Class 17: Inference about Proportions (Text: Section 8.1)

Quantitative and Categorical Variables

So far (in Chapters 6 and 7), we have made confidence intervals and done hypothesis tests for *means* of quantitative variables. Now we look at categorical variables, where we are interested in *proportions*.

Distribution of Proportions: Recall the Central Limit Theorem:

The sampling distribution of proportions $\hat{p}$ in a sample of size n is

- Approximately normal, for $n > 30$
- Mean is p, the population proportion
- Standard error is $SE = \sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion

We assume: $n > 30$ and $np \geq 10$ and $n(1-p) \geq 10$ if we make these approximations. That is, sample size must be large enough and p cannot be too small or too close to 1.

HYPOTHESIS TESTS

For a null hypothesis $H_0: p = p_0$, use

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Example: Testing the Malaria Vaccine

During the Mozambique trial of the potential malaria vaccine,[1] the effect of the drug was measured on:

- The number of children infected
- The length of time until infection

In a previous class, we looked at the length of time till infection; now we look at the proportion of children infected.

Ex: Are the variables quantitative? Or categorical?
Number getting infected: Categorical;
Length of time: Quantitative

Ex: What kind of *p*-value would you expect to see if the drug were **not** effective?
Large, above 5% (or above the significance level)
What kind of *p*-value would you expect to see if the drug **were** effective?
Small, below 5% (or below significance level)

[1] "Efficacy of the RTS,S/AS202A vaccine against *Plasmodium falciparum* infection and disease in young African children: randomized controlled trial" by P. Alonso et al, *The Lancet*, Oct 16, 2004.

Ex: Without the drug, the rate of severe malaria infection in the area of the study was 34.9 children per 1000. Of 745 children given the drug, 11 got severe malaria during the course of the study. Does this data suggest that the drug reduces the rate of severe malaria infections?
Note that 34.9/1000 = 0.0349 = 3.49%. This is the infection rate for the untreated population.
Let p be the proportion of treated children in the population who get sick if we treated all of them.
Step 1:
Null hypothesis: We assume drug does not work: $H_0: p = p_0 = 0.0349$
Proportion of treated population getting sick = Proportion of untreated population getting sick = 0.0349.
Alternate hypothesis: One sided because we want to know if the infection rate is reduced.
$H_a: p < p_0 = 0.0349$ Proportion of treated group getting sick is less than 0.0349.
Step 2:
Find $\hat{p}$, then the standard error of $\hat{p}$, and then z:

$$\hat{p} = \frac{11}{745} = 0.0148$$

Assuming H_0, we know that the standard error of $\hat{p}$ is

$$\text{SE} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.0349(1-0.0349)}{745}} = 0.00672.$$

(Note: We use 0.0349 in calculation of SE because we are assuming the null hypothesis is true.)
By the Central Limit Theorem: Since $\hat{p}$ is *normally* distributed (rather than having the t-distribution), we calculate z:

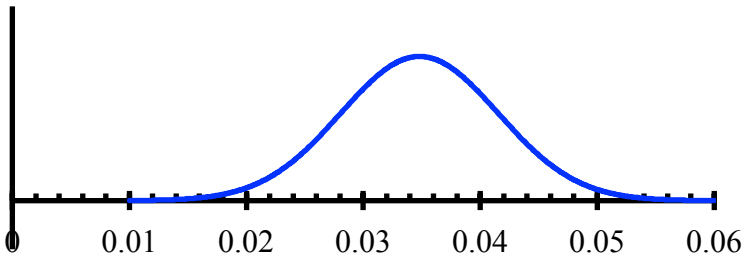
$$z = \frac{0.0148 - 0.0349}{0.00672} = -2.99.$$

Step 3:
The p-value is the probability of getting a sample with a more extreme proportion than 0.0148. The p-value of -2.99 is $p = 0.0014 = 0.14\%$.
Step 4: If null hypothesis were true, it is very unlikely we would see an infection rate as low as 0.0148. So we decide null hypothesis is most likely ***not*** true. We ***reject*** the null hypothesis and conclude the infection rate for the treated population is ***lower*** than for the rest of population. Thus we reject the null hypothesis at the 5% significance level (because 0.14% < 5%)
Ex: For the preceding example, what conclusion can we draw at the 1% significance level?
We reject the null hypothesis at the 1% significance level (because 0.14% < 1%). Thus

- The experiment provides evidence, significant at the 5% and the 1% levels, that the drug reduces infection.
- The results are significant both with $p < 0.05$ and with $p < 0.01$.

This is the kind of evidence the FDA (Food and Drug Administration) wants before authorizing a drug.

Distribution of Sample Proportions:
Mean 0.0349, Std Dev 0.00672



Ex: For the preceding example, interpret the p-value.
If the drug didn't work, there is less than a 1% chance we'd see the results we did.

