Contigency Tables

Dr. Mutua Kilai

Department of Pure and Applied Sciences

2024-01-25

Introduction

- Let X and Y be two categorical variables, X with I categories and Y with J categories.
- Classifications of subjects on both variables have IJ combinations.
- The responses (X, Y) of a random subject have a probability distribution.
- A rectangular table having I rows and J columns displays the distribution. The cells of the table represent the IJ outcomes.
- When the cells contain frequency counts, the table is called a contingency table or cross-classification table.

Probability Structure

- Let π_{ij} denote the probability that (X, Y) occurs in the cell in row i and column j.
- The probability distribution $\{\pi_{ij}\}$ is the joint distribution of X and Y
- The marginal distribution are the row and column totals that result from summing the joint probabilities denoted by $\{\pi_{i+}\}$
- Probabilities sum to one:

$$\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1$$

- The probabilities $\{\pi_{1|i}, ..., \pi_{j|i}\}$ form conditional distribution of Y at category i of X

Example

Consider the treatment and placebo example illustrated in this table

	Treatment	Placebo
Improvement	99	60
No improvement	211	242

- Calculate the marginal and conditional probabilities.

Marginal Probabilities

- The marginal probability is the chance of an event, but completely ignores the influence/effect of other factors.

$$p_I = \frac{159}{612} = 0.259, p_{NI} = \frac{463}{612} = 0.741$$

Joint Probabilities

	Treatment	Placebo
Improvement	99 $(\frac{99}{310} = 0.32)$	60 $(\frac{60}{302}) = 0.2$
No improvement	211 $(\frac{211}{310}) = 0.68$	242 $(\frac{242}{302}) = 0.8$

We see that 32% percent of the Mindoxil group saw an increase in hair growth, whereas only 20% of Placebo group saw an increase. These are called conditional probabilities, because they are conditional on a treatment.

Conditional Probabilities

- $$P(A|B) = \frac{P(A\&B)}{P(A)}$$

- $$P(I|T) = \frac{P(I\&T)}{P(I)} = \frac{0.32}{0.62}$$

Chi-Square test of association

- The chi-squared test tests the hypothesis that there is no relationship between two categorical variables.
- It compares the observed frequencies from the data with frequencies which would be expected if there was no relationship between the variables.
- The Null hypothesis is stated as:

$$H_0 :$$

There is no association between the two categorical variables

- The alternative hypothesis is given as:

$$H_a$$

: There is association between the two categorical variables.

Expected Values and Test Statistic

- The expected frequencies are computed as:

$$E_{ij} = \frac{Row\ \ total_i \times Column\ \ Total_i}{Grand\ \ Total}$$

- The calculated χ^2 statistic is given as:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  where O_{ij} observed cell value and E_{ij} expected cell frequency.

- The statistic follows a χ^2 with $(r-1) \times (c-1)$ degrees of freedom

Example

Electronic devices are made on three production lines. Records are kept of faults found on devices made on each line. Faults are classified as "electronics", "power supply" or "mechanical". The data are as follows.

	Production line		
	1	2	3
Electronic	13	33	15
Power supply	7	4	11
Mechanical	18	10	14

Test the hypothesis that there is no association between production line and type of fault. Use the 5% level of significance.

Solution

2. Observed frequencies:

	Production Line			
	1	2	3	Total
Electronic	13	33	15	61
Power supply	7	4	11	22
Mechanical	18	10	14	42
Total	38	47	40	125

Expected frequencies, e.g. $61 \times 38/125 = 18.544$.

	Production Line			
	1	2	3	Total
Electronic	18.544	22.936	19.520	61
Power supply	6.688	8.272	7.040	22
Mechanical	12.768	15.792	13.440	42
Total	38.000	47.000	40.000	125

Test statistics

$$W = \sum \frac{(O-E)^2}{E} = \frac{(13-18.544)^2}{18.544} + \cdots + \frac{(14-13.440)^2}{13.440} = 15.860.$$

Degrees of freedom: $(3-1) \times (3-1) = 4$.

Critical value: $\chi^2_4(5\%) = 9.488$.

The test statistic is significant at the 5% level. We reject the null hypothesis and conclude that there is an association between fault type and production line. In particular there seems to be an excess of electronic faults on Line 2.

In R

```r
# Creating a contingency table

data <- matrix(c(13,33,15,7,4,11,18,10,14), nrow = 3,
               byrow = TRUE)

colnames(data) <- c("1", "2", "3")

rownames(data) <- c("Electronic", "Power supply",
                    "Mechanical")

# Performing the chi-square test of independence

chisq.test(data)
```

Thank You!