Statistical inference (MVE155/MSG200)

Categorical data analysis

Set-up

Categorical data consist of observations which can be divided into different categories/groups, each observation belonging to one category.

We are interested in a situation, where we have two categorial factors,

- A with I categories
- B with J categories.

Two (three) tests:

- (goodness-of-fit test (Chapter 4))
- test of homogeneity
- test of independence

Questions/tests

Example: We would like to investigate whether the preferred color of a cell phone depends on gender.

Two questions/tests:

- Test of homogeneity: We have independent random samples from J multinomial distributions and would like to know whether the distributions are the same. Concerning the example above, we may decide to ask 100 women and 100 men about the color of their phone and test, whether the proportions of the different colors (color distributions) are similar in the two groups. Color is a random variable.
- Test of independence: We have one random sample of the population and we would like to know whether there is a relationship between the two categorical factors. Concerning the example, we may ask a store celling cell phones which color phones they have sold to each of the genders. In this case, both color and gender are random variables.

Test of homogeneity

Below, we have the observed counts from J independent random samples (factor B) in I categories (A):

	Sample 1	Sample 2	$\cdots$	Sample J	Total counts
Category a_1	c_{11}	c_{12}	$\cdots$	c_{1J}	c_1
Category a_2	c_{21}	c_{22}	$\cdots$	c_{2J}	c_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Category a_I	c_{I1}	c_{I2}	$\cdots$	c_{IJ}	c_I
Sample sizes	n_1	n_2	$\cdots$	n_J	n

I.e. we have random samples from J multinomial distributions

$$(C_{1j}, ..., C_{Ij}) \sim Mn(n_j; \pi_{1|j}, ..., \pi_{I|j}), \quad j = 1, ..., J,$$

where

$$\pi_{i|j} = \mathrm{P}(A = a_i | B = b_j) = \frac{\mathrm{P}(A = a_i, B = b_j)}{\mathrm{P}(B = b_j)} =: \frac{\pi_{ij}}{\pi_{.j}}.$$

Test of homogeneity

We test the hypothesis

$$H_0 : \pi_{i|j} = \pi_{i\cdot} \quad \text{for all} \quad (i, j).$$

The ML estimates for the sample proportions are $\hat{\pi}_i = c_i/n$, $i = 1, ..., I$, leading to the expected cell counts

$$e_{ij} = n_j \hat{\pi}_i = c_i n_j / n.$$

We use the test statistic

$$x^2 = \sum_{i=1}^{I} \sum_{j=i}^{J} \frac{(c_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{I} \sum_{j=i}^{J} \frac{(c_{ij} - c_i n_j/n)^2}{c_i n_j/n}$$

and reject the null hypothesis with large values of x^2. The stochastic version of the test statistic has the distribution χ^2_{df} under H_0 with

$$df = J(I-1) - (I-1) = (I-1)(J-1).$$

Example

The following data were collected from 101 patients with Hodgkin's disease and 107 controls without the disease to compare the percentages of tonsillectomy (removal of tonsils) in these two groups:

	Hodgin's (A)	Control ($\bar{A}$)	Total
Tonsillectomy (E)	67	43	110
No tonsillectomy ($\bar{E}$)	34	64	98
Total	101	107	208

Example (continues)

The noll hypothesis is

H_0 : proportions (tonsillectomy/not) are the same in the two groups

and the alternative hypothesis

H_1 : proportions are not the same in the two groups

Example (continues)

The observed (left) and expected (right) counts ($I = J = 2$)

	A	$\bar{A}$	T
E	67	43	110
$\bar{E}$	34	64	98
T	101	107	208

	A	$\bar{A}$
E	53.14	56.59
$\bar{E}$	47.59	50.41

The test statistics (χ^2-distributed with (2-1)(2-1)=1 degrees of freedom under H_0) takes the value

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 14.26.$$

The observed value 14.26 is much larger than the critical value 5.02 from the χ^2_1-distribution using the significance level 5%. We can reject H_0 and conclude that tonsillectomy was more common in the group with Hodgin's disease than in the control group.

Test of independence

Below, we have the observed counts from a single random sample of size n:

	b_1	b_2	$\cdots$	b_J	Total
a_1	c_{11}	c_{12}	$\cdots$	c_{1J}	c_1
a_2	c_{21}	c_{22}	$\cdots$	c_{2J}	c_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
a_I	c_{I1}	c_{I2}	$\cdots$	c_{IJ}	c_I
Total	n_1	n_2	$\cdots$	n_J	n

I.e. we have a single sample from the multinomial distribution

$$(C_{11}, ..., C_{IJ}) \sim Mn(n; \pi_{11}, ..., \pi_{IJ}).$$

We test the hypothesis

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j} \quad \text{for all pairs} \quad (i, j).$$

Test of independence

The ML estimates for the sample proportions are

$$\hat{\pi}_{i.} = \frac{c_i}{n}, \quad \hat{\pi}_{.j} = \frac{n_j}{n}$$

leading to the expected cell counts

$$e_{ij} = n\hat{\pi}_{ij} = n\hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{c_i n_j}{n}.$$

We use the same test statistic as before, namely

$$x^2 = \sum_{i=1}^{I}\sum_{j=i}^{J} \frac{(c_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{I}\sum_{j=i}^{J} \frac{(c_{ij} - c_i n_j/n)^2}{c_i n_j/n}$$

and reject the null hypothesis with large values of x^2. The stochastic version of the test statistic has the distribution χ^2_{df} under H_0 with

$$df = IJ - 1 - (I-1) - (J-1) = (I-1)(J-1).$$

Matched pairs design: example

In another study concerning the relationship between tonsillectomy and Hodgkin's disease, data were collected from 85 patients with Hodgkin's disease and 85 controls without the disease resulting in the following data:

	Hodgin's	Control	Total
Tonsillectomy	41	33	74
No tonsillectomy	44	52	96
Total	85	85	170

A χ^2 (homogeneity) test was conducted leading to the p-value 0.215 (test statistic gets the value 1.53 resulting in the p-value 1-pchisq(1.53,1) in R). Therefore, null hypothesis cannot be rejected, i.e. Hodgin's disease seems not to be related to the removal of tonsils.

