Chapter 7 Notes and elaborations for STAT 141-Introduction to Statistics

Assignment:

For section 1 of chapter 7 you may stop reading on page 377. I won't be asking about proper sample sizes, nor rounding rules. Do read all of section 7.2. Be sure you know when to use a z-confidence interval and when to use a t-confidence interval. Read up to 393 in section 7.3. We won't cover 7.4.

Do the following exercises:
7.1: 1, 5, 7 – 11 odd, 17 – 21 odd.
7.2: 1, 3, 4, 5, 11-15 odd
7.3: 1 – 11 odd, add #2 as well.

In the following notes, we begin with a general discussion of confidence intervals. We then discuss the constructions of specific confidence intervals, namely, we construct confidence intervals for the mean as well as confidence intervals for proportions.

Prerequisite knowledge: you'll need to understand what an interval is and order-of-operations for calculating the intervals correctly.

7.0 Confidence Intervals

They are quite messy to derive (that is, to make the general theory of how they work, etc.), but with a little practice they are not hard to construct. That is, the part I have to do to explain them is a bit mind-numbing, but the part you have to do is very formulaic, and actually pretty easy once you get the hang of it.

What on earth is a confidence interval? Well, confidence intervals go something like this.

Suppose that you have a super-giant container of jellybeans, and you can't see into it. You press a button, it mixes the jellybeans up really good, and then it pops one of the jellybeans out (I had a big ridiculous story for this, but decided that this example doesn't need any llamas.)

Unfortunately for me, I don't like those buttered popcorn jellybeans. Every once in awhile I get one. So I do an experiment. I push the button 100 times, and of those 100 trials I get exactly 23 buttered popcorn jellybeans. What have I learned?

Well, I know nothing about how many jellybeans are in the machine. But, I can surmise that about 23% of the jellybeans are buttered popcorn flavored. This is my best estimator of the true proportion of buttered popcorn jellybeans. If you were to repeat the experiment a couple of times you most likely wouldn't get exactly 23, but something pretty close, because the number of observed buttered popcorn jellybeans is actually random.

Before constructing a confidence interval, we need to decide on a confidence level. Most of the time, the confidence level is given to you by your boss or your stats professor, so you don't actually pick it. But typical choices are 95% and 90%. With a couple of computations (which we will learn in a bit), I get these:

The 90% confidence interval is: 23% ± 6.92% which is the same as the interval: (16.08%, 29.92%).
The 95% confidence interval is: 23% ± 8.25% which is the same as the interval: (14.75%, 31.25%).

So what does this mean? Well, a 90% confidence interval is constructed so that 90% of the time that you make one, the true value of the parameter will be in the interval. That is, if I were to repeat over and over the experiment of drawing 100 jelly beans and I made the 90% confidence interval with each experiment, I would expect 90% of the time that my interval contains the true proportion of buttered popcorn jellybeans in the container. In other words, if I ran the experiment 100 times, I would have 100 confidence intervals, and I would expect 90 of those intervals to contain the true proportion of buttered popcorn jellybeans.

Notice that you don't get something for nothing. A 95% confidence interval will 'capture' the true proportion more often that the 90% confidence interval does. . .but it is a larger interval.

Stop if you aren't completely digesting what I wrote. Re-read it. Understanding what a confidence interval is important.

7.0 Confidence Intervals, more details

You aren't expected to be able to reproduce any of the rest of this section, but I am obligated to show you how and why a confidence interval works. You will have to be able to determine which type of confidence interval to use in a given situation, including the possibility that you don't have a known way to form a confidence interval, and then, if possible, find the confidence interval. Here goes:

We have a population, and a sample of size n taken from that population, usually because we are interested in some parameter θ. For example, we may be interested in μ (the population mean) or σ^2 (the population variance); in these two cases, $\theta = \mu$ and $\theta = \sigma^2$, respectively. So, at times, we will speak in general of a parameter θ, but keep in mind that θ represents a specific population parameter.

Generally, we are given α, where $0 < \alpha < 1$, and we construct a confidence interval based on α. In this context, α is a probability (so it is an area) The interval that we construct is written in the form (L, U). Here, L and U are numbers where L = "lower limit" and an U = "upper limit". The main idea is to construct this interval (L, U) so that $P(L< \theta < U) = 1- \alpha$ (this is actually the confidence level - we spoke about it above). When we find L and U, we are actually finding them so that the probability that θ lies between L and U is $1- \alpha$.

