numbers, "Normality", and Independence

For the purposes of the large number tests for the mean and the variance test, large means greater than 30 data points. As more data points are taken, the average of those points tends to become more "normal" in a probabilistic sense.[1]

The requirement for the small sample mean tests and the is that the data are approximately "normal". Several simple tests for normality are described in Section 6.2. Without large numbers, which here means 30 data points or more, this condition is much more important in ensuring the accuracy of the small sample tests. In the case of non-normal data, please see the tests described in Chapter 6. Normality is also important in testing for the variance.

As described in the chapter on experimental design, independence of all data points and between groups of data points is required for these tests to be accurate.

4.2 Some Important Statistics and Notation

Before testing we must define certain quantities and notation. We begin with the mean, which is the most commonly tested for statistic.

Definition 1 (sample mean, $\bar{X}$). *The mean will be denoted by $\bar{X}$ and is defined as the average of all obtained values:*

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \tag{4.1}$$

[1] This is because of the *Central Limit Theorem*, which is covered in undergraduate probability courses.

where $x_1, x_2, \ldots, x_n$ are the n values obtained in the experiment. While $\bar{X}$ is the sample mean, the mean itself (which is what is used in the null hypothesis) is written with the greek letter μ (mu).

The other most important statistic is the sample variance. The sample variance is given by the following formula:

Definition 2 (sample variance, S)**.** *The sample variance is*

$$S = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + (x_3 - \bar{X})^2 + \ldots + (x_n - \bar{X})^2}{n-1}. \tag{4.2}$$

It is important to note that you do not divide by n, but by $n-1$, unlike the sample mean. When testing for the variance, the assumed value in the null hypothesis is not written with S, but is instead written as σ^2 (sigma squared).

Various probabilities will also need to be calculated to complete the tests below. This can either be done using the statistical programming package R, or using the various websites that are given with each test. For more information on how to install and use R, please see Appendix A.

4.3 Large-Sample Tests for the Mean

The mean, as noted above, is the statistic for which it is most common to test. Two tests are presented here: the first is a test with only one set of data, and the second is for two independent sets of data. These tests are only suitable for use if there are 30 points of data or more. When there are two sets of data, each set must have 30 or more data points.

4.3.1 One-Sample Test

1. Form a null hypothesis. In other words, suppose that the mean $\mu = \mu_0$ for some specific value μ_0.
2. Form an alternate hypothesis. This is either that the true mean is either greater than μ_0, less than μ_0, or simply different from μ_0.
3. Choose a significance level, α. Usually 5%.
4. Calculate $\bar{X}$ using the formula in equation 4.1 above.
5. Calculate S using the formula in equation 4.2 above.
6. Calculate the test statistic

$$z = \frac{\bar{X} - \mu_0}{\sqrt{S/n}}. \tag{4.3}$$

7. To calculate the p-value for this particular value of z, go to http://www.changbioscience.com/stat/ztest.html. If your test is that the true mean is simply different than μ_0, choose "Two-tailed p-value" from the drop-down menu above the "calculate p" button. Otherwise, leave it at "One-tailed p-value". Enter the value of z you calculated in the previous step in the field to the right of "z=(y-mu)/(sigma/sqrt(n))". Press the "Calculate p" button. Multiply the resultant number by 100%. To use R, see Subsection A.2.1.

8. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 3. *Archie has 36 full grown, male box turtles with an average weight of 330 grams. He decides to switch their food for two months to the new Grow-Big Turtle Chow and see if they gain or lose weight. He records the change in each turtle's weight (in grams) at the end of the two months. He finds the following data:*

5	7	4	8	−1	5	7	2	3
3	−1	4	1	5	9	2	−6	−5
7	12	−15	5	14	17	3	0	−2
2	4	1	0	0	4	−3	1	6

Archie's null hypothesis is that the turtles' average weight will not change. In other words, that the average weight change will be 0. His alternate hypothesis is that the turtles will change weight. In other words, that the average turtle would have gained or lost weight. He chooses to test at the $\alpha = 5\%$ *significance level.*

He finds that $\bar{X} = 3$, *i.e., that the male box turtles gained on average 3 grams over the 2 months of eating the Grow-Big Turtle Chow. Calculating the sample variance yields* $S \approx 32.57$. *Using 4.3, Archie finds the z value of*

$$z = \frac{3 - 0}{\sqrt{32.57/36}} = 3.154.$$

Since Archie wanted only to test for a different weight, he chooses the two-tailed test on the website and finds a value of 0.001610. Multiplying by 100% gives 0.16%. This is less than the $\alpha = 5\%$ *significance that Archie chose at the beginning of his analysis. He therefore rejects the null hypothesis and accepts his alternate hypothesis that the Grow-Big Turtle Chow does lead to weight change in the turtles.*

Being an enterprising and enthusiastic science student, Archie changes his alternate hypothesis to be that the Grow-Big Turtle Chow actually increases the weight of the turtles. Choosing the one-sided test instead, he finds a p-value of 0.00805%, such that he accepts the alternate hypothesis that the new turtle food increased the weight of his turtles at the 5% level.

4.3.2 Two-Sample Test

This two-sample test can only be used with independent data. In other words, the two groups of data should come from entirely different sources. For example, if you wanted to test the weights of frogs at 1 month and 2 months old, and the frogs used in each sample are the same, this test does not apply. In such a case, simply subtract the weights from each other and use the one sample test from above. An example of how to use this test will be provided after the test procedure. Remember, both sample populations must have at least 30 data points.

Finally, we will let $x_1, x_2, \ldots, x_{n_x}$ be the n_x data points from the first group of data with $\bar{X}$ being its average value and S_x its sample variance. Let $y_1, y_2, \ldots, y_{n_y}$ be the n_y data points from the second group of data with $\bar{Y}$ being its mean and S_y its sample variance.

1. Form a null hypothesis. To do this, let D be the expected difference between $\bar{X}$ and $\bar{Y}$ (without regard for what $\bar{X}$ and $\bar{Y}$ are found to be). This is often that there should be no difference, in which case $D = 0$. An example is provided below for clarification.

2. Form an alternate hypothesis. This is either that the true difference should be greater than D, less than D, or simply different than D.

3. Determine a significance level, α. Usually 5%.

4. Calculate $\bar{X}$ and $\bar{Y}$ using equation 4.1 above.

5. Calculate S_x and S_y using equation 4.2 above.

6. Calculate the test statistic
$$z = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{S_x}{n_x} + \frac{S_y}{n_y}}}.$$

7. To calculate the p-value for this particular value of z, go to http://www.changbioscience.com/stat/ztest.html. If your test is that the true difference is simply different than D, choose "Two-tailed p-value" from the drop-down menu above the "calculate p" button. Otherwise, leave it at "One-tailed p-value". Enter the value of z you calculated in the previous step in the field to the right of "z=(y-mu)/(sigma/sqrt(n))". Press the "Calculate p" button. Multiply the resultant number by 100%. To use R, see Subsection A.2.1.

8. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 4. *Sophie wants to study the effects of listening to different types of music while studying. She makes two hour long mixes, one of classical music and one of hip-hop. She also makes a study sheet of basic algebra facts and determines that it will take about half an hour to study the sheet. She gives the study two groups of students, one of 30 students and one of 32 students, none of whom know anything about algebra. She tells the first group to study while listening to the classical music on their mp3 players, and the second group to listen to her hip-hop mix. She gives all the students the same quiz about the study sheets* *the next day. The students who listened to classical music (the X set) received the following scores out of 100:*

81	83	84	85	83	82	82	75	83	83
74	78	84	81	72	79	84	89	84	81
87	89	82	85	83	88	78	88	86	79

while the students who listened to hip-hop (the Y set) while studying received the scores:

80	82	90	87	87	86	85	97
87	84	82	87	90	92	89	98
79	89	99	99	83	99	89	94
89	79	83	97	94	97	80	83

Sophie's null hypothesis is that there will be no difference between the scores, although she suspects that this will not be the case. In other words, her null hypothesis is that $D = 0$*. Her initial alternate hypothesis is that the students who listened to the classical music while studying will score on average higher than the students who listened to classical music. However, she noticed that the mean score of the classical music listening students was actually lower than the mean score for the hip-hop listening students. She therefore must accept the null hypothesis in this case. Instead, Sophie decides to use the alternate hypothesis that hip-hop students score higher than the classical music students. Surprised at her results, she decides to be rigorous and use a* $\alpha = 0.5\%$ *significance level.*

She calculates the classical music listening student mean to be $\bar{X} = 82.5$ *while the hip-hop students' average score was* $\bar{Y} = 88.6$*. The sample variances were found to be* $S_x = 16.3$ *and* $S_y = 40.8$*. The z value she finds is*

$$z = \frac{82.5 - 88.6 - 0}{\sqrt{\frac{16.3}{30} + \frac{40.8}{32}}} = -4.52.$$

She finds a p-value of approximately 0.0000309%. This value is lower than her chosen α *level of significance, so she rejects the null hypothesis and accepts the alternate hypothesis that the students who listened to hip-hop scored higher than those that listened to classical music.*

Example 5. *Suppose Troy wants to study how long two different types of batteries last in a given brand of headlamp. He chooses batteries brands X and Y. He buys 32 batteries of brand X and 40 batteries of brand Y. Advertisements for brand X claim that their batteries last on average two hours longer than leading competitors. Troy decides to make this his null hypothesis, that the batteries of brand X will last on average two hours longer than the batteries of brand Y. In other words, that* $D = 2$*. His alternate hypothesis is that D is less than two. In other words, that the batteries of brand X will not perform 2 hours longer than those of brand Y. He decides to test at the* $\alpha = 5\%$ *level. After testing, he has the following data for the type X batteries:*

18.5	27.2	24.1	20.7	22.6	24.5	21.8	22.0
26.2	17.8	21.1	23.2	19.0	21.5	25.1	24.2
21.7	19.6	18.4	25.7	24.0	21.8	17.7	21.5
21.7	19.4	23.9	24.2	22.6	17.7	24.1	21.5

The data for the Y batteries are:

22.2	21.5	16.3	19.8	21.5	22.9	19.3	19.9	19.4	17.3
20.6	19.5	18.2	19.9	17.2	18.7	21.0	18.2	20.6	23.8
16.9	19.4	22.6	21.0	21.5	21.8	20.8	25.5	17.6	21.7
13.8	21.1	18.5	23.5	19.2	17.0	20.0	20.8	20.4	18.0

Troy calculates that $\bar{X} = 22.03$ and $\bar{Y} = 19.97$ and that $S_x = 6.804$ and $S_y = 5.195$. Using these values, he calculates the test statistic using the equation in part 6 of the outlined method to be 0.1046. He finds that the p-value to be 45.8%. Troy was testing at the 5% level such that he cannot reject the null hypothesis and concludes that the batteries of type X really do last 2 hours longer than those of type Y.

4.4 Small-Sample Tests for the Mean

Unfortunately, while having 30 different data points for one or possibly two groups is preferable, it is not always possible. There are still tests available in this situation. The catch is that these tests are only accurate when the data are approximately normal, as described in Section 6.2. Please consult that section before beginning to use these tests. The two tests presented here are merely modifications of the tests used above. Please refer to the large sample tests for the full, guided instructions.

4.4.1 One-Sample Test

The test procedure below only includes the modifications to the procedure for the one-sample test.

6. The statistic is called t instead of z, but the same calculation is used.

7. Go to `http://www.danielsoper.com/statcalc/calc08.aspx`. Where it says "t-Value", enter the value from step 6. Where it says "Degrees of Freedom", enter the total number of data points minus one ($n - 1$). Press "calculate" and record the value multiplied by 100%. To use R, see Subsection A.2.2.

The degrees of freedom are particularly important here. The more degrees of freedom, the less likely it is to get a very positive or very negative value of t. As n approaches 30 and larger, the t and z statistics will have nearly identical values. However, the t-test is always approximate, while the z-test can be considered exact if the data are normal, and the exact mean and exact variance (as opposed to the sample variance we calculate) are somehow known.

4.4.2 Two-Sample Test

The two-sample test, like the one-sample test, is a modification of the earlier large sample tests. However, the modifications are slightly more extensive.

5. After calculating S_x and S_y, we now calculate S_p as follows:

$$S_p = \frac{(n_x - 1)S_x + (n_y - 1)S_y}{n_x + n_y - 2}.$$

6. The statistic, as in the one-sample test, is called t instead of z. It is calculated by

$$t = \frac{\bar{X} - \bar{Y} - D}{\sqrt{S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}.$$

7. Go to http://www.danielsoper.com/statcalc/calc08.aspx. Where it says "t-Value", enter the value from step 6. Where it says "Degrees of Freedom", enter the total number of data points minus two ($n_x + n_y - 2$). Press "calculate" and record the value multiplied by 100%. To use R, see Subsection A.2.2.

4.5 Tests for the Variance

The two variance tests presented here assume normal data, as do the two small sample tests. The two-sample tests, as with the earlier ones, assume that the two populations are independent as well as both normal.

4.5.1 One-Sample Test

1. Form a null hypothesis. In other words, suppose that the variance $\sigma^2 = \sigma_0^2$ for some specific value σ_0^2.

2. Form an alternate hypothesis. This is either that the true mean is either greater than σ_0^2, less than σ_0^2, or simply different from σ_0^2.

