Stats 9.2: Still another statistic/parameter pair!

So far we know about confidence intervals and hypothesis tests for the following parameters:

μ, p, & μ_1-μ_2

Which means we are missing one: CI or HT for p_1-p_2

Let's line them up this way:

μ	p
μ_1-μ_2	p_1-p_2

The top two are called "1 sample tests/intervals." We are comparing either μ or p to some suspected value, or maybe we are comparing to some historical value. The other two are called "2 sample tests/intervals." We are collecting samples from two different populations and checking to see if those populations are the same.

Example: The average age of ENC students has always been 22.4, but we don't think that is true this year.
This is a 1-sample test/interval for μ. Can you see that we have this historical value of 22.4, and we want to see if that is still true? We are comparing our current population to history.

Example: This coin should be fair, but I don't think it is.
This is a 1-sample test/interval for p. Can you see that we know the proportion of heads should be 0.5, but we don't think that's true.

Example: The average IQ at ENC is the same as at MIT.
This is a 2-sample test/interval for μ, that is μ_1-μ_2. We are comparing the means of two independent populations.

Example: The proportion of Republicans in Quincy is the same as the proportion in Weymouth.
This is a 2-sample test/interval for p, that is p_1-p_2. We are comparing the proportions in two independent populations.

So now we attack our problems this way:

Is it a confidence interval problem or a hypothesis test problem?
Is it an average problem or a proportion problem?
Are we looking at one sample compared to history, or two samples compared to one another?

We need to construct our statistic to work with our parameter p_1-p_2. I hope it doesn't surprise you that the statistic will be $\hat{p}_1$-$\hat{p}_2$. That is, we will take a sample proportion from both groups and subtract them to estimate p_1-p_2.

Here's the theory, but you won't have to repeat this to me:

We remember (maybe) that the statistic $\hat{p}$ is ND(p, $\sqrt{pq/n}$), under the right conditions. (The conditions were: $np \geq 5$ and $nq \geq 5$.)

So, under the same conditions, $\hat{p}_1$ is ND(p_1, $\sqrt{p_1q_1/n_1}$) and $\hat{p}_2$ is ND(p_2, $\sqrt{p_2q_2/n_2}$) .

We saw recently that you can subtract two normal distributions, and you get another normal distribution. The new mean is the difference of the old means, and the new variance is the sum of the old variances. All of that adds up to:

$\hat{p}_1$-$\hat{p}_2$ is ND(p_1-p_2, $\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}$)

We are almost ready for our confidence interval formula and our hypothesis testing statistic. But a couple details. In both cases we will need to calculate the standard deviation in that formula just above, and we will need p_1, q_1, p_2 & q_2. But we don't have those values, and if we would, we wouldn't need statistics. So:

a) When we do the confidence interval, we will make the best substitution for those parameters that we can, which means we will use $\hat{p}_1$, $\hat{q}_1$, $\hat{p}_2$ and $\hat{q}_2$. So we plug those into the standard deviation, and now we can set up our confidence interval formula.

$$\hat{p}_1-\hat{p}_2 \pm z \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

It is z (not t) because only $\overline{X}$ turns into t, not $\hat{p}$.

(Next page to see what we do for hypothesis tests.)

b) When we do the hypothesis test, we will also need to substitute for p_1, q_1, p_2 & q_2, but this time we have a little extra information that will help us. Since our H_o will always be $p_1=p_2$, and we calculate our test statistic assuming that H_o is true, we incorporate that idea into our formula. That is, we assume both groups of data are calculating the same proportion, and we "pool" the data together and calculate just a single "pooled estimate", called $\hat{p}$, for the common value of p_1 & p_2. If we know $\hat{p}_1$ comes from a/n_1, and $\hat{p}_2$ comes from b/n_2, then the pooled estimate $\hat{p} = (a+b)/(n_1+n_2)$. If we don't know those things, we can find $\hat{p}$ using $\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$. Now our test statistic will be $z_{\hat{p}_1 - \hat{p}_2} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}}$

End of the theory. Beginning of the examples.