Notice that α is specifying the error level. So $\alpha = 0.35$ or 35% means a 65% confidence interval.

Let's recap the previous paragraph: if θ is my parameter of interest and α is given, then we need to find L and U so that

Equation 7.0.0 $$P(L< \theta < U) = 1- \alpha.$$

Common choices for α are 0.10, 0.05, 0.025, and 0.01 which correspond to 90%, 95%, 97.5% and 99% confidence intervals, respectively. Recall that to find the area between two numbers on the standard normal curve, we find the z-scores (standardized scores) and calculate the area accordingly. Recall also that there is a one-to-one correspondence between area and probability. Here's an how it goes for confidence intervals.

Example 7.0.0

Consider the standard normal distribution. Find the confidence intervals corresponding to

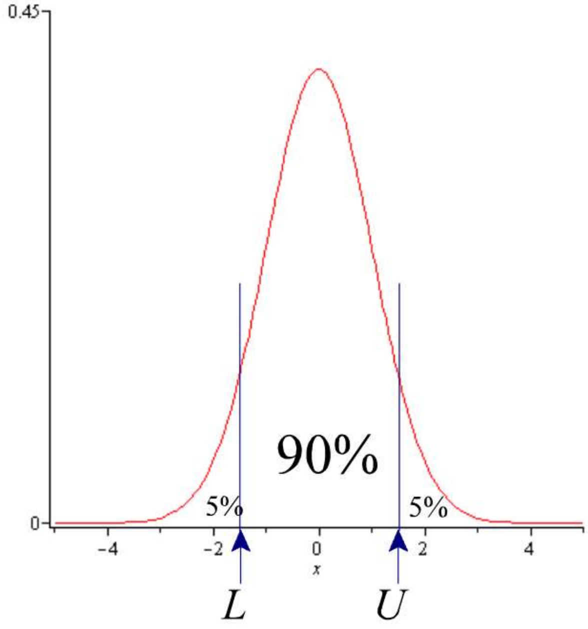
(a) $\alpha = 0.10$, (b) $\alpha = 0.05$, (c) $\alpha = 0.025$, (d) $\alpha = 0.01$.

Have Table E handy, we're going to need it. Here's what you should be thinking: we have a standard normal random variable *X* and we are building confidence intervals to capture *X* with a certain probability.

(a) For $\alpha = 0.10$, we have a confidence level of 90%. This means we need to find *L* and *U* so that

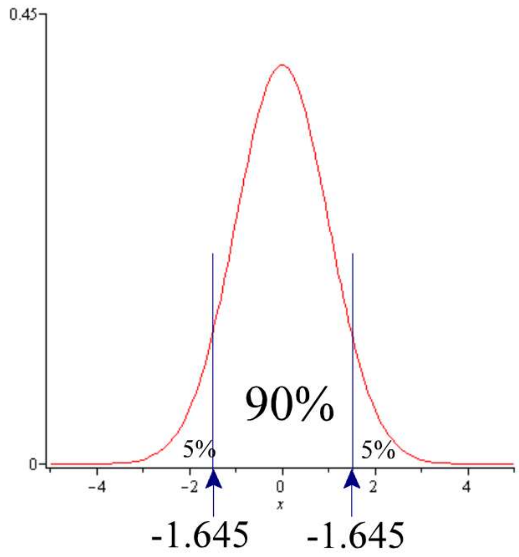
$$P(L < X < U) = 0.90.$$

Let's look at a picture:



We need to find *L* and *U*. Well, since this is the standard normal curve, we use Table E. We need to find the scores that match the areas in the picture. So, we find *L* by looking in the table for the number 0.05; this is the area under the curve up to *L*. The number *L* is the corresponding score.

Looking in Table E, I can find 0.0495, which has a matching score of -1.64, and 0.0505, which has a matching score of -1.65. So, we just take *L* to be -1.645. Now, you can spend time looking up the score that matches 0.95 (this will be *U*), or you can just use the symmetry of the curve and see that *U* = 1.645. So now our picture looks like this:

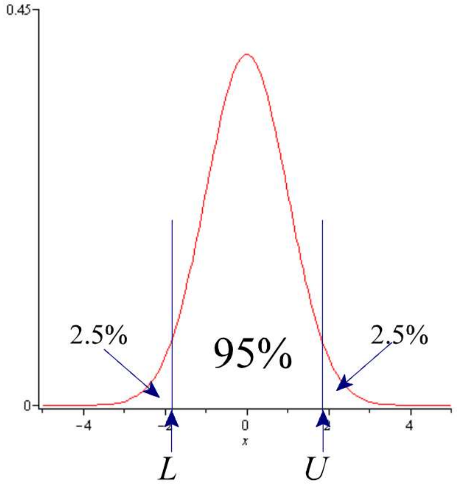


We have found that P(-1.645 < *X* < 1.645) = 0.90, which is what we set out to do. Thus, a 90% confidence interval for *X* is (-1.645, 1.645). This can be written as ±1.645 (and often is).