3. Choose a significance level, α. Usually 5%.

4. Calculate $\bar{X}$ using the formula in equation 4.1 above.

5. Calculate S using the formula in equation 4.2 above.

6. Calculate the test statistic

$$\chi^2 = \frac{(n-1)S}{\sigma_0^2} \tag{4.4}$$

7. To calculate the p-value for this particular value of χ^2, go to http://www.danielsoper.com/statcalc/calc11.aspx. Enter the value from the previous part where it says "Chi-Square Value". Where it says "Degrees of Freedom", enter $n - 1$, the total number of data points minus one. Multiply the resultant number by 100%. If your hypothesis is that the true variance will be less than the variance from the null hypothesis, use 100% minus the p-value instead of just the p-value. Do the same thing if testing for the variances being simply different and the resultant number is above 50%. In either case, multiply by 2. To use R, see Subsection A.2.3.

8. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 6. *Simon wants to test the variance in the scores of a gifted student's math scores. Data provided by the student's high school shows that the average student over many years and many classes has math scores with a variance of 225 (this is a standard deviation of 15, which is the square root of the variance). The gifted student has the following 40 test scores taken from all of her high school math courses:*

91	91	94	90	96	96	96	94	95	93
87	95	98	86	95	86	91	91	89	95
98	96	96	91	86	91	91	93	96	93
94	91	98	86	97	92	91	94	83	99

Simon calculates the variance in her scores to be 15.1. The chi-square he calculates for the student's scores is 2.3. With the 39 degrees of freedom, the p-value for this is nearly indistinguishable from 0%[2] *such that Simon rejects the null hypothesis, and accepts the alternate hypothesis that the gifted student's scores vary less than that of the average student.*

4.5.2 Two-Sample Test

As with the tests for the mean, let σ_x^2 be the true variance of the first population and let σ_y^2 be the true variance of the second population.

1. There is only one null hypothesis with this test, which is that $\sigma_x^2 = \sigma_y^2$.

2. Form an alternate hypothesis. This is either that $\sigma_x^2 > \sigma_2^2$ or $\sigma_x^2 < \sigma_y^2$.

3. Calculate S_x and S_y using equation 4.2 above.

4. Calculate the test statistic
$$\mathsf{f} = \frac{S_x}{S_y}.$$

5. To calculate the probability of this value f, go to `http://www.danielsoper.com/statcalc/calc07.aspx`. Enter the value of f found above where it says "F-Value", the number of values in the first population minus one ($n_x - 1$) where it says "Numerator Degrees of Freedom", and the number of values in the second population minus one ($n_y - 1$) where it says "Denominator Degrees of Freedom". The test requires that $n_x - 1$ is less than $n_y - 1$. Press "Calculate" and record the value. Multiply the resultant number by 100%. If the test is for the alternate hypothesis that $\sigma_x^2 > \sigma_y^2$, keep the value as is. If the test is for the hypothesis that $\sigma_x^2 < \sigma_y^2$, then subtract the value from 100%. This is the p-value for the statistic f calculated above. To use R, see Subsection A.2.4.

[2]If the value returned is 1 or 0 with no difference after many decimal places, the value is still approximate, but close enough to the exact value that the computer cannot tell the difference.

6. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 7. *Kaylee wants to test how strawberries coming from two different plants vary in mass. The first plant is grown under using the "ideal" soil as provided by a plant nursery. The second is grown from the soil from the garden of Kaylee's high school. Kaylee's null hypothesis is that the strawberries from both plants will have the same variance in mass. Her alternate hypothesis is that the strawberries grown using the school's soil will vary more in size, i.e., that* $\sigma_x^2 < \sigma_y^2$*, and she chooses to test at the 5% level. She picks strawberries three days after she determines them to be ripe.*[3] *She picks 20 from the plant grown in ideal soil and 24 from the plant grown in the school's soil. The plant in the ideal soil produces strawberries with the following masses in grams:*

26.8	24.0	20.4	30.8	23.3	25.8	18.8	22.3	25.6	26.0
23.5	24.3	18.8	17.4	28.2	29.1	16.1	22.9	25.7	24.5

The strawberries grown from the school soil have the masses below:

11.4	20.1	28.3	20.2	25.4	18.6	16.3	11.4	19.4	28.8	18.9	18.4
20.0	16.3	24.0	17.7	11.3	15.1	19.0	24.6	16.3	18.7	6.9	21.4

Kaylee calculates S_x *to be 15.1 and* S_y *to be 28.2 with 19 and 23 degrees of freedom, respectively.* f *is then 0.535. The p-value is then found to be* $100\% - 91.5\% = 8.5\%$*. Since this value is greater than the 5% value of* α*, Kaylee maintains the null hypothesis and rejects the alternate hypothesis. She therefore concludes that the strawberries grown in the ideal soil have the same variance in mass as those grown in the school's soil. Note that the variances are actually quite different, but with the small sample sizes involved, the 5% level is too rigorous to conclude that the variances are different.*

[3] How could she standardize this part of the procedure to make the process less dependent on how she determines strawberries to be ripe?

Chapter 5

Other Types of Data

5.1 Count Data

Count data is a type of data that is often found in tables of percentages or numbers, and often involves categorizing data. Opinion and political polls are often count data.

There are two common tests for this type of data. The first tests whether the percentage of data in any of several categories is different from the predicted percentage. This is called a goodness of fit test, as it tests whether data fits a certain pattern. It is commonly used in genetics to see if Punnett squares are accurately capturing the effects of genes. The second test applies when there are two overall categories each with at least two subcategories, and the test is to see whether the two large categories are independent or not. In other words, whether belonging in a particular subcategory of the first big category affects to which subcategory of the other big category a data point belongs. For example, a social scientist might see if being male or female (the two larger categories) is independent of wealth (divided into several subcategories, perhaps tax brackets).

Both of these tests are best seen through examples, more of which are provided below.

5.1.1 Test for Goodness of Fit

1. Separate the data into k categories.

2. Form a null hypothesis. Namely, for each of the k categories choose the percentage of the sample that should belong in it. In other words $p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$ where p_k is the probability of being in the kth category, and p_{k0} is the predicted probability. Since every data point must be in some category, $p_{10} + p_{20} + \ldots + p_{k0}$ should equal 1.

3. The alternate hypothesis is that at least one of the percentage value is different from its predicted value in the null hypothesis.

4. Choose a significance level, α. Usually 5%.

5. Observe the data. Let O_i be the observed number of data points belonging to the ith category. Note that this is not a percentage.

6. Let $E_i = p_{i0}n$ where n is the total number of data points in all the categories. E_i is the expected number of data points in the ith category.

