MS-C1620 Statistical inference

5 Distribution tests

Pekka Pere

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2023–2024
Period III–IV

Contents

- In statistics, assumptions on the underlying distribution are done all the time.
- Assumptions on normally distributed observations are made particularly often, especially with classical statistical methods.
- Many statistical methods become ineffective or even give false results if their assumptions do not hold.
- The tails of the distribution are at focus if abnormalities need to be screened (talented individuals, potential illnesses, risky customers, products not meeting the standards, outliers in general *etc*).
- Distribution of the data may be of interest as such.
- It may be thus important to carry out tests about the distribution of the data.

Contents

Normality testing

Normality testing, assumptions

Assume that $x_1, x_2, ..., x_n$ are i.i.d. observed values of a random variable x.

Normality testing, hypotheses

H_0 : Random variable x is normally distributed.

H_1 : Random variable x is not normally distributed.

Bowman-Shenton normality test

Bowman-Shenton normality test, test statistic

- The Bowman-Shenton (Jarque-Bera) normality test is a function of skewness and kurtosis,

$$BS = n(\frac{\hat{\gamma}^2}{6} + \frac{\hat{\kappa}^2}{24}),$$

  where $\hat{\gamma}$ is the sample skewness coefficient and $\hat{\kappa}$ is the sample kurtosis coefficient discussed in lecture 1.
- The test tests whether the skewness and kurtosis of the data-generating distribution match with the normal distribution.
- If the observed skewness or kurtosis values differ significantly from the skewness and/or kurtosis values of the normal distribution (0 and 0), the test statistic gets large values.

Bowman-Shenton normality test

Bowman-Shenton normality test, test statistic

- If n is large, then under H_0 the test statistic BS follows approximately $\chi^2(2)$ distribution.
- The expected value of the test statistic under H_0 is approximately 2 and **large values** of the test statistic suggests that the null hypothesis H_0 is false.

Note that the Bowman-Shenton test is suitable only for large sample sizes.

Rank plot / Quantile-quantile (Q-Q) plot

- Let $y_1 \leq y_2 \leq \cdots \leq y_n$ be the data points $x_1, x_2, \ldots, x_n$ ordered from the smallest one to the largest one.
- Let q_i be the $i/(n+1)$ quantile from the standard normal distribution $\mathcal{N}(0,1)$ and plot the pairs $(q_i, y_i), i = 1, 2, \ldots, n$.
- If the observations x_i do come from a normal distribution, then the points (q_i, y_i) should approximately lie on a line.
- If the points do not lie on a line, there is evidence of non-normality.
- The plot can be used in detecting skewness of a distribution and in finding outliers.

Shapiro-Wilk normality test

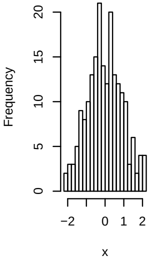
Shapiro-Wilk normality test, test statistic

- The Shapiro-Wilk normality test statistic is the squared value of the Pearson sample correlation coefficient calculated from the rank plot points $(q_i, y_i), i = 1, 2, \ldots, n$.
- The null distribution of the test statistic is complicated and the test is usually performed with statistical software.
- **Small** values of the test statistic suggest that the assumption of normality does not hold. **Large** values of the test statistic are in line with the null hypothesis.

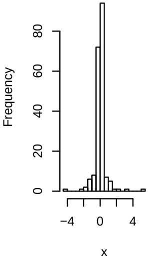
The Shapiro-Wilk normality test requires a large sample size. Skewness or thin tails tend to increase and fat tails to decrease the power of the test.

Q-Q plot quiz

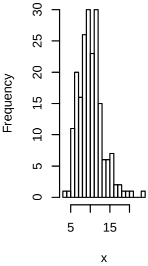
Which of the following histograms corresponds to each of the Q-Q plots on the next slide? (the answers are given on the next slide after that)



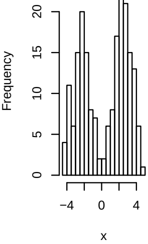
A



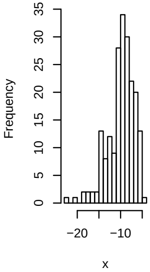
B



C

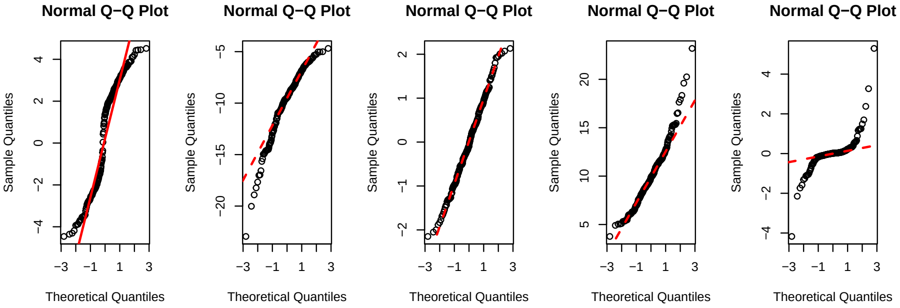


D



E

Q-Q plot quiz



1 2 3 4 5

Q-Q plot quiz, answers

- The correct pairs are (histogram, qqplot):

