Chapter 23: Inferences About Means

Enough Proportions!

We've spent the last two units working with proportions (or qualitative variables, at least)—now it's time to turn our attentions to quantitative variables.

For qualitative variables, the parameter (when there was one) was the population proportion. Now, there are two parameters—the population mean and the population variance. Inference for variance is beyond the scope of this course, so we'll only concern ourselves with the mean.

The statistic that estimates the population mean is the sample mean. Naturally, this is a random variable…and as such, we need to know something about the sampling distribution of this statistic.

Some Theory About the Sample Mean

A Reminder

Recall from a previous chapter that the sampling distribution of $\bar{x}$ has mean $\mu_{\bar{x}} = \mu_x$, variance $\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n}$, and an approximately normal shape under certain circumstances. We now need to revisit this idea with an eye towards reality.

The key for the mean is that the equation is true; that the equals sign was correct…it actually doesn't affect our (upcoming) calculations to *not* know the value of μ_x.

The variance equation was only valid if the sample was small relative to the population (less than 10%). That continues to be true, and also continues to be the least of our worries. The bigger issue is that we *need* this value for our upcoming calculations.

How often are we going to know the variance of the population?

Never!

A Complication

We're certainly not going to stop and cry about this! There is a way around it—in fact, we've encountered the problem before. A few chapters ago, we let go of the parameter p and started using the statistic $\hat{p}$ in its place.

Logically, then…now that we don't have the value of σ_x^2 (or σ_x), what should we use in its place?

The statistic, of course! Let's replace σ_x with s_x. Just like before, we'll start calling this thing the **standard error**: $SE_{\bar{x}} = \frac{s_x}{\sqrt{n}}$.

Previously, switching $\hat{p}$ for p had no effect on the shape of the distribution because $\hat{p}$ was an unbiased estimator of p. Wouldn't it be cool if s_x was an unbiased estimator of σ_x?

Alas, it isn't. The larger the sample, the smaller the variation; thus, the value of s_x will typically get smaller as the sample size increases…right up until the point that the sample is the population, and s_x becomes σ_x.

What you should take from that is that σ_x is typically smaller than s_x. One direct result of this is that $SE_{\bar{x}}$ will typically be larger than $\sigma_{\bar{x}}$.

The other direct result of this concerns the Central Limit Theorem. It said that the sampling distribution approached normal with standard deviation $\frac{\sigma_x}{\sqrt{n}}$. We are now replacing that existing standard deviation with one that is larger…that means that the shape of the distribution will be different!

Another Complication

The Central Limit Theorem says that $\frac{x-\mu_x}{\frac{\sigma_x}{\sqrt{n}}}$ has an approximately standard normal shape; what kind of shape does $\frac{x-\mu_x}{\frac{s_x}{\sqrt{n}}}$ have?

Fortunately, a very smart guy figured this out a long time ago. He derived a new distribution, which he called Student's *t* (be sure to read your textbook for the full story). This distribution looks a lot like the standard normal, but with fatter tails—and the shape changes as the sample size increases! We saw this before in our study of Chi Square, and we saw how the idea was handled in terms of the graph: **degrees of freedom**. It turns out that the degrees of freedom for the *t* distribution are (for now) $n-1$.

Finding Probabilities

You need to be able to find probabilities for a *t* distribution. Happily, this skill is identical to finding probabilities for a Chi Square distribution!

When using the chart, first find the degrees of freedom down the left hand side. Next, find the spot where the given statistic value ought to lie. Then, look up to find the right hand area. Finally, make sure that you actually answer the question that was asked (this may involve symmetry and the complement).

When using the calculator, know that tcdf() works identically to chisqcdf()!

Examples

[1.] Find $P(t>2)$ if $df=15$.

I get an exact answer of 0.03197, and a chart answer between 0.025 and 0.05.

					Right Tail Area		
df	0.2500	0.2000	0.1500	0.1000	0.0500	0.0250	0.0200
1	1.0000	1.3764	1.9626	3.0777	6.3138	12.7062	15.8945
2	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	4.8487
3	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	3.4819
4	0.7407	0.9410	1.1896	1.5332	2.1318	2.7764	2.9985
5	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	2.7565
6	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	2.6122
7	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.5168
8	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.4490
9	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.3984
10	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.3593
11	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.3281
12	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.3027
13	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.2816
14	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.2638
15	0.6912	0.8662	1.0735	1.3406	1.7531	2.1314	2.2485

Figure 1 - T Table Excerpt for Example 1

[2.] Find $P(t < 2.5)$ if $df = 20$

I get an exact answer of 0.9894, and a chart answer between 0.975 and 0.99. You must use the complement for this one, since the question asked for left hand area but the chart only gives right hand area.

