Estimating Bivariate Association

Statistical Research Methods I

Seongsoo Choi (최성수)

This week: estimating bivariate association

- Association of two categorical variables
  - χ^2 test
  - Odds ratio
- Association of two continuous variables
  - Covariance
  - Correlation

Two Categorical Variables

Contingency Table

- The table displaying the number of subjects observed at all combinations of possible outcomes for the two (categorical) variables
- Example:

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	495	590	272	1357
Males	330	498	265	1093
Total	825	1088	537	2450

Political party identification and gender (GSS)

Percentage comparisons: conditional probabilities

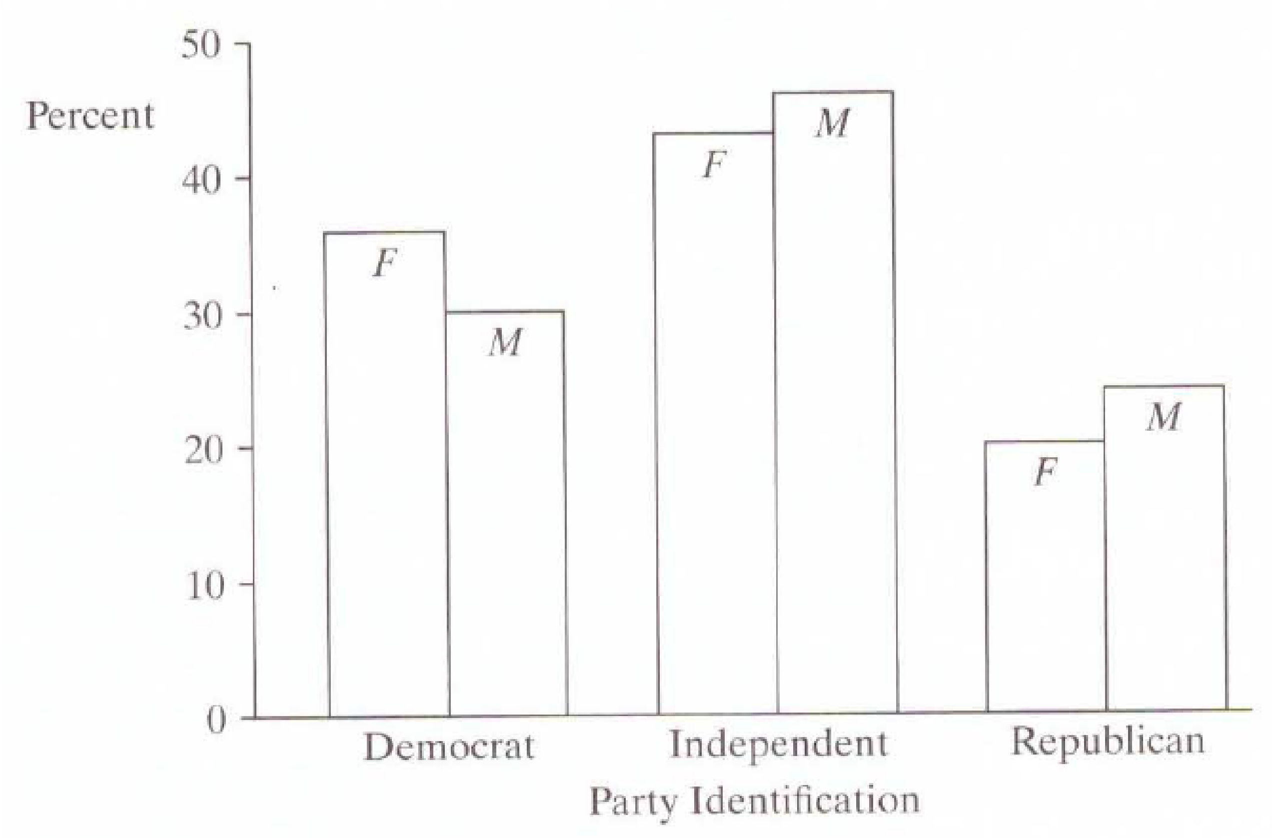
- Row percentage / column percentage
- No difference in direction or order: Which one should be columns, which one should be rows
- Recall three probabilities
  - conditional
  - joint
  - marginal

Conditional probabilities

	Party Identification				
Gender	Democrat	Independent	Republican	Total	n
Females	36%	43%	20%	100%	1357
Males	30%	46%	24%	100%	1093

Political party identification and gender (GSS)

- What does this table show?



Political party identification and gender (GSS)

Dependence and independence

- Two categorical variables are *statistically independent*
  - if the population conditional distributions on one of them are identical at each category of the other
  - The variables are *statistically dependent* if the conditional distributions are not identical

	Party Identification			
Ethnic Group	Democrat	Independent	Republican	Total
White	3500 (35%)	4000 (40%)	2500 (25%)	10,000 (100%)
Black	350 (35%)	400 (40%)	250 (25%)	1000 (100%)
Hispanic	875 (35%)	1000 (40%)	625 (25%)	2500 (100%)

Party identification and racial/ethnic groups

Chi-squared (χ^2) test of independence

- We would like to figure out whether the observed sample association between two categorical variables (e.g., gender and party identification) would hold even in the population

- That is, we would like to draw a statistical inference from the sample contingency table

- Then, what is the null hypothesis?

