Hypothesis Tests
Tests for Categorical Data

DASC 512

Overview

- Multinomial Experiments
- Chi-squared test for one-way multinomial experiments
- Contingency test for two-way multinomial experiments

Multinomial Experiments

In week 2, we talked about the binomial experiment

- Experiment of n iid Bernoulli trials
- There are 2 possible mutually exclusive outcomes to each trial: success (1) or failure (0)
- The probability of success is p

Multinomial Experiments

The multinomial experiment extends that concept into multiple classes

- Experiment of n iid trials
- There are k possible mutually exclusive outcomes to each trial (classes, categories, or cells)
- The probability of each outcome is p_i where $\sum_i p_i = 1$
- The random variable of interest is the vector $(n_1, n_2, \dots, n_k)$ of cell counts, the number of observations of each class

One-way Table

A one-way (single variable) table is used to summarize outcomes from a multinomial experiment. It looks like a frequency table.

For each of the categories of the multinomial variable (represented by cells in the table), counts are listed that result from the experiment.

Classes are compared by the proportions of counts in each cell to the expected counts, according to some hypothesized proportions.

Multinomial Experiment?

Grade	B	B+	A-	A	Total
	82	103	128	97	**410**

Assumptions

- Type of data – Qualitative
- Randomization – Data gathered randomly (iid)
- Population distribution – Assumes observed proportions are normally distributed (using the normal approximation to the binomial)
- Sample size – Sample size must be large. $np_{i,0} > 5$ for each i.

Hypotheses

Null hypothesis: Every proportion is equal to some known value.

$$H_0: p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_k = p_{k,0}$$

Alternative hypothesis: At least one proportion is different

$$H_a: p_i \neq p_{i,0} \text{ for some } i \in \{1,2,\ldots,k\}$$

Test Statistic

Assuming observed proportions are normally distributed, the sum of squared deviations divided by the expected number of observations is χ^2 distributed.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i}$$

where

$$E_i = n_i p_{i,0}$$

Conclusion

- Reject the null hypothesis if

$$\chi^2 > \chi^2_{ISF}(\alpha, df = k - 1)$$

Example

Let's revisit our multinomial example. Are the probabilities of each grade equal?

$$H_0: p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$$

$$H_a: p_i \neq \frac{1}{4}, \text{for some } i \in \{1,2,3,4\}$$

Grade	**B**	**B+**	**A-**	**A**	**Total**
	82	103	128	97	**410**

Example

Test Statistic:

$$E_i = np_{i,0} = \frac{410}{4} = 102.5 \text{ for all } i$$

$$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - E_i)^2}{E_i} = \frac{(82 - 102.5)^2}{102.5} + \cdots + \frac{(97 - 102.5)^2}{102.5} = 10.74$$

$$\chi^2_{ISF}(0.05, df = 3) = 7.81$$

Grade	B	B+	A-	A	Total
	82	103	128	97	**410**

Example 2

What if we had hypothesized proportions?

$$H_0 : p_1 = \frac{1}{5}, p_2 = \frac{3}{10}, p_3 = \frac{3}{10}, p_4 = \frac{1}{5}$$

$$H_a : p_i \neq p_{i,0} \text{ for some } i$$

Grade	B	B+	A-	A	Total
	82	103	128	97	**410**

Two-Way Contingency Table

Two-way contingency tables (aka cross-tabs) classify data from a multinomial experiment with two qualitative variables.

This is used to examine relationships between the qualitative variables with respect to the experiment.

We use R_i to denote row i and C_j to denote column j. n_{ij} is the cell count.

	Grades		
Program	B	A	Total
Masters	141	161	305
PhD	44	61	105
Total	185	225	410

Hypotheses (Contingency Test)

Null hypothesis: The classifications are independent.

Alternative hypotheses: The classifications are dependent.

In this example, the null hypothesis is that Masters and PhD students get the same proportions of As and Bs. Alternative: they do not.

	Grades		
Program	B	A	Total
Masters	141	161	305
PhD	44	61	105
Total	185	225	410

Expected Cell Count

If the 2 factors are independent,

$$P\left(R_i \cap C_j\right) = P(R_i)P\left(C_j\right)$$

So, the expected cell count assuming independence is

$$E_{ij} = np_{R_i}p_{C_j} = n\left(\frac{R_i}{n}\right)\left(\frac{C_j}{n}\right) = \frac{R_iC_j}{n}$$

As before, summing squared deviations between actual and expected proportions allows us to use the χ^2 distribution.

Assumptions

- Type of data – Qualitative
- Randomization – Data gathered randomly (iid)
- Population distribution – Assumes observed proportions are normally distributed (using the normal approximation to the binomial)
- Sample size – Sample size must be large. Expected cell count must be at least 5 for every cell.

Test Statistic

Again assuming observed proportions are normally distributed,

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

where

$$E_{ij} = \frac{R_i C_j}{n}$$

Conclusion

- Reject the null hypothesis if

$$\chi^2 > \chi^2_{ISF}\big(\alpha, df = (r-1)(c-1)\big)$$

Example

A university wants to see if grades vary by degree type.

H_0: Grades do not vary by degree type.

H_a: Grades do vary by degree type.

	Grades		
Program	B	A	Total
Masters	141	161	305
PhD	44	61	105
Total	185	225	410

Example

$$\chi^2 = 0.54$$
$$\chi^2(0.05, df = 1) = 3.84$$
$$p = 0.463$$

We fail to reject the null hypothesis that grades vary by degree type.

	Grades		
Program	B	A	Total
Masters	141	161	305
PhD	44	61	105
Total	185	225	410

Recap

- Multinomial Experiments
- Chi-squared test for one-way multinomial experiments
- Contingency test for two-way multinomial experiments

Congratulations!

We are now half-way through the course! This is the end of the material covered by the mid-term exam.