[3.] Find $P(t < -1.75)$ if $df = 5$

I get an exact answer of 0.0703, and a chart answer between 0.05 and 0.1. You must use symmetry for this one, since the chart only uses positive values of *t*.

[4.] Find $P(t > -3)$ if $df = 6$

I get an exact answer of 0.988 and a chart answer between 0.975 and 0.99. You must use symmetry and the complement for this one.

A Confidence Interval for the Mean

So…how does this affect our procedures for confidence intervals?

The Formula

$\bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$ with $df = n - 1$.

The Conditions

The *t* procedures require that the sample was obtained randomly, that the sample is small enough (the 10% condition), and the variable has a normal distribution in the population.

Yet Another Complication

As before, we will often assume that the sample was obtained randomly. Also, I'll keep the second condition in mind, but I'll rarely mention it.

The last condition will almost certainly fail! Fortunately, the *t* procedures are what we call **robust**. That means that they still give reasonably accurate results even when the conditions are violated. This does *not* mean that we will simply plow ahead and forget about the condition—rather, it means that what we need to check is going to be slightly different from one problem to another.

If the population is normal, then we are good to go. If the population is approximately normal, or not very non-normal, then the robust nature of the *t* procedures will allow us to continue. If the population is clearly (or incredibly) non-normal, then we'll only be able to continue if the sample size is large (because as the sample size increases, the closer we get to a situation where the Central Limit Theorem kicks in).

…but how can we know anything about the population? How can we determine if it is OK to continue if we don't have the whole population to look at?

Think, *think*! What have we done in the past cases?

We replaced the population information with sample information.

Thus, if we can't look at the shape of the population, we should instead look at the shape of the sample!

The Solution

For small samples (say, less than 15), we need a fairly normal population—or, in terms of the sample, there cannot be any clear indication of skewness.

For slightly larger samples (say 15 through 40), we need a population that isn't too skew—so we need to see a sample that isn't too terribly skew. No, you can't make that any more precise! Deal with it.

For larger samples, we almost don't care what the population looks like—so we could have almost any amount of skew in the sample.

How are you going to decide if there is skew? If you have data, then you'll need to graph the data—and that means that you'll have to draw the graph as part of your answer. If you don't have the data, then you're going to have to make an assumption about the population. Make sure that you don't assume too much! Only assume as much as is needed in order to move forward with the procedure.

In all cases, outliers are an issue…*in reality*. As far as AP is concerned, outliers should not stop you from performing the procedure.

The Conditions (AP Exam Version)

Alas, what has appeared in the scoring rubrics on the AP Exam isn't exactly what I've described—specifically with regards to the shape requirement. Here's what is typically expected:

We need the sample to be a random sample from the population, and either a normally distributed population or a large sample size.

All of the scoring rubrics result in either a very small sample size (in which case you should check to see if the sample shows any sign of skew, or assume something about the population) or a quite large sample size (in which case the requirement is met). I haven't seen any with a medium-sized sample where you must either see a not-too-skew sample or assume a not-too-skew population.

So…I'm going to go ahead and work examples the way I'd like to see them in class. Be aware that the same answers on the AP Exam might not receive full credit.

Example

[5.] During an study on car safety, the braking distance (feet) was measured for a car traveling at several different speeds. The data are as follows:

Table 1 - Braking Distances for Example 5

2	10	4	22	16	10	18	26	34	17	28	14	20	24	28	26	34
26	36	60	80	20	26	54	32	40	32	40	50	42	56	76	84	36
32	48	52	56	64	66	54	70	92	93	120	85	46	68	46	34	

Construct a 95% confidence interval for the population mean braking distance for these cars.

This calls for a one sample *t* interval for the true mean. This requires that the sample was obtained randomly and that the population variable is normally distributed. I'll have to assume that the sample was obtained randomly. I don't know anything about the population distribution, but with a sample size of 50 I'll almost certainly be able to continue regardless of how the sample distribution looks. I'll take a look anyway…

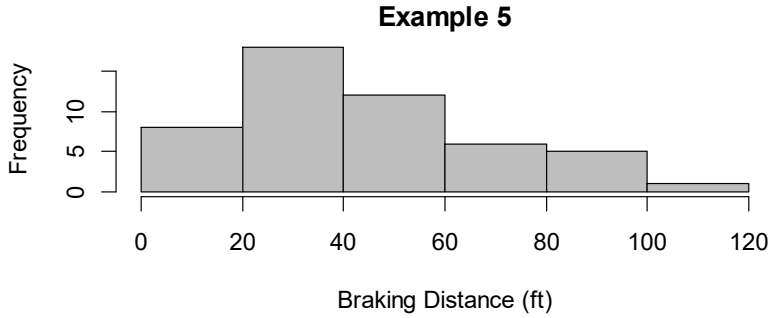

Figure 2 - Histogram for Example 5

Nothing in the sample indicates a problem—I should be able to continue.