Matched pairs design: example

Further investigation revealed that the 85 controls were collected by choosing a sibling of the same gender and without the disease of each of the 85 patients with Hodgkin's disease. Therefore, a matched pairs design had been used.

The test of homogeneity assumes independent samples, one from the population with Hodgkin's disease and one from the population without the disease.
$\rightarrow$ The χ^2-test that was performed is not valid.

Matched pairs design: example

An appropriate analysis would be to treat the data in the form

		Sibling	
		No tonsillectomy	Tonsillectomy
Hodgin's	No tonsillectomy	37	7
	Tonsillectomy	15	26

This data are a sample of size 85 from a multinomial distribution with four cells with the corresponding probabilities

			Total
	π_{11}	π_{12}	$\pi_{1.}$
	π_{21}	π_{22}	$\pi_{2.}$
Total	$\pi_{.1}$	$\pi_{.2}$	1

Matched pairs design: example

The null hypothesis would then be that the probabilities of tonsillectomy and no tonsillectomy are the same among the patients with Hodgin's disease and within siblings, i.e. that $\pi_{1.} = \pi_{.1}$ and $\pi_{2.} = \pi_{.2}$ or

$$H_0 : \pi_{12} = \pi_{21}.$$

This leads to McNemar's test (Chapter 7, comparing population proportions when data are paired) with the test statistic

$$\frac{(c_{12} - c_{21})^2}{c_{12} + c_{21}} = \frac{(7 - 15)^2}{7 + 15} = 2.91,$$

where the stochastic version is approximatively χ^2_1-distributed. The p-value is

$$\mathrm{P}(X^2 \geq 2.91 | H_0) \approx 2(1 - \Phi(\sqrt{2.91})) = 0.09.$$

Odds ratio

We have a random event A and the probability of it $\mathrm{P}(A)$ which is between 0 and 1 and let A^c be the complementary event of A. The odds of the event A is defined as

$$\text{odds}(A) = \frac{\mathrm{P}(A)}{\mathrm{P}(A^c)} = \frac{\mathrm{P}(A)}{1 - \mathrm{P}(A)}.$$

Then,

$$\mathrm{P}(A) = \frac{\text{odds}(A)}{1 + \text{odds}(A)}$$

and when $\mathrm{P}(A)$ is small, $\mathrm{P}(A) \approx \text{odds}(A)$. We can also define the conditional odds for A given B as

$$\text{odds}(A|B) = \frac{\mathrm{P}(A|B)}{\mathrm{P}(A^c|B)} = \frac{\mathrm{P}(AB)}{\mathrm{P}(A^cB)}.$$

Odds ratio

Finally, the odds ratio of a pair of events (A, B) is defined as

$$\Delta_{AB} = \frac{\text{odds}(A|B)}{\text{odds}(A|B^c)} = \Delta_{BA} = \frac{1}{\Delta_{AB^c}}$$

Interpretation of the odds ratio:

- The events A and B are independent if and only if the odds ratio is 1, i.e. the odds of one event are the same in either the presence or absence of the other event.
- If the odds ratio is greater than 1, then A and B are associated (correlated) in the sense that, compared to the absence of B, the presence of B raises the odds of A, and symmetrically the presence of A raises the odds of B.
- If the odds ratio is less than 1, then A and B are negatively correlated, and the presence of one event reduces the odds of the other event.

Odds ratio: case-control example

	Hodgin's (A)	Control ($\bar{A}$)	Total
Tonsillectomy (E)	67	43	110
No tonsillectomy ($\bar{E}$)	34	64	98
Total	101	107	208

The (first) homogeneity test rejected the null hypothesis of no relationship between tonsillectomy and Hodgin's disease. How strong is the observed relationship?
$\rightarrow$ odds ratio

Let A be the event of having Hodgin's disease and E the event having had tonsillectomy. Then, the odds ratio is

$$\Delta_{AE} = \frac{\text{odds}(A|E)}{\text{odds}(A|\bar{E})} = \frac{\text{odds}(E|A)}{\text{odds}(E|\bar{A})} = \frac{67 \cdot 64}{43 \cdot 34} = 2.93,$$

i.e. tonsillectomy seems to increase the odds for the onset of Hodgkin's disease by factor 2.93.

Three sampling designs (Hudgin's example)

- A single random sample from the entire population. Since the disease is rare, the sample size should be very large to guarantee that a large enough number of individuals with Hodgin's disease would be included.
- A prospective study: Take a sample of a fixed size from the population, where the tonsils have been removed, and from the population, where they have not been removed and check, in each sample, how many individuals have Hodgin's disease. Even here, it can be difficult to have enough individuals with Hodgin's disease included in the samples.
- A retrospective study: Take a sample of a fixed size from the population with Hodgin's disease and from the sample without the disease and find out which ones had in the past had tonsillectomy.