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:

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?

Ex: What kind of p-value would you expect to see if the drug were **not** effective?

What kind of p-value would you expect to see if the drug **were** effective?

[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?

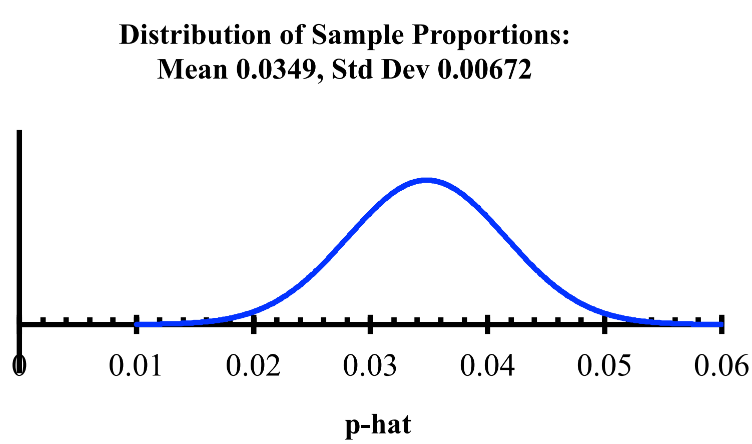
Step 1:

Step 2:

Step 3:

Step 4:

Ex: For the preceding example, what conclusion can we draw at the 1% significance level?



Ex: For the preceding example, interpret the p-value.

Ex: What does the data tell us about whether the drug changes (instead of reduces) the length of time to get sick?

CONFIDENCE INTERVALS

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

The margin of error = The value of z depends on what confidence level we are using.

What is the problem with using this formula?
What do we do?

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.

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

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

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.

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