Two sample: Binary Response

ST551 Lecture 21

Charlotte Wickham

2017-11-08

Two sample: Binary Response

Setting: two **independent** samples

$Y_1, \ldots, Y_n$ i.i.d from Bernoulli(p_Y)
$X_1, \ldots, X_m$ i.i.d from Bernoulli(p_X)

Parameter: Difference in population proportions $p_Y - p_X$

$p_Y = E(Y_i) = P(Y_i = 1)$
$p_X = E(X_i) = P(X_i = 1)$

As a contingency table

Represent resulting data in a 2 x 2 contingency table:

	0	1	**Total**
Y_i	a	b	**n = a+b**
X_i	c	d	**m = c+d**
Total	**a+c**	**b+d**	**m + n**

Two sample: Binary Response - Alternate view

Setting: two **independent** samples

$$(Y_1, G_1), (Y_2, G_2), \ldots, (Y_n, G_n), (Y_{n+1}, G_{n+1}), \ldots, (Y_{n+m}, G_{n+m})$$

where G is a binary *grouping* variable which indicates which population the observation came from:

$$G_i = \begin{cases} 0, & \text{observation from } Y \\ 1, & \text{observation from } X \end{cases}$$

As a contingency table - Alternate view

Represent resulting data in a 2 x 2 contingency table:

	$Y_i = 0$	$Y_i = 1$	**Total**
$G_i = 0$	$n_{11} = a$	$n_{12} = b$	$\mathbf{n = a+b = } R_1$
$G_i = 1$	$n_{21} = c$	$n_{22} = d$	$\mathbf{m = c+d = } R_2$
Total	$\mathbf{a+c = } C_1$	$\mathbf{b+d=} C_2$	$\mathbf{a + b + c + d = N}$

Two views are equivalent

If we are interested in the response variable given the group.

- I sample 40 OSU graduate students and 20 OSU undergraduate students:
  - Y_i = graduate student, did you vote in 2016? $i = 1, \ldots, 40$
  - X_i = undergraduate student did you vote in 2016? $i = 1, \ldots, 20$
- I sample 60 OSU students and record:
  - Y_i = did you vote in 2016?, $i = 1, \ldots, 60$
  - G_i = student's level (0 = graduate, 1 = undergraduate), $i = 1, \ldots, 60$

Inference focuses on:

Comparing $P(Y_i = 1)$ and $P(X_i = 1)$ - first view
Comparing $P(Y_i = 1|G_i = 0)$ and $P(Y_i = 1|G_i = 1)$ - second view

Ways to compare two proportions

$Y_1, \ldots, Y_n$ i.i.d from Bernoulli(p_Y)
$X_1, \ldots, X_m$ i.i.d from Bernoulli(p_X)

Typical null hypothesis: $H_0 : p_Y = p_X$)

Difference in population proportions: $p_Y - p_X$

- $H_0 : p_Y - p_X = 0$

Relative risk: p_Y / p_X

- $H_0 : p_Y / p_X = 1$

Odds ratio: $\frac{p_Y}{1-p_Y} / \frac{p_X}{1-p_X}$

- $H_0 : \frac{p_Y}{1-p_Y} / \frac{p_X}{1-p_X} = 1$

Example

(From `class_data`, assuming you are like some random sample from a larger population)

G_i: **Do you prefer cats or dogs?** Y_i: **Did you east breakfast this morning?**

```
##         ate_breakfast
## cat_dog no yes
##    cats  6   9
##    dogs  6  14
```

Your turn: Fill in the table margins

Estimates

Probability of eating breakfast, given you prefer cats:

$$p_Y = P(Y_i = 1 | G_i = 0) = \frac{P(Y_i = 1 \,\&\, G_i = 0)}{P(G_i = 0)}$$

Estimate

$$\hat{p}_Y = \frac{b/N}{R_1/N} = \frac{9}{15} = 0.6$$

Probability of eating breakfast, given you prefer dogs:

$$p_X = P(Y_i = 1 | G_i = 1) = \frac{P(Y_i = 1 \,\&\, G_i = 1)}{P(G_i = 1)}$$

Estimate

$$\hat{p}_X = \frac{d/N}{R_2/N} = \frac{14}{20} = 0.7$$

Estimates

Difference in proportions

$$\hat{p}_Y - \hat{p}_X = 0.6 - 0.7 = -0.1$$

Relative Risk

$$\frac{\hat{p}_Y}{\hat{p}_X} = \frac{0.6}{0.7} = 0.86$$

Odds Ratio

$$\frac{\hat{p}_Y}{1-\hat{p}_Y} / \frac{\hat{p}_X}{1-\hat{p}_X} = \frac{0.6}{1-0.6} / \frac{0.7}{1-0.7} = 0.64 = \frac{bc}{ad}$$

Two sample Z-test of proportions

(Comes from considering proportion as mean and looking at two sample Z-test)

Null hypothesis: $H_0 : p_Y = p_X$

$$Z = \frac{\hat{p}_Y - \hat{p}_X}{\sqrt{\hat{p}_c(1 - \hat{p}_c)\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

where $p_c = \frac{(np_Y + mp_X)}{n+m} = \frac{b+d}{N}$

When null is true Z has a N(0, 1) distribution.

Confidence interval for difference in proportions

$(1-\alpha)100\%$ CI:

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

Like in one sample case, binomial test and CI may not agree because they use different estimates of the variance of the difference in sample proportions.

Your Turn

$$p_c = \frac{(np_Y + mp_X)}{n+m} = \frac{b+d}{N}$$

What is p_c for our table?

```
##          ate_breakfast
## cat_dog no yes Sum
##    cats  6   9  15
##    dogs  6  14  20
##    Sum  12  23  35
```

Example: Z-stat

$$Z = \frac{\hat{p}_Y - \hat{p}_X}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{-0.1}{\sqrt{0.66(1-0.66)(\frac{1}{15}+\frac{1}{20})}} = -0.62$$

Compare to $z_{1-\alpha/2} = 1.96$

p-value (for two sided alternative) `2*(1 - pnorm(abs(z)))`$=0.54$

95% confidence interval:

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

$$= -0.1 \pm \sqrt{\frac{0.6(1-0.6)}{15} + \frac{0.7(1-0.7)}{20}}$$

$$= (-0.26, 0.06)$$

Pearson's Chi-squared Test

$H_0 : p_Y - p_X = 0$

$$X = \sum_{j,k=1,2} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}$$

$O_{jk} = n_{jk}$

$E_{jk} = \frac{R_j C_k}{N}$

If null is true, X has χ^2_1 distribution

Example: Chi-squared test

```
##         ate_breakfast
## cat_dog no yes Sum
##    cats  6   9  15
##    dogs  6  14  20
##    Sum  12  23  35
```

```
##         no   yes
## cats 5.14  9.86
## dogs 6.86 13.14
```

E.g $\frac{15 \times 12}{35} = 5.14$

Example: Chi-squared test

$$X = \frac{(6 - 5.14)^2}{5.14} + \frac{(9 - 9.86)^2}{9.86} + \frac{(6 - 6.86)^2}{6.86} + \frac{(14 - 13.14)^2}{13.14}$$
$$= 0.38$$

Compare to $\chi^2_1(1 - \alpha) = 3.84$

p-value: `(1 - pchisq(X, df = 1)`$=0.54$

Summary

Pearson's Chi-squared test for homogeneity of proportions across groups is equivalent (i.e. results in the same p-value) to the Z-test for proportions (when there are two groups).

$X = Z^2$