Math 316 Summary of Ideas

Section 1.1 gives us an idea of what statistics is all about: making an inference (estimate) about a population—a group of people, a certain type of plant, etc. We make that inference based on information we gather—often the measurement of a certain variable—from a sample from that population. One important issue is what sort of information and evidence we want to gather and how to gather it. This is discussed in Section 1.2. Section 1.3 discusses a very important (and often very difficult) aspect of gathering the information: how to get a truly random sample which accurately represents the population from which the sample is taken.

There is always uncertainty in statistics. The only way to know with *100% certainty* some information about an entire population is to gather that information from the *entire* population. Thus no matter what inferences or conclusions we do or do not make about a population, there is always the possibility that we are making a mistake. Connected to uncertainty is probability, the study of how likely something is to happen. Section 2.1 gives a nice summary of and introduction to what probability is. In Section 2.2 we are introduced to distributions. A frequency distribution consists of the possible outcomes and the frequency with which each outcome occurred. Generally, more useful than frequency is relative frequency: the fraction of the time each outcome *does* occur. (In Chapter 3 we see the related idea of probability distribution: the possible outcomes and what fraction of the time each outcome—or a particular range of outcomes—*should* occur. In Section 2.3 we read about some of the possible distributions—some examples are shown in Figure 2.2.14 on page 36—and we are introduced to the idea that area under the curve = probability (or relative frequency). In Section 2.3 median and mean are discussed. In Section 2.4 we learn about boxplots. Section 2.5 covers some miscellaneous ideas, including scatterplots, and numerical vs. categorical data. In Section 2.6 we learn of measures of dispersion (how spread out a collection of data is, i.e. how much variation there is in the data), in particular, variance and standard deviation. Some useful intuition: standard (typical, average) deviation (how different something is) is essentially the average distance from each of the values in the data to the mean of the data. Regardless of the distribution, the majority of all data is within 1 standard deviation of the mean, most of all data is within 2 standard deviations, and nearly all data is within 3 standard deviations of the mean. In Figure 4.3.5 on page 127 we see more precise percentages for normally distributed data. In Section 2.7 we see that if we have a collection of Y values where $Y = aX + b$, we will have $\mu_Y = a\mu_X + b$ and $\sigma_Y = a\sigma_X$. Other transformations are discussed in Section 2.7. These aren't too important to what we do in this class, but they can occasionally be important in some real life situations. Finally, in Sections 2.8 and 2.9 we get back to what statistics is all about: making inferences about one or more

populations based on information we get from samples. Notation is helpful to do mathematical analysis as well as to keep track of what we are talking about. For instance, see Table 2.9.1 on page 78.

In Chapter 3 we learn about <u>probability</u>. In Chapter 3.2 is the simple idea that the probability of an event occurring is simply how many times should that event occur out of the total number of possible outcomes. If there is no theoretical prediction of the probability, the next best thing to predict what should happen is simply what has happened in the past; that is, the probability of an event happening in the future is what fraction of the time it has occurred in the past. An important notion of probability is that the more you repeat an experiment, the closer the actual relative frequency will be to the probability (the expected relative frequency) of that event. We see this in Figure 3.2.1 on page 87. <u>Probability trees</u> are quite handy for visualizing and organizing the possible outcomes and determining probabilities. Example 3.2.11 on page 91 is simple, yet quite interesting and important. In Section 3.3, some basic rules of probability are given and <u>conditional probability</u> is introduced:

$$\Pr\{E_2|E_1\} = \frac{\#\ of\ ways\ both\ E_1\ and\ E_2\ can\ occur}{\#\ of\ ways\ \ E_1\ \ can\ occur} = \frac{\frac{\#\ of\ ways\ both\ E_1\ and\ E_2\ can\ occur}{total\ \#\ of\ possible\ outcomes}}{\frac{\#\ of\ ways\ \ E_1\ \ can\ occur}{total\ \#\ of\ possible\ outcomes}} = \frac{\Pr\{E_1 and\ E_2\}}{\Pr\{E_1\}}.$$

