VOTRE LIEN AVEC CE QUI COMPTE — CONNECTS YOU TO WHAT MATTERS

Comparing Proportions

ADM 2304

Winter 2016

Comparing Proportions

There are two different situations:

-independent proportions from different samples (section 12.8);

-dependent proportions from the same sample (not covered in text).

Independent Samples

In general, we test:

$$H_0: p_1 - p_2 = p_0 \text{ against } H_a: p_1 - p_2 \neq p_0 ,$$

p_1 and p_2 are the population proportions in the two populations. If we have random samples from distinct populations, then this guarantees independent samples.

Note: the hypotheses *do not specify values for* p_1 *or* p_2, only a value p_0 for the difference.

Look at (p_1-hat – p_2-hat)

- If H_0 is true, then (**p_1-hat – p_2-hat**), the "observed" difference between the two sample proportions, estimates $\mathbf{p_1 - p_2 = p_0}$
- If both samples are large, then
  (**p_1-hat – p_2-hat**) has a normal distribution, with mean $\mathbf{p_0}$.
- What is its standard deviation?

Variance of (X – Y) ?

- Recall that

$$\text{Var}(X \pm Y) = \text{Var } X + \text{Var } Y \mp 2*\text{Cov}(X,Y)$$

where [Cov(X,Y) = SD(X)SD(Y)corr(X,Y)]

- If X and Y are independent, their covariance is zero, and

$$\text{Var}(X \pm Y) = \text{Var } X + \text{Var } Y;$$

- The square root law is:

$$\text{SD}(X \pm Y) = \sqrt{(\text{Var } X + \text{Var } Y)}.$$

Ref. SDVW p. 271

Notation

- Suppose the two sample sizes are n_1 & n_2, with X_1 & X_2 respondents in the class of interest, respectively.
- The two sample proportions are

$$p_1\text{-hat} = X_1 / n_1 \text{ , \&}$$
$$p_2\text{-hat} = X_2 / n_2$$

SD and SE

Since the sample proportions are independent, their difference has a standard deviation $SD(p_1\text{-hat} - p_2\text{-hat})$

$= \sqrt{[Var(p_1\text{-hat}) + Var(p_2\text{-hat})]}$

$= \sqrt{[p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2]}$

Since p_1 and p_2 are unknown, in general we use $\mathbf{p_1\text{-hat}}$ and $\mathbf{p_2\text{-hat}}$ to estimate them and to calculate the

$SE(p_1\text{-hat} - p_2\text{-hat})$.

Pooled Proportion

In the special case where the hypothesized difference p_0 is 0, we first "pool" the two samples to obtain a (combined) "pooled" estimate of the common proportion $p = p_1 = p_2$, using

$$\text{p-bar}_{\text{pooled}}$$

$$= (n_1 * p_1\text{-hat} + n_2 * p_2\text{-hat}) / (n_1 + n_2)$$

$$= (X_1 + X_2) / (n_1 + n_2)$$

Test Statistic

The standardized test statistic is:

$$\text{z-stat} = \frac{[(p_1\text{-hat} - p_2\text{-hat}) - p_0]}{SE(p_1\text{-hat} - p_2\text{-hat})}$$

where $SE(p_1\text{-hat} - p_2\text{-hat}) =$

$$\sqrt{[p_1\text{-hat} * q_1\text{-hat} / n_1 + p_2\text{-hat} * q_2\text{-hat} / n_2]}$$

for nonzero p_0, and

$$\sqrt{[p\text{-bar}_{pooled} * q\text{-bar}_{pooled} * (1/n_1 + 1/n_2)]}$$

when $p_0 = 0$.

Example

- Suppose we want to compare the proportion of Canadians (in 1994 versus 2013) who want to abolish the Senate.

