STAT 231

December 5, 2016.

Office hours: This week.

Tuesday: 2 - 3
Wednesday: 2 - 3
Thursday 2 - 4

} M3, 2017

Roadmap

- Step-by-Step method of solving contingency table problems.
- Statistical significance vs. Practical Significance
- Correlation vs. Causation
  - → Blocking
  - → Randomization
- Examples.

Tests For Independence of Categorical Variables

Given: Data of units divided into

Categories	Math. B	Non-Math. B^c	
Left A	y_{11}	y_{12}	$y_{11}+y_{12}$ 40.
Right A^c	y_{21}	y_{22}	$y_{21}+y_{22}$ 60.
	$y_{11}+y_{21}$ = 50	$y_{12}+y_{22}$ = 50	n 100

Objective: To test whether there is an association between A and B.

Step 1: Construct the expected frequency table.

© M NM.

L	e_{11}	e_{12}
R	e_{21}	e_{22}.

$$e_{ij} = \frac{r_i \times c_j}{n} \Rightarrow$$

where r_i = # sum of row i

c_j = sum of column j

Step 2:

Compute the value of your test. statistic

$$\lambda = 2 \sum_i \sum_j y_{ij} \log \frac{y_{ij}}{e_{ij}}$$

Step 3: Compute the p-value.

p-value: $P(\Lambda \geq \lambda)$

$$\Lambda \sim \chi^2_{(a-1)(b-1)}$$

a = # of rows

b = # of columns

For this problem,

$$\Lambda \sim \chi^2_1$$

So the p-value can be calculated directly.

	M	NM.	
L	30 e_{11}	70 e_{12}	100
R	50 e_{21}	50 e_{22}	100
	80	120	200

Question: Find a 95% C.I for proportion of Left Handed Math majors

$$\hat{\pi} = 30/100 = 0.3$$

C.I for Bin = $\hat{\pi} \pm z^{\alpha} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

Some final points

- For χ^2 test-statistic, if the df = 1, we should use z^2, if df = 2 we should use Exp(2)
- Goodness of fit tests. μ
  we have to make sure that $\left.\begin{matrix} y_j \geq 5 \\ n \geq 50 \end{matrix}\right\}$

$X_1, \ldots X_n \sim \text{Poi}(\theta)$

	Freq
0	10
1	15
2	10
3	7
4	5
5	3
≥ 6	3

Sometimes, we can merge the categories to make the freq ≥ 5

STATISTICAL SIGNIFICANCE VS PRACTICAL SIGNIFICANCE

Example

Difference in premiums paid by smokers and non-smokers.

$$\mu_1 - \mu_2 \geq 5$$

S N

Sometimes, a result might be st statistically significant but not have any policy implications → No PRACTICAL SIGNIFICANCE

CORRELATION $\not\Rightarrow$ CAUSATION.

$$Y = \alpha + \beta X + R$$

$$R \sim G_R(0, \sigma)$$

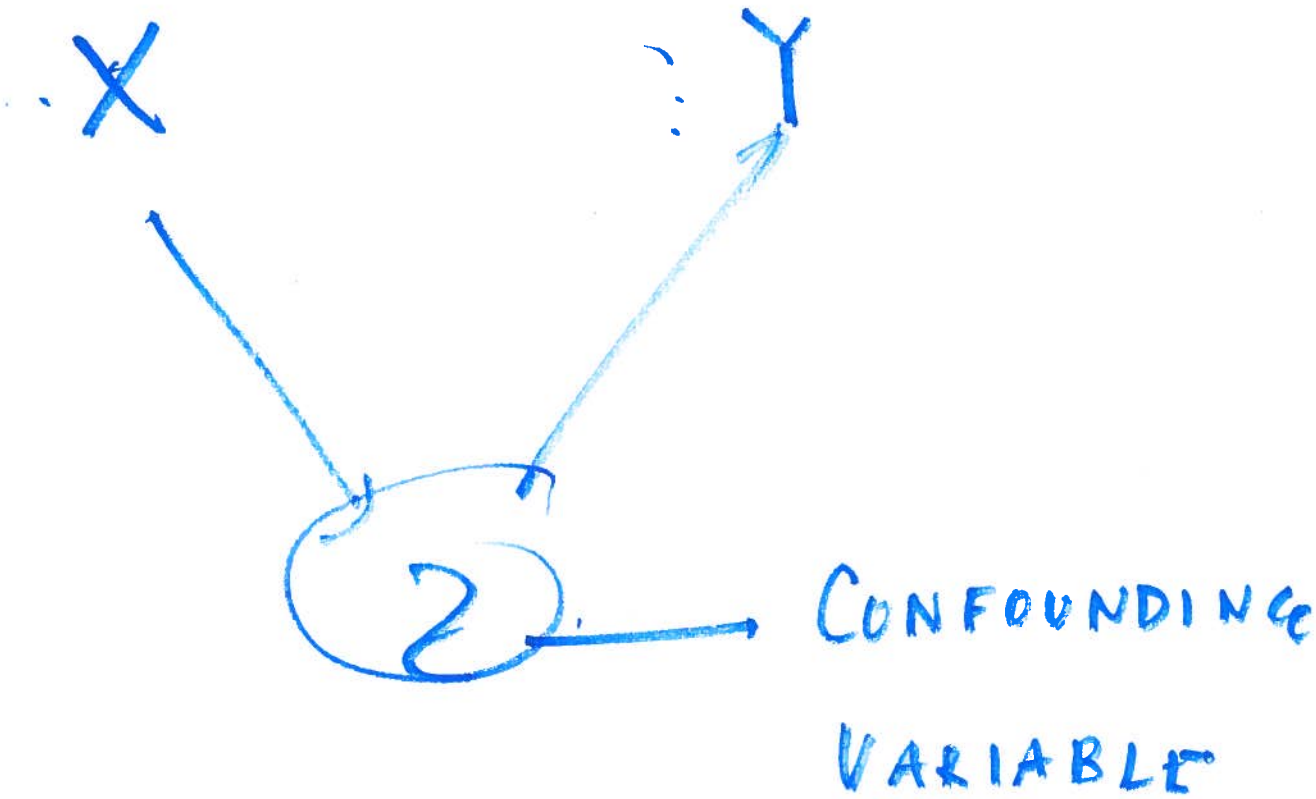
$$H_0 : \beta = 0$$

The test is rejected

$\Downarrow$

X and Y are correlated, but it does not imply that X causes Y.

X "causes" Y if all other things equal, a change in X causes a change in the distribution of Y.



Solutions

- Blocking: Collect data with the value of all confounding variables fixed

Difficult to verify all confounding variables.

- Randomization: We divide the sample into two random groups.

⇒ the samples are equivalent in all ways but X and Y.