7. Calculate the test statistic:

$$Q^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots + \frac{(O_k - E_k)^2}{E_k}.$$

8. To calculate the p-value for this particular value of Q^2, go to http://www.danielsoper.com/statcalc/calc11.aspx. Enter the value from the previous part where it says "Chi-Square Value". Where it says "Degrees of Freedom", enter $k-1$, the total number of categories (not data points) minus one. "Press calculate". Multiply the resultant number by 100%.

9. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 8. *Danny wants to test how lonely seniors in his high school feel versus how the lower classmen feel. He passes out a short survey to everyone and asks them to describe how lonely they feel on a scale from 1 to 5, with 5 signifying not at all lonely and 1 meaning very lonely. After surveying students from each of the freshman, sophomore, and junior classes, Danny obtains the following results:*

1	*2*	*3*	*4*	*5*
10%	*6%*	*30%*	*27%*	*27%*

Danny takes these percentages to be his null hypothesis for how the data should look for the senior class. How many people are in this data does not matter, as he assumes these percentages to be true independent of sample size. In other words, he lets $p_{10} = .1$, $p_{20} = .06$, $p_{30} = .3$, $p_{40} = .27$, *and* $p_{50} = .27$. *The alternate hypothesis is that the senior will deviate in at least one of these categories. He wants to test at the* $\alpha = 5\%$ *level. Danny collects data from 150 seniors and finds the following percentages:*

1	*2*	*3*	*4*	*5*
12%	*4%*	*28%*	*24%*	*32%*

(In other words, $O_1 = 18, O_2 = 6, O_3 = 42, O_4 = 36, O_5 = 48$*) He calculates the test statistic to be*

$$Q^2 = \frac{(18 - (.1)(150))^2}{(.1)(150)} + \frac{(6 - (.06)(150))^2}{(.06)(150)} + \frac{(42 - (.3)(150))^2}{(.3)(150)} + \frac{(36 - (.27)(150))^2}{(.27)(150)} + \frac{(48 - (.27)(150))^2}{(.27)(150)} = 3.69.$$

The p-value for this value of Q^2 *with* $5 - 1 = 4$ *degrees of freedom is 45% such that Danny rejects the null hypothesis, and accepts that the seniors differ in how they rate their loneliness from the other students.*

5.1.2 Test for Independence

1. Divide the data into a table. Let r be the number of rows and let c be the number of columns.

2. The null hypothesis is that the two large categories are independent.

3. The alternate hypothesis is that the two large categories are dependent.

4. Choose a significance level, α. Usually 5%.

5. Observe the data. Let O_{ij} be the number of data points in row i and column j of the table. For example, O_{23} would be the number of observations in the second row and third column. Let N be the total number of observations.

6. Let m_i be the number of observations in the ith row and let n_j be the number of observations in the jth column. Calculate $E_{ij} = \frac{m_i n_j}{N}$ for each entry of the data table.

7. Calculate $Q_{ij}^2 = \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ for every entry in the table.

8. The test statistic, Q^2 is the sum of the Q_{ij}^2 for each entry in the table.

9. To calculate the p-value for this particular value of Q^2, go to http://www.danielsoper.com/statcalc/calc11.aspx. Enter the value from the previous part where it says "Chi-Square Value". Where it says "Degrees of Freedom", enter $(r-1)(c-1)$, the number of rows minus 1 multiplied by the number of columns minus 1. "Press calculate". Multiply the resultant number by 100%.

10. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 9. *Deanna wants to see if the number of Advanced Placement (AP) classes taken is independent of gender. She divides the number of classes into three categories: 0-2, 3-5, and 6+. Her null hypothesis is that gender and number of AP classes is independent while the alternate hypothesis is that gender and number of AP classes taken is not independent. She decides to test at the 1% level for an extra rigorous test, even though 5% is more common. She finds the following data for the previous year's graduated seniors:*

	0-2	*3-5*	*6+*	*Row Total*
Female	*115*	*30*	*5*	*150*
Male	*121*	*23*	*3*	*147*
Column Total	*237*	*53*	*8*	*297*

Therefore $O_{11} = 115$, $O_{12} = 30$, $O_{13} = 5$, $O_{21} = 121$, $O_{22} = 23$, $O_{23} = 3$. *Also, the row totals are* $m_1 = 150$ *and* $m_2 = 147$ *while the column totals are* $n_1 = 237$, $n_2 = 53$, *and* $n_3 = 8$. *Finally, the total number of students, N is 297. Deanna then calculates that*

$$E_{11} = \frac{(150)(237)}{297} = 119.7 \quad E_{12} = \frac{(150)(53)}{297} = 26.8 \quad E_{13} = \frac{(150)(8)}{297} = 4.0$$
$$E_{21} = \frac{(147)(237)}{297} = 117.3 \quad E_{22} = \frac{(147)(53)}{297} = 26.2 \quad E_{23} = \frac{(147)(8)}{297} = 4.0$$

Deanna, still calculating, now sees that

$$
\begin{array}{lll}
Q_{11}^2 = \frac{(115-119.7)^2}{119.7} = 0.185 & Q_{11}^2 = \frac{(30-26.8)^2}{26.8} = 0.382 & Q_{11}^2 = \frac{(5-4.0)^2}{4.0} = 0.250 \\
Q_{11}^2 = \frac{(121-117.3)^2}{117.3} = 0.117 & Q_{11}^2 = \frac{(23-26.2)^2}{26.2} = 0.391 & Q_{11}^2 = \frac{(3-4.0)^2}{4.0} = 0.250
\end{array}
$$

Adding these values up, she finds that the test statistic is $Q^2 = 1.575$. *This table has* $(2-1)(3-1) = 2$ *degrees of freedom. The p-value Deanna finds is 45% such that she accepts the null hypothesis that gender and number of AP classes taken are independent of each other.*

5.2 Testing for Correlation

A common question in science is whether two variables are correlated. In other words, if one of the variables changes, is it likely that there was a change in the other variable. A couple examples of sets of variables that might be correlated are SAT scores and collegiate GPA, hours of study and success on a test, and hours of Star Trek watched per week and ability to quote Mr. Spock.

5.2.1 Correlation versus Causation

Statisticians' favorite phrase might be *"Correlation does not imply causation."* Causation means that if one event happens, a second event either will or will not happen as a direct influence of the other event. Correlation merely implies that if the first event occurs, then the second event is either more likely or less likely to occur. For example, eating ice cream and drowning on the same day are correlated events. However, avoiding ice cream will not help anyone avoid drowning since eating ice cream does not cause drowning. The correlation is there because eating ice cream is an action usually taken on warm days. Swimming is also most frequently done on warm days, such that eating ice cream and swimming, and hence drowning, are correlated events.

