Lecture Notes, Set 3

This set of notes covers estimation and confidence intervals, two of our three main review topics from Math 218. The notes are more detailed than what I went over in class. As usual, you are only responsible for the concepts covered in lecture.

Estimation and Sampling

Next we will review the topics of sampling and point estimation. It is worth briefly reiterating the distinction between a population and a sample. Usually, there is some characteristic, or **parameter**, of the population in which we are particularly interested. For now think of that parameter as the population mean. Generally, we cannot measure the variable of interest for everyone in the population, and so do not know the population mean. Instead, we take a sample and use its characteristics to make **inferences** about the population. The first thing we do is to **estimate** the population parameter in which we are interested.

Suppose we are interested in the population mean, μ, but have only a sample. What is our "best guess" for μ? Naturally, it is the sample mean, $\bar{X}$. Similarly, if we are interested in the population variance, σ^2, our best guess is the sample variance, s^2. The sample mean and sample variance are called **point estimates** or **sample estimates** of the corresponding population parameters.

Naturally, we would like our estimates of the population parameters to be "good" estimates. What exactly does this mean? There are several things we would like to be true. First, we would like our estimate to be **right on average.** We know, for instance, that the sample mean will not always equal the population mean–that would be too much to hope for. However, we will be pretty unhappy if the sample mean is consistently too high (or too low). For instance, suppose we are measuring the weight of people in the United States, and that the true mean weight is μ=160 lbs. It would be bad if our estimate for μ were always greater than 160 lbs. We want it to "typically" or "on average" be 160 lbs, even though any particular sample will be a bit above or a bit below. An estimate of a population parameter is said to be **unbiased** if on average it comes out to the true population value. For example, the sample mean is unbiased if $E(\bar{X}) = \mu$. We will see later that this is the case. Similarly, the sample standard deviation is unbiased, that is $E(s^2) = \sigma^2$. This is precisely the reason we divide by n-1 rather than n in the sample variance–we want an unbiased estimator. Another thing that would be "good" would be for our estimators to be very stable, or equivalently to have a low variance. Why is this important? If our estimate has a low variance, that means it is never very far from its typical value. If it is unbiased its typical value is the true population value. Thus an unbiased estimate with low variance should always be very close to the true population value. This is definitely desirable. We would not be very happy in the example given above if our estimate of the mean weight sometimes produced values of 80 lbs and sometimes produced values of 300 lbs, even if it averaged out to 160 lbs. We want our estimate to be close to the true value.

In order for our estimates to have the desired characteristics, we need to have a "good" or "representative" sample. What should such a sample be like? The most important thing is that the sample be **random**. In a **simple random sample** every possible sample of size n has the same chance of being chosen. For instance, if you have five numbers in a hat, and you draw a sample of size two, every pair of two numbers is equally likely to be drawn (provided the numbers were well mixed). This is called **sampling without replacement.** Sampling without replacement has a disadvantage–it turns out that the individual draws from the hat are not independent. This should make sense. If you choose one of the five numbers the first time, you know you aren't going to get it the second time. Another way to do random sampling is to make sure that members of the sample are chosen **independently** and according to the **same distribution.** One example is **sampling with replacement.** If you a draw a sample of size two from the hat, with replacement, you put the first number back in the hat after you draw it. Then knowing what the first number was tells

you nothing about the second. Each number has a probability of 1/5th on each draw. As another example suppose you want a random sample of size two of outcomes of the roll of a die. You roll the die twice. The rolls are independent, and each time each face of the die has the same chance of coming up as it did in the previous roll. Note that this is true even if the die is loaded. Suppose the probability of a 1 is 1/2, and the probability of all other values is 1/10. Since one has a 1/2 chance of coming up on the first roll, it still has a 1/2 chance of coming up on the second roll, even though the chance of getting a 2 is different than the chance of getting a 1. This is what I mean by "the members of the sample should come from the same distribution." Each member of this population (1,2,3,4,5,6) does not have the same chance of being selected on a single roll. However, the chance of rolling any particular value stays the same from roll to roll. There are many other sampling techniques some of which we may discuss later. How you do the sampling changes what you can say about your parameter estimates. For now, we will mostly assume samples are independent and "identically distributed", the second option described above.