This is the formula the book gives on page 96. In Section 3.4 we learn of <u>density curves</u>, a function that shows what the possible outcomes are and the relative frequencies of those outcomes. See Example 3.4.2 on page 100. Section 3.5: a <u>random variable</u> is simply a value that varies (variable) and that is random (it varies from measurement to measurement). Notation: typically the book uses a capital letter when talking about the random variable and a lower case letter when talking about a specific value that variable might have. For example, $\Pr\{Y = y_2\} = \frac{1}{3}$ means that there is a probability of 1/3 that the value of variable Y will be y_2. Related to this: often a Greek letter is used for the population and an English letter is used for the sample. For example, the *population* standard deviation is σ (the Greek letter for 's') and the *sample* standard deviation is s. We end the chapter in Section 3.6 with the binomial distribution in which there are two possible outcomes (a few examples: a free throw is either made or not; a child is either iron deficient or not; an animal is either male or not). The formula corresponding to the probability of a particular outcome is on page 110, where p is the probability of "success" for an individual trial:

$$\Pr\{j\ successes\ of\ n\ trials\} = {}_nC_j \cdot p^j(1-p)^{n-j}\ .$$

Note that $p^j(1-p)^{n-j}$ is the probability of having j successes and $n-j$ failures in some *specific* order, and there are ${}_nC_j$ different ways (different orderings) of an outcome having j successes and $n-j$ failures, so multiplying $p^j(1-p)^{n-j}$ by ${}_nC_j$ is of course

the same as simply adding $p^j(1-p)^{n-j}$ up ${}_nC_j$ times, since there are ${}_nC_j$ different ways of having j successes and $n-j$ failures. In 3.7 they discuss estimating the probability p of success from the given data.

In Chapter 4 we learn about the normal distribution. *Normal* is another word for *typical* or *common*. Section 4.1 introduces the main idea of normal distribution. Many real life collections of data are approximately normal. There are two parameters that define a particular normal distribution; that is, there are two parameters that make one normal distribution different from another: the mean (where the data are centered) and the standard deviation (how spread out the data are). See Figure 4.2.2 on page 125 for a comparison of three normal distributions. Quite important is the idea of a standard (or Z) value or score, introduced in Section 4.3. This type of value comes up over and over throughout statistics: how different is our measurement from what we expected to get, relative to how much variation there is in the data. For example, with

$$Z = \frac{Y-\mu}{\sigma}$$

we compute how different a value Y is from what (on average) is expected μ, relative to how spread out the rest of the data is, as measured by standard variation σ. Another way to look at this comes from solving for Y, which leads to $Y = \mu + Z \cdot \sigma$. That is, Z is how many standard deviations the given value Y is from the mean μ. For normally distributed data, Table 3 tells us what fraction of the data a particular value is greater than. For example, a value with a Z score of 1.57 is larger than .9418 (94.18%) of the other data. (Again, this is assuming that the data are perfectly normally distributed.) We can also pretty easily compute what fraction of the data is between two Z values. See, for instance, Figure 4.3.9 on page 129. In Section 4.4 we learn how to determine whether data are normally distributed (or more accurately, how normally distributed the data are). We basically are looking at whether the data are spread out in the way that normally distributed data should be.

Chapter 5 is all about samples. First, Section 5.1 points out something obvious, but important to realize: every sample from a particular population will be different from (even if similar to) every other sample. The sampling distribution for a given population consists of the possible sample means and what fraction of the time those sample means would occur. In other words, if we could take lots and lots (infinitely many) samples from the same population (that is, if we carried out a meta study), the sampling distribution would be the collection (distribution) of all of the sample means. As discussed in Section 5.2, we would expect that most of those samples would have a mean approximately equal to the population mean, but there would still be some with mean somewhat different