5.2.2 The Correlation Coefficient

When determining the degree of correlation between events, the data must be ordered pairs: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Given this data, a correlation coefficient[1] can be calculated. The formula is given by

$$
r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2\right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2\right]}}. \tag{5.1}
$$

If you wish to use R to calculate the correlation coefficient, see Subsection A.2.5. This formula is huge and intimidating, but broken up into steps is not difficult to calculate. Follow along with the example below to calculate it with several steps and without too much difficulty.

[1] If you have not seen summation notation, $\sum_{i=1}^n x_i = x_1 + x_2 + \ldots x_n$, $\sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n$, $\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \ldots + x_n^2$, etc.

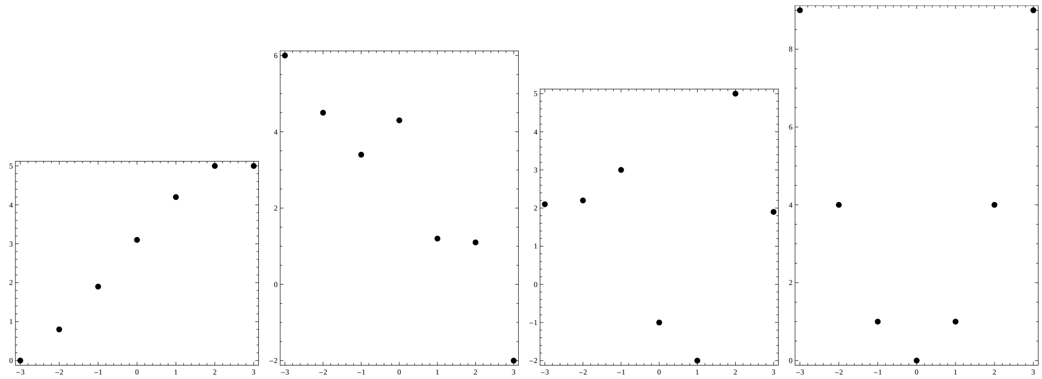

Figure 5.1: The correlation between x and y in the graphs, from left to right: $0.98, -0.93, 0.17, 0$.

Though the formula for calculating the correlation coefficient r is complicated, r is always between -1 and 1. If $r = 0$, then x and y are not correlated, (which only says the variables are not related in a linear fashion). If r is close to 1, the variables are positively correlated. This means that as the values of x increase, so do the values of y. If r is close to -1, then the variables are inversely correlated; as x increases, y decreases in value. The plots in figure 5.1 show data with different correlation coefficients. Note that in the graph furthest to the right, the points actually lie on the parabola $y = x^2$, yet the correlation is 0. The absence of correlation does not imply that there is no relationship between the x and y variables, just that the relationship is not linear, i.e., that the value do not increase together or one increases while the other decreases.

5.2.3 Simple Linear Regression

Since correlation describes the linear relationship between two variables, it is natural to try to fit a line that describes the relationship between correlated variables (plus some error). This line can then be used to predict the y value from an x value. The process of finding this line is called linear regression. When the relationship is only between one predicted variable (y) and one independent variable (x), it is called simple linear regression. Linear regression finds the line that minimizes the distance between the line and each of the observed ordered pairs.

Given data points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), we want a linear equation of the form $y = mx + b$, where m is the slope and b is the y-intercept. To calculate the coefficients, follow these steps:

1. Calculate $\bar{X}$ and $\bar{Y}$, the averages of the x and y values.
2. Calculate $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{X})^2$.
3. Calculate
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}).$$

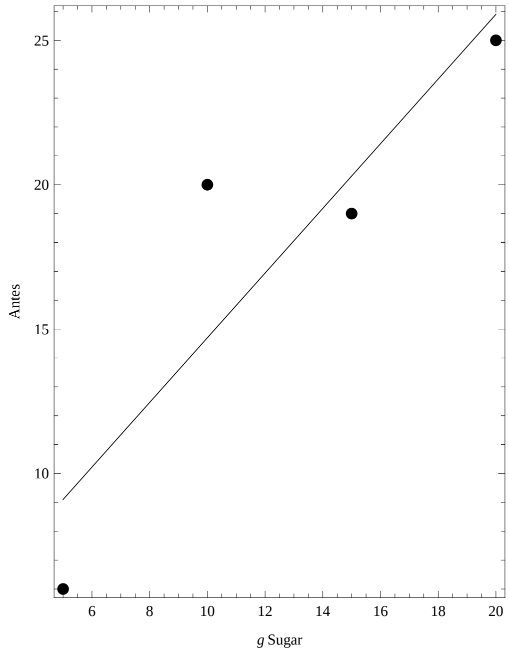

Figure 5.2: Plot of Pablo's data with best fit line.

An alternate but equivalent formula for S_{xy} is

$$S_{xy} = \sum_{i=1}^{n} x_i y_1 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right).$$

4. Finally, $m = \frac{S_{xy}}{S_{xx}}$ and $b = \bar{Y} - m\bar{X}$, and $y = mx + b$ is the linear model that is closest to describing the relationship between the observed x and y variables.

Microsoft Excel and most graphing calculators are able to calculate best fit lines.

Example 10. *Pablo wants to see how ants respond to different amounts of sugar. He finds an anthill and places 5, 10, 15, and 20 grams of sugar equally distant from the anthill in different directions from it. He watches for an hour and counts the number of ants that show up. He finds that 6 ants come to the 5 grams of sugar, 20 to the 10 grams of sugar, 19 to the 15, and 25 to the 20 grams.*

He calculates the correlation first:

$$\sum_{i=1}^{4} x_i y_i = 5 \cdot 6 + 10 \cdot 20 + 15 \cdot 12 + 20 \cdot 25 = 910$$
$$\sum_{i=1}^{n} x_1 = 5 + 10 + 15 + 20 = 50$$
$$\sum_{i=1}^{n} y_1 = 6 + 20 + 19 + 25 = 70$$
$$\sum_{i=1}^{n} x_i^2 = 25 + 100 + 225 + 400 = 750$$
$$\sum_{i=1}^{n} y_i^2 = 36 + 400 + 361 + 625 = 1422$$
$$r = \frac{4 \cdot 1015 - 50 \cdot 70}{\sqrt{(4 \cdot 750 - 50^2)(4 \cdot 1422 - 70^2)}} = 0.892$$

The correlation is positive, as Pablo expected, and relatively strong. Now, Pablo wants to calculate a best fit line for the data and graph it. He finds that

$$\bar{X} = \frac{5 + 10 + 15 + 20}{4} = 12.5$$
$$S_{xx} = (5 - 12.5)^2 + (10 - 12.5)^2 + (15 - 12.5)^2 + (20 - 12.5)^2 = 125$$
$$\bar{Y} = \frac{6 + 20 + 19 + 25}{4} = 17.5$$
$$S_{xy} = (5 - 12.5)(6 - 17.5) + (10 - 12.5)(20 - 17.5)$$
$$+ (15 - 12.5)(19 - 17.5) + (20 - 12.5)(25 - 17.5) = 140$$
$$m = \frac{140}{125} = 1.12$$
$$b = 17.5 - 1.12 \cdot 12.5 = 3.5$$

Therefore the best fit line is $y = 1.12x + 3.5$, where y is the number of ants and x is the amount of sugar in grams. A graph of the data and the best fit line is included in figure 5.2 on the next page.