The Mean and Variance of $\bar{X}$

In this section we concentrate on the sample mean, which is our best estimate of the population mean, μ. It is important to realize that $\bar{X}$ is a random variable! It is random, because it depends on which sample you choose. If I choose two different samples, I will (probably) get two different values of $\bar{X}$. If $\bar{X}$, the mean of a sample of size n, is a random variable, then it must come from some population. What is that population? It is the population of all possible samples of size n from the *original* population. Don't get confused. There are two different populations here! Since $\bar{X}$ is a random variable, it must also have a distribution–it takes on certain values with certain probabilities. From that information, we could compute the mean and variance of $\bar{X}$.

Example: Suppose our original population consists of five businesses, with lifetimes of 2, 4, 6, 8, and 10 respectively. I can calculate the population mean and variance by our old formulas:

$$\mu = E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{5}[2+4+6+8+10] = 6$$

$$\sigma^2 = Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{5}[(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2] = 8$$

Now suppose that I am interested in samples of size n=2 from this population as drawn with replacement. What are all the possible samples? There are 25, all equally likely, if I described the draws as (first-draw, second-draw): They are (2,2), (2,4), (2,6), (2,8), (2,10), (4,2), (4,4) (4,6), (4,8), (4,10), (6,2), (6,4), (6,6), (6,8), (6,10), (8,2), (8,4), (8,6), (8,8), (8,10), (10,2), (10,4), (10,6), (10,8), and (10,10). Note that the there are two pairs with a 2 and a 4, (2,4) and (4,2) etc., but only pair of the form (2,2) etc. The corresponding sample means are 2, 3, 4, 5, 6, 5, 6, 7, 3, 4,...., 9, 10. Thus I can construct a table of all the possible values of the sample mean for samples of size two, with their associated probabilities:

x	2	3	4	5	6	7	8	9	10
$P(X=x)$	.04	.08	.12	.16	.2	.16	.12	.08	.04

From this table we can easily compute the mean and variance of $\bar{X}$:

$$E(\bar{X}) = \sum_x xP(\bar{X}=x) = 2\times.04+3\times.08+4\times.12+5\times.16+\cdots+10\times.04 = 6$$

$$Var(\bar{X}) = \sum_x (x-E(\bar{X}))^2 P(\bar{X}=x) = (2-6)^2\times.04+(3-6)^2\times.08+(4-6)^2\times.12+\cdots+(10-6)^2\times.04 = 4$$

Notice something interesting. The expected value of $\bar{X}$ turned out to be 6, the same as the true population mean. This is always the case for the two kinds of random sampling described in the first section.

Important Result: If $X_1, X_2, \ldots, X_n$ represents a random sample from a population with mean μ then the expected value of the sample mean is also μ. In other words, $E(\bar{X}) = \mu$.

Optional Proof: This is just a consequence of the rules for adding expectations:

$$E(\bar{X} = E(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\sum \mu = \frac{1}{n} n\mu = \mu$$

Example: Suppose we are interested in the weights of packages coming off then production line at a sugar processing plant. The production line fills the packages to a mean weight of $\mu = 10$ pounds. Each day, a quality control engineer takes a random sample of nine packages produced in the previous day and computes the sample mean $\bar{X}$ of the nine package. If the engineer does this for 400 days, he or she would have collected 400 sample means. These 400 sample means could themselves be added and divided by 400 to obtain their mean. The above result suggests that the 400 sample means would have a mean close to 10 pounds. If we continued taking sample means forever we would get exactly 10 pounds.

It turns out that the standard deviation of the sample mean is related to the standard deviation of the original population, just as the expected value of the sample mean is related to the population. Suppose that the population in question is infinite (or else so large that it doesn't matter whether sampling is done with or without replacement) and that the elements of the sample $X_1, X_2, \ldots, X_n$ are independent. Then we get the following:

Important Result 2: If $X_1, X_2, \ldots, X_n$ are independent samples from a population with variance σ^2 then

$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \quad and \quad SD(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Optional Proof: These rules are a consequence of the addition rules for variances under independence:

$$Var(\bar{X}) = Var(\frac{1}{n}\sum X_i) = \frac{1}{n^2}\sum Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

These formulas should make intuitive sense. They say that the bigger the sample, the smaller the variability of the sample mean. But the bigger the sample, the more sure we are that the sample mean is close to the true mean–if the sample were so big we had the whole population, we would be completely sure that the sample mean equalled the population mean. In this case, their would be no variability in the sample mean.