from the population mean. In Figure 5.2.1 we see an example of a sampling distribution of the sample mean $\bar{Y}$. Theorem 5.2.1 describes how the statistics (the mean and standard deviation) of the sample means relate to the statistics of the original population. Figure 5.2.6 gives a nice example of how given a certain population (shown in (a)) what a particular sample of size $n = 25$ might look like—the values are in Table 5.2.3. In Figure 5.2.7 are several other possible $n = 25$ samples from the same population, each with its own mean. Notice each sample is like a miniature version of the original population: same basic shape (in this case, normally distributed), and approximately the same mean and standard deviation. If we were to take lots and lots (infinitely many) of $n = 25$ samples then the distribution of those samples means would be as given in Figure 5.2.6(c). Notice the mean in (c) is the same as the mean in (a) and the standard deviation in (c) is the standard deviation in (a) divided by $\sqrt{25}$, where 25 is the sample size. These values are summarized in Table 5.2.2 on page 156. Let me remind you that most samples would have a mean near the population mean of (500 for this example), but occasionally there can be an unusual sample, like that in Figure 5.2.7(c). One of the big questions is: given a single sample (we usually take only a single sample from a population), how likely is it that the population is what we thought it would be? (For example, often we are interested in the population mean.) For instance, consider any one the eight samples in Figure 5.2.7. How likely are we to believe that the population mean is 500 given that one particular sample? The Central Limit Theorem in Section 5.3 says that no matter what the distribution of the original population, the sampling distribution will become more and more normal the larger the sample is. This is actually the main reason that the normal distribution is so important: it is not because many populations themselves are normally distributed; it's because their sample distributions would be approximately normal. The population whose distribution is given in Figure 5.3.3 is not normal. Suppose we took lots and lots (infinity many) samples of size $n = 4$. The resulting distribution of the sample means is the first plot in Figure 5.3.4. Suppose that we did this again, but with larger sample sizes. The distributions of the sample means for the different sample sizes are shown in the other plots. We see that the larger the sample size, the more normally distributed the sample means would be. Shifting gears a bit, Section 5.4 shows us that a <u>binomial distribution</u> is approximately normal—the larger the sample size, the more normally distributed it is. Since a normal distribution includes a mean and a standard deviation, we are told what the mean and standard deviation are for binomially distributed data. These are summarized in Theorem 5.4.1 on page 164. The <u>continuity correction</u> is a minor adjustment that makes our use of the normal distribution to approximate a binomial distribution more accurate.

Chapter 6 is finally a more detailed discussion of how to use *a sample* to make inferences (estimates) about *the population*—this is the main purpose of this course. Section 6.1 tells us the obvious: our best estimate for the population mean μ and standard deviation σ are the sample mean $\bar{y}$ and standard deviation s. Of course, there is always some uncertainly, since we don't have the entire population, just a small part of it in the sample. The <u>standard error</u> is a measure of how uncertain we are about the sample statistics (such as the sample mean $\bar{y}$) as estimates for population statistics (such as the population mean μ). The two main things that affect uncertainty are (1) variation within the sample data (which is measured by the sample standard deviation s) and (2) the sample size n. In general, the standard error will be smaller if (1) there is *less variation* in the data and/or (2) the *sample size* is *larger*. Both (1) and (2) make us happy. In Section 6.3, this uncertainty shows up in the confidence interval we find in estimating μ using $\bar{y}$:

$$\mu = \bar{y} \pm t_{\alpha/2} SE \ .$$

The formula for the standard error SE varies depending on what we are estimating: a population mean, the difference between two population means, a population proportion, etc. See the Summary of Formulas and Tests for the formulas (that I created) for various ways SE is computed. The t value comes from <u>Student's t distribution</u>, from Table 4. We're familiar with this confidence interval formula $\mu = \bar{y} \pm t_{\alpha/2} SE$, but where does it come from? Formula 6.3.1 on page 183 and the subsequent lines show how we go from

The probability that a sample mean would be within 1.96 standard deviations is 0.95

to

The probability that the mean would be within 1.96 standard errors of the sample mean is 0.95

We would use $z = 1.96$ from Table 4 for a 95% confidence interval if we knew that the sample means were perfectly normally distributed. Since we can't count on that, Student developed his t distribution (Student is the name of the fellow who came up with the idea), which is reflected in <u>Table 4</u>. Let's make an important distinction: Table 3 is used when we have perfectly normal distributed *population* data and we are making predictions about a *sample*. Table 4 is used when (1) we have a *sample* from an approximately normally distributed population or else a large enough sample (say $n = 20$ or 30) so that we can assume approximately normal distribution in the sample distribution and (2) we are taking a sample and making inferences/estimates about the *population*. As the Central Limit Theorem states, for infinitely large samples, the sample distribution would be perfectly normal (of course this can't happen in real life, but we often have large enough samples for which it essentially happens), thus the bottom row of Table 4 with $df = \infty$ corresponds to the values in Table 3. In Table 3 we use a z value to find an area, and in Table 4 for a given desired area we find the corresponding t value. Now back to the confidence interval $\mu = \bar{y} \pm t_{\alpha/2} \cdot SE$. Things that will make a more precise (narrower, i.e. the margin of error $t_{\alpha/2} \cdot SE$ smaller) estimate include:

