Inference on proportions and contingency tables

Anna Helga Jónsdóttir
Sigrún Helga Lund

Prepared at the University of Iceland

July 2020

Main topics:

1. Inference on the ratio of a population
2. Inference on the ratio of two populations
3. Inference on the ratio of more populations
4. Contingency tables

Estimate of the ratio of a population

In this section we will discuss confidence intervals and hypothesis tests for one ratio p that describes the ratio of subjects within a population that have a particular value of a categorical variable.

Bernoulli trial and the binomial distribution

Bernoulli trial

Every trial in a group of repeated trials is classified as a **Bernoulli trial** if the following holds:

1. Every trial has only two possible outcomes (positive and negative).
2. The probability of a positive outcome are the same in every trial.
3. The outcomes are indipendant.

The number of positive outcomes in n Bernoulli trials follows the **binomial distribution** with the parameters n and p, written $X \sim B(n, p)$, where p is the probability of a positive outcome.

Estimate of the ratio of a population

The ratio of the population, denoted p, is estimated with the sample proportion:

$$\widehat{p} = \frac{x}{n}$$

where x is the number of measurements that receive the corresponding outcome and n is the size of the sample.

Normal approximation

- When certain criteria is met, the binomial distribution is similar to the normal distribution.
- Then we can use methods that assume the characteristics of the normal distribution to make inference on random variables that in deed are binomially distributed.
- That is called to apply a **normal approximation**

When can one use normal approximation?

If $n\hat{p}$ and $n(1-\hat{p})$ are greater then 15, the normal approximation can be used to make inference on the proportion of a binomial distribution.

Confidence interval

Confidence interval for the ratio of a population

If the criteria for using the normal approximation is met, the lower bound for p can be calculated with:

$$\widehat{p} - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

and the upper bound with:

$$\widehat{p} + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

where $\widehat{p} = \frac{x}{n}$ and $z_{1-\alpha/2}$ is in the standardized normal distribution table

The null hypothesis

- The null hypothesis in this section tests the hypothesis that the ratio of the sample, p is equal to a certain value that we call p_0.
- The null hypothesis is written $H_0 : p = p_0$.
- If the test is two-sided we can conclude that the ratio p differs from p_0.
- If it is one sided we can only conclude that p is either greater or less than p_0, depending on the case.

Hypothesis test for the ratio of a population

Hypothesis test for the ratio of a population

If the criteria for using the normal approximation are met, the following hypothesis test can be used The null hypothesis is

$$\mathrm{H}_0 : p = p_0$$

The test statistic is

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

where X is the number of successful experiments and n is the size of the sample.

If the null hypothesis is true, the test statistic follows the standardized normal distribution, or $Z \sim N(0, 1)$.

Alternative hypothesis for the ratio of a sample.

Alternative hypothesis for the ratio of a sample.

The alternative hypothesis along with the rejection areas are shown below.

Alternative hypothesis	Reject H_0 if:
$H_1 : p < p_0$	$Z < -z_{1-\alpha}$
$H_1 : p > p_0$	$Z > z_{1-\alpha}$
$H_1 : p \neq p_0$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

Inference on the ratio of two populations

We often want to compare the ratios of a certain value of a categorical variable in two populations.

We denote the ratios in the two populations with p_1 and p_2 and estimate them with

$$\widehat{p}_1 = \frac{x_1}{n_1}, \qquad \widehat{p}_2 = \frac{x_2}{n_2}$$

where x_1 and x_2 are the number of successful outcomes in the two samples.

Criteria for normal approximation

A normal approximation can be used if $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ are all greater than 15

Confidence interval for the ratio of two populations

Confidence interval for the ratio of two populations

If the criteria for using the normal approximation are met, the lower bound for the difference p_1 and p_2 can be calculated with:

$$\widehat{p_1} - \widehat{p_2} - z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

and the upper bound with:

$$\widehat{p_1} - \widehat{p_2} + z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}$$

where $\widehat{p_1} = \frac{x_1}{n_1}$, $\widehat{p_2} = \frac{x_2}{n_2}$ and $z_{1-\alpha/2}$ is in the standardized normal distribuiton table

The null hypothesis

- The hypothesis test in this section tests the null hypothesis that the ratios in the two populations are equal.
- The null hypothesis is written $H_0 : p_1 = p_2$.
- If the hypothesis test is two sided we draw the conclusion that the ratios are different if we reject the null hypothesis.
- It is one sided we can only conclude that one ratio is greater than the other or vice verse, depending on the case.