Chapter 6

Testing non-Normal Data

6.1 What is Normal?

Many of the statistical tests considered in this booklet require either that the data be normal or that there be enough data to for certain tests to be good approximations. So far, we have neglected to cover when data can be considered normal.

Data which are normal follow what is often called a "bell-curve", which is pictured in figure 6.1. The bell curve was first described by Carl Friedrich Gauss (1777-1855), and is sometimes called a "Gaussian" curve in his honor. But what does this bell curve mean? The height of the curve describes the probability of getting particular values. However, it is not quite that straightforward. The bell curve runs the entire length of the x-axis, from $-\infty$ to ∞, and is always positive. If you pick two values, say 1 and 2, and look at the curve just between these values, the area between the curve and the x-axis is the probability that a random value from normal data is between 1 and 2. Since probabilities are always between 0 and 1, and the probability of getting any value at all is 1, the area underneath the entire curve, from $-\infty$ to ∞ is 1. The curve can be thought of like a continuous histogram. In fact, the histogram coming from normal data should look like a bell curve.

This type of data is incredibly important, as it can be used very well for approximations of certain characteristics of non-normal data, as was seen in testing for the mean of large samples in Section 4.3. Also, normality is common, and thus the assumption that data are normal is not usually a bad one.

6.2 Testing Data for Normality

There are a variety of ways to test if data follow a normal distribution that range from simple, visual tests to more rigorous statistical analyses. We will stick to visual tests, since they are simpler and most of the statistical tests only require the data to be approximately normal. These tests are best used to find data that are highly non-normal.

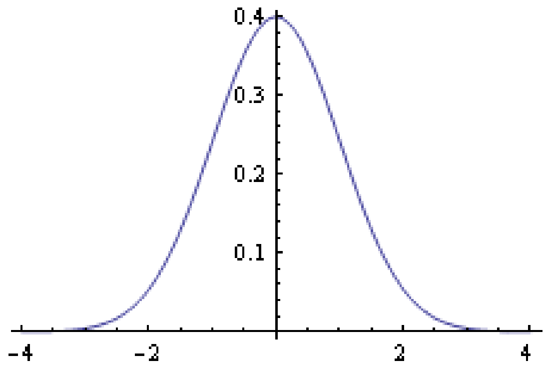

Figure 6.1: Graph of the *Probability Density Function* of the Normal or Gaussian Distribution. Also called a "bell-curve".

6.2.1 Histograms

The simplest way to test for normality is to make a histogram. The process of making a histogram is as follows:

1. Find the range in your data, i.e., look for the difference between the maximum and minimum values.

2. Divide the range into any number of evenly spaced intervals, usually 5 or 10 will suffice.

3. Count the number of data points that belong in each interval.

4. Make a bar graph for the number of data points in the intervals. The intervals should be on the x-axis, and the bars' length depends on the number of values that fall in that interval. To do this with R, see Subsection A.2.6

5. Visually examine the histogram to see if it looks like a bell curve.

Though not a very exact test, it is simple and can easily find data that are not normal. However, it is not very useful for very small numbers of data points, since one or two outliers can highly change the shape of the histogram. But with even 20 data points, the test can be effective.

6.2.2 Normal-Score Plots

A second test is to make a *normal-score plot.* This test is somewhat more rigorous than the histogram, and is, accordingly, slightly more difficult to perform.

1. Label the n data values as $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, where $x_{(1)}$ is the lowest value and $x_{(n)}$ is the largest.

2. We now need to find the n values that divide the x-axis underneath the standard normal bell curve into $n+1$ regions of equal probability (or area under the curve). To do this, we can look for the values of x where the area from $-\infty$ to x is $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$. This

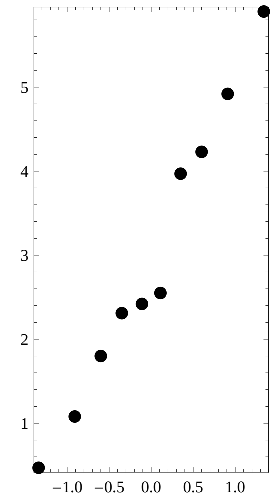

Figure 6.2: Normal score plot for Samantha's data.

can be done at http://sampson.byu.edu/courses/z2p2z-calculator.html. Go to the bottom under the heading "Calculate z from cumulative probability p". Type in each of $\frac{1}{n+1}, \ldots, \frac{n}{n+1}$ where it says "Given cumulative probability p" and record the output values as $z_1, z_2, \ldots, z_n$.

3. Plot on a graph the points $(z_1, x_{(1)}), (z_2, x_{(2)}), \ldots, (z_n, x_{(n)})$.

4. If the data are approximately in a straight line, then the data are approximately normal. To determine how linear the progression is, you can calculate the correlation coefficient r, as described in Section 5.2. The closer it is to 1, the closer the data are to being normal.

Example 11. *Samantha wants to wants to see if her small data sample is approximately normal. She has 10 data points which, in ascending order, are*

$$\begin{array}{lllll} x_{(1)} = 0.47 & x_{(2)} = 1.08 & x_{(3)} = 1.80 & x_{(4)} = 2.31 & x_{(5)} = 2.42 \\ x_{(6)} = 2.55 & x_{(7)} = 3.97 & x_{(8)} = 4.23 & x_{(9)} = 4.93 & x_{(10)} = 5.90 \end{array}$$

She needs to calculate 10 normal scores, for the probabilities for $1/11, 2/11, \ldots, 10/11$. *The normal scores turn out to be*

$$\begin{array}{lllll} z_{(1)} = -1.34 & z_{(2)} = -0.91 & z_{(3)} = -0.60 & z_{(4)} = -0.35 & z_{(5)} = 0.11 \\ z_{(6)} = 0.11 & z_{(7)} = 0.35 & z_{(8)} = 0.60 & z_{(9)} = 0.91 & z_{(10)} = 1.34 \end{array}$$

The ordered pairs that Samantha needs to plot are $(-1.34, 0.47)$, $(-0.91, 1.08)$, $(-0.60, 1.80)$, $(-0.35, 2.31)$, $(-0.11, 2.42)$, $(0.11, 2.55)$, $(0.35, 3.97)$, $(0.60, 4.23)$ $(0.91, 4.937)$ $(1.34, 5.90)$. *Plotting these points shows that they do seem to be mostly in a line, as seen in figure 6.2. Calculating the correlation coefficient between the x values and their corresponding normal scores shows the correlation to be an incredibly high 0.987, such that Samantha concludes that her data are at least approximately normal.*

6.3 Tests for non-Normal Data

We need a new test for non-Normal data. If the probability distribution of the data were known, as the normal probability distribution is known or at least approximated in the other tests, then we could devise tests that were exact. However, as there are many distributions, often with incredibly complex formulae, this is impractical. However, we can devise tests that do not assume that the data come from any type of probability distribution. To do this, we first need to introduce one new statistic.

