Applied Statistics

for Computer Science BSc, Exam

Probability Theory and Mathematical Statistics

for Computer Science Engineering BSc, Term grade

István Fazekas

University of Debrecen

2020/21 fall

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

Main topics

1. Probability theory
2. Statistics

Mathematical tools: combinatorics, calculus

Computer tool: Matlab

Book:

Yates, Goodman:

Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers

Lecture 12

Non-parametric tests

Parametric and non-parametric tests

Parametric tests are used when the underlying distribution is known, but the parameters are not known. The underlying distribution is usually the normal distribution. The z-test and the t-test are parametric tests. The analysis of variance is also a parametric statistical method.

Non-parametric statistical methods are usually distribution-free, that is they do not rely on the assumption that the data are drawn from a given parametric family of probability distributions.

Non-parametric tests

Some important goals of non-parametric statistics are the following.

1. Testing the goodness of fit, that is testing whether a sample is drawn from a given distribution. Methods:

Chi-square test for goodness of fit

Kolmogorov-Smirnov one-sample test

2. Testing homogeneity: testing whether two samples are drawn from the same distribution. Methods:

Chi-square test of homogeneity

Kolmogorov-Smirnov two-sample test

MannWhitney U or Wilcoxon rank sum test

Sign test

3. Testing independence: testing whether two samples are independent or measuring dependence of two samples. Methods:

Chi-square test of independence

Spearman's rank correlation coefficient

The chi-squared test. Multinomial distribution

Let $A_1, .., A_r$ be a partition of the sample space.
Let $p_i = P(A_i) > 0$, $i = 1, \ldots, r$.
Repeat the experiment N-times independently.
Denote by X_i the number of occurrences of A_i.
Then the joint distribution of $(X_1, \ldots, X_r)$ is multinomial with parameters N and $p_1, \ldots, p_r$, that is

$$P(X_1 = k_1, \ldots, X_r = k_r) = \frac{N!}{k_1! \ldots k_r!} p_1^{k_1} \ldots p_r^{k_r},$$

where $k_1, \ldots, k_r$ are non-negative integers with $k_1 + \cdots + k_r = N$.
Obviously, X_i is binomial with parameters N and p_i, that is

$$\begin{aligned} P(X_i = k_i) &= \binom{N}{k_i} p_i^{k_i} (1 - p_i)^{N - k_i}, \quad k_i = 0, \ldots, N, \\ \mathbb{E}X_i &= Np_i, \quad \mathrm{Var}X_i = Np_i(1 - p_i). \end{aligned}$$

Multinomial distribution...

Then

$$\mathrm{Cov}(X_i, X_j) = -Np_ip_j, \qquad i \neq j.$$

So $X = (X_1, \ldots, X_r)^\top$ is a random vector with expectation

$$\mathbb{E}X = N(p_1, \ldots, p_r)^\top$$

variance matrix $\mathrm{Var}X = NS$, where

$$S = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \ldots & -p_1p_r \\ -p_2p_1 & p_2(1-p_2) & & -p_2p_r \\ \ldots & & \ddots & \\ -p_rp_1 & -p_rp_2 & & p_r(1-p_r) \end{pmatrix}.$$

Let $Z = N^{-1/2}(X - \mathbb{E}X)$. It converges in distribution to $\mathcal{N}_r(0, S)$ as $N \to \infty$.

Multinomial distribution...

Theorem.
Let $X_1, \ldots, X_r$ be multinomial with parameters N and $p_1, \ldots, p_r$, where $p_1 > 0, \ldots, p_r > 0$. Then

$$\sum_{j=1}^{r} \frac{(X_j - Np_j)^2}{Np_j}$$

converges in distribution to $\mathcal{X}^2_{r-1}$ as $N \to \infty$.

The chi-squared test. Testing goodness of fit

Let $A_1, \ldots, A_r$ be a partition of the sample space.
Assume $p_1 > 0, \ldots, p_r > 0$ are given with $\sum_{i=1}^{r} p_i = 1$.
Decide if the null-hypothesis

$$H_0 : P(A_i) = p_i, \quad i = 1, \ldots, r,$$

is true.
Repeat the experiment containing the events $A_1, \ldots, A_r$ N-times independently. The event A_i occurred k_i-times out of the N repetitions. Calculate

$$\mathcal{X}^2 = \sum_{i=1}^{r} \frac{(k_i - Np_i)^2}{Np_i}$$

By our previous theorem, if H_0 is true, then the distribution of $\mathcal{X}^2$ is asymptotically $\mathcal{X}^2_{r-1}$.

The chi-squared test. Testing goodness of fit...

The structure of the above statistic $\mathcal{X}^2$ is

$$\sum \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}}.$$

So $\mathcal{X}^2$ is small if and only if the observed values are close to the values which are expected if H_0 is true.
So we accept H_0 if $\mathcal{X}^2$ is small.
If $\mathcal{X}^2$ is larger than the critical value, then we reject H_0.
For given level α the critical value is obtained from a table of $\mathcal{X}^2_{r-1}$ distribution.
The $\mathcal{X}^2$ test is applicable for large N.
If there are s unknown parameters, then we estimate them from the sample. But then the degree of freedom will be $r - 1 - s$.

Testing goodness of fit...

Example. Test if a coin is fair. So

$$H_0 : P(H) = \frac{1}{2}, \quad P(T) = \frac{1}{2}$$

Toss the coin 200 times. We obtained 110 heads, and 90 tails.

$$\mathcal{X}^2 = \frac{(110 - 200 \cdot 0.5)^2}{200 \cdot 0.5} + \frac{(90 - 200 \cdot 0.5)^2}{200 \cdot 0.5} = 2.$$

The critical value at significance level 90% from the table of $\mathcal{X}^2_{2-1} = \mathcal{X}^2_1$ distribution is 2.71.
As now $\mathcal{X}^2 < 2.71$, so we accept H_0.