(b) For α = 0.05, we have a confidence level of 95%. This means we need to find *L* and *U* so that

$$P(L < X < U) = 0.95.$$

Let's look at a picture:



We need to find *L* and *U*. Again, we use Table E. We need to find the scores that match the areas in the picture. So, we find *L* by looking in the table for the number 0.025; the matching score is -1.96. So, we just take *L* to be -1.96. Now, you can spend time looking up the score that matches 0.975 (this will be *U*), or you can just use the symmetry of the curve and see that *U* = 1.96.

Thus, a 95% confidence interval for *X* is (-1.96, 1.96) or ±1.96.

(c) For α = 0.025, we have a confidence level of 97.5%. This means we need to find *L* and *U* so that

$$P(L < X < U) = 0.975.$$

Following the same process as in (a) and (b), we put an area of α/2 in each of the tails, i.e., put 1.25% in each tail. The corresponding scores are ±2.33. Thus, a 97.5% confidence interval for *X* is (-2.33, 2.33) or ±2.33.

(d) For α = 0.01, we have a confidence level of 99%. This means we need to find *L* and *U* so that

$$P(L < X < U) = 0.99.$$

Following the same process as in (a) and (b), we put an area of α/2 in each of the tails, i.e., put 0.5% in each tail. The corresponding scores are ±2.575. Thus, a 99% confidence interval for *X* is (-2.575, 2.575) or ±2.575.

Some very important standardized scores for us will be ± 1.645, ± 1.96, ± 2.33, and ± 2.575. This is because they correspond to the most commonly used confidence levels.

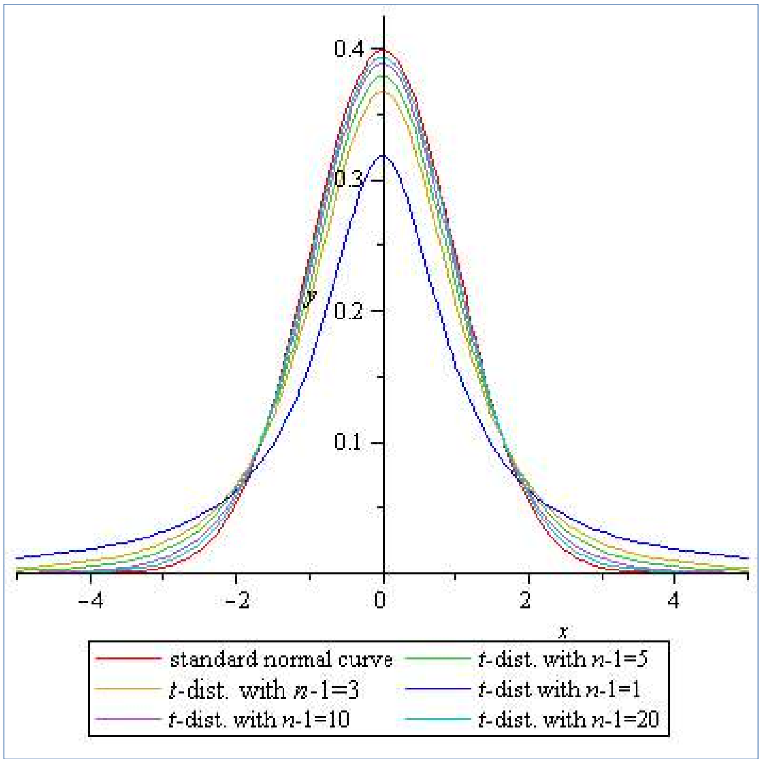
Note also that when you raise the confidence level, the intervals get wider.

Of course, it is not always so simple to construct a confidence interval because our population is not always distributed *N*(0,1). The way we find *L* and *U* changes depending on the sample size, the type of parent distribution, the parameter(s) we wish to know more about, and what is already known about the distribution. We will focus only on a few methods of finding confidence intervals for means and proportions.