6.3.1 The Median

Instead of the mean, which can only be properly estimated or tested for if the probability distribution is known, the median can be predicted without knowing the probability distribution. This is because the mean, as well as being the average value, is the expected value of numbers drawn randomly from a probability distribution.

The mean, however, is the "middle" of the distribution, in that it is the point where 50% of the values are less than or equal to it, and 50% of the values are greater than or equal to it. Though this appears to have the same problem as the mean, it does not. More than 50% of values can be less than the mean. For example, the mean of $1, 1, 1, 2, 3$, and 70 is 13, and only one of the six values is greater than the mean and the other five are less than the mean. However, 50% of the values will always be less than or equal to the median and 50% of the values will always be greater than or equal to the median. For the values above, the median is 1.5, the average of 1 and 2, since there are an even number of values and we require that 3 values be on both sides of the median.

To calculate the median, find the value such that half of the values are greater than the median and half are fewer. If there are an even number of values, the median is the average of the $(n/2 - 1)$th and $(n/2 + 1)$th values, as above.

6.3.2 Sign Test for One Sample

Before using this test, the critical assumption is that the number of data points n is greater than 10. This test is only an approximation and the approximation is not good with fewer values.

1. Form a null hypothesis. In other words, suppose that the true median of the data is m_0 for some value m_0. This value should not be the median value of the given data, or the test will always confirm the null hypothesis as true.

2. Form an alternate hypothesis. The three possible null hypotheses are that the true median is greater than m_0, less than m_0, or simply not equal to m_0.

3. Choose a significance level α. Usually 5%.

4. Let n^+ be the number of data points that are greater than m_0. If any data points are equal to m_0, remove them from the data set and reduce the size of n accordingly. In order for the approximation to remain valid, n must still be greater than 10.

5. Calculate the test statistic
$$z = \frac{2n^+ - n}{\sqrt{n}}.$$

6. To calculate the p-value for this particular value of z, go to http://www.changbioscience.com/stat/ztest.html. If your test is that the true mean is simply different than m_0, choose "Two-tailed p-value" from the drop-down menu above the "calculate p" button. Otherwise, leave it at "One-tailed p-value". Enter the value of z you calculated in the previous step in the field to the right of "z=(y-mu)/(sigma/sqrt(n))". Press the "Calculate p" button. Multiply the resultant number by 100%. To use R, see Subsection A.2.1

7. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 12. *Rory wants to test the effect of high pitch sound on concentration. He found a previous project involving a memory test. It involved attempting to memorize a list of 50 words in 5 minutes. The data from that experiment showed a median of 22 words memorized with an average of 21.5 words. He performs the same experiment with 20 new participants, but while they are trying to memorize the words, he uses a sonic device to play a high pitched buzz for 10 seconds at the end of every minute. He finds the following data:*

37	18	20	16	20	15	20	13	32	17
36	30	23	31	11	29	19	13	25	21

After looking at a histogram of the data, Rory decides that the data are not approximately normal. He decides to use the median test instead of testing for the mean. His null hypothesis is that the true median of the data is 22 versus the alternate hypothesis that the students who heard the high pitched noise would memorize fewer words, and thus have a lower median. He tests at the $\alpha = 10\%$ level.

Since none of the participants memorized exactly 22 words, David does not remove any of the values from his observed data. 8 of the 20 data points are greater than 22 such that $n^+ = 8$. The test statistic z is then $\frac{2\cdot 8-20}{\sqrt{20}} = -0.89$. The p-value for this value of z is approximately 18.7%. Since Rory was testing at the 10% level, he cannot reject the null hypothesis that the students subjected to the high pitched noise had a more difficult time memorizing words than the students from the original study.

6.3.3 Median Test for Two Sets of Data

As with the other tests for two sets of data, the data must be independent. In other words, if the data sets are before and after, the sets are not independent of each other. (In this case subtract one from the other and use the one-sample test above.) The other assumption in this test is that both n_1, the number of data points in the first set, and n_2, the number of data points in the second set, are both greater than 5.

1. The null hypothesis is that $m_1 = m_2$, where m_1 is the true median of the first sample and m_2 is the true median of the second sample.

2. Form an alternate hypothesis. The three possible alternate hypotheses are that $m_1 < m_2$, $m_1 > m_2$, and that $m_1 \neq m_2$.

3. Choose a significance level α. Usually 5%.

4. Let m be the median of all of the data points. Remove any values that are equal to m. Lower n_1 and n_2 if necessary after removing values. Let $n = n_1 + n_2$ be the total number of remaining values. (n_1 and n_2 should both still be greater than 5.)

5. Let N_a be the number of the n values that are larger than m, and let N_b be the number of of the values less than m

6. Let N_{1a} be the number of values from the first set that are above m, and let N_{1b} be the number of values from the first set that are below m. Define N_{2a} and N_{2b} similarly for the second set of data.

7. Calculate $E = \frac{N_a n_1}{n}$ and $S = \frac{N_a n_1 n_2 N_b}{n^2(n-1)}$.

8. Calculate the test statistic
$$z = \frac{N_{1a} - E}{\sqrt{S}}.$$

9. To calculate the p-value for this particular value of z, go to http://www.changbioscience.com/stat/ztest.html. If your test is that the true medians are simply different, choose "Two-tailed p-value" from the drop-down menu above the "calculate p" button. Otherwise, leave it at "One-tailed p-value". Enter the value of z you calculated in the previous step in the field to the right of "z=(y-mu)/(sigma/sqrt(n))". Press the "Calculate p" button. Multiply the resultant number by 100%. To use R, see Subsection A.2.1

10. If the p-value is less than α, reject the null hypothesis and accept the alternate hypothesis. If the p-value is greater than α, you must continue to accept the null hypothesis.

