

Soci708 – Statistics for Sociologists

Module 8 – Inference for Proportions¹

François Nielsen

University of North Carolina
Chapel Hill

Fall 2009

¹Adapted from slides for the course Quantitative Methods in Sociology (Sociology 6Z3) taught at McMaster University by Robert Andersen (now at University of Toronto)

Inference for Proportions

- ▶ So far we have looked only at how we can make inferences about population means
- ▶ Similar techniques can be used to make inferences about population proportions
- ▶ Recall that a sample proportion is calculated as follows:

$$\hat{p} = \frac{\text{count of successes in the sample}}{\text{total observations in the sample}}$$

Here \hat{p} denotes a sample proportion and p is the population proportion.

- ▶ As with the case for means, we use the sampling distribution to make inferences about population proportions

Sampling Distribution of a Sample Proportion

- ▶ The sampling distribution of a sample proportion behaves in a manner similar to the sampling distribution of the sample mean
- ▶ As we saw earlier, the sampling distribution of a sample proportion has the following characteristics:
 1. The sampling distribution of \hat{p} becomes *approximately normal* as the sample size increases
 2. The *mean* of the sampling distribution of \hat{p} is p
 3. The standard deviation of the sampling distribution of \hat{p} is:

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

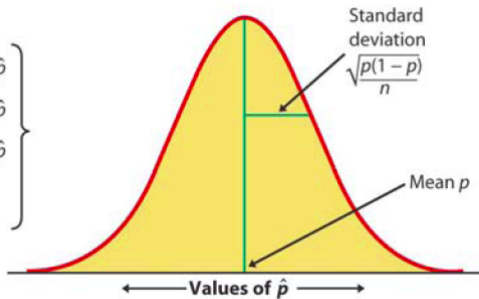
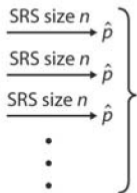
Assumptions for Inference about Proportions

1. We assume a *simple random sample*
 2. The normal approximation and the formula for the standard deviation hold only when the sample is *no more than 1/10 the size of the population*
 3. The sample size must be sufficiently large in relation to p :
 - ▶ np and $n(1 - p)$ must both be at least 10
 - ▶ This suggests, then, that the normal approximation is most accurate when p is close to .5 and least accurate when $p = 0$ or $p = 1$
- ▶ When these criteria are met, we can replace the unknown standard deviation of the sampling distribution of \hat{p} with its *standard error*

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$



Population proportion p of successes



Confidence Intervals and Hypothesis Tests for Proportions

- ▶ *CIs for proportions* take the usual form:

$$\text{Estimate} \pm z^* \times SE$$

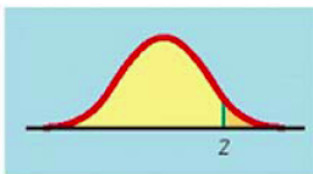
- ▶ Since we are assuming a normal distribution, we use a critical z value:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

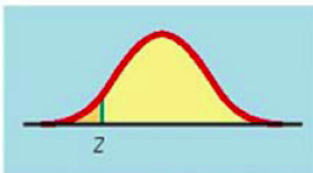
- ▶ Here z^* is the upper $(1 - C)/2$ standard normal critical value – i.e., we look to Table A in Moore et al. (2009)
- ▶ We test the hypothesis $H_0 : p = p_0$ by computing the z statistic:

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

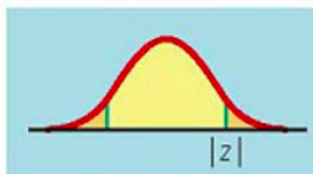
$$H_a: p > p_0$$



$$H_a: p < p_0$$



$$H_a: p \neq p_0$$



Example of a Confidence Interval

- ▶ Imagine that we have a SRS of 2500 Canadians. We ask whether the respondent has lived abroad for at least 1 year. We count $X = 187$ “Yes” answers. We want to estimate the population proportion p with 99% Confidence.
- ▶ We have the following information from our sample:

$$\hat{p} = \frac{187}{2500} = .075 \qquad n = 2500$$

$$z^* = 2.576 \quad (\text{for 99\% CI})$$

- ▶ Substituting this information into the formula, we get:

$$\begin{aligned}\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} &= .075 \pm 2.576 \sqrt{\frac{(.075)(.925)}{2500}} \\ &= .075 \pm .014 \\ &= .061 \text{ to } .089\end{aligned}$$