Chi-squared (χ^2) test of independence

- We would like to figure out whether the observed sample association between two categorical variables (e.g., gender and party identification) would hold even in the population
- That is, we would like to draw a statistical inference from the sample contingency table
- Then, the null hypothesis is:
  - H_0: The two variables are statistically independent
- Can we reject this null hypothesis?

Expected frequencies for independence

- We can imagine a hypothetical contingency table where frequencies across cells satisfy independence (that is, the null hypothesis)
- This hypothetical table has the same row and column marginal totals as the observed frequencies but at the same time satisfy independence. They are called *expected frequencies* (f_e, note the notation for observed frequences is f_o)

$$f_e = \frac{\text{row total} \times \text{column total}}{\text{sample total}}$$

Excercise

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Female	495 ( )	590 ( )	272 ( )	1357
Male	330 ( )	498 ( )	265 ( )	1093
Total	825	1088	537	2450

Expected frequencies are in parentheses

Excercise

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Female	495 (456.9)	590 (602.6)	272 (297.4)	1357
Male	330 (368.1)	498 (485.4)	265 (239.6)	1093
Total	825	1088	537	2450

Expected frequencies are in parentheses

The Pearson statistic for testing independence

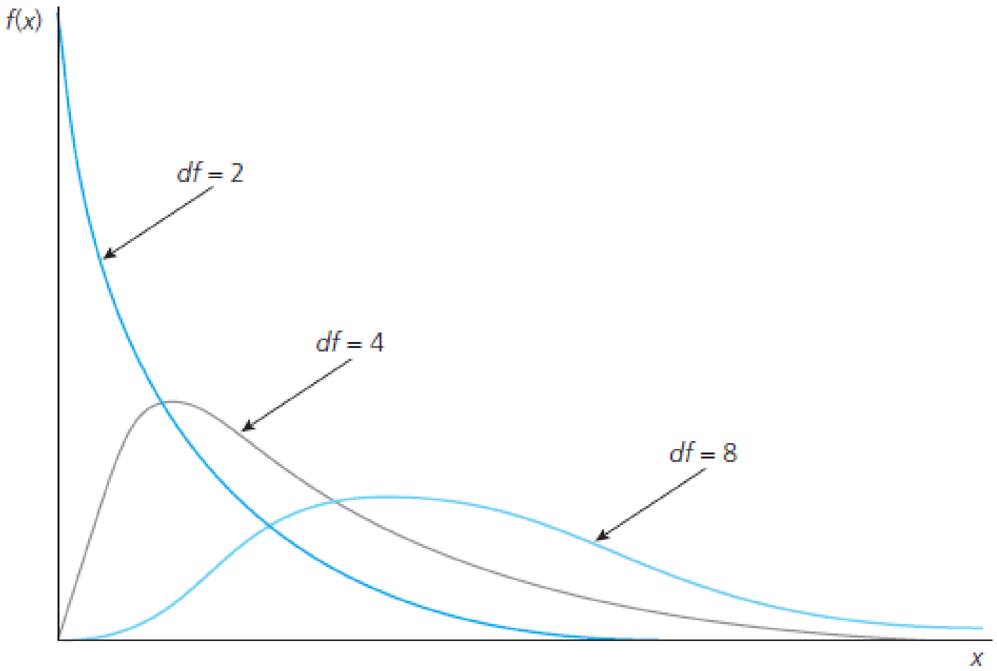
$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- When H_0 is true (independence), f_o and f_e tend to be close for each cell, and χ^2 is relatively small
- If H_0 is false, at least some f_o and f_e values tend not to be close, leading to large $(f_o - f_e)^2$ values and a large test statistic
- The larger χ^2 value, the greater the evidence against H_0: independence
- The Pearson statistic χ^2 follows the chi-squared probability distribution where the degrees of freedom (df) is

$$df = (r-1)(c-1)$$

Chi-squared (χ^2) distribution

- Sensitive to the degrees of freedom (df): skewed to the right when df is small but more bell-shaped as df increases
- $X \sim \chi^2_n$ where n is df, $X = \sum_{i=1}^{n} Z_i^2$, Z is a standard normal RV



The Chi-squared (χ^2) test

- Using the Chi-squared (χ^2) distribution table, find the corresponding p-value (or the range of the p-value)
- If the p-value is small enough (conventionally, smaller than 0.05), the null hypothesis can be thought to be highly implausible and so can be rejected
  - We found evidence of statistically significant association between two variables
- If the p-value is not small enough (e.g., greater than 0.05), we fail to reject (*not accept* or *verify*) the null hypothesis ()
  - We did not find statistical evidence that two variables are systematically associated
  - We can say that the observed sample association is a product of chance rather than a product of their true association

Excercise: Find χ^2 statistic and p-value

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Female	495 (456.9)	590 (602.6)	272 (297.4)	1357
Male	330 (368.1)	498 (485.4)	265 (239.6)	1093
Total	825	1088	537	2450

Expected frequencies are in parentheses

- Are gender and party identification statistically significantly associated?