The Central Limit Theorem and the Distribution of $\bar{X}$

Now we know the mean and variance of the sample mean. It would be nice if we actually knew the **distribution** of the sample mean. (This is sometimes known as the "sampling distribution of the mean.") Remarkably, it turns out that provided your sample is large enough, the sample mean, $\bar{X}$, always has a normal distribution. This result is called the **Central Limit Theorem.**

The Central Limit Theorem has two forms which you should have learned in Math 218. We will only use version 1 in this class:

Central Limit Theorem 1: If $X_1, X_2, \ldots, X_n$ are independent and drawn from the same distribution with mean μ and variance σ^2, and n is large then $\bar{X} = \frac{1}{n}\sum X_i$ is approximately normally distributed with mean μ and variance σ^2/n. Furthermore,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Central Limit Theorem 2 (Optional): If $X_1, X_2, \ldots, X_n$ are independent and drawn from the same distribution with mean μ and variance σ^2, and n is large then $\sum X_i$ is approximately normally distributed with mean $n\mu$ and variance $n\sigma^2$ Furthermore,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$$

Note: The means and variances in the two versions of the Central Limit Theorem follow completely from what was done in the previous section. What is remarkable is that averages and sums of large numbers of independent random variables are always normal. By "n is large" we usually mean n=15-30, although it depends on how skewed the original distribution of the $X_i's$ is. Remember that n is the size of the sample used to compute $\bar{X}$. What the CLT basically says is that if we picked a random sample of size n, calculated $\bar{X}$, picked another sample of size n and calculated its $\bar{X}$, and repeated the process many times then the histogram of the collection of $\bar{X}'s$ would look approximately hump-shaped or normal. Of course in real life, we only get one sample and one $\bar{X}$. However, the CLT allows us to make probability calculations about that $\bar{X}$ which are extremely useful.

Note: The two versions of the Central Limit Theorem are really the same. You can get the second from the first by multiplying the numerator and denominator of $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ by n.

Note: In order to use one of the versions of the Central Limit Theorem, there are three things you must check: **(1)** The quantity in whose distribution you are interested is a **sample mean** or the **sum** of elements of a sample **(2)** The members of the sample are **independent** and all come from the **same distribution** **(3)** The sample size, n, is **sufficiently large**–usually n=15-20 will do if the original distribution is not too skewed. If these three things apply then your sample mean or sum is **approximately normal by the Central Limit Theorem.** If the original distribution was **exactly normal** then the sample mean or sum will also be **exactly normal,** irrespective of how large n is. This follows from our cool fact about adding normal random variables. The mean in this case will be μ (or $n\mu$ for a sum) whether or not members of the sample are independent. If the members of the sample are independent, the variance will be σ^2/n (or $n\sigma^2$ for a sum).

To get a better grip on the Central Limit Theorem, let us do some examples:

Example 1: Consider the distribution for inspection times at a border patrol station. A sign says the mean inspection time is $\mu = 8$ minutes and the standard deviation is $\sigma = 6$ minutes. A group of n=64 minorities pass through the border patrol station. Their average inspection time is 10 minutes. The leader of the group suspects discrimination. In order to test this she computes the probability of getting an average inspection time of 10 minutes or more for a group of size 64 if the true mean and standard deviation are as stated. Do you think there is discrimination?

Solution: Even though the original distribution is probably not normal (think about why!), by the Central Limit Theorem, version 1, the sample mean is approximately normal assuming we treat the 64 minorities as independent passers through the inspection station. (This may be a bad call of course, especially since they came in a group!) The probability in question is

$$P(\bar{X} \geq 10) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{10 - \mu}{\sigma/\sqrt{n}}) = P(Z \geq \frac{10 - 8}{6/8}) = P(Z \geq 2.67) = .5000 - .4962 = .0038$$

In other words, there is less than a 1% chance of the sample mean for our group being this high if the true mean and standard deviation are as given. The group leader has reason to be suspicious.

