Recitation, Week 10

Ye Wang

New York University

POL-850

Spring 2018

Outline

1. Review: Comparison of means
2. Chi-square test
3. Other measures of association (λ, Cramer's V, and γ)

Review

Which method of comparison to use? It depends on **the type of the variable**

		Y		
		Nominal	Ordinal	Interval
X	Nominal	Cross-tab		Comparison of means
	Ordinal			
	Interval			

Review

The procedures of hypothesis testing (comparison of means):

1. Data at hand

Review

The procedures of hypothesis testing (comparison of means):

1. Data at hand
2. The **significance level** (e.g. 0.05)

Review

The procedures of hypothesis testing (comparison of means):

1. Data at hand
2. The **significance level** (e.g. 0.05)
3. **Test-statistic**

Review

The procedures of hypothesis testing (comparison of means):

1. Data at hand
2. The **significance level** (e.g. 0.05)
3. **Test-statistic**
4. **p-value**

Review

The procedures of hypothesis testing (comparison of means):

1. Data at hand
2. The **significance level** (e.g. 0.05)
3. **Test-statistic**
4. **p-value**
5. **Compare** our p-value with the significance level

   If p-value$<$0.05, then we reject H_0

Review

1. Data at hand

	Treated	Control
Mean	2.1 years	2 years
σ	0.4 years	0.6 years
N	100	100

$$H_0 \text{ is } \mu_T = \mu_C$$
$$H_A \text{ is } \mu_T \neq \mu_C$$

Review

1. Data at hand

	Treated	Control
Mean	2.1 years	2 years
σ	0.4 years	0.6 years
N	100	100

$$H_0 \text{ is } \mu_T = \mu_C$$
$$H_A \text{ is } \mu_T \neq \mu_C$$

2. The **significance level**: let's choose the 0.05 significance level

Review

3. **Test-statistic**: then we can calculate the test statistic:

$$z = \frac{\overline{x}_T - \overline{x}_C}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} = \frac{2.1 - 2}{\sqrt{\frac{0.4^2}{100} + \frac{0.6^2}{100}}} = 1.39$$

4. **p-value**: we get the corresponding p-value

$$Pr[Z < -1.39 \text{ or } Z > 1.39] = 0.1646$$

Review

3. **Test-statistic**: then we can calculate the test statistic:

$$z = \frac{\overline{x}_T - \overline{x}_C}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} = \frac{2.1 - 2}{\sqrt{\frac{0.4^2}{100} + \frac{0.6^2}{100}}} = 1.39$$

4. **p-value**: we get the corresponding p-value

$$Pr[Z < -1.39 \text{ or } Z > 1.39] = 0.1646$$

5. **Compare** our p-value with the 0.05 significance level

$$0.1646 > 0.05$$

Review

3. **Test-statistic**: then we can calculate the test statistic:

$$z = \frac{\overline{x}_T - \overline{x}_C}{\sqrt{\frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}}} = \frac{2.1 - 2}{\sqrt{\frac{0.4^2}{100} + \frac{0.6^2}{100}}} = 1.39$$

4. **p-value**: we get the corresponding p-value

$$Pr[Z < -1.39 \text{ or } Z > 1.39] = 0.1646$$

5. **Compare** our p-value with the 0.05 significance level

$$0.1646 > 0.05$$

We fail to reject the null hypothesis at the 0.05 significance level

Chi-square test

- When Y is nominal/ordinal variable, we have a cross-tab

Chi-square test

- When Y is nominal/ordinal variable, we have a cross-tab
- In this case, our hypothesis testing boils down to the question below

Chi-square test

- When Y is nominal/ordinal variable, we have a cross-tab
- In this case, our hypothesis testing boils down to the question below

Does observed dispersion of cases differ from that expected under null hypothesis?

Chi-square test

- When Y is nominal/ordinal variable, we have a cross-tab
- In this case, our hypothesis testing boils down to the question below

Does observed dispersion of cases differ from that expected under null hypothesis?

- Technically, we use χ^2 **(chi-square)** for the hypothesis testing

Chi-square test

- When Y is nominal/ordinal variable, we have a cross-tab
- In this case, our hypothesis testing boils down to the question below

Does observed dispersion of cases differ from that expected under null hypothesis?

- Technically, we use χ^2 **(chi-square)** for the hypothesis testing
- Then we get the **p-value** and then **compare**!

Chi-square test

- When Y is nominal/ordinal variable, we have a cross-tab
- In this case, our hypothesis testing boils down to the question below

Does observed dispersion of cases differ from that expected under null hypothesis?

- Technically, we use χ^2 **(chi-square)** for the hypothesis testing
- Then we get the **p-value** and then **compare**!
  (cf. z-statistic or t-statistic for the comparison of means)

The familiar example again

Our question is: "Does **gender** have something to do with the **response** to the statement?"