- ▶ We are 99% confident that between 6.1% and 8.9% of Canadians have lived abroad

Another Example of a Confidence Interval

Obama v. McCain – Gallup Poll of 22 Oct 2008

- ▶ The Gallup Poll reports Obama 51%, McCain 45%, Other or undecided 6%; $n = 2788$ registered voters; margin of error is $\pm 2\%$
- ▶ For a 95% CI we have $z^* = 1.960$
- ▶ Focusing on Obama support and substituting into the formula, we get:

$$\begin{aligned}\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= .51 \pm 1.960 \sqrt{\frac{(.51)(.49)}{2788}} \\ &= .51 \pm 0.019 \\ &= .491 \text{ to } .529\end{aligned}$$

- ▶ We are 95% confident that Obama support is between 49% and 53% of registered voters
- ▶ Note that declared margin of error of ± 2 is slightly conservative

Example of a Significance Test

- ▶ Using an earlier example, we now test whether the percentage of Canadians who lived abroad differed from 5%
- ▶ We chose an $\alpha = .01$ and thus need a critical value of $z^* = 2.576$ (this is a two-tailed test!)

$$H_0 : p = .05$$

$$H_a : p \neq .05$$

- ▶ Substituting the known information into the formula we get:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.075 - .05}{\sqrt{\frac{.05(1-.05)}{2500}}} = 5.735$$

- ▶ Since the z-statistic is significantly larger than the critical value z^* , we can reject H_0

One-sample Test for Proportion in R

```
> # in R  
> prop.test(187, 2500, p=.05, alternative="two.sided",  
            conf.level=.99, correct=FALSE)
```

1-sample proportions test without continuity correction

```
data: 187 out of 2500, null probability 0.05  
X-squared = 32.3705, df = 1, p-value = 1.274e-08  
alternative hypothesis: true p is not equal to 0.05  
99 percent confidence interval:  
 0.06234433 0.08950663  
sample estimates:  
      p  
0.0748
```

```
> # Alternatives are "two.sided", "greater" (p1>p2),  
    and "less" (p1<p2)
```

One-sample Test for Proportion in Stata

```
. * in Stata  
. prtesti 2500 187 .05, level(99) count
```

```
One-sample test of proportion                                x: Number of obs =    2500
```

Variable	Mean	Std. Err.	[99% Conf. Interval]	
x	.0748	.0052614	.0612476	.0883524

```
      p = proportion(x)                                z =    5.6895  
Ho: p = 0.05
```

```
Ha: p < 0.05  
Pr(Z < z) = 1.0000
```

```
Ha: p != 0.05  
Pr(|Z| > |z|) = 0.0000
```

```
Ha: p > 0.05  
Pr(Z > z) = 0.0000
```

Sample Size for Desired Margin of Error

- ▶ Just as was the case for inference for means, when collecting data it can be important to choose a sample size large enough to obtain a desired margin of error
- ▶ The margin of error is determined by:

$$m = z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- ▶ We must guess the value of \hat{p} with p^* . We can use either a pilot study or use the conservative estimate of .5 (this will give the largest possible margin of error).
- ▶ Sample size can then be calculated as follows:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

Sample Size for Desired Margin of Error

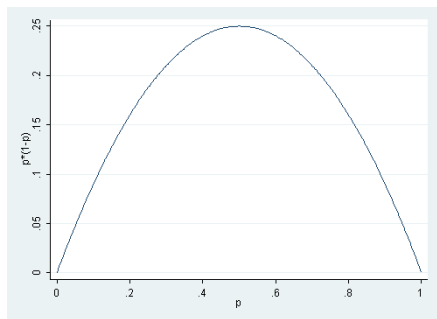
An Example

- ▶ A public opinion firm wants to determine the sample size needed to estimate the proportion of adults in North Carolina holding a variety of opinions with 95% confidence with a margin of error of $\pm 3\%$
- ▶ Since there are several opinion questions, the firm wants a sample size sufficient to insure the desired margin of error in the worst case scenario – i.e., when $p = 0.5$
- ▶ Applying the formula for n on the previous slide, the firm calculates:

```
> # in R  
> (1.96/0.03)^2*0.5*(1-0.5)  
[1] 1067.111
```
- ▶ Thus the firm will need a sample of $n = 1,068$ respondents

Sample Size for Desired Margin of Error

Why $p = .5$ is used to calculate sample size when true p is unknown



- . * in Stata
 - . twoway function y=x*(1-x),
range(0 1) xtitle('p')
ytitle('p*(1-p)')
- $p = .5$ corresponds to the maximum variance $p(1 - p)$, hence the largest (most conservative) estimate of n needed to achieve the desired margin of error