Example 2: A commuter plane carries 100 passengers. Each passenger brings baggage with a mean weight of $\mu = 50$ pounds and a standard deviation of $\sigma = 10$ pounds. The plane can carry a total baggage weight of 5200 pounds. During the holiday season, the plane is booked completely full. The airline wants to know the probability that the passengers bring more baggage than the plane can carry.

Solution1 (Optional): We are interested in $S = \sum_{i=1}^{100} X_i$, the total weight of the bags brought by the customers. We want to know $P(S \geq 5200)$. S is a **sum** of **n=100 independent** random variables, each with the same distribution. (This last is a slight assumption–the weights of baggage brought by individual passengers may not be independent if they are traveling together, but let's assume they are.) Therefore, the second version of the Central Limit Theorem applies, and S has an approximately normal distribution with mean $E(S) = n\mu = 100 * 50 = 5000$ pounds and standard deviation $SD(S) = \sqrt{n\sigma^2} = \sqrt{100 * 100} = 100$ pounds. We can therefore compute probabilities for S using a Z-score:

$$P(S \geq 5200) = P(\frac{S - n\mu}{\sigma\sqrt{n}} \geq \frac{5200 - n\mu}{\sigma\sqrt{n}}) = P(Z \geq \frac{5200 - 5000}{100}) = P(Z \geq 2) = .5000 - .4772 = .0228.$$

Note that we could have said the probability was approximately 2.5% since we were just finding the probability that a normal distribution was more than two standard deviations above its mean. I got the exact answer using the normal table.

Solution 2: There is a second approach to this problem using the first version of the Central Limit theorem. It is equivalent to ask whether the total weight of the passengers' baggage exceeds 5200 pounds, or whether the average weight of their baggage exceeds 52 pounds. The sample mean of the passenger's baggage has, by the first version of the Central Limit Theorem, approximately a normal distribution with mean $E(\bar{X}) = \mu = 50$ and standard deviation $SD(\bar{X}) = \sigma/\sqrt{n} = 10/\sqrt{100} = 1$. Thus the probability that the mean baggage weight of 100 passengers exceeds 52 is given by

$$P(\bar{X} \geq 52) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{52 - \mu}{\sigma/\sqrt{n}}) = P(Z \geq 2) = .0228$$

just as before.

The above examples illustrate the Central Limit Theorem as it applies to the sample mean. Next, I give two further examples involving the use of the Central Limit Theorem for proportions. In order to use the CLT for proportions, we have to understand why it applies. What is a sample proportion? It is the number of "successes" in n independent trials each with a "yes or no" outcome. Suppose we let X_i be the random variable that is 1 if the ith trial is a success, and 0 if the ith trial is a failure. Then $X = \sum_{i=1}^{n} X_i$ is just the number of successes. (Of course, X has the binomial distribution with parameters n and p.) Furthermore, the sample proportion of successes is just $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. **In other words, the sample proportion is really a sample average!** Thus we can apply the central limit theorem. Of course to do that we need the mean and variance of the $X_i's$. Fortunately, this is easy. Each X_i is Bernoulli so $E(X_i) = p$ and $SD(X_i) = \sqrt{p(1-p)}$. Using these facts, if n is large, the Central Limit Theorem, Version 1, says that $\hat{p}$ is approximately normally distributed with mean p and standard deviation $\sqrt{p(1-p)/n}$. The Central Limit Theorem 2 says that $X = \sum X_i$ is also approximately normally distributed with mean np and standard deviation $\sqrt{np(1-p)}$. This is the basis of the "normal approximation to the binomial distribution" which you may have learned in Math 218. Let's do an example to see how this works.

Example 1–Proportion version: Suppose 80% of a school's accounting class pass the CPA exam on the first try. If a randomly chosen 100 students sit the exam, what is the chance that 85% or more pass on the first try?