Table: US would be better off if we stay home

	Male	Female	Total
	Observed	Observed	
Agree	1089 38.69	1062 35.46	2151 37.02
Disagree	1726 61.31	1933 64.54	3659 62.98
Total	2815 100	2995 100	5810 100

Group exercise 1

H_0: the gender has nothing to do with the answer

What should be the expected frequency (percent) in each cell, should the null hypothesis hold?

Table: US would be better off if we stay home

	Male		Female		Total
	Observed	Expected	Observed	Expected	
Agree	1089 38.69	? ?	1062 35.46	? ?	2151 37.02
Disagree	1726 61.31	? ?	1933 64.54	? ?	3659 62.98
Total	2815 100		2995 100		5810 100

The familiar example again

Table: US would be better off if we stay home

	Male		Female		Total
	Observed	Expected	Observed	Expected	
Agree	1089 38.69	1042.11 37.02	1062 35.46	1108.75 37.02	2151 37.02
Disagree	1726 61.31	1772.88 62.98	1933 64.54	1886.25 62.98	3659 62.98
Total	2815 100		2995 100		5810 100

Chi-square test

- How to calculate the chi-square test statistic?

$$\chi^2 = \sum_{n=1}^{N} \frac{(f_0 - f_e)^2}{f_e}$$

- where
  f_0: the observed frequency
  f_e: the expected frequency given H_0
  N: the total number of cells
- Plugging the values into the formula yields $\chi^2 = 6.4791$
- The degrees of freedom is (# of rows-1)(# of columns-1)= 1

Chi-square test

- Therefore, the corresponding p-value is 0.011 (STATA result)
- What would be our conclusion from this result?

Chi-square test

- Therefore, the corresponding p-value is 0.011 (STATA result)
- What would be our conclusion from this result?

(1) Reject the null or (2) fail to reject the null?

Chi-square test

- Therefore, the corresponding p-value is 0.011 (STATA result)
- What would be our conclusion from this result?

(1) Reject the null or (2) fail to reject the null?

- Actually, we **should have chosen** our criteria–the significance level at the beginning of our analysis

Chi-square test

- Therefore, the corresponding p-value is 0.011 (STATA result)
- What would be our conclusion from this result?

(1) Reject the null or (2) fail to reject the null?

- Actually, we **should have chosen** our criteria–the significance level at the beginning of our analysis
- Suppose we had chosen the 0.05 significance level,

Chi-square test

- Therefore, the corresponding p-value is 0.011 (STATA result)
- What would be our conclusion from this result?

(1) Reject the null or (2) fail to reject the null?

- Actually, we **should have chosen** our criteria–the significance level at the beginning of our analysis
- Suppose we had chosen the 0.05 significance level, then (1) We say our result is **statistically significant** at the 0.05 significance level
- Would our results have still been significant if we had chosen a 0.01 significance level instead?

Measures of association

- So far only the **existence** of **relationship** b/w X and Y

Measures of association

- So far only the **existence** of **relationship** b/w X and Y
- When Y is interval variable, z- or t-statistic

Measures of association

- So far only the **existence** of **relationship** b/w X and Y
- When Y is interval variable, z- or t-statistic
- When Y is ordinal/nominal variable, χ^2 statistic

Measures of association

- So far only the **existence** of **relationship** b/w X and Y
- When Y is interval variable, z- or t-statistic
- When Y is ordinal/nominal variable, χ^2 statistic

But we do also care about the **strength** of **association**

Measures of association

- So far only the **existence** of **relationship** b/w X and Y
- When Y is interval variable, z- or t-statistic
- When Y is ordinal/nominal variable, χ^2 statistic

But we do also care about the **strength** of **association**

- In other words, if any, how strong the association is?
- For that matter, we use λ, Cramer's V, and γ

Measures of association

Table: Measures of association

Statistic	When	Min.	Meaning	Max.	Meaning
λ	Nom-Nom	0	None	1	**Perfect**
Cramer's V*	Nom-Nom	0	None	1	**Perfect**
γ	Ord-Ord*	-1	Perfect -	1	Perfect +

NB: Cramer's V can take a value **[-1, 1]** for 2x2 tables.
NB2: Can use γ when have dichotomous ordinal variable pair.

Lambda

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1}$$

- where
  $\varepsilon_1 = N_{\text{total}} - N_{\text{mode}}$
  $\varepsilon_2 = \sum_{\text{for all catagories}} (N_{\text{category}} - N_{\text{mode for category}})$

  ε_1: Error without knowledge (independent variable)
  ε_2: Error with knowledge (independent variable)

Lambda

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1}$$

- where
  $\varepsilon_1 = N_{\text{total}} - N_{\text{mode}}$
  $\varepsilon_2 = \sum_{\text{for all catagories}} (N_{\text{category}} - N_{\text{mode for category}})$

  ε_1: Error without knowledge (independent variable)
  ε_2: Error with knowledge (independent variable)