(1) a *smaller* SE, already mentioned half-a-page above, and (2) a *smaller* $t_{\alpha/2}$ value, which occurs if the sample size n is larger and/or the confidence level is *lower*. It's a trade-off: *lower* confidence $\Leftrightarrow$ *smaller* $t_{\alpha/2}$, or equivalently, *higher* confidence $\Leftrightarrow$ *larger* $t_{\alpha/2}$. Also notice in Table 4 that for a given confidence level, *larger* sample $\Rightarrow$ *smaller* $t_{\alpha/2}$. Consider the interval given at the bottom of page 187. It doesn't make sense to say that there is a 95% chance the population mean μ is in that interval, since it is either in the interval or it is not. A more appropriate interpretation: if the population mean were whatever the sample mean $\bar{y}$ is, then there is a 95% probability the sample mean would be a value within the given interval. So for the example just mentioned, there is a 95% probability that the sample mean would be between 31.4 and 34.2. It's a subtle distinction. To conclude 6.3, one-sided confidence intervals are discussed. In Section 6.4 we learn how to estimate what sample size would be needed to ensure the margin of error $t_{\alpha/2} \cdot SE$ is smaller than a given tolerance (i.e. how much error or uncertainty we can live with). This is a big issue: how large does our sample size need to be in order to achieve a particular result, for example, a particular confidence interval width? Section 6.5 discusses when the formulas given can be used: (1) a sufficiently large sample and (2) sample data that are independent. The first is easy to determine (but not necessarily easy to make happen in real life), the second is a little tougher to be sure of. In Sections 6.6 and 6.7 we first learn about determining whether two populations have different means. We take a sample from each population and determining how different the two samples are from each other. The more different the two samples are, the more certain we are that the populations are different. There are three ways to do this, all of which are essentially equivalent. Two involve test statistics, the third involves a confidence interval, which is what is covered at this stage, in Section 6.7. Remember that the standard error is a measure of how much uncertainty there is about how reliably the sample can be used to estimate or make an inference about the population. There are two ways to compute the standard error for the difference in two sample means:

$$SE_{\bar{Y}_1-\bar{Y}_2} = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

or

$$SE_{pooled} = \sqrt{\frac{s_{pooled}^2}{n_1} + \frac{s_{pooled}^2}{n_2}} = \sqrt{s_{pooled}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where s_{pooled}^2 is a <u>weighted average</u> of s_1^2 and s_2^2:

$$s_{pooled}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}.$$

Equivalently, s^2_{pooled} is really just the mean square (denoted MS in later chapters) of the difference between each value and its sample mean: for each sample, the variance is $s^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$, so

$$(n-1)s^2 = (n-1)\frac{\sum(y_i - \bar{y})^2}{n-1} = \sum(y_i - \bar{y})^2$$

(in later chapters denoted as the sum of squares SS), thus

$$s^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_1-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{\sum(y_i - \bar{y})^2 \, for\ sample\ 1 + \sum(y_i - \bar{y})^2 \, for\ sample\ 2}{(n_1-1)+(n_2-1)}$$

The first version of the standard error $SE_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{SE_1^2 + SE_2^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is what we use most of the time. The pooled version shows up again in Chapter 11 when doing ANOVA testing. Of course for the confidence interval we need a $t_{\alpha/2}$ value in $\mu = \bar{y} \pm t_{\alpha/2} SE_{\bar{Y}_1 - \bar{Y}_2}$, and for the $t_{\alpha/2}$ value from Table 4, we need df. Formula 6.7.1 on page 211 is a little complicated, and the value is usually just given to us. Note that $\min(n_1 - 1, n_2 - 1) \le df \le (n_1 + 1) + (n_2 - 1)$.