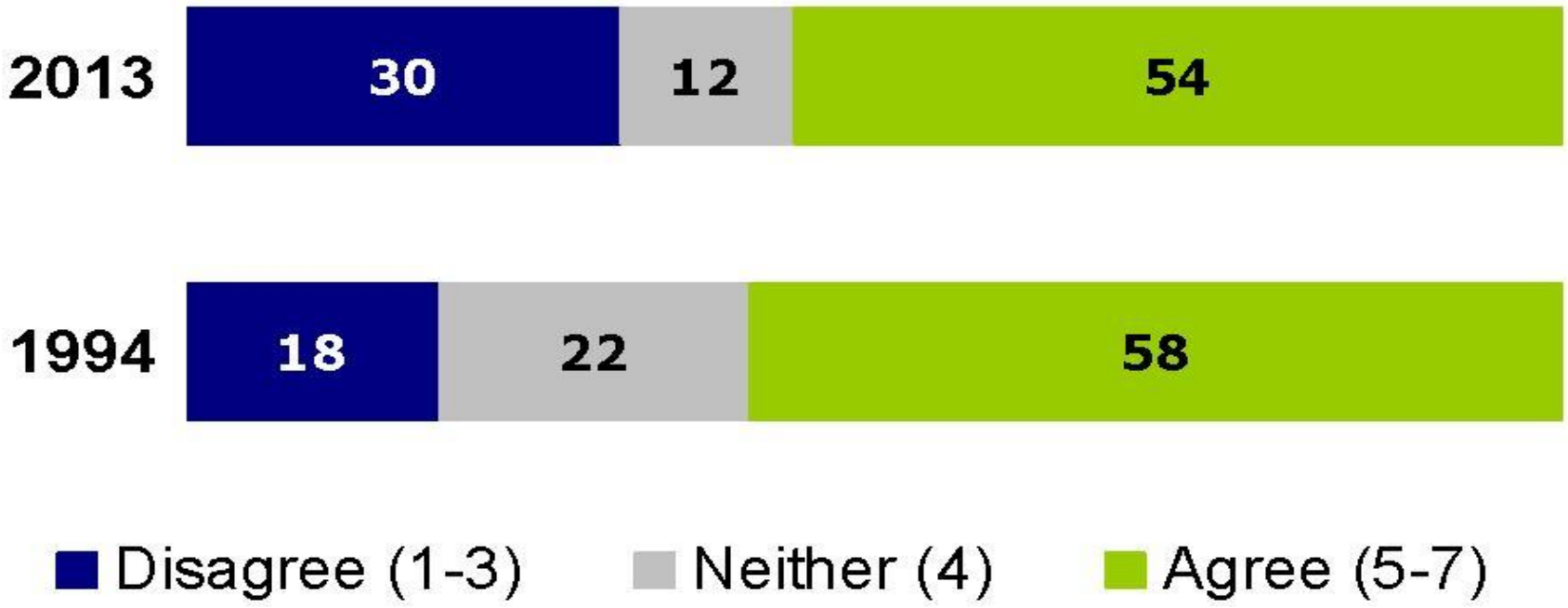
Hypotheses

- Test whether there is a difference in population proportions, at the .05 level of significance.
- We have $H_0: p_1 - p_2 = p_0$
  $H_a: p_1 - p_2 \neq p_0$

  where p_0 is zero.
- This is a 2-sided test since $p_1 - p_2 \neq 0$ means either $p_1 - p_2 < 0$ or $p_1 - p_2 > 0$.

Support for senate abolition

Q. *Please rate the extent to which you agree or disagree with the following statement:* ***I think that the Canadian Senate should be abolished immediately***



BASE: Canadians; most recent data point October 26-29, 2013 (n=1,377)

http://www.ekospolitics.com/index.php/2013/10/stephen-harper-plumbing-record-lows-on-trust-direction-and-approval/

Calculations

- We have p_1-hat = .58 and p_2-hat = .54 .
- We are given $n_2 = 1377$.
- Suppose $n_1 = 1000$.
- Since the hypothesized difference is zero, we pool the two sample proportions to get:

  $p\text{-bar}_{pooled}$ $= (1000*.58 + 1377*.54)/2377$

  $= 0.557$

- Note: the pooled proportion is between 0.54 and 0.58.

Rejection region

- For a 2-sided test, we reject the null hypothesis if the z-stat is too big or too small.
- We divide the α level of significance into two and we reject the null hypothesis if

$$\text{z-stat} < -z_{\alpha/2} \text{ or } \text{z-stat} > z_{\alpha/2} \ ,$$

or equivalently,

$$|\text{z-stat}| > z_{\alpha/2}.$$

Confidence Interval

- The 100(1-α)% two-sided confidence interval for (p_1 - p_2) is:

  (p_1-hat – p_2-hat)

  $$\pm z_{\alpha/2} * SE(p_1\text{-hat} - p_2\text{-hat})$$

  where

  $SE = \sqrt{[p_1\text{-hat}*q_1\text{-hat} / n_1 + p_2\text{-hat}*q_2\text{-hat} / n_2]}$

  (we do not calculate a pooled proportion.)

- We reject the null hypothesis if the CI does not cover p_0 .

Large Sample Condition

- If n_1p_1 , n_2p_2 , $n_1(1-p_1)$, and $n_2(1-p_2)$, are all at least 10, then the test statistic has a standard normal distribution.
- *In practice, we can only check whether*

  $X_1 = n_1*p_1\text{-hat}$, $X_2 = n_2*p_2\text{-hat}$,

  $n_1 - X_1 = n_1*q_1\text{-hat1}$, *and* $n_2 - X_2 = n_2*q_2\text{-hat}$

  are all at least 10.

