Inference on two proportions

Lecture 22

03/13/2013

Are the proportions the same?

- Are men or women more likely to go to college?
- Do smokers get lung cancer more often than non-smokers?
- Are minority applicants les likely to be given interviews?
- Are college educated voters more likely you support Barack Obama than voters without a college degree?

Two Independent Samples

What issue is most important to you?

- 5409 voters
  - Economy is most important
  - 2867 support Obama

$$\hat{p}_1 = \frac{2867}{5409} = 0.53$$

- 859 voters
  - War in Iraq is most important
  - 507 voted for Obama

$$\hat{p}_2 = \frac{507}{859} = 0.59$$

- Is there a difference in the support of Obama between these two groups?

Statistical Model

$$X_1 \sim \mathit{Binomial}(n_1, p_1)$$
$$X_2 \sim \mathit{Binomial}(n_2, p_2)$$

- n_1 and n_2 are known
- p_1 and p_2 are unknown

Null and Alternative Hypotheses

- Null = proportions are the same

$$H_0 : p_1 = p_2$$

- Three possible Alternatives

$$\begin{aligned} H_\alpha &: \; p_1 \neq p_2 \quad (two - sided) \\ H_\alpha &: \; p_1 > p_2 \quad (one - sided) \\ H_\alpha &: \; p_1 < p_2 \quad (one - sided) \end{aligned}$$

Example: Hypothesis

- Null:

  The proportions supporting Obama is the same in each sample

$$H_0 : p_1 = p_2$$

- Alternative:

  The proportions are not the same.

$$H_\alpha : p_1 \neq p_2$$

Estimation

- Some algebra on our hypothesis

$$H_0 : p_1 - p_2 = 0$$
$$H_\alpha : p_1 - p_2 \neq 0$$
$$H_\alpha : p_1 - p_2 > 0$$
$$H_\alpha : p_1 - p_2 < 0$$

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

Comparing Sample Proportions

- Economy sample: $n_1 = 5409,\ X_1 = 2867$

$$\hat{p}_1 = \frac{2867}{5409} = 0.53$$

- Iraq sample: $n_2 = 859,\ X_2 = 507$

$$\hat{p}_2 = \frac{507}{859} = 0.59$$

$$\hat{p}_2 - \hat{p}_2 = -0.06$$

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

Assumptions

- $np > 10$ and $n(1-p) > 10$
- The two samples are independent.
- Under the Null, expectation is 0.
- Standard Deviation?

Sums of independent random variables

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2)$$

If H_0 is true...

$$X_1 + X_2 \sim Binomial(n_1 + n_2,\ p),$$

where $p = p_1 = p_2$.

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

In our example

$$\bar{p} = \frac{2867 + 507}{5409 + 859} = 0.5383$$

Standard Deviation

$$
\begin{aligned}
sd(\hat{p}_1 - \hat{p}_2) &= \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} \\
&= \sqrt{p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]} \\
&\approx \sqrt{\bar{p}(1-\bar{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}
\end{aligned}
$$

Test Statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

where

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

Test Statistic

In the previous example...

$$\hat{p}_1 - \hat{p}_2 = -0.06$$

$$\bar{p} = 0.5385$$

$$z = \frac{-0.06}{\sqrt{0.5383(1 - 0.5383)\left[\frac{1}{5409} + \frac{1}{859}\right]}} = -3.28$$

Conclusion

- *p*-value for a *two-sided test*

$$P = 2\mathbb{P}(Z < -3.28) = 2(0.0005) = 0.001$$

- Reject the null at level 0.05.
- There is a significant difference vetweenj the two samples. The voters most interested in the war in Iraq were more likely to vote for Obama.

Bias

- Sample doesn’t represent population:
  - Generalizations are no longer valid.
  - Conclusions may no longer be true

Sources of Bias

- Selection Bias
  - Problem in sampling scheme
  - Difference between population of interest and effective population
- Non-response Bias
  - Subjects don’t answer
  - Skip questions
- Response Bias
  - Subjects lie
  - Interviewer effect

Why those internet polls are worthless

Self-selected sample

- More passionate = More likely to respond
- Minority opinion - more passion
- Selection bias