In Chapter 7 we learn about doing <u>hypothesis testing</u> (which in a way, we've already learned a bit about, since we could actually do hypothesis testing using a confidence interval, which we learned about in Chapter 6). The first version of hypothesis testing is regarding whether two populations have different means. The more different the samples are from each other, the more likely we are to conclude that the populations are different. Reminder: the only way to know with 100% certainty something about an entire population is to get that information from the entire population, not just a sample. Thus no matter what inferences or conclusions we do or do not make about a population (such as comparing two populations means), there is always the possibility that we are making a mistake. If we decide to reject the null hypothesis (for example, that the population means are the same), then there is the possibility that we are making a <u>Type I Error</u>. If we decide that there is not enough difference in the sample means to be sufficiently certain that the populations are different (that is, if the test statistic is not sufficiently large), then there is a possibility that we are making a <u>Type II Error</u>. These error types are discussed in Section 7.3. In Section 7.2 we learn about the first type of <u>test statistic</u>, t_s, which we would use if the two populations are approximately normal and/or the sample sizes are large (say 20 or 30 or more). The intuition that we see for the first time in this section is true in every test through the rest of the book: the *more different* the two samples are, the *larger* the test statistic will be, or conversely the larger the test statistic, the more different the two samples are. There are three things that make us more certain that two populations do actually have different means, that is,

three things that result in a larger test statistic: (1) *larger* difference between the sample means, (2) *smaller* variation (standard deviation) within each sample, and (3) *larger* sample sizes. Another idea that we first learn in Section 7.2 but that carries into each of the tests that we learn about is that each test statistic has a corresponding P value. How we find the P value depends on what sort of test we are doing (different tables in the back of the book relate to different tests), and these different tests depend on what we are testing (Different population means? Are the proportions of some characteristic in a single population as expected? etc.) and the assumptions we are making (Normally distributed data? Paired data? etc.). The P value is the likelihood that we would get samples this different (i.e. as different as these two samples) are even more different if the null hypothesis were really true. More precisely, P is the likelihood of getting a sample with the given test statistic or larger if the null hypothesis were actually true. Thus is also described as the likelihood we are making a Type I Error: that the null hypothesis is actually true but we are deciding it is false based on the samples. The level of significance α is the maximum amount of Type I Error we can stand to live with—it is the level of risk we are willing to take. The confidence level is $1-\alpha$, e.g. if $\alpha = 0.05$ then the confidence level is 95%. In Section 7.2 we are reminded that in real life we use technology to do a lot of the work, including to find a P value for a given test statistic. In 7.3 we see the connection between the t test and a confidence interval. The four possible results, including Type I and Type II Errors are summarized in Table 7.3.2. Again, no matter what our conclusion is (whether we reject H_0 or not), there is always the possibility that we are making a mistake, since are using just a sample, rather than the entire population(s). In 7.4, association (two things happen to occur at the same time) and causation (one thing actually causes another) are discussed, as are observational studies (the researcher doesn't do anything—he/she merely observes) and experimental studies (the researcher intervenes or manipulates the study to see what happens). An experimental unit is one (hence the word "unit") of the objects in the sample, e.g. a person or a plant or an animal. We can do one tailed tests, first mentioned in 7.5, if we have legitimate reason to believe that there is a certain direction in which the alternative hypothesis might be true (for example, $H_A: \mu_1 > \mu_2$ or $H_A: \mu_1 < \mu_2$, rather than the non-directional $H_A: \mu_1 \neq \mu_2$), based on prior information. We can also have one-sided confidence intervals. When a direction is possible for an alternative hypothesis, then we can do a one tailed test. The P value for a one tailed test is half of the P value for a two-tailed (also known as non-directional) test. As described in 7.6, significant evidence means that there is enough evidence (for example, two sample means are different enough) to conclude that the null hypothesis is false and the alternative is true (for example, that the two population means are different). Importance means that that the difference actually matters. Table 7.6.3 on page 273 is a nice summary of some possible scenarios. Effect size corresponds to how different the two samples are: how much *effect* does the thing we're interested in have. For example, how much effect does a weight

