Statistics ST2334 Topic 9: Inferences Concerning Proportions (Textbook Chap 10a)

2011/2012 Semester 2

Confidence Intervals and Hypothesis Tests for One or Two Proportions. Population Size vs Sample Size.

Proportions

- We are interested in the proportions or percentages of a specified category.
  - What proportion of people in Singapore are smokers?
  - What proportion of people commute by bus?
  - What percentage of prime time television viewer are female?
- To estimate it, as we have done previously, is to take a random sample, and look at the sample proportion.
- And just as before, we should be able to infer confidence intervals, error bounds, conduct hypothesis tests, etc.

In more technical terms,

- In general, we have a population, where some proportion p of them have some desired trait. Let's call these guys good. That is, $100p\%$ of the population is good.
- For now, let's assume the population is large.
- We take a sample of size n from this population.
  - Let Y_i be 1 if the ith observation is good and 0 otherwise.
  - Then $Y_1, \ldots, Y_n$ are independent Bernoulli trials with success probability p.
  - And we have $E(Y_i) = p$ and $Var(Y_i) = p(1-p)$
  - Note that $\sum_{i=1}^{n} Y_i$ is the number of good observations and hence $\bar{Y}$ is the sample proportion, which we call $\hat{p}$.
- So using the sample proportion $\hat{p}$ to estimate the population proportion p is really just using a sample mean $\bar{Y}$ to estimate the population mean $E(Y_i)$.

Same old story?

- If n is large, we know CLT kicks in, and that $\hat{p} = \bar{Y}$ is normal with $E(\hat{p}) = p$ and $Var(\hat{p}) = p(1-p)/n$. So,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

- The big difference between this and what we had before is that in prior cases, σ^2 was either known or estimated. Here, however, it depends on p, the very quantity that we are trying to make inferences about!
- However, since n is large, $\hat{p}$ should be reasonably close to p, we simply plug in $\hat{p}$ and have

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx N(0,1)$$

Confidence Interval for p

- From above, we have

$$P(|\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}| < z_{\alpha/2}) \approx 1-\alpha$$

- Rewriting this we get

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1-\alpha$$

- So a $(1-\alpha)$ CI for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: Energy Saving Awareness

If 36 of 100 persons interviewed are familiar with the tax incentives for installing certain energy-saving devices, construct a 95% confidence interval for the corresponding true proportion p.

- n is large, so our CI expression is valid.
- $\hat{p} = \frac{36}{100} = 0.36$. Hence the 95% CI for p is

$$0.36 \pm z_{\alpha/2}\sqrt{\frac{0.36(1-0.36)}{100}} = [0.266, 0.454]$$

Maximum Estimation error and Sample size

- The error when we use $\hat{p}$ as an estimator of p is given by $|\hat{p} - p|$
- Again using the (approximately) normal distribution, we can assert with probability $1 - \alpha$ that

$$|\hat{p} - p| \leq z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- So with probability $1 - \alpha$ the maximum estimation error is

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- If E is specified in advance, we need n below to achieve the error

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Example: Public Transport Satisfaction

In a sample survey conducted in a large city, 136 of 400 persons answered yes to the question of whether their city's public transportation is adequate. With 99% confidence, what can we say about the maximum error if the sample proportion is used as an estimate of the true proportion? if we want the maximum estimation error to be 0.05 with the same confidence, what is the appropriate sample size?

- $$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.575\sqrt{\frac{0.34 * 0.66}{400}} = 0.061$$
- $E = 0.05$,
$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2 = 595.16$$

  596 samples are needed.

Population Size vs Sample Size

- Suppose City A has population size 1 million, and City B has population size 2 million.
- We want to conduct the transport satisfaction survey in both cities, and wish to have the same maximum estimate of error.
- Do we then have to take double the sample size in City B than in City A?
- No! Notice the expression for E and sample size needed n we worked out does not depend on the population size.
- We would have got the same numbers all else being equal (the true proportions are similar).
- This is unintuitive, but true! The precision of a sample survey is determined by the sample size, and have little to do with the population size.