Calculations

- Using the pooled p-bar of 0.557,

  $Z = (0.58 - 0.54) / \sqrt{[(.557*.443)(1/1000+1/1377)]}$

  $= 0.04 / .0206 = 1.94$

- *The 95% CI is:*

  $0.04 \pm 1.96 * \sqrt{(.58*.42/1000 + .54*.46/1377)}$

  $= 0.04 \pm 1.96 * .0206$

  $= 0.04 \pm .0404 = (-0.0004, 0.08)$

Results

- Since $|1.94| \leq 1.96 = z_{.05/2}$, we do not reject the null hypothesis at the 0.05 level;
- We make the same decision based on the 95% CI since *(-.0004, 0.08)* covers zero.
- The p-value is $P(|Z| > 1.94)$
  $= 2*P(Z>1.94) = 2*0.026 = 0.052.$
- Conclude there is insufficient evidence to show a real change in the support for abolishing the Senate from 1994 to 2013.

Using Minitab

Select “2 Proportions” under “Basic Statistics” .

The data can consist of two different text or quantitative values in separate columns or in one column with subscripts in second column, or summarized.

There is an option to check

“use pooled estimate of p for test” (Minitab 14)

or to choose

“separate proportions” or “pooled estimate”.

Minitab output

Test and CI for Two Proportions

```
Sample    X     N  Sample p
1       580  1000  0.580000
2       744  1377  0.540305

Difference = p (1) - p (2)
Estimate for difference:  0.0396950
95% CI for difference:  (-0.000661920, 0.0800519)
Test for difference = 0 (vs not = 0):
      Z = 1.92    P-Value = 0.054
```

Same Sample Comparison

- We are interested often in comparing two proportions from the same sample.
- For example, political polling usually compares one candidate with another or one party with another, using sample proportions from the ***same*** sample.

Proportions are not independent

- Using the same sample, we want to compare the proportion who support one party versus those who support another party. Obviously, the sample proportions are not independent because the support for one party tends to be inversely related to the support for another party.
- To account for the covariance or correlation between the two sample proportions, we add a *covariance* term to the standard error calculation of the estimated difference.

Variance of a difference between two correlated proportions

- In general,

  Var (***p_1-hat – p_2-hat***)

  = Var (***p_1-hat***) + Var(***p_2-hat***)

  – 2 Cov(***p_1-hat***, ***p_2-hat***)

  Here the covariance is negative, based on a negative correlation:

  $$\text{Cov}(p_1\text{-hat}, p_2\text{-hat}) = - p_1 * p_2 / n$$

New Standard Error

We calculate the usual z-statistic:

$$\text{z-stat} = \frac{[(p_1\text{-hat} - p_2\text{-hat}) - p_0]}{SE(p_1\text{-hat} - p_2\text{-hat})}$$

except $SE(p_1\text{-hat} - p_2\text{-hat})$ is calculated by:

$$\sqrt{[p_1\text{-hat}*q_1\text{-hat}/n + p_2\text{-hat}*q_2\text{-hat}/n + 2 * p_1\text{-hat}*p_2\text{-hat}/n]}$$

Example

In a survey of “best leader”, test at the 0.01 level of significance whether Rona Ambrose’s support can be shown to be more than 0.05 higher than Tom Mulcair’s support:

H_0 : P(Ambrose) - P(Mulcair) $\leq$ 0.05

H_a : P(Ambrose) - P(Mulcair) > 0.05

Here a 5% difference is deemed to be of practical significance.

Comparing Liberal and NDP

- The latest results are Trudeau 53%, Ambrose 28%, and Mulcair 17%, based on an assumed sample size of 1000.

Data and Calculations

- The standard error of the difference between the two *dependent* proportions is:

  $\sqrt{((0.28*0.72 + 0.17*0.83+2*0.28*0.17)/1000)} = 0.021$ and

  $Z = [(0.28-0.17) - 0.05] / .021 = 2.86.$

- We reject the null hypothesis since z = 2.86 is greater than $z_{.01} = 2.326$ (the p-value is $P(Z>2.86) = 0.0021$).
- Therefore, we conclude that Ambrose's support is more than 0.05 higher than that of Mulcair, for best leader.

The proportions were not independent

- If we had used the independent samples formula incorrectly, then the standard error would have been:

$\sqrt{(0.28*0.72/1000+0.17*0.83/1000)}$ =0.0185,

and $Z = (0.06)/0.0185 = 3.24 > z_{.01}$

- We would make the same decision to reject the null hypothesis, but with a smaller p-value of P(Z>3.24)=0.0006.