Solution: Let p be the true proportion of students who pass the bar on their first attempt. In this problem, p = .80. We are talking about a group of 100 students taking the exam, so n=100. We want to know the probability that $\hat{p} \geq .85$, that is that the sample proportion we observe will be more than .85. We can use the Central Limit Theorem, version 1, since the sample proportion is really a sample mean. The standardized sample proportion will be approximately normally distributed. Since the standard deviation of the sample proportion is $\sqrt{p(1-p)/n}$ we have

$$P(\hat{p} \geq .85) = P(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq \frac{.85 - p}{\sqrt{p(1-p)/n}}) = P(Z \geq \frac{.85 - .80}{\sqrt{(.80)(.20)/100}}) = P(Z \geq 1.25) = .5 - .3944 = .1056$$

Example 2–Binomial version (Optional): Let X be the number of students in the sample of 100 who pass the bar. The expected number of students is just np = 100*.80 = 80. The standard deviation of the number of students who pass is $SD(X) = \sqrt{np(1-p)} = \sqrt{100 * .80 * .20} = 4$. If 85% or more of the students in the group pass the exam, that is the same as 85 or more students passing. Thus we want to calculate $P(X \geq 85)$. Since n is large, by the Central Limit Theorem, version 2, X is approximately normally distributed. This is because X can be thought of as a sum. We get

$$P(X \geq 85) = P(\frac{X - np}{\sqrt{np(1-p)}} \geq \frac{80 - np}{\sqrt{np(1-p)}}) = P(Z \geq \frac{85 - 80}{4}) = P(Z \geq 1.25) = .1056$$

Confidence Intervals

The next topic of these notes is **confidence intervals.** We have often talked in the past about a range of values into which some fraction of a data set or distribution must fall. In particular Chebychev's Rule tells us that at least 75% of any data set (or distribution) must lie within two standard deviations of the mean, and at least 89% of the data (or distribution) must lie within three standard deviations of the mean. Similarly, if the data or distribution is approximately normal around 68% percent of the data (distribution) lies within one standard deviation, 95% lies within two standard deviations, and 99.7% lies within three standard deviations of the mean. Another way of saying, for example, that 95% of a distribution lies within two standard deviations of the mean is to say that we are 95% sure or *confident* that the next data point we select from that distribution will lie within the interval determined by going two standard deviations to either side of the mean. Thus this interval is a "95% confidence interval."

Of course, we are generally interested not so much in where the next data point selected from our distribution will lie, but in what is the true value of the mean of the distribution or population. After all, we usually only have a sample. Our best guess, or estimate, of the population mean is the sample mean, but we would like to know what is a reasonable range of values that the true mean could take on. It seems fairly natural to go two standard deviations to either side of the sample mean. Let us see why this makes sense. Let $X_1, X_2, \ldots, X_n$ be a sample whose members are independent and come from a common distribution with mean μ and variance σ^2. Then provided n is large, by the Central Limit Theorem, version 1, we know that $\bar{X}$ is approximately normal with expected value $E(\bar{X}) = \mu$ and variance $Var(\bar{X}) = \sigma^2/n$. Thus using the bell-shaped rule we are 95% sure that the sample mean, $\bar{X}$ will be within 2 of its SD's or $2\sigma/\sqrt{n}$ of the true mean, μ. Intuitively, this suggests that any value within $2\sigma/\sqrt{n}$ to either side of $\bar{X}$ is a plausible value for μ.
We can actually show mathematically, using Z-scores, that the range of values selected above has a 95% chance of including the true mean. We calculate as follows, taking advantage of symmetry:

$$\begin{aligned} Confidence &= P(\bar{X} - 2\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 2\sigma/\sqrt{n}) \\ &= P(-2\sigma/\sqrt{n} \leq \mu - \bar{X} \leq 2\sigma/\sqrt{n}) \\ &= P(-2\sigma/\sqrt{n} \leq \bar{X} - \mu \leq 2\sigma/\sqrt{n}) \\ &= P(-2 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 2) \\ &= .95 \end{aligned}$$

All of this follows simply by appropriate manipulation of the inequalities. We conclude that there is a 95% chance that the interval $[\bar{X} - 2\sigma/\sqrt{n}, \bar{X} + 2\sigma/\sqrt{n}]$ **contains the true mean.** We call this interval a **95% confidence interval for μ.** Note that it is the confidence interval which is random, NOT the population mean. The population mean, μ is a fixed number. The confidence interval depends on $\bar{X}$ which in turn depends on the particular random sample selected. So what does it really mean to say that $\bar{X} \pm 2\sigma/\sqrt{n}$ is a 95% confidence interval? It means that if we took 100 random samples of size n, and formed the corresponding confidence intervals, that about 95% of those confidence intervals would contain the true mean.