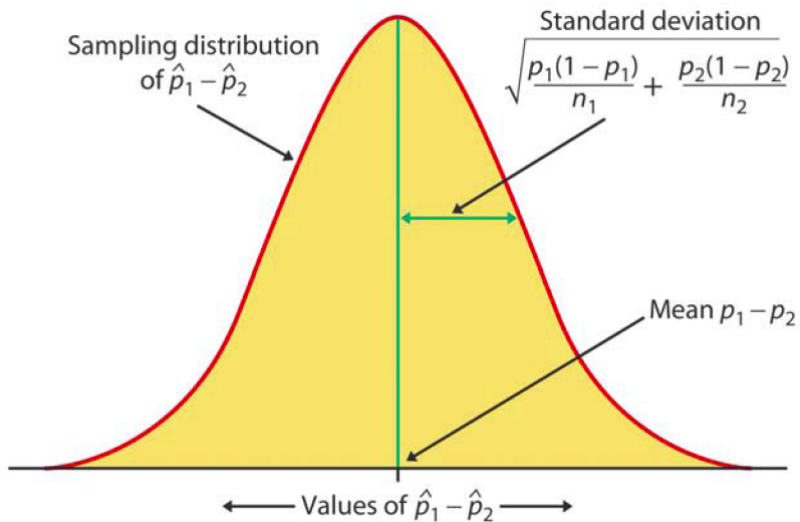
Comparing Two Proportions: Confidence Intervals

- ▶ Again as in the case with means, it is often of interest to compare two populations
- ▶ The *confidence interval* for comparing two proportions is given by:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* SE_{\hat{p}_1 - \hat{p}_2}$$
$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

As usual, z^* is the upper $(1 - C)/2$ standard normal critical value

- ▶ This confidence interval can be used when the populations are at least 10 times as large as the samples and counts of success in both samples is 5 or more



Comparing Two Proportions: Significance Tests

- ▶ Significance tests for the differences between two proportions also follow a similar pattern to the tests for difference in means
- ▶ To test $H_0 : p_1 = p_2$ we calculate the z statistic as follows:

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

Here \hat{p}_1 is for sample one; \hat{p}_2 is for sample two; \hat{p} (without the subscript) is the *pooled sample proportion*:

$$\hat{p} = \frac{\text{count of successes in both samples combined}}{\text{total observations in both samples combined}}$$

Two-Sample Test for Proportions

Example: Frequency of Left-Higher Ridge Count in Right-Handers



- ▶ Higher ridge count on fingers of left hand is a measure of body asymmetry
- ▶ Perhaps related to differential development of hemispheres of the brain – thus differing between men and women
- ▶ From Kimura, Doreen (1999, Figure 12.2 p. 167)

Figure 12.2

Example of a loop fingerprint pattern. A line is drawn from the triradial point at the left to the core point at the right. The ridge count is the number of lines between the triradial and core points. In this case it is 14.

Two-Sample Test for Proportions

Example: Frequency of Left-Higher Ridge Count in Right-Handers (2)

- ▶ Data from Kimura (1999, Table 12.2 p.169):

Frequency of Left-higher Ridge Count in Right-handers

	Left-higher	Not Left-higher	Total
Women	23	73	96
Men	20	134	154

- ▶ We have the following information from the data:

Women (1)

$$n_1 = 96$$

$$X_1 = 23$$

$$\hat{p}_1 = \frac{23}{96} = .240$$

Men (2)

$$n_2 = 154$$

$$X_2 = 20$$

$$\hat{p}_2 = \frac{20}{154} = .130$$

Two-Sample Test for Proportions

Example: Frequency of Left-Higher Ridge Count in Right-Handers (3)

- ▶ Using the formula for the confidence interval for the difference of two proportions we have:

$$\begin{aligned} & (\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\ & (.240 - .130) \pm 1.960 \sqrt{\frac{.240(1 - .240)}{96} + \frac{.130(1 - .130)}{154}} \\ & .110 \pm 1.960 \times .05133 = 0.0094 \text{ to } 0.2106 \end{aligned}$$

- ▶ We conclude with 95% confidence that the difference in proportion left-higher between women and men is between 0.9% and 21.1%
 - ▶ As this interval does not include zero we can also conclude that the difference is significant at the .05 level