- If $\varepsilon_2 \approx 0$, then $\lambda \approx 1$

Lambda

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1}$$

- where
  $\varepsilon_1 = N_{\text{total}} - N_{\text{mode}}$
  $\varepsilon_2 = \sum_{\text{for all catagories}} (N_{\text{category}} - N_{\text{mode for category}})$

  ε_1: Error without knowledge (independent variable)
  ε_2: Error with knowledge (independent variable)

- If $\varepsilon_2 \approx 0$, then $\lambda \approx 1$
  (i.e. knowledge significantly reduces the existing error)

Lambda

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1}$$

- where
  $\varepsilon_1 = N_{\text{total}} - N_{\text{mode}}$
  $\varepsilon_2 = \sum_{\text{for all catagories}} (N_{\text{category}} - N_{\text{mode for category}})$

  ε_1: Error without knowledge (independent variable)
  ε_2: Error with knowledge (independent variable)

- If $\varepsilon_2 \approx 0$, then $\lambda \approx 1$
  (i.e. knowledge significantly reduces the existing error)

- If $\varepsilon_1 \approx \varepsilon_2$, then $\lambda \approx 0$

Lambda

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1}$$

- where
  $\varepsilon_1 = N_{\text{total}} - N_{\text{mode}}$
  $\varepsilon_2 = \sum_{\text{for all catagories}} (N_{\text{category}} - N_{\text{mode for category}})$

  ε_1: Error without knowledge (independent variable)
  ε_2: Error with knowledge (independent variable)

- If $\varepsilon_2 \approx 0$, then $\lambda \approx 1$
  (i.e. knowledge significantly reduces the existing error)

- If $\varepsilon_1 \approx \varepsilon_2$, then $\lambda \approx 0$
  (i.e. knowledge barely reduces the existing error)

Group exercise 2

Let's calculate λ with the same example but slightly **different** data

Hint: what's our best guess about the response **without knowing gender**? How about **with knowing gender**?

Table: US would be better off if we stay home

	Male	Female	Total
	Observed	Observed	
Agree	1089 38.69	1933 64.54	3022 52.01
Disagree	1726 61.31	1062 35.46	2788 47.99
Total	2815 100	2995 100	5810 100

Lambda

- Without gender columns, our best guess about the response is "Agree" $\varepsilon_1 = 5810 - 3022 = 2788$
- With gender columns, $\varepsilon_2 = (2815 - 1726) + (2995 - 1933) = 1089 + 1062 = 2151$
- Therefore, $\lambda = \frac{2788-2151}{2788} = 0.2285$

Table: US would be better off if we stay home

	Male	Female	Total
	Observed	Observed	
Agree	1089 38.69	1933 64.54	3022 52.01
Disagree	1726 61.31	1062 35.46	2788 47.99
Total	2815 100	2995 100	5810 100

Key takeaways

You should be able to answer these questions:

1. How do we test the existence of association?

Key takeaways

You should be able to answer these questions:

1. How do we test the existence of association?
   we use test statistics and corresponding p-values
2. When comparing (group) means? (which test statistic?)

Key takeaways

You should be able to answer these questions:

1. How do we test the existence of association?
   we use test statistics and corresponding p-values
2. When comparing (group) means? (which test statistic?)
   z-statistic (σ known) or t-statistic (σ unknown)
3. Cross tabulations? (which test statistic?)

Key takeaways

You should be able to answer these questions:

1. How do we test the existence of association?
   we use test statistics and corresponding p-values
2. When comparing (group) means? (which test statistic?)
   z-statistic (σ known) or t-statistic (σ unknown)
3. Cross tabulations? (which test statistic?)
   chi-square statistic
4. How to measure the strength of association?

Key takeaways

You should be able to answer these questions:

1. How do we test the existence of association?
   we use test statistics and corresponding p-values
2. When comparing (group) means? (which test statistic?)
   z-statistic (σ known) or t-statistic (σ unknown)
3. Cross tabulations? (which test statistic?)
   chi-square statistic
4. How to measure the strength of association?
   λ, Cramer's V, and γ
5. What does $\lambda = V = 1$ mean? How about $\gamma = 1$?

Key takeaways

You should be able to answer these questions:

1. How do we test the existence of association?
   we use test statistics and corresponding p-values
2. When comparing (group) means? (which test statistic?)
   z-statistic (σ known) or t-statistic (σ unknown)
3. Cross tabulations? (which test statistic?)
   chi-square statistic
4. How to measure the strength of association?
   λ, Cramer's V, and γ
5. What does $\lambda = V = 1$ mean? How about $\gamma = 1$?
   Perfect association and perfect **positive** association.

STATA cheat sheet

How to adopt them in STATA?

The **existence** of association

- Comparison of means (t statistic): `ttest Y, by(X)`
- Cross-tab (χ^2): `tab Y X, chi`

The **strength** of association

- λ: manually or `lambda Y X` (You have to install the package)
- Cramer's V: `tab Y X, V`
- γ: `tab Y X, gamma`

A tip: take advantage of `help command!` (e.g. `help ttest`)