$$(A, 3); (B, 5); (C, 4); (D, 1); (E, 2).$$

Contents

Multinomial distribution

Consider a random experiment which has k mutually exclusive outcomes and which is run independently n times.

Let the vector $\mathbf{y} = (y_1, \ldots, y_k)$ contain the observed frequencies of the k outcomes.

The distribution of $\mathbf{y}$ is known as the multinomial distribution, the generalization of the binomial distribution into more than two outcomes.

Multinomial distribution

The random vector $\mathbf{y} = (y_1, \ldots, y_k)$ follows the multinomial distribution with index n and parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ if its probability mass function is,

$$P(\mathbf{y}) = \frac{n!}{y_1! y_2! \cdots y_k!} \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k},$$

where

$$\sum_{j=1}^{k} y_j = n \quad \text{and} \quad \sum_{j=1}^{k} \pi_j = 1.$$

Then the statistic

$$\sum_{j=1}^{k} \frac{(y_j - n\pi_j)^2}{n\pi_j},$$

follows approximately the $\chi^2(k-1)$-distribution for large n. Here $n\pi_j$ are the expected frequencies of the outcomes.

χ^2 goodness-of-fit test

The χ^2 goodness-of-fit test uses the multinomial distribution to tests whether the distribution of a random variable x is some particular, arbitrary distribution.

Goodness-of-fit tests, assumptions

Assume that $x_1, x_2, \ldots, x_n$ are i.i.d. observed values of a random variable x.

Goodness-of-fit tests, hypotheses

H_0 : Random variable x follows the distribution F_x (with or without unknown parameters).

H_1 : Random variable x does not follow the distribution F_x.

χ^2 goodness-of-fit test

- Categorize the n observations into k categories.
- Calculate the frequencies $O_1, \ldots, O_k$, where O_j is the observed frequency of the jth category (note that $\sum_{j=1}^{k} O_j = n$).
- Let π_j be the probability that, under the null hypothesis, the random variable x belongs gets a value belonging to the jth category.
- Calculate the expected frequencies $E_j = n\pi_j$ of the k categories (note that $\sum_{j=1}^{k} \pi_j = 1$ and $\sum_{j=1}^{k} E_j = n$).
-

Now, under the null hypothesis, the random vector $(O_1, \ldots, O_k)$ follows the multinomial distribution with index n and parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ and expected category frequencies $(E_1, \ldots, E_k)$.

χ^2 goodness-of-fit test

χ^2 goodness-of-fit test, test statistic

- The test statistic,

$$\chi_g^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

  follows, for large n, under H_0 approximately the $\chi^2(k-1-e)$-distribution, where e is the number of estimated parameters (see the salary example below).
- The expected value of the test statistic under H_0 is approximately $k-1-e$ and **large** values of the test statistic suggest that the null hypothesis H_0 does not hold.
- Note that a very small value of the test statistic could be an indicator of *overfitting*.

χ^2 goodness-of-fit test, example with unknown parameters

- Consider testing whether the monthly salary of the Finns follows a normal distribution.
- Select randomly n Finns and document their salaries.
- The null hypothesis is that the observations come from a normal distribution with an unknown expected value and an unknown variance.

χ^2 goodness-of-fit test, example with unknown parameters

1. Estimate the unknown parameters (μ and σ^2) from the sample.
2. Discretize the continuous salary variable into k categories.
3. Calculate the observed category frequencies $O_1, \ldots, O_k$.
4. Calculate the category probabilities for the estimated normal distribution, for example,
   $\ldots, \mathbb{P}(1900 < X \leq 2000), \mathbb{P}(2000 < X \leq 2100), \ldots$
5. Calculate the expected category frequencies $E_1, \ldots, E_k$.
6. Calculate the test statistic. Under the null hypothesis the test statistic approximately follows $\chi^2(k-1-e) = \chi^2(k-3)$-distribution, where k is the number of categories and we estimated $e = 2$ parameters (μ and σ^2).
7. Calculate the p-value and based on that either reject or do not reject the null hypothesis.

χ^2 homogeneity test

The χ^2 homogeneity test is used to assess whether multiple samples come from the same distribution.