Two-Sample Test for Proportions

Example: Frequency of Left-Higher Ridge Count in Right-Handers (4)

- ▶ To directly test the hypothesis $H_0 : p_1 = p_2$ we calculate the z statistic as follows:

$$\begin{aligned} z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \\ &= \frac{.240 - .130}{\sqrt{.172(1 - .172) \left(\frac{1}{96} + \frac{1}{154} \right)}} = 2.2415 \end{aligned}$$

- ▶ Here \hat{p}_1 is for sample one; \hat{p}_2 is for sample two; \hat{p} (without the subscript) is the *pooled sample proportion*

$$(23 + 20)/(96 + 154) = .172$$

Two-Sample Test for Proportions

Example: Frequency of Left-Higher Ridge Count in Right-Handers (5)

- ▶ The two-sided p-value $P(|Z| > 2.2415)$ corresponding to the alternative hypothesis $H_a : p_1 \neq p_2$ is 0.025.
 - ▶ We conclude once again that the proportion left-higher differs significantly between women and men at the $\alpha = .05$ level
- ▶ If we wanted to test the one-sided alternative $H_a : p_1 \geq p_2$ we would obtain the one-sided p-value $P(Z > 2.2415)$ by dividing the two-sided p-value .025 by 2, obtaining 0.0125
 - ▶ We conclude the one-sided alternative hypothesis $H_a : p_1 > p_2$ is also significant at the .05 level
 - ▶ Software packages often give only the two-sided p-value, which has to be divided by 2 to obtain the one-sided p-value

Two-Sample Test for Proportions in R

Left-Higher Ridge Count: Two-Sided Test

```
> # in R
> left.higher <- c(23, 20) # women lh, men lh
> subjects <- c(96, 154) # n women, n men
> prop.test(left.higher, subjects, correct=FALSE)
```

```
      2-sample test for equality of proportions without continuity
      correction
```

```
data: left.higher out of subjects
X-squared = 4.9982, df = 1, p-value = 0.02537
alternative hypothesis: two.sided
95 percent confidence interval:
0.009170057 0.210256350
sample estimates:
  prop 1    prop 2 
0.2395833 0.1298701

> sqrt(4.9982)
[1] 2.235665
```

Two-Sample Test for Proportions in R

Left-Higher Ridge Count: One-Sided Test

```
> # in R
> left.higher <- c(23, 20) # women lh, men lh
> subjects <- c(96, 154) # n women, n men
> prop.test(left.higher, subjects, alternative="greater", correct=FALSE)
```

```
      2-sample test for equality of proportions without continuity
      correction
```

```
data: left.higher out of subjects
X-squared = 4.9982, df = 1, p-value = 0.01269
alternative hypothesis: greater
95 percent confidence interval:
0.02533473 1.00000000
sample estimates:
  prop 1    prop 2 
0.2395833 0.1298701
```

Two-Sample Test for Proportions in Stata

Left-Higher Ridge Count: One-Sided & Two-Sided Tests

```
. * in Stata
. * syntax is 'prtesti n1 p1 n2 p2''
. * or 'prtesti n1 X1 n2 X2, count''
. prtesti 96 23 154 20, count
```

Two-sample test of proportion

x: Number of obs = 96
y: Number of obs = 154

Variable	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
x	.2395833	.0435631			.1542013 .3249654
y	.1298701	.0270886			.0767775 .1829628
diff	.1097132	.0512985			.0091701 .2102564
	under Ho:	.0490742	2.24	0.025	

diff = prop(x) - prop(y)

z = 2.2357

Ho: diff = 0

Ha: diff < 0

Pr(Z < z) = 0.9873

Ha: diff != 0

Pr(|Z| < |z|) = 0.0254

Ha: diff > 0

Pr(Z > z) = 0.0127

Two-Sample Test for Proportions in Stata

Left-Higher Ridge Count: Equivalence of z Test for Proportions & χ^2 Test

```
. * in Stata
. * 'prtesti 96 23 154 20, count' is equivalent to using
. * the tabi command with the original contingency table:
. tabi 23 73\ 20 134, chi2
```

row	col		Total
	1	2	
1	23	73	96
2	20	134	154
Total	43	207	250

Pearson chi2(1) = 4.9982 Pr = 0.025

```
. * take the square root of the chi-squared
. display sqrt(4.9982)
2.2356654
. * note it is the same z as obtained with prtesti!
. * this is a glimpse of the next topic
```

Next week:

- ▶ Inference for crosstabs