Analysis of Categorical Data

Outline

- The multinomial experiment
- Pearson's Chi-square statistic
- Testing specified cell probabilities
- Contingency tables: a two-way classification

The multinomial experiment

- The experiment consists of n identical trials
- The outcome of each trial falls into one of k categories
- The probability that the outcome of a single trial falls into a particular category is p_i ($0 \leq p_i \leq 1$, $\sum p_i = 0$) and remains constant from trial to trial.
- Trials are independent.
- The experimenter counts the observed number of outcomes in each category, O_1, O_2, …, O_n with $\sum O_i = n$.

Pearson's Chi-square Statistic

$$X^2 = \sum \frac{(O_i - E_i)^2}{\mathrm{E}_i} = \sum \frac{(O_i - np_i)^2}{np_i}$$

- X^2 has **chi-square probability distribution** in repeated sampling.
- If the expected cell counts E_i are different to the observed cell counts O_i => X^2 is large => **use right-tailed statistical test** and look for unusually large value of the test statistic X^2.
- **Important.** Expected cell counts E_i should be larger than 5. If not,
    - Increase sample size n
    - Combine smaller cells

The goodness-of-fit test (Testing specified cell probabilities)

- H_0: The specified multinomial probabilities p_1, p_2, …, p_k are **true**
- Ha: The specified multinomial probabilities p_1, p_2, …, p_k are **false** (i.e. observations don't follow these probabilities)

- Test statistic: $X_t^2 = \sum \frac{(O_i - E_i)^2}{\mathrm{E}_i}$, where $E_i = np_i$, with a degree of freedom of $\mathrm{df} = (k-1)$

- If the value of X_t^2 is very large, i.e. $\mathrm{P}(X^2 > X_t^2) < \alpha$, reject the null hypothesis H_0.

The goodness-of-fit test (Testing specified cell probabilities)

Example 1. A researcher sent a rat into a ramp the end of which divides into 3 different doors of three different colours. The number of times the rat enters each door is in the table below.

Does the rat have a preference for one of the three doors?

	Door		
	Green	Red	Blue
Observed counts	20	39	31

The goodness-of-fit test (Testing specified cell probabilities)

Example 2. The proportions of blood phenotypes A, B, AB, and O in a population are 41%, 1%, 4%, and 45%, respectively. A random sample of 200 people were selected from the population, whose blood phenotypes are presented in the below table. Test the goodness of fit of these blood phenotypes proportions.

	A	B	AB	O
Observed	89	18	12	81

Contingency tables: A two-way classification

- Observed counts are structured along 2 dimensions (variables).
- An important/common interest is in examining the relationship between the two variables.
- In other words, is one method of classification dependent on the other method of classification?

	Shift			
Type of Defects	1	2	3	Total
A	15	26	33	74
B	21	31	17	69
C	45	34	49	128
D	13	5	20	38
Total	94	96	119	309

	No Vaccine	One Shot	Two Shots	Total
Flu	24	9	13	46
No Flu	289	100	565	954
Total	313	109	578	1000

The Chi-square test of independence

- H_0: the two methods of classification are **independent**
- Ha: the two methods of classification are **dependent**

- Test statistic: $X_t^2 = \sum \frac{\left(O_i - E_{ij}\right)^2}{\mathrm{E_{ij}}}$, where $E_{ij} = np_{ij} = n\frac{r_i}{n}\frac{c_j}{n}$, with a degree of freedom of $\mathrm{df} = (r-1)(c-1)$

- If the value of X_t^2 is very large, i.e. $\mathrm{P}(X^2 > X_t^2) < \alpha$, reject the null hypothesis H_0.

The Chi-square test of independence

Example 3. Does the data presented in the table below provide enough evidence to conclude that there is a dependence between defect types and shifts?

	Shift			
Type of Defects	1	2	3	Total
A	15	26	33	74
B	21	31	17	69
C	45	34	49	128
D	13	5	20	38
Total	94	96	119	309

The Chi-square test of independence

Example 4. The below table presents results from a survey conducted to evaluate the effectiveness of a new flu vaccine.

Is there enough evidence to conclude that the vaccine was successful in reducing the number of flu cases?

	No Vaccine	One Shot	Two Shots	Total
Flu	24	9	13	46
No Flu	289	100	565	954
Total	313	109	578	1000