χ^2 homogeneity test, assumptions

We observe a total of r samples such that the samples are independent and the observations within a single sample are i.i.d. Assume that the sample $i \in \{1, \ldots, r\}$ has n_i observations.

χ^2 homogeneity test, hypotheses

H_0 : The samples come from the same distribution F_x.

H_1 : The samples do not come from the same distribution.

χ^2 homogeneity test, observed frequencies

- Categorize all observations into k categories.
- Calculate the frequencies O_{ij}, $i \in \{1, 2, \ldots, r\}$, $j \in \{1, 2, \ldots, k\}$, where O_{ij} is the observed frequency of the observations of the sample i in category j.

	1	2	$\cdots$	k	sum
1	O_{11}	O_{12}	$\cdots$	O_{1k}	n_{1+}
2	O_{21}	O_{22}	$\cdots$	O_{2k}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	O_{r1}	O_{r2}	$\cdots$	O_{rk}	n_{r+}
sum	n_{+1}	n_{+2}	$\cdots$	n_{+k}	n

χ^2 homogeneity test, expected frequencies

- Let $\hat{\pi}_{+j} = n_{+j}/n$ be an estimate of the proportion of the jth category under H_0 (under the null hypothesis the probability of the category j is the same for each sample i).
- Calculate the expected frequencies under the null, $E_{ij} = n_{i+}\hat{\pi}_{+j}$.

	1	2	$\cdots$	k	sum
1	E_{11}	E_{12}	$\cdots$	E_{1k}	n_{1+}
2	E_{21}	E_{22}	$\cdots$	E_{2k}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	E_{r1}	E_{r2}	$\cdots$	E_{rk}	n_{r+}
sum	n_{+1}	n_{+2}	$\cdots$	n_{+k}	n

χ^2 homogeneity test

χ^2 homogeneity test, test statistic

- The test statistic,

$$\chi_h^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

  follows, for large n, under H_0 approximately the $\chi^2((r-1)(k-1))$ distribution.

- Under H_0 the expected value of the test statistic is approximately $(r-1)(k-1)$ and **large** values of the test statistic suggest that the null hypothesis H_0 is false.

χ^2 test of independence

χ^2 test of independence is used to study if two random variables (factors) are stochastically independent.

χ^2-test of independence, assumptions

We observe an i.i.d. random sample of size n and the observations are divided into r classes with respect to a factor A and into k classes with respect to a factor B.

χ^2-test of independence, hypotheses

H_0 : The variables A and B are independent.

H_1 : The variables A and B are not independent.

χ^2 test of independence, observed frequencies

- Let n_{i+} be the frequency of the observations in class i of the factor A and let n_{+j} be the frequency of the observations in class j of the factor B.
- Let O_{ij} be the observed frequency of the observations that are in class i of the factor A and in class j of the factor B.

	1	2	$\cdots$	k	sum
1	O_{11}	O_{12}	$\cdots$	O_{1k}	n_{1+}
2	O_{21}	O_{22}	$\cdots$	O_{2k}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	O_{r1}	O_{r2}	$\cdots$	O_{rk}	n_{r+}
sum	n_{+1}	n_{+2}	$\cdots$	n_{+k}	n

χ^2 test of independence, expected frequencies

- Let $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$. Under the null hypothesis of independence the probability to fall in to the cell (i, j) can be estimated to be $\hat{\pi}_{i+}\hat{\pi}_{+j}$.
- The estimated expected frequencies under the null are

$$E_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}\hat{\pi}_{+j} = n_{+j}\hat{\pi}_{i+}.$$

	1	2	$\cdots$	k	sum
1	E_{11}	E_{12}	$\cdots$	E_{1k}	n_{1+}
2	E_{21}	E_{22}	$\cdots$	E_{2k}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	E_{r1}	E_{r2}	$\cdots$	E_{rk}	n_{r+}
sum	n_{+1}	n_{+2}	$\cdots$	n_{+k}	n

χ^2 test of independence

χ^2-test of independence, test statistic

- The test statistic,

$$\chi_I^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

  follows, for large n, under H_0 approximately the $\chi^2(r-1)(k-1)$-distribution.
- The expected value of the test statistic under H_0 is approximately $(r-1)(k-1)$ is and **large** values of the test statistic suggest that the null hypothesis is false

The homogeneity test and the test of independence

- Note that the χ^2 test of independence and χ^2 homogeneity test have their test statistics and the degrees of freedom calculated identically.
- However, the tests apply to different situations:
  - If the group sizes of one of the factors are pre-determined, one can not speak of the independence of the factors (since one of them has its frequencies fixed), and the correct interpretation is via the χ^2 homogeneity test.
  - If only the overall sample size n is fixed and the observations are allowed to freely fall into the categories with respect to both factors, the correct interpretation is via the χ^2-test of independence.