loss drug actually have, i.e. how different is the one sample which received the drug from the control group sample which did not? Larger effect means the two samples are more different (less overlap between the samples), which in general means we are more likely to conclude the populations are different. However, even with small effect size (small difference between sample, i.e. large overlap between the two samples), larger sample sizes can still allow us to conclude that the two populations are different. We see this fact in Table 5. In 7.7 <u>power</u> is introduced. Level of confidence is the probability that we would correctly *not reject* the null hypothesis when the null hypothesis is *true*. Similarly, power is the probability that we would correctly *reject* the null hypothesis when the null hypothesis is *false*. For example, a confidence level of 95% and power of 90% means that there is up to a $\alpha = 0.05$ (5%) chance we will make a Type I Error and up to a $\beta = 0.10$ (10%) chance of making a Type II Error. We see in Table 5 that *larger power* and/or *higher significance level* and/or *smaller effect* all lead to the need for a *larger sample* size. In 7.9 the meaning of P is further discussed. Some ways of thinking of P are on page 287. In 7.10 is our first introduction to the fact that the t test is not always the appropriate test. We learn of the Wilcoxon-Mann-Whitney Test, which is used if (1) we don't know anything about the population (that is, we don't know that it is normally distributed) and (2) the sample is small. Remember that if the sample is large, regardless of the distribution of the population, we can use the t test.

In Chapter 8 we begin learning about several other types of tests, as summarized in the Summary of Formulas and Tests. In pretty much every situation, we compute a test statistic which measures how different the samples are from what the null hypothesis H_0 describes. There are many ways that the test statistic is computed, depending on what test is being used. But in every case, there are a few recurring themes: *larger* samples make us *more* confident about whatever it is we will conclude, *more* variation within each population's sample makes us *less* certain about making a conclusion, and in general, the *more* different the two (or more) samples are, the *larger* the test statistic and the *smaller* the P value, that is, the *less* likely this sample could have come from the population(s) if the null hypothesis were actually true. Chapter 8 deals with whether two population means are different, using <u>paired</u> data. There are a few different tests we use for paired data, depending on whether we assuming normal distribution. If we can assume approximately normal distribution and/or we have sufficiently larger sample sizes, then we can use the paired-sample t Test. If not, we can use the Sign Test or (by doing a bit more work) the Wilcoxon Signed-Rank Test. At the beginning of 8.3, the idea of pairing is discussed. Pairing makes our test more powerful. That is, if the two populations are actually different, it is more likely that we would actually end up concluding they are different (rejecting H_0 and accepting H_A). However, if there is pairing and we ignore that pairing, our test is less powerful. Even worse than ignoring paring is treating

unpaired data as if it were paired. The moral of the story: if data is paired, treat it as paired. If not, then do not treat it as paired.

Chapter 9 deals first with categorical (rather than numerical) data for the case of *one population*, *two categories* (in Section 9.1). This is also known as dichotomous observations. (A *dichotomy* is the splitting of a whole into two non-overlapping parts. Some "splittings" are more obvious or clear than others. Examples: male or female, diseased or not, old or young, etc.) We are interested in the proportion of the sample/population that is in one of the two categories (e.g. male). We sometimes refer to this as "success." (This is related to binomial distributions.) Notation for proportion: p is population proportion, $\hat{p}$ is sample proportion, and $\tilde{p}$ is the Wilson-Adjusted Sample Proportion (a slight modification of $\hat{p}$). $\tilde{p}$ is used in computing a test statistic or a confidence interval, both of which also involve standard error. As always, the standard error decreases as sample sizes increases. In 9.2 we work with confidence intervals at 95% confidence, and in 9.3 we generalize the same idea to other confidence levels. So what is covered in 9.2 is simply a special case of what is covered in 9.3. One thing a little different than what we've been used to is that for proportions we use z values rather than t values. This is because the binomial distribution is essentially normal. In Section 9.4 we work with *one population* with *multiple categories* (red or blond or dark hair; etc.). This is our first experience with the Chi-Square Goodness-of-Fit Test: we are testing how well the given proportions in each category match what we had expected. As usual, the more different what you observe is from what is expected, the larger the test statistic χ_s^2, the smaller P (from Table 9), and the more likely we are to conclude that the population is different from what we expected it to be.

In Chapter 10 we generalize this idea with categorical data to *two or more populations* with *multiple categories*. Table 10.5.4 is a nice example of this. We still perform a Chi-Square test. If there is one population, then each expected value is the fraction of the sample specified in the null hypothesis, like in HW 9.4.1 (the ratio of 12:3:1). If there are multiple populations, then the expected values are computed based on the given data, like in Table 10.5.3. Section 10.7 is included in Chapter 10 (rather than in Chapter 9, which deals with proportions, like in 9.2) since it involves two populations rather than one. In 10.8 the special case of *paired* data from *two populations* with *two categories* is discussed.