In doing so, we will be employing a distribution called Student's *t*-distribution (or just *t*-distribution). This distribution is characterized by the parameter *n* - 1, which we call the "degrees of freedom". The Student *t*-distribution is very close to normal when the degrees of freedom, *n* - 1, are large (so the CLT applies) but it is significantly different from normal when the degrees of freedom is small (refer to image below). Why we must use this in certain situations, and proof that it works, are well beyond the scope of this class, so you'll just have to trust me on this. On the right is a graph of the normal distribution plotted against the graphs of t-distributions with varying degrees of freedom.

BTW, it is interesting to note why this distribution is called "Student *t*." See the History and Etymology section of the Student t distribution here:
http://en.wikipedia.org/wiki/Student's_t-distribution .

Make sure that you know how to read the *t*-distribution table in your book. It is very similar to reading a normal table, but if you have any difficulty, let me know immediately.



Example 7.0.1

Put simple example here of reading *t*-table.

7.1 Confidence Intervals for the Mean

In this section, we discuss how to construct confidence intervals for the mean in four different situations. Before doing so, I'd like to say a little something about sample size (again).

We will, in general, assume that when the sample size is larger than 30, that the sample average is distributed like a normal random variable via the CLT. It is simply for convenience. In many situations it isn't a safe assumption, and in some others 30 is complete overkill. We assume this for the sake of learning the associated concepts. Following are the four cases we will consider:

Case 1. Our sample is from a normal population, and it's variance is *known*.
Case 2. Our sample is from a normal distribution, it's variance is *unknown*, and $n \geq 30$.
Case 3. Our sample is from a "well-enough behaved" distribution and $n \geq 30$.
Case 4. Our sample is from a normal population, it's variance is *unknown*, and $n < 30$.

Case 1. Our sample is from a normal population, and it's variance is *known*.

The confidence interval is given by:

$$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The smaller number is the lower limit and the larger number is the upper limit. Note this expression can also be written as the interval:

$$\left(\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \ , \ \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

Example 7.1.0

Suppose we take a sample of $n = 49$ from a normal population and it is known that the population variance is $\sigma^2 = 16$. We compute a sample mean of = 10. Note that these are the conditions of Case 1 above. Find the confidence interval for the population mean for α = 0.05 and also for α = 0.01.

a. $\alpha = 0.05$

We need to find $z_{0.025}$. Finding the z value closest to having an area of 0.025 is -1.96. So, using the formula we we have:

$$10 \pm 1.96 \frac{4}{\sqrt{49}} = 10 \pm 1.12$$

As an interval this is (8.88 , 11.12). This is the 95% confidence interval.

So, our confidence interval for μ at level 1-0.05 = 0.95 is (8.88, 11.12). What does this mean? Many people like to say that this means there is a 95% chance that the population mean falls within the interval (8.88, 11.12). But, this is actually a somewhat vaguely incorrect statement. A more correct statement would be that we have 0.95 confidence that the interval (8.88,11.12) captures the true population mean, μ. The distinction is subtle and is beyond the scope of this class, but for those interested, technically, it means if we were to sample the same population an infinite number of times in the same manner, the resulting confidence intervals would capture the true mean 95% of the time.

b. $\alpha = 0.01$

We need to find $z_{0.005}$. From an example above, we get that $z_{0.005} = -2.575$, so using our formula, we get

$$10 \pm 2.575 \frac{4}{\sqrt{49}} = 10 \pm 1.47$$

As an interval this is (8.53 , 11.47). This is the 99% confidence interval.

So, a 99% confidence interval for μ is (8.53, 11.47). Notice again that raising our confidence level made the interval larger. This is the price we pay for higher confidence.

Case 2. Our sample is from a normal distribution, it's variance is *unknown*, and n is large. ($n \geq 30$)

Since *t*-distributions with many degrees of freedom, i.e., with *n* large enough, are very close to being normal, we can use the theory as given in Case 1 but with the sample variance used instead. I will not do an example of this since it would follow Example 7.1.0 exactly with σ replaced by s, the sample variance.

The endpoints of the interval are given by:

$$\overline{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Case 3. Our sample is from a "well-enough behaved" distribution and $n \geq 30$.

Since our parent distribution is "well-enough behaved," the sample mean is close to being normally distributed and so the confidence interval can be taken as in Case 1 if the variance is known or Case 2 if the variance is unknown. (Every distribution that has a finite variance will be "well-behaved" for *n* large enough although it is sometimes much larger than 30.)

