Association Between Two Categorical Variables (Ch. 11)

Basics of Contingency Tables

Contingency tables give us a way to study the relationship between two **categorical** variables. (e.g. relationships between gender and autoimmune disease, smoking and coronary heart disease, etc.).

- Data is observed counts (i.e. # of observations in each group).
- We present the data in the contingency table.
- We compare groups using percentages that are conditional probabilities.
- We also have χ^2 tests for **independence** of 2 categorical variables.

A, B
$P(A|B) = \frac{P(AB)}{P(B)}$

Assumptions for Contingency Table Inference:

- SRS – sample random, representative of population of interest.
- Minimum of 5 observations per cell (True assumption: min 5 expected counts per cell. In practice, we go for 5obs)

Example: Does money make people happy? The General Social Survey (GSS) collected data on 1993 individuals who were classified into three categories for family income: Above Average, Average, and Below Average. They surveyed these individuals for happiness level: Not Too Happy, Pretty Happy, or Very Happy. The contingency table gives the following observed counts.

	Happiness Level			
Family Income	Not Too Happy	Pretty Happy	Very Happy	**Total**
Above Average	26	233	164	423
Average	117	473	293	883
Below Average	172	383	132	687
Total	315	1089	589	1993

$\frac{315}{1993} \times 423$

$\frac{1089}{1993} \times 883$

2 categorical variables: FAM Income, Happiness level.

explanatory VAR: Income

response VAR: Happiness

Assumptions: • SRS – GSS ✓ American adults. Year?

• 5 obs per cell ✓
(5 = min)

A, B $P(A|B) = \frac{P(AB)}{P(B)}$ A = {get 2} B = {get an even number}

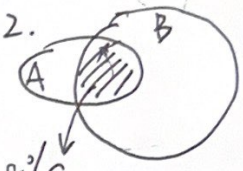
$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$

Conditional Probabilities

A' = {get 2 or 3} $P(A'|B) = \frac{P(A'B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{1}{3}$

→ to see relationship

compute conditional prob. for each income level (EXP VAR)



$\frac{P(A)}{P(B)}$ $\frac{P(C)}{P(B)}$ X Y

- What % of ~~PPL Above Ave~~ Average Income ppl are VERY HAPPY $\frac{164}{423} = 0.388 = 38.8\%$
- . . . . . Average Income ppl are VERY HAPPY $\frac{293}{883} = 33\%$
- . . . . Below Average Income ppl are VERY HAPPY. $\frac{132}{687} = 19\%$

Joint prob. $P(X=x, Y=y)$ $P(A,B)$

marginal prob. $P(A)$ $P(B)$

- What % of VERY HAPPY ppl have Below Avg Income: $\frac{132}{589} = 22\%$

Marginal Probabilities — looking at one variable at a time.

% ppl are NOT TOO HAPPY : $\frac{315}{1993} = 15\%$
pretty HAPPY : $\frac{1089}{1993} = 55\%$
very HAPPY : $\frac{589}{1993} = 30\%$
} Marginal Distribution of Happiness

% ppl have Income Above Avg : $\frac{423}{1993} = 21\%$
Avg : $\frac{883}{1993} = 44\%$
Below Avg : $\frac{687}{1993} = 34\%$
} Marginal Distribution of Income

Prob. for INTERSECTION

% of ppl are not too Happy AND have Above Avg Income: $\frac{26}{1993} = 1.3\%$

Test for Independence for Two Categorical Variables

- **Observed** counts are given in the contingency table.
- **Expected counts** for each cell are computed as follows:

$$\text{exp counts} = \frac{\text{row tol} \times \text{col tol}}{\text{big tol}}$$

Our procedure for testing independence is as follows:

- **Hypotheses:** H_0: Two categorical variables are independent (NOT ASSOCIATED)

  H_a: " " " " are NOT independent (ASSOCIATED)

- **Test Statistic:**

$$TS: \chi^2 = \sum \frac{(obs - exp)^2}{exp}$$

- **p-value:**

- **Conclusions:**

- **DF**

Cautions:

- Small p-value -> **strong evidence** of association, NOT evidence of a **strong association**
- Large p-value -> not enough evidence of association, NOT proof the variables are independent
- Just because quantitative data can be turned into categories does NOT mean you should use χ^2 tests for everything!