Hypotheses Concerning One Proportion

- The test of null hypothesis that a proportion equals some specified constant is widely used in sampling inspection, quality control, and reliability verification.
- Null hypothesis $H_0 : p = p_0$

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0,1)$$

-

Null hypothesis	Alternative hypothesis	Reject null hypothesis if
$H_0 : p = p_0$	$H_1 : p < p_0$	$z < -z_\alpha$
$H_0 : p = p_0$	$H_1 : p > p_0$	$z > z_\alpha$
$H_0 : p = p_0$	$H_1 : p \neq p_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

Example: Detonating Explosives

In a study designed to investigate whether certain detonator used with explosives in coal mining meet the requirement that at least 90% will ignite the explosive when charged, it is found that 174 of 200 detonators function properly. Test the null hypothesis p=0.9 again the alternative p< 0.9 at the 0.05 level of significance.

- Step 1. Hypotheses: $H_0 : p = 0.9$ versus $H_1 : p < 0.9$
- Step 2. Level of significance: $\alpha = 0.05$
- Step 3. test statistic

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

Criterion: Reject the null hypothesis if $Z < -1.65$

Example: Detonating Explosives

- Step 4. calculate the z value

$$z = \frac{(174/200 - 0.9)}{\sqrt{0.9 * 0.1/200}} = -1.41$$

- Step 5. Do not reject the null hypothesis.
  (p-value: $P(Z < -1.41) = 0.079$)
- Notice that we did not plug in $\hat{p}$ to estimate the variance. This is because, in the hypothesis test setting, we are assuming H_0 is true, that is we assume p is known and is equal to p_0.

Notation in the book

- I have used Y_i earlier to denote the Bernoulli random variables.
- Another (equivalent) way to derive the results is to consider $X = \sum_{i=1}^{n} Y_i$.
- X is of course Binomial(n, p), and we use the Normal approximation to Binomial (which is a special case of CLT). So X is normal and hence $\hat{p} = X/n$ is also normal.
- In this notation, X is the random variable representing the number of "good" observations in your sample. (p is the proportion of your population that is "good")

Two Proportions

Suppose we have two populations with proportion p_1 and p_2 respectively. We are interested in the difference between p_1 and p_2.

- Just as we did for two means, we have

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- By the CLT, we have

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \approx N(0,1)$$

CI for the difference $p_1 - p_2$

Plugging in $\hat{p}_1, \hat{p}_2$, we have

$$P\left(\left|\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}\right| < z_{\alpha/2}\right) \approx 1 - \alpha$$

Rearranging,

$$P\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < (p_1 - p_2) < \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) = 1 - \alpha$$

Hence, the $(1 - \alpha)$ CI for $(p_1 - p_2)$ is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Hypothesis test for two proportions

- Say we are interested in testing the hypothesis

$$H_0 : p_1 = p_2$$

- In this case, we do not have p_1 or p_2 explicitly. However, we can estimate p_1 and p_2 better by the pooled estiamtor

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

- so under H_0,

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0,1)$$

Hypothesis test for two proportions

- Thus, the rejection region (with significance level α) and p-values are listed below for different alternatives

H_1	rejection region	p-value
$p_1 > p_2$	$z > z_\alpha$	1 - F(z)
$p_1 < p_2$	$z < -z_\alpha$	F(z)
$p_1 \neq p_2$	$\lvert z\rvert > z_{\alpha/2}$	2F(-$\lvert z\rvert$)

Example: Assembling Tractors

A study shows that 16 of 200 tractors produced on one assembly line required extensive adjustments before they could be shipped, while the same was true for 14 of 400 tractors produced on another assembly line.

- Find the large sample 95% confidence interval for $p_1 - p_2$.

$$\hat{p}_1 = 0.08, \hat{p}_2 = 0.035$$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= 0.08 - 0.035 \pm 1.96\sqrt{\frac{0.08 \cdot 0.92}{200} + \frac{0.035 \cdot 0.965}{400}}$$

$$= [0.003, \quad 0.087]$$

Example: Assembling Tractors

- At the 0.01 level of significance, does this support the claim that the second production line does superior work?
  1. $H_0 : p_1 = p_2$, v.s. $H_1 : p_1 > p_2$
  2. Level of significance: $\alpha = 0.01$
  3. Criterion: Reject the null hypothesis if $Z > 2.33$, where

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

  4. Calculations:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{16 + 14}{200 + 400} = 0.05$$

     Substituting into the Z statistic, we have $Z = 2.38$
  5. Decision: Since $Z > 2.33$, we reject the null hypothesis and conclude that the true proportion of tractors requiring extensive adjustments is greater for first assembly line than for the second.