Example 1. We want to know if the proportion of voters who approve of the governor's job in Quincy is the same as that in Weymouth. In a random sample of 100 voters in Quincy, 68 approved of the job performance of the governor. From 80 voters in Weymouth, 49 approved. Construct a 95% confidence interval

Solution: My comments first. We see this is a confidence interval problem. We can see that it is a proportion problem, and we can see that it is comparing two independent groups, so it's a 2-sample problem. We see that $\hat{p}_1 = 68/100 = 0.68$ and $\hat{p}_2 = 49/80 = 0.613$. Since it is 95%, and we are using z, the z-value is 1.96. So:

95% CI for $p_1 - p_2$.

$$\hat{p}_1 - \hat{p}_2 \pm z \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} \Rightarrow (0.68 - 0.613) \pm 1.96 \sqrt{\frac{0.68x0.32}{100} + \frac{0.613x0.387}{80}}$$

$$\Rightarrow 0.067 \pm 1.96x0.072 \quad \Rightarrow \quad 0.067 \pm 0.141 \quad \Rightarrow \quad [-0.074, 0.208]$$

A follow-up question: Are the towns the same or not?
Remember that every value inside the confidence interval is a reasonable value for the parameter. Since 0.0 is in the interval, it is a reasonable value for $p_1 - p_2$, and $p_1 - p_2 = 0$ means the proportions are the same. So we do indeed conclude that the towns are the same (in this respect).

Example 2. Adverse side effects are always a concern when testing and trying new medicines. Placebo-controlled clinical studies were conducted in patients 12 years of age and older who were receiving "once-a-day" doses of Allegra, a seasonal allergy drug. The following results were published in the April 2005 edition of Reader's Digest. Test if there is a difference between the two groups in the proportion of patients reporting headaches.

Allegra group	$n_1 = 283$	30 reporting headaches
Placebo group	$n_2 = 293$	22 reporting headaches

Solution. My comments first. We see this is a hypothesis testing problem. We can see that it is a proportion problem, and we can see that it is comparing two independent groups, so it's a 2-sample problem. Since this is a z problem, and we will choose $\alpha = 0.05$, then we will get $z_C = 1.96$. $\hat{p}_1 = 30/283 = 0.106$ and $\hat{p}_2 = 22/293 = 0.075$. Also, $\hat{p} = (30+22)/(283+293) = 52/576 = 0.090$.

HT for p_1-p_2

H_o: $p_1 - p_2 = 0$
H_1: $p_1 - p_2 \neq 0$

$\alpha = 0.05$ (We choose this because one was not given.)

DR: If $z_{\hat{p}_1-\hat{p}_2} > 1.96$ or $z_{\hat{p}_1-\hat{p}_2} < -1.96$ then rej H_o.

$$z_{\hat{p}_1-\hat{p}_2} = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\hat{p}\hat{q}}{n_1}+\frac{\hat{p}\hat{q}}{n_2}}} = \frac{(0.106-0.075)-(0)}{\sqrt{\frac{0.09x0.91}{283}+\frac{0.09x0.91}{293}}} = 0.031/0.024 = 1.29$$

Fail to reject H_o.

The two proportions are the same.

Homework on the next page

Homework for this section

Section 9.2: # 11,12,14,17,18,23
Notice that some of these problems are hypothesis tests, and others are confidence intervals. Make sure you take the correct approach.

Your responses to me:

On #11 & 12, there is no calculating to do. Look at the given display, and draw your conclusion directly from that. (Look at the p-value.) Email your conclusion to me.

On #14 & 17, email the $z_{\hat{p}_1-\hat{p}_2}$ value to me, as well as your two conclusions.

On #18 & 23, email me the final confidence interval.

If you want me to look over anything else, you can scan/photograph your work and email it to me, but that is not required.

If you feel like this approach to our class is not working well for you, please let me know. Tell me what you think might be better for you.