Example 13. *Karen devises a test as follows: She makes 30 flash cards with color words on them, such as orange, green, and blue. Each of the words is colored a different color from what the words says. For example,* BLUE, GREEN, PINK. *She gives the words to two groups of ten people. The first group consists of third graders, and the second group is high school freshman. The goal is to read every word correctly, with only one attempt given per word and no more than 5 seconds can be spent on one word. The data from the first group is*

7 9 9 11 17 17 17 18 19 20

The freshman recorded the scores

9 9 15 16 16 23 23 23 25 25

Neither group of data appears normal and there are not enough points to use the large number test, so Karen decides to test for the median instead of the mean. Her null hypothesis is that both groups have the same median, and her alternate hypothesis is that the freshman group *will have a higher median. She tests at the* 5% *level. The median of the data is 17, so she throws these three values out of the first group (which still leaves it with 7 points). Of the remaining data points, Karen finds the following:* $N_a = 8$, $N_b = 9$, $N_{1a} = 4$, $N_{1b} = 3$, $N_{2a} = 5$, *and* $N_{2b} = 5$. *She calculates that* $E = \frac{8 \cdot 7}{17} = 3.29$ *and* $S = \frac{8 \cdot 7 \cdot 10 \cdot 9}{17^2 \cdot 16} = 1.09$. *The test statistic is then* $z = \frac{4-3.29}{\sqrt{1.09}} = 0.68$. *The p-value is 0.25. This is higher than the* α *value of* 5% *such that Karen accepts the null hypothesis that the median third grader and median freshman can read the same number of words correctly in the test.*

Appendix A

Installing and Using R

As the provided websites could be unavailable when needed or even removed permanently, each section has instructions on how to perform the test using the statistical package R. The package is free, and fairly powerful. This appendix includes both installation instructions and references to the proper commands for running the tests in this booklet.

A.1 Installation

1. Go to `http://www.r-project.org/`
2. On the left side of the page, under "Download, Packages", click on the link marked CRAN.
3. Click on a link for any of the mirror sites (I used Case Western Reserve University)
4. Under "Download and Install R", click on the appropriate link based on your operating system, and follow the provided instructions.

A.2 Instructions for Tests

A.2.1 Z Tests

1. If the alternate hypothesis is that the true value is greater than what is assumed in the null hypothesis (z is positive), type `1-pnorm(z)`, where `z` is the value found for the test, and hit enter. Multiply by 100% to get the p-value.
2. If the alternate hypothesis is that the true value is less than what is assumed in the null hypothesis (z is negative), type `pnorm(z)`, where `z` is the value found for the test, and hit enter. Multiply by 100% to get the p-value.
3. If testing only for difference between values, follows the instructions in 1 if $z < 0$; follow the instructions in 2 if $z > 0$. Multiply the resultant value by 2. Multiply by 100% to get the p-value.

Links:

- Test for Percentage Likelihood - One Set of Data: Subsection 3.3.1
- Test for Percentage Likelihood - Two Sets of Data: Subsection 3.3.2
- Large Sample Test for Mean - One Set of Data: Subsection 4.3.1
- Large Sample Test for Mean - Two Sets of Data: Subsection 4.3.2
- Test for the Median - One Set of Data: Subsection 6.3.2
- Test for the Median - Two Sets of Data: Subsection 6.3.3

A.2.2 T Tests

1. If the alternate hypothesis is that the true value is greater than what is assumed in the null hypothesis (t is positive), type `1-pt(t,df)`, where `t` is the value found for the test and `df` is the degrees of freedom, and hit enter. Multiply by 100% to get the p-value.

2. If the alternate hypothesis is that the true value is less than what is assumed in the null hypothesis (z is positive), type `pt(t,df)`, where `t` is the value found for the test and `df` is the degrees of freedom, and hit enter. Multiply by 100% to get the p-value.

3. If testing only for difference between values, follows the instructions in 1 if $t < 0$; follow the instructions in 2 if $t > 0$. Multiply the resultant value by 2. Multiply by 100% to get the p-value.

Links:

- Small Sample Test for Mean - One Set of Data: Subsection 4.4.1
- Small Sample Test for Mean - Two Sets of Data: Subsection 4.4.2

A.2.3 Chi-Squared Tests

1. If the alternate hypothesis is that the true value is greater than what is assumed in the null hypothesis, type `1-pchisq(chi2,df)`, where `chi2` is the value found for the test and `df` is the degrees of freedom, and hit enter. Multiply by 100% to get the p-value.

2. If the alternate hypothesis is that the true value is less than what is assumed in the null hypothesis, type `pchisq(chi2,df)`, where `chi2` is the value found for the test and `df` is the degrees of freedom, and hit enter. Multiply by 100% to get the p-value.

3. If testing only for difference between values, follows the instructions in 1 if $\chi^2 >> 0$; follow the instructions in 2 if $\chi^2 \approx 0$. Multiply the resultant value by 2. Multiply by 100% to get the p-value.

Links:

- Test for Variance - One Set of Data: Subsection 4.5.1
- Test for Goodness of Fit: Subsection 5.1.1
- Test for Independence: Subsection 5.1.2

A.2.4 F Tests

1. If the alternate hypothesis is that $\sigma_x^2 < \sigma_y^2$ (you should have $S_x < s_y$), type `1-pf(f,dfx,dfy)`, where `f` is the value found for the test, `dfx` is the degrees of freedom for the X population, and `dfy` is the degrees of freedom for the y populion, and hit enter. Multiply by 100% to get the p-value.

2. If the alternate hypothesis is that $\sigma_x^2 > \sigma_y^2$ (you should have $S_x > s_y$), type `pf(f,dfx,dfy)`, where `f` is the value found for the test, `dfx` is the degrees of freedom for the X population and, `dfy` is the degrees of freedom for the y populion, and hit enter. Multiply by 100% to get the p-value.

Links:

- Test for Variance - Two Sets of Data: Subsection 4.5.2

A.2.5 Correlation

1. To enter the data, type `x<-c(x1,x2,x3,...,xn)`, where `x1,x2,...,xn` are the n data points, and hit enter.

2. To enter the data, type `y<-c(y1,y2,y3,...,xn)`, where `y1,y2,...,yn` are the n data points, and hit enter.

3. Type `cor(x,y)` and hit enter to find the correlation between `x` and `y`, where `x1` is paired with `y1`, `x2` is paired with `y2`, etc.

Links:

- Correlation: Subsection 5.2.2

A.2.6 Histograms

1. To enter the data, type `x<-c(x1,x2,x3,...,xn)`, where `x1,x2,...,xn` are the n data points, and hit enter.

2. Type `hist(x)` and hit enter. The histogram should display on the right side of the screen.

3. For more options with labeling the histogram, typing `help(hist)` should bring up a webpage with all the options for the histogram explained, though somewhat densely.

Links:

- Histograms: Subsection 6.2.1

Bibliography

[1] CT Tsokos KM Ramachandran, *Mathematical Statistics with Applications*, Academic Press, 2009.

[2] Sheldon Ross, *A First Course in Probability*, eighth ed., Prentice Hall, 2009.