In Chapter 11 is our first encounter with working with more than two populations. While the details are bit more complex than earlier, the main idea is the same: just as

$$t_s = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{Y}_1 - \bar{Y}_2}}$$

is "what is the difference <u>between</u> the samples" divided by "what is the variation <u>within</u> the samples." The same is true with ANOVA, in its three forms, in which

$$F_s = \frac{MS(Between)}{MS(Within)} \;\; or \;\; F_s = \frac{MS(Interaction)}{MS(Within)}.$$

I'll not expect you to compute these values—they'll be given. But you should have some intuition about how things change when a certain value in the data is smaller or larger. In <u>One-Way ANOVA</u>, we are interested in how different levels in one factor/treatment affects whatever we are measuring. See the big box on page 454. Section 11.3 helps us understand how the total variation of data (the total sum of the squares $\sum\left(y_{ij} - \bar{\bar{y}}\right)^2$ for each observed value y_{ij}) is partly from the variation between the groups and partly within each group. Section 11.4 tells us what to do with these values: find F_s, then find the corresponding P in Table 10, etc. In One-Way ANOVA <u>with blocking</u>, there are two different factors, one of them being how we organized or blocked the experimental units, which we want to remove or adjust for. In <u>Two-Way ANOVA</u>, there are two factors that affect whatever we are measuring, and we are interested in how each factor affects things, as well as how the interaction between those two factors might affect things. In Two-Way ANOVA, we always check the interaction first, and if

$$F_s = \frac{MS(Interaction)}{MS(Within)}$$

is significant, then the test ends: we reject the null (that there is no difference due to the factors") and accept the alternative (that there is a difference due to the factors). We wouldn't check the individual factors, as the strong interaction between them makes it unclear which of the two factors is having an effect on whatever is being measured. If the interaction is not significant, then we do check both

$$F_s = \frac{MS(Factor)}{MS(Within)}$$

values to see if Factor 1 and/or Factor 2 has an effect. A rule of thumb is that more parallel lines in the plots (as in Figure 11.7.3) means less effect of interaction of the two factors, and less parallel lines means more interaction between the two factors is affecting things (as in Figure 11.7.4).

The main idea of Chapter 12 is fitting models (functions, such as straight lines) to data. This is also known as <u>regression</u>. We mostly focus on straight lines (so we are doing <u>linear regression</u>), but we see in Section 12.8 there are other types of functions we can fit to data. We don't do that in this course. There are two main types of data: lots of Y values for each of a few given X values, as in Figure 12.1.1 (random subsampling), or simply lots of pairs of values, as in Figure 12.1.2 (bivariate random sampling). In either case, we can determine what line $Y = b_0 + b_1 X$ would best fit the sample data and how well the line would fit the data, as measured by the <u>correlation coefficient</u> r, where $-1 \leq r \leq 1$. The closer r is to 1 or -1, the more linear the data are, that is, the more the data are linearly correlated. As usual, the values found from the sample are our best estimates for the same values for the population. As is typical, there is some uncertainly in the estimate of b_1 for β_1, so we can find a confidence interval for β_1 using b_1 rather than simply saying $\beta_1 \approx b_1$. The formula for the population line $\mu_{Y|X} = \beta_0 + \beta_1 X$ reminds us that this line doesn't predict a specific Y value for a given X value: it predicts what the *average* Y value would be for a given X value. This is illustrated in Figure 12.4.1. Associated with this is the standard deviation of the error (how much difference there is in the predicted values $\hat{y}_i$ vs. the measured values y_i for each given value of x_i): s_e for the sample and σ_e for the population. About 2/3 of the measured data is within one standard deviation of the regression line, as seen in Figure 12.3.8, and similiarly for two and three standard deviations, as first discussed way back in Figure 4.3.5. Finally, one thing that we see a few times in this chapter is how data points farther away from the main cluster of data can affect things. This is one of the difficult-to-answer questions when working with data: do we use outliers (probably) or do we simply ignore them (probably a little dangerous)?

Chapter 13 gives a nice summary of the various tests we've covered. I would read it through once as you study for the final.