The chi-squared test. Testing independence

Let $\mathcal{A} = \{A_1, A_2, \ldots, A_r\}$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_s\}$ be two partitions of the sample space.
Test the independence of $\mathcal{A}$ and $\mathcal{B}$, that is

$$H_0 : P(A_i \cap B_j) = P(A_i)P(B_j), \quad i = 1, \ldots, r, \quad j = 1, \ldots, s.$$

If the probabilities $p_i = P(A_i)$, $q_j = P(B_j)$ are known, then we have to check

$$H_0 : P(A_i \cap B_j) = p_i q_j, \quad i = 1, \ldots, r, \quad j = 1, \ldots, s.$$

As here

$$\{A_i \cap B_j : \quad i = 1, \ldots, r, \quad j = 1, \ldots, s\}$$

is again a partition of the sample space and

$$\{p_i \cdot q_j : \quad i = 1, \ldots, r, \quad j = 1, \ldots, s\}$$

is a given distribution, so we can follow the chi-squared test for goodness of fit. The limit of the test statistic will be $\mathcal{X}^2_{rs-1}$.

Testing independence. Contingency table

Contingency table (also known as a cross tabulation) displays the frequencies of the variables.
In applications p_i and q_j are not known. So we shall estimate them.
Repeat the experiment N-times. Denote by k_{ij} the frequency of $A_i \cap B_j$.

	B_1	B_2	$\dots$	B_s	$\sum$
A_1	k_{11}	k_{12}		k_{1s}	$k_{1.}$
A_2	k_{21}	k_{22}		k_{2s}	$k_{2.}$
$\vdots$					
A_r	k_{r1}	k_{r2}		k_{rs}	$k_{r.}$
$\sum$	$k_{.1}$	$k_{.2}$		$k_{.s}$	N

Testing independence. Contingency table...

On the margins of the table we have

$$k_{i.} = \sum_{j=1}^{s} k_{ij} \quad \text{(the frequency of } A_i\text{)}$$

$$k_{.j} = \sum_{i=1}^{r} k_{ij} \quad \text{(the frequency of } B_j\text{)}$$

We estimate the probabilities by relative frequencies:

$$p_i = P(A_i) \quad \text{is estimated by} \quad \frac{k_{i.}}{N}$$

$$q_j = P(B_j) \quad \text{is estimated by} \quad \frac{k_{.j}}{N}$$

Testing independence. Contingency table...

So the observed value of $A_i \cap B_j$ is k_{ij}, and its expected value (under H_0) is

$$N \cdot \frac{k_{i.}}{N} \cdot \frac{k_{.j}}{N} = \frac{k_{i.} k_{.j}}{N}$$

So the $\mathcal{X}^2$ statistics is

$$\mathcal{X}^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(k_{ij} - \frac{k_{i.} k_{.j}}{N}\right)^2}{\frac{k_{i.} k_{.j}}{N}}.$$

The unknown parameters are $p_1, \ldots p_r$, $q_1, \ldots, q_s$.
Because of $\sum p_i = 1$ and $\sum q_j = 1$ we have to estimate $r - 1$ values of p_i and $s - 1$ values of q_j.
So if H_0 is true, then $\mathcal{X}^2$ is asymptotically $\mathcal{X}^2_{rs-1-(r+s-2)} = \mathcal{X}^2_{(r-1)(s-1)}$.

Testing independence. Contingency table...

Example. 2000 boys doing sports were involved in a research. 150 of them were gymnasts, 1500 played football, and 350 played basketball. The heights of the boys were measured. The results are in the following contingency table. Decide if the type of the sport is independent of the height.

sport \ height	< 170	$170 - 190$	$190 <$	$\sum$
gymnast	100	50	0	150
footballer	400	1000	100	1500
basketball player	50	100	200	350
$\sum$	550	1150	300	2000

Testing independence. Contingency table...

$$\mathcal{X}^2 =$$

$$= \frac{\left(100 - \frac{550 \cdot 150}{2000}\right)^2}{\frac{550 \cdot 150}{2000}} + \frac{\left(50 - \frac{1150 \cdot 150}{2000}\right)^2}{\frac{1150 \cdot 150}{2000}} + \frac{\left(0 - \frac{300 \cdot 150}{2000}\right)^2}{\frac{300 \cdot 150}{2000}} +$$

$$+ \frac{\left(400 - \frac{550 \cdot 1500}{2000}\right)^2}{\frac{550 \cdot 1500}{2000}} + \frac{\left(1000 - \frac{1150 \cdot 1500}{2000}\right)^2}{\frac{1150 \cdot 1500}{2000}} + \frac{\left(100 - \frac{300 \cdot 1500}{2000}\right)^2}{\frac{300 \cdot 1500}{2000}} =$$

$$+ \frac{\left(50 - \frac{550 \cdot 350}{2000}\right)^2}{\frac{550 \cdot 350}{2000}} + \frac{\left(100 - \frac{1150 \cdot 350}{2000}\right)^2}{\frac{1150 \cdot 350}{2000}} + \frac{\left(200 - \frac{300 \cdot 350}{2000}\right)^2}{\frac{300 \cdot 350}{2000}} =$$

$$= 700.72$$

The degree of freedom of the chi-square distribution is $(r-1)(s-1) = 4$.

The value of 700.72 is so large that we reject independence at any usual significance level.