Ex: What does the data tell us about whether the drug changes (instead of reduces) the length of time to get sick?
The test is now two-sided: The alternate hypothesis depends on what statement we are testing:
If we want to know if the drug reduces the length of time, then the alternate hypothesis is:
H_a: Proportion of treated group getting sick < 0.0349
If we want to know if the drug changes the length of time, then the alternate hypothesis is:
H_a: Proportion of treated group getting sick $\neq 0.0349$.
For this new null hypothesis, the p-value is $2(0.14\%) = 0.28\%$, so the conclusion is the same. It changes the length of time to get sick significantly.

CONFIDENCE INTERVALS

Since $\hat{p}$ has the standard normal distribution, the confidence interval for p, the population proportion, is

$$\left(\hat{p} - z\sqrt{\frac{p(1-p)}{n}}, \quad \hat{p} + z\sqrt{\frac{p(1-p)}{n}}\right)$$

The margin of error $= z\sqrt{\frac{p(1-p)}{n}}$. The value of z depends on what confidence level we are using.

What is the problem with using this formula? We don't know the value of p.
What do we do?

- Use $\hat{p}$ instead of p,…..we hope close
- Use $p = 0.5$ ...…

Estimating the Margin of Error with $p = 0.5$
Using 0.5 instead of p gives an estimate for the margin of error which is slightly too large, but usually very close. This has several advantages:

- The margin of error can be estimated before the sample is taken
- The same margin of error can be used for several questions derived from the same sample.
- It's a safe estimate because it is as large as, or larger than, the real one.

Ex: In 2003, a February 24-26 CNN poll of 1004 Americans found that 59% supported sending troops to Iraq. Find a 90% confidence interval for the proportion of Americans who supported sending troops to Iraq.

The standard error is

$$SE_{\hat{p}} = \sqrt{0.59(1-0.59)/1004} = 0.0155 \text{ or } SE_{\hat{p}} = \sqrt{0.5(1-0.5)/1004} = 0.0158.$$

For 90% confidence, we use $z = 1.645$, so, with the first ME, the confidence interval is

$$(0.59 - 1.645\,(0.0155, 0.59 + 1.645(0.0155)) = (0.564, 0.616) = (56.4\%\ 61.6\%)$$

Ex: Find the margin of error for a 95% confidence interval.

For 95% confidence, $z = 1.96$, so using $p = 0.59$, we have

$$\text{Margin error} = z \cdot SE_{\hat{p}} = 1.96 \cdot 0.0155 = 0.03 = 3\%$$

Or using $p = 0.5$, we have

$$\text{Margin error} = z \cdot SE_{\hat{p}} = 1.96 \cdot 0.0158 = 0.03 = 3\%$$

Notie Polls are usually quoted to be accurate $\pm 3\%$; this means the margin of error is 3%.

Ex: What sample size gives a margin of error of 1% for a 95% confidence interval? (Same data.)

Let the sample size be n. Then

$$ME = z\sqrt{\frac{p(1-p)}{n}}$$

$$0.01 = 1.96\sqrt{\frac{p(1-p)}{n}}$$

Now we either use $p = \hat{p} = 0.59$ or we use $p = 0.5$. Then

$$0.01 = 1.96\sqrt{\frac{0.59(10.59)}{n}}$$

$$n = \left(\frac{1.96}{0.01}\right)^2 0.59(1-0.59) = 9292.8 \approx 9293.$$

Or

$$0.01 = 1.96\sqrt{\frac{0.5(10.5)}{n}}$$

$$n = \left(\frac{1.96}{0.01}\right)^2 0.5(1-0.5) = 9604.$$

Thus a sample size of 9604 will work—and possibly smaller.

Notice the using $p = 0.5$ gives the most conservative estimate.

Ex: A reporter stated 2/3 population were in favor of sending troops to Iraq. Does this poll (59% in sample of 1004) provide support provide support this assertion? Use a 10% significance level and a confidence interval.

We are interested in a 10% significance level and the test is two-sided, so we use the 90% confidence interval, which is $(56.4\%\ 61.6\%)$. Since $2/3 = 66.7\%$ is not in the interval, the true proportion is unlikely (less than 10% chance) to be 66.7%. Thus the poll does not provide evidence of in support of the reporter's claim.

Ex: Use a hypothesis test to answer the previous question.

Step1: Hypotheses: $H_0 : p = 2/3$ and $H_a : p \neq 2/3$.

Step 2: Test statistic:

$$z = \frac{0.59 - 2/3}{\sqrt{\frac{\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)}{1004}}} = -5.15$$

Step 3: Table gives: P-value is < 0.0002 (smallest in table); Calculator gives: P-value = 0.0000001

Step 4: We reject the null hypothesis since the P-value is small. There is evidence that against the fact that 2/3 of the population was in favor of sending troops.