Odds

- Note that the χ^2 statistic is not a measure of the association between two variables
- A popular measure for the association between two categorical variables is the odds ratio

$$\text{Odds} = \frac{Pr(\text{success})}{Pr(\text{failure})}$$

$$Pr(\text{success}) = \frac{\text{Odds}}{\text{Odds} + 1}$$

- When odds=3, a success is three times as likely as a failure; when odds=0.75, a success is 75% more likely than a failure
- When odds=1?

Odds ratio

- In 2 by 2 contingency table defined by two dichotomous categorical variables, odd ratio is the ratio of odds in one group to odds in the other group

	Race of Victim		
Race of Offender	White	Black	Total
White	2509 a	409 b	2918
Black	189 c	2245 d	2434

- Odds ratio $\theta = \frac{\text{Odds for white offenders}}{\text{Odds for black offenders}} = \frac{6.13}{0.0842} = 72.9$
- Odds ratio $\theta = \frac{a/b}{c/d} = \frac{a \times d}{b \times c}$

Properties of odds ratio

- The odds ratio takes the same value regardless of the choice of response variable
- When the success probabilities are identical in the two rows, $\theta = 1$ (no association)
- When $\theta > 1$, the odds of success are higher in row 1 than in row 2 (positive association)
- When $\theta < 1$, the odds of success are higher in row 1 than in row 2 (negative association)
- Odds ratio is not sensitive to marginal distributions: odds ratio captures a bivariate association that is not influenced by differences in the margins
- In a $r \times c$ contingency table, we can get $(r-1) \times (c-1)$ odd ratios

Two Continuous Variables

Covariance as a measure of bivariate association

- Say, for two continuous variables, X and Y, $\mu_x = E(X)$ and $\mu_y = E(Y)$
- Recall that $Var(X) = \sigma_X^2 = E[(X - \mu_X)^2]$
    - in a sample, $Var(X) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$
- The covariance of X and Y is: $Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$
    - in a sample, $Cov(X, Y) = \frac{\sum_{i=1}^{N}[(X_i - \bar{X})(Y_i - \bar{Y})]}{N-1}$

Properties of covariance

- Covariance indicates how a deviation of one variable from its mean is associated with a deviation of the other from the mean
- Covariance measures the amount of linear dependence between two continuous variables
  - A positive value indicates that two variables move in the same direction
  - A negative value indicates that two variables move in opposite directions
- If X and Y are independent, $Cov(X, Y) = 0$ (note that the reverse is not necessarily true)
- $Cov(a, b) = 0$ where a and b are constants
- $Cov(aX + b, cY + d) = acCov(X, Y)$

The Pearson correlation coefficient

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{\sigma_{XY}}{\sigma_x \sigma_y}$$

- The correlation coefficient is the value we get from dividing the covariance by the product of the standard deviations of the two variables
- $Corr(X,Y) = 1$: complete positive linear dependency
- $Corr(X,Y) = -1$: complete negative linear dependency

Properties of the correlation coefficient

- The correlation coefficient is the standardized version of covariance
  - All correlation coefficient ranges between -1 and 1: $-1 \leq Corr(X, Y) \leq 1$
  - If $Corr(X, Y) = 0$ or equivalently $Cov(X, Y) = 0$, there is no linear relationship between X and Y (or the two variables are uncorrelated)
  - But note that $Corr(X, Y) = 0$ doesn't necessarily mean $X \perp Y$ (independence)
    - $Corr(X, Y) = 0$ means X and Y are *linearly independent*
    - They can be dependent in nonlinear way
- $Corr(aX + b, cY + d) = Corr(X, Y)$ if $ac > 0$
- $Corr(aX + b, cY + d) = -Corr(X, Y)$ if $ac < 0$

Excercise

	Change in math score	Change in reading score
Chris	1	3
Jamal	-2	2
Jieun	3	4
Yingyao	0	6
Sho	3	0

- Find the covariance and correlation coefficient of the two variables
- Are changes in math and reading scores correlated? Explain how they are associated.

Excercise

	Change in math score	Change in reading score
Chris	1	3
Jamal	-2	2
Jieun	3	4
Yingyao	0	6
Sho	3	0

- $\bar{X} = 1$ and $\bar{Y} = 3$
- $Cov(X, Y) = \frac{-4}{4} = -1$
- $SD(X) = \frac{18}{4} = \sqrt{4.5} = 2.1213$ and $SD(Y) = \frac{20}{4} = \sqrt{5} = 2.2361$
- $Corr(X, Y) = \frac{-1}{(2.1213)\times(2.2361)} = -0.2108$