So, if the parent distribution is "well-enough behaved," we use either formula from case 1 or the one from case 2, depending on if the population standard deviation is known.

Case 4. Our sample is from a normal population, it's variance is *unknown*, and $n < 30$.

We must use the *t*-distribution when the variance is unknown and the sample is small. We will not derive this; just know that both endpoints of the interval are given by the formula

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \quad \text{where } t_{\alpha/2} \text{ has n – 1 degrees of freedom.}$$

$t_{\alpha/2}$ with $n - 1$ degrees of freedom is the score from the *t*-table that corresponds to the area α/2 with *n*-1 degrees of freedom, and *s* is the sample standard deviation.

Example 7.1.1

Suppose our sample is from a normal population and we do not know the population variance. Our sample size is $n = 25$, and we compute a sample mean of 10 and a sample standard deviation of 2. Find the following confidence intervals for the population mean:

a. 90% confidence interval

Since we are looking for a 90% confidence interval, our $\alpha = 0.10$, which means $\alpha/2 = 0.05$. Our degrees of freedom is $n - 1 = 24$, so we need to find $t_{0.05}$. From a *t*-distribution table, we get that $t_{0.05} = 1.711$ and using our formula for a t confidence interval, we have:

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 10 \pm 1.711 \frac{2}{\sqrt{25}} = 10 \pm 0.6844$$

So, a 90% confidence interval for μ is (9.3156, 10.6844).

b. 95% confidence interval

Since we are looking for a 95% confidence interval, our $\alpha = 0.05$, which means $\alpha/2 = 0.025$. Our degrees of freedom is $n - 1 = 24$, so we need to find $t_{0.025}$. From a *t*-distribution table, we get that $t_{0.025} = 2.064$. Doing the same as in part a, we have:

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 10 \pm 2.064 \frac{2}{\sqrt{25}} = 10 \pm 0.8256$$

So, a 95% confidence interval for μ is (9.1744, 10.8256). Note again that, for higher confidence, we have a larger interval.

7.2 Confidence Intervals for Proportions

We assume here that the population consists of 0's and 1's. A situation where we are counting successes and failures. Each trial, we either succeed with probability *p* or we fail with probability 1-*p*. We wish to construct a confidence interval for *p*, the true proportion of successes. Recall that the number of successes is a binomial random variable.

Again, our best estimator of *p* is the sample mean but instead of calling it , we call it . The central limit theorem applies pretty quickly provided *np* > 5, and *n*(1-*p*) > 5 (this is a standard rule of thumb) and so if this is the case, we use the *z*-table.

A 1- α confidence interval for *p* is given by

$$\hat{p} \ \pm\ z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{where} \quad \hat{p} \quad \text{is the percentage of successes.}$$

This formula results from the facts that binomial random variables with sufficiently large n are approximately normally distributed with mean *np* and variance *np*(1-*p*). We will not derive this here, though the derivation is fairly straightforward.

Aside, not needed for an introductory statistics class.
In many books (including yours), you will find mention of the *continuity correction*. This is simply a numerical correction to account for the fact that we are approximating a discrete probability function (the binomial) with a continuous probability function (the normal). To employ the continuity correction, just subtract 0.5/*n* from the lower limit *L*, and add it to the upper limit *U*. This makes for a larger confidence interval, but it will be more accurate (not much more, just more). With the continuity correction, we get

$$\hat{p} \ \pm\ \left(z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}+\frac{0.5}{n}\right)$$

I won’t ask you to use this.

Example 7.2.0

A sample of 400 people are given a drug, and 278 report that their symptoms are relieved. Find a 90% confidence interval for the true proportion of the population who experience relief of symptoms due to this drug.

This is a confidence interval for a proportion. The data consists of people that have relieved symptoms or not. So the confidence interval will be give by:

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

The table value is 1.645. But what is $\hat{p}$? It is 278/400 = 0.695 or 69.5%

So our confidence interval is:

$$\hat{p} \ \pm \ z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ = \ 0.695 \ \pm \ 1.645 \sqrt{\frac{0.695(1-0.695)}{400}}$$

$$\approx \ 0.695 \pm 1.645 * 0.02302$$

$$\approx 0.6571, 0.7329$$

So our 90% confidence interval for the true population proportion is given by (0.6571, 0.7329).