

Mathematical Statistics

A summary by Bernhard Kabelka,
based on the lecture by Prof. Grill in 2002/03

Version 1.02 (March 15, 2004)

The author would like to stress that

- (1) this summary was created without the knowledge or cooperation of Prof. Grill,
- (2) although every effort has been made in order to supply an accurate and complete summary, mistakes cannot be ruled out (if you spot a mistake, please inform the author, preferably by sending an email to bernhard@kabelka.net),
- (3) reading this summary does not make up for attending the lectures or reading the script.

The latest version of this file is available at:

<http://fsmat.at/~bkabelka/math/stochast/download/stat.pdf>

<http://fsmat.at/~bkabelka/math/stochast/download/stat.ps.gz>

Contents

1	Introduction	1
2	Basic Theory of Estimation	2
2.1	Point Estimation	2
2.2	Confidence Intervals	3
3	Testing	5
3.1	The Neyman-Pearson and the Likelihood Ratio Test	5
3.2	Unbiased Tests	6
3.3	Special tests for the normal distribution	6
4	Analysis of Variance	7
4.1	The Fisher-Cochran Theorem	7
4.2	One-Way Analysis of Variance	8
4.3	Two-Way Analysis of Variance	9
5	Linear Regression	9
5.1	Simple Linear Regression	9
5.2	Multiple Linear Regression	11
5.3	Other Functional Relations	11
6	The χ^2-Family of Tests	12
6.1	The χ^2 -Goodness-of-Fit Test	12
6.2	The χ^2 -Test of Independence	13
6.3	The χ^2 -Test of Homogeneity	13
7	The Kolmogorov-Smirnov Test	14
7.1	One-Sample Test	14
7.2	Two-Sample Test	14
	Appendix	15
A	Probability Distributions	15
B	German Translations of Technical Terms	16

1 Introduction

The purpose of mathematical statistics is the determination of properties of a (usually large) population based on a so-called random sample. What we do is to pick one individual “at random” (that means that each individual has the same chance of being chosen) and record the value of data (e. g. length, height, weight, ...) associated with this individual. This value X is a random variable whose distribution is the (relative) frequency distribution of that value of data within the population. If we repeat this process n times, we get a random sample (X_1, \dots, X_n) .

Sampling can be done with or without replacement which means that one certain individual can be chosen more than once or only once respectively. The former leads to a sample (X_1, \dots, X_n) with independent, identically distributed (i. i. d.) random variables X_1, \dots, X_n . In the latter case X_1, \dots, X_n are not independent. However, if the population is large enough with respect to the sample size, there is almost no difference between these two methods, so we can assume that sampling is always done with replacement.

A **random sample** of size n from distribution P is a sequence (X_1, \dots, X_n) of i. i. d. random variables with common distribution P . n is called the **sample size** and X_i an **observation**.

Given a random sample, we would like to make some statement about the underlying distribution which is made possible by the **Glivenko-Cantelli theorem**: For a sequence (X_1, \dots, X_n, \dots) of i. i. d. random variables with common distribution F we define the **empirical distribution function** as

$$F_n(x) := \frac{|\{i \leq n \mid X_i \leq x\}|}{n}$$

Then, F_n converges to F uniformly with probability one.

If the distribution of X can be characterized by one or more real numbers (**parameters**), we speak of **parametric statistics**, otherwise we speak of **non-parametric statistics**.

If f is a function from \mathbb{R}^n to \mathbb{R}^d and (X_1, \dots, X_n) is a random sample, $T = f(X_1, \dots, X_n)$ is called a **statistic**.

Important statistics are

$$\begin{aligned} \bar{X}_n &:= \frac{1}{n} \cdot \sum_{i=1}^n X_i && \text{(sample mean)} \\ S_n^2 &:= \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X}_n)^2 && \text{(sample variance)} \end{aligned}$$

A statistic T is called **sufficient** for the parameter θ , if the conditional distribution of (X_1, \dots, X_n) given T does not depend on θ .

An important criterion for sufficiency is the following: Let $(P_\theta, \theta \in \Theta)$ be a **parametric family** of distributions dominated by a measure μ , and $f_\theta = \frac{dP_\theta}{d\mu}$. The statistic $T = T(X_1, \dots, X_n)$ is sufficient for θ if the so-called **likelihood-function**

$$L(x_1, \dots, x_n, \theta) = f_\theta(x_1) \cdot \dots \cdot f_\theta(x_n)$$

admits a decomposition

$$L(x_1, \dots, x_n, \theta) = g(T, \theta) \cdot h(x_1, \dots, x_n)$$

where $h(x_1, \dots, x_n)$ does not depend on θ .

2 Basic Theory of Estimation

2.1 Point Estimation

An **estimator** is a sequence $\hat{\theta} = \{\hat{\theta}_n \mid n \in \mathbb{N}\}$ of statistics, where $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$.

An estimator is called

- **(weakly) consistent** if $\hat{\theta} \rightarrow \theta$ in probability with respect to P_θ .
- **strongly consistent** if $\hat{\theta} \rightarrow \theta$ with probability one.
- **unbiased** if $\mathbb{E}_\theta(\hat{\theta}_n) = \theta$.
- **asymptotically unbiased** if $\mathbb{E}_\theta(\hat{\theta}_n) \rightarrow \theta$.
- **efficient** if it is unbiased and has the smallest variance among all unbiased estimators.

For example, \bar{X}_n is an unbiased, strongly consistent estimator of the expectation μ of the underlying distribution, and S_n^2 is an unbiased, strongly consistent estimator of the variance σ^2 .

There are two well-known methods for calculating an estimator:

(1) **Method of moments:**

Let us suppose that the expectation is a function f of the parameter θ that has a continuous inverse f^{-1} . So we have $\theta = f^{-1}(\mathbb{E}_\theta(X))$. If we replace $\mathbb{E}_\theta(X)$ by its estimator \bar{X}_n , we get $\hat{\theta}_n = f^{-1}(\bar{X}_n)$.

If there is more than one parameter, use the higher moments $\mathbb{E}_\theta(X^k)$ and replace them by \bar{X}_n^k .

(2) **Likelihood method:**

Simply use the value of θ that maximizes the likelihood-function as an estimator. This estimator is called **maximum likelihood estimator**.

The **Cramér-Rao theorem** provides a lower bound for the variance of an unbiased estimator: Let X be a random variable with distribution P_θ , where $\theta \in \Theta$ is a real parameter and Θ is supposed to be an interval. Moreover, the density $f_\theta(x)$ should be twice differentiable with respect to θ , and both $|f'|$ and $|f''|$ should be bounded above uniformly in θ by an integrable function, that means

$$\left| \frac{\partial}{\partial \theta} f_\theta(x) \right|, \left| \frac{\partial^2}{\partial \theta^2} f_\theta(x) \right| \leq g(x) \quad \forall \theta \in \Theta \quad \text{with} \quad \int g(x) d\mu(x) < \infty$$

Furthermore, let $\hat{\theta}$ be an unbiased estimator of θ . Then

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

where

$$I(\theta) = \mathbb{E} \left(\left(\frac{\partial \log f_\theta(X)}{\partial \theta} \right)^2 \right) = -\mathbb{E} \left(\frac{\partial^2 \log f_\theta(X)}{\partial