With 95% confidence and 49 degrees of freedom, $t^* = 2.01$.

The interval is $\bar{x} \pm t^* \frac{s_x}{\sqrt{n}} = 42.98 \pm 2.01 \frac{25.77}{\sqrt{50}} = (35.656, 50.304)$

I am 95% confident that the population mean braking distance is between 35.656 feet and 50.304 feet.

A Hypothesis Test for the Mean

This will follow the same pattern as the other tests we've learned.

The Hypotheses

We'll assume that the parameter (μ) has some specific value (μ_0). The alternative will be one of the three inequalities. Be sure to explicitly define the parameter!

$H_0 : \mu = \mu_0$

$H_a : \mu \boxed{?} \mu_0$

The Conditions

The conditions here are the same as for the interval—the sample was obtained randomly, that the sample is small enough (the 10% condition), and the variable has a normal distribution in the population.

As was the case before, that last condition will fail. You'll need to graph the data or make an appropriate assumption in order to continue.

Be sure to read my earlier comments about how what I'm writing might not be exactly what is expected on the AP Exam!

The Mechanics

$t = \dfrac{\overline{x} - \mu_x}{\dfrac{s_x}{\sqrt{n}}}$ with $df = n - 1$. Calculate the *p*-value the using the *t*-distribution much as you did for one sample proportion tests. Be sure to explicitly state the level of significance that you will be using.

The Conclusion

The conclusion is much the same as it was in previous procedures!

If [null hypothesis] then I can expect to find [probability statement] in [*p*-value] of repeated samples. Since [$p < \alpha$ / $p > \alpha$], this occurs [too rarely / often enough] to attribute to chance at the [α] level; it is [significant / not significant], and I [reject / fail to reject] the null hypothesis. [conclusion in context—make a statement about the alternate hypothesis].

Example

[6.] The girth (diameter; measured in inches) of 31 black cherry trees was measured. The data are as follows:

Table 2 - Cherry Tree Data for Example 6

8.3	8.6	8.8	10.5	10.7	10.8	11.0	11.0	11.1	11.2	11.3
14.2	14.5	16.0	16.3	17.3	17.5	17.9	18.0	18.0	20.6	
12.9	13.3	13.7	13.8	11.4	11.4	11.7	12.0	14.0	12.9	

Do these data provide evidence that the population mean girth is different from 12 inches?

I'll let μ represent the population mean girth of a Cherry tree.

$H_0 : \mu = 12$ (the population mean girth is 12 inches)

$H_a : \mu \neq 12$ (the population mean girth is not 12 inches)

This calls for a one sample *t* test for the population mean. This requires that the sample was obtained randomly and that the population variable is distributed normally.

I'll have to assume that the sample was obtained randomly. I don't know how the population is distributed, but with a sample size of 31 I should be able to continue in almost any case…I'll go ahead and look at the data anyway.

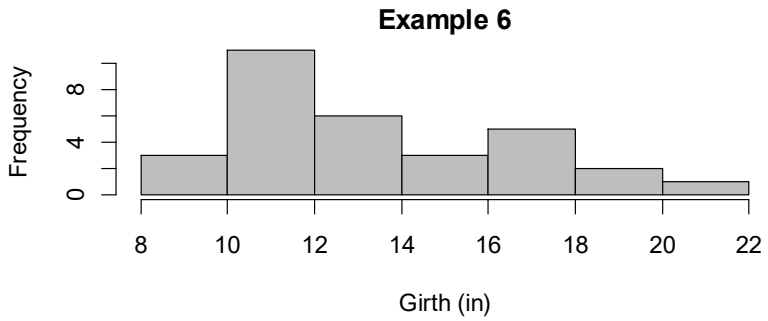

Figure 3 - Histogram for Example 6

The skew here is OK—I can continue.

I'll use $\alpha = 0.05$.

$t = \dfrac{\overline{x} - \mu_x}{\dfrac{s_x}{\sqrt{n}}} = \dfrac{13.248 - 12}{\dfrac{3.138}{\sqrt{31}}} = 2.215$. With $df = 30$, $2P(t > 2.215) = 0.0345$.

If the population mean girth is 12 inches, then I can expect to find a sample with a mean girth less than 10.75 inches or greater than 13.248 inches in about 3.45% of samples.

Since $p < \alpha$, this occurs too rarely to attribute to chance at the 5% level. This is significant; I reject the null hypothesis.

The data do provide evidence that the population mean girth is different from 12 inches.