Hypothesis test for the ratio of two populations

Hypothesis test for the ratio of two populations

If the criteria for using the normal approximation are met, the following hypothesis test can be used:
The null hypothesis is:

$$\mathrm{H}_0 : p_1 = p_2$$

The test statistic is:

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\sqrt{\widehat{P}(1-\widehat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where} \quad \widehat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

If the null hypothesis is true the test statistic follows the standardized normal distribution, or $Z \sim N(0, 1)$.

The alternative hypothesis

The alternative hypothesis

The alternative hypothesis along with their rejection areas are shown below:

Alternative hypothesis	Reject H_0 if:
$\mathsf{H}_1 : p_1 < p_2$	$Z < -z_{1-\alpha}$
$\mathsf{H}_1 : p_1 > p_2$	$Z > z_{1-\alpha}$
$\mathsf{H}_1 : p_1 \neq p_2$	$Z < -z_{1-\alpha/2}$ or $Z > z_{1-\alpha/2}$

Chi squared test

- The hypothesis in last section can be generalized such that it compares the ratio of more than two populations.
- Then one cannot use methods based on normal proximation, but so called chi-squared tests are used (χ^2-test).
- The method can also be used when comparing the ratios of two populations, but only if the alternative hypothesis is two-sided.
- Then the Chi-squared test statistic and the Z-statistic be the same.

The null hypothesis

- The hypothesis test in this section tests whether the ratios of c populations are all equal.
- It is written $H_0 : p_1 = p_2 = ... = p_c$.
- If it is rejected we can conclude that the ratio are not all equal.
- That does not mean that they are all different!
- The hypothesis test does not say which of the ratios differ from the other.
- More evolved methods are used to do so, which are not taught in this lecture.

Tables for chi-squared próf

Töflur fyrir chi-squared próf

Þegar framkvæma á chi-squared test er gott að búa til þrjár töflur:

- Table 1: Contains the observed frequency in the investigation, denoted with o.
- Table 2: Containes the expected frequency in the investigation, denoted with e. The values are calculated by multiplying the sums for the corresponding column and row and divide by the total number of measurements. All values in this table need to be greater than 5 for the test to be valid.
- Table 3: Contains the tribute to the test statistic, calculated with $\frac{(o-e)^2}{e}$. Finally all the values in Table 3 are added together to calculate the value of the test statistic (see next slide).

Chi-squared test for ratios

Chi-squared test for ratios

The hypothesis are:

$$H_0 : p_1 = p_2 = ... = p_c$$

$$H_1 : \text{the ratios are not all equal}$$

The test statistic is:

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

where r is the number of rows, c is the number of columns, o is the observed frequency and e is the expected frequency.

If the null hypothesis is true, the test statistic follows the χ^2-distribution with (r - 1) $\cdot$ (c - 1) degrees of freedom. The null hypothesis is rejected if $\chi^2 > \chi^2_{1-\alpha,((r-1)\cdot(c-1))}$.

Contingency tables

- In the previous section we saw how to compare the ratio of a categorical variable in different populations.
- In this section we will see how to compare two categorical variables when the measurements are made in the same population.
- We use contingency tables to do so and the tests answer the question whether the two variables are independent or not.
- The test statistic is the same as for the chi-squared test and all the calculations are the same.
- The hypothesis are stated in a different manner.

Contingency tables

Contingency tables

Contingency tables are used to investigate whether two categorical variables are independent or not. The hypothesis are

$$\begin{aligned} H_0 &: \text{The variables are independent} \\ H_1 &: \text{The variables are dependent} \end{aligned}$$

The test statistic is

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

where r is the number of rows, c is the number of columns, o is the observed frequency and e is the expected frequency.
If the null hypothesis is true, the null hypothesis follows the χ^2-distribution with $(r$ - 1) $\cdot$ $(c$ - 1) degrees of fredom. The null hypothesis is rejected if $\chi^2 > \chi^2_{1-\alpha,((r-1)\cdot(c-1))}$.