The more times we repeated the experiment, the closer we would get to 95%.

There is nothing special about a 95% confidence interval for the mean. We can get a confidence interval with any percentage of certainty. Define the α-level **critical value** of a standard normal distribution to be the value z_α such that

$$P(Z \geq z_\alpha) = \alpha$$

To get a confidence interval which has a $100(1-\alpha)\%$ chance of including the mean, by symmetry we need to leave out $\alpha/2\%$ in each tail. Thus our interval will take the form $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$. (Note that there are two uses of α, one in defining the critical values, and one in defining the confidence level for the confidence intervals. Don't get them mixed up!) Let's double check that the formula for the $100(1-\alpha)\%$ CI is right:

$$\begin{aligned} P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}) &= P(-z_\alpha\sigma/\sqrt{n} \leq \mu - \bar{X} \leq z_{\alpha/2}\sigma/\sqrt{n}) \\ &= P(-z_{\alpha/2}\sigma/\sqrt{n} \leq \bar{X} - \mu \leq z_{\alpha/2}\sigma/\sqrt{n}) \\ &= P(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) \\ &= 1 - \alpha/2 - \alpha/2 \\ &= 1 - \alpha \end{aligned}$$

Example: A company fills packages of cement. From past experience, they know that the mean weight μ of their bags changes over time as the machine wears down. However, the standard deviation of the weight of the bags seems to remain constant at $\sigma = 3$. Suppose the company needs the mean weight of its bags to be at least 150 pounds. Suppose they take a sample of n=25 bags and find that their average weight is $\bar{X} = 151$ pounds. Are they sure that the mean weight of their bags is above 150 pounds?

Solution: We want to know what range of values contains the true mean. This calls for computation of a confidence interval. Let's compute a 90% confidence interval and a 95% confidence interval. To do this, we simply need to know $z_{.05} = 1.645$ and $z_{.025} = 1.96$. The two confidence intervals are

$$\bar{X} \pm z_{.05}\sigma/\sqrt{n}$$
$$\bar{X} \pm z_{.025}\sigma/\sqrt{n}$$

or

$$151 \pm (1.645)(3)/(5)$$
$$151 \pm (1.96)(3)/(5)$$

Thus the 90% confidence interval is [150.01, 151.99], and the 95% confidence interval is [149.82, 152.18]. Note that the first confidence interval contains only values above 150 pounds, whereas the second interval contains a few values below 150 pounds. Which interval is "right"? They are both right. Which interval you choose depends on how sure you want to be that your interval includes the true mean. The first interval has the advantage that it is narrower and hence seems to pin the mean down better. However, it has a higher chance of not including the mean. The second interval is more likely to contain the mean, but as a result is slightly wider. You always encounter this tradeoff. A huge confidence interval will almost always contain the mean, but is not very useful. What value you select for α should depend on context. It is often helpful to compute more than one confidence interval to get an idea of where the cutoffs are. In this case, the company can be quite sure that their true mean is above 150 pounds. There is only about a 5% chance that this is not so.

Example: A company fills boxes of cereal. From past experience, they know that the standard deviation of the weight of the bags seems to remain constant at $\sigma = 5$ ounces. The packages are advertised to weigh 64 ounces. The company quality control manager takes a sample of n=100 randomly selected boxes and finds

that their average weight is $\bar{X} = 65$ ounces. Can the manager be sure that the mean weight of the boxes is above 64 ounces?

