MS-C1620 Statistical inference

4 Inference for binary data

Jukka Kohonen

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2021–2022
Period III–IV

Contents

Binary observations

In many applications the observations are binary.

- Something is true/false.
- Something happened/did not happen.
- Someone belongs/does not belong to a group.

Such observations are conveniently coded as 0/1-valued *indicator variables*.

Recall that if we have a iid sample of binary observations, their distribution is necessarily the *Bernoulli distribution*.

Bernoulli distribution

A random variable x has the *Bernoulli distribution* with the *success probability* θ if,

$$\mathbb{P}(x = 1) = \theta \quad \text{and} \quad \mathbb{P}(x = 0) = 1 - \theta.$$

The expected value and variance of x are

$$\begin{aligned} \mathbb{E}(x) &= \theta \\ \mathrm{Var}(x) &= \theta(1 - \theta). \end{aligned}$$

Note that the Bernoulli distribution has only a single parameter to estimate (no separate "variance parameter").

The sum of n i.i.d. Bernoulli random variables with the success probability θ has the binomial distribution with the parameters n and θ.

(Often the success probability is called p, but here we use θ to emphasize that it is a parameter to estimate, and to avoid confusion with p-values in hypothesis testing.)

Contents

Approximate confidence interval

Central limit theorem can be used to obtain a confidence interval for the success probability θ of a Bernoulli distribution.

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from the Bernoulli distribution with the success probability/expected value θ.

For large n, a level $100(1-\alpha)\%$ confidence interval for the success probability θ is obtained as

$$\left(\hat{\theta} - z_{\alpha/2}\frac{\sqrt{\hat{\theta}(1-\hat{\theta})}}{\sqrt{n}}, \hat{\theta} + z_{\alpha/2}\frac{\sqrt{\hat{\theta}(1-\hat{\theta})}}{\sqrt{n}}\right),$$

where $\hat{\theta}$ is the observed proportion of successes and $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

One-sample proportion test

To test whether the success probability of a Bernoulli distribution equals some pre-specified value, we employ one-sample proportion test.

One-sample proportion test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from a Bernoulli distribution with the success probability θ.

One-sample proportion test, hypotheses

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0.$$

One-sample proportion test

One-sample proportion test, test statistic

- The test statistic,

$$C = \sum_{i=1}^{n} x_i,$$

  follows the binomial distribution with parameters n and θ_0 under H0.
- Under H_0, the test statistic has $\mathrm{E}[C] = n\theta_0$ and $\mathrm{Var}(C) = n\theta_0(1-\theta_0)$ and both large and both **large** and **small** values of the test statistic suggest that the null hypothesis H_0 is false.

The distribution of the test statistic C is tabulated and statistical software calculate exact p-values of the test.

Asymptotic one-sample proportion test

If the sample size is large, then under the null hypothesis H_0 the standardized test statistic,

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}},$$

where $\hat{\theta}$ is the unbiased estimator $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ of the parameter θ, follows approximately the standard normal distribution.

The approximation is usually accurate enough if $n\hat{\theta} > 10$ and $n(1-\hat{\theta}) > 10$. For smaller samplea one should use the exact distribution of the test statistic C.

Contents

Two-sample proportion test

The one-sample proportion test can be seen as the equivalent of t-test when the normal distribution is replaced by the Bernoulli distribution.

As with t-test, a two-sample version readily follows and in two-sample proportion test parameters of two independent Bernoulli-distributed samples are compared.

Two-sample proportion test, assumptions

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample from a Bernoulli distribution with the success probability θ_x and let $y_1, y_2, \ldots, y_m$ be an i.i.d. sample from a Bernoulli distribution with the success probability θ_y. Furthermore, let the two samples be independent.

Two-sample proportion test, hypotheses

$$H_0 : \theta_x = \theta_y \quad H_1 : \theta_x \neq \theta_y.$$

Two-sample proportion test

Two-sample proportion test, test statistic

- Calculate the sample proportions

$$\hat{\theta}_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\theta}_y = \frac{1}{m}\sum_{i=1}^{m} y_i, \quad \hat{\theta} = \frac{n\hat{\theta}_x + m\hat{\theta}_y}{n+m}.$$

- The test statistic

$$Z = \frac{\hat{\theta}_x - \hat{\theta}_y}{\sqrt{\hat{\theta}(1-\hat{\theta})\left(\frac{1}{n}+\frac{1}{m}\right)}},$$

  follows *for large n* under H_0 the standard normal distribution.
- Values **far from zero** (positive or negative) suggest that the null hypothesis H_0 is false.

The normal approximation is usually good enough if $n\hat{\theta}_x > 5$, $n(1-\hat{\theta}_x) > 5$, $m\hat{\theta}_y > 5$ and $m(1-\hat{\theta}_y) > 5$.

Frequency tables

Assuming a “paired binary sample”, the previous test is no longer valid.

id	X	Y
1	0	1
2	0	0
3	0	1
4	1	1
⋮	⋮	⋮

This kind of data is conveniently represented in a contingency table (aka cross tabulation).

	$Y = 0$	$Y = 1$
$X = 0$	173	40
$X = 1$	65	53

Inference from contingency tables is discussed next time.

Contents

Lecture quiz

A lecture quiz to determine what you have learned thus far!

Answer the following questions on your own or in small groups.

Lecture quiz

Question 1

Consider the following random sample: 5, -4, -2, 2. Calculate the following sample quantities:

1. Sample mean
2. Sample standard deviation
3. Sample median
4. Sample median absolute deviation
5. Sample range
6. Signs of the sample points
7. Ranks of the sample points
8. Signed ranks of the sample points with respect to distance to 0.

Lecture quiz

Question 2

Give concrete examples when you would/would not use the following measures of location:

1. Sample mean
2. Sample median
3. Mode

Question 3

Give concrete examples when you would/would not use the following measures of scatter:

1. Standard deviation
2. Median absolute deviation
3. Sample range

Lecture quiz

Question 4

What does it mean in practice if:

- The confidence interval of a parameter is narrow
- The significance level of a test is set to small value
- The p-value of a test is high
- Type I error occurs in a statistical test
- Type II error occurs in a statistical test

Lecture quiz

Question 5

How would you visualize the following samples:

- The heights of the male and female students attending a course.
- The exam points (0-24) on a large course.
- The proportions of faulty products produced by 5 different production lines.
- Stock prices of 3 companies over some time interval
- The monthly salaries and postal codes of adults living in Helsinki area.

Lecture quiz

Question 6

The following plots show the distributions of the test statistics of **A.** t-test, **B.** sign test, **C.** signed rank test for the null hypothesis of zero location, with a sample of $n = 10$ points from from the standard normal distribution. Which plot is from which test?