Solution: We want to know what range of values contains the true mean. This calls for computation of a confidence interval. Let's compute a 90% confidence interval and a 95% confidence interval. To do this, we simply need to know $z_{.05} = 1.645$ and $z_{.025} = 1.96$. The two confidence intervals are

$$\bar{X} \pm z_{.05}\sigma/\sqrt{n}$$
$$\bar{X} \pm z_{.025}\sigma/\sqrt{n}$$

or

$$65 \pm (1.645)(5)/(10)$$
$$65 \pm (1.96)(5)/(10)$$

Thus the 90% confidence interval is [64.18, 65.82], and the 95% confidence interval is roughly [64, 66].

Confidence Intervals: σ Unknown

In real life, the population standard deviation will not be known. All we will have is a sample, and a sample standard deviation, s. How do we compute confidence intervals in this case? The answer, of course, is that we just replace σ by s in the formula and go on our merry way. However, there is one other important difference. If we use s instead of σ, our "Z-score" will be

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Unfortunately this no longer has a standard normal distribution, so we can not use the critical values $z_{\alpha/2}$ in our confidence intervals anymore. Fortunately, it is possible to work out the distribution of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$. Because s is random as well as $\bar{X}$, this quantity is more variable and is said to have "heavier tails." Its distribution is called the **t distribution**. Because it is based on the sample variance, the t distribution has associated with it **degrees of freedom.** When you look at a **t table** you will see the same picture as for the Z distribution except that there will be a row in the table for each number of degrees of freedom, and a column for each of the critical values. Just as $P(Z \geq z_\alpha) = \alpha$ so $P(T \geq t_\alpha) = \alpha$. Note also that as n gets very large, the t distribution gets very close to the normal distribution. This is because for large n, s will be extremely close to σ. Your t table does not have degrees of freedom above 120. Past that point, you may simply use a Z value.

To find a confidence interval when the population standard deviation is unknown, you simply replace σ by s, and $z_{\alpha/2}$ by $t_{\alpha/2,n-1}$. The number of degrees of freedom associated with the t critical value is n-1 where n is the size of your sample. We have n-1 degrees of freedom for all the same reasons that we used n-1 when computing the sample standard deviation. Thus our confidence interval formula becomes:

$$\bar{X} \pm t_{\alpha/2,n-1}s/\sqrt{n}$$

Example: Twenty homeowners are surveyed to learn how much they are paying each month on their mortgages. The average mortgage price in the sample is $\bar{X} = \$800$ and the standard deviation of mortgage prices is $s = \$60$ per month. Find a 98% confidence interval for the true mean mortgage payment in this community.

Solution: Since we have only a sample standard deviation we must use the t distribution rather than the Z distribution. We have n=20, so we use 19 degrees of freedom. The critical value is $t_{.01,19} = 2.539$. The resulting confidence interval is

$$800 \pm (2.539)(60)/\sqrt{20})$$

Thus the 98% confidence interval is [766, 834].

Confidence Intervals For Proportions

There is one last important kind of confidence interval–the confidence interval for a proportion. We have seen in previous lectures, that the best estimate of a population proportion, p, is the sample proportion $\hat{p}$. It turns out that $E(\hat{p}) = p$ and $Var(\hat{p}) = \frac{p(1-p)}{n}$. Furthermore, since the sample proportion is really an average, if n is large, the Central Limit Theorem applies, and the sample proportion is approximately normally distributed. We form confidence intervals for proportions just as for means:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

There are two key points to note. First, we use a Z critical value rather than a t critical value. Secondly, we use $\hat{p}$ in the standard deviation part of the formula because we don't know the true p. It may seem funny that because of this we don't use a t distribution. It turns out that in this case, the Z distribution works well enough.

Example: One hundred voters are surveyed to learn whether they support Candidate A. Of those surveyed, 64 support Candidate A. Can the candidate be pretty certain he will win (i.e. that more than 50% of voters support him)?

Solution: We compute a 95% confidence interval for the true proportion of voters who support Candidate A. (That choice is arbitrary. Try using other confidence levels for practice!) We know that $\hat{p} = .64$, $n = 100$, and $z_{.025} = 1.96$. Thus our confidence interval is

$$.64 \pm (1.96)\sqrt{\frac{(.64)(.36)}{100}}$$

Thus the 95% confidence interval is [.546, .734]. As long as these voters are representative of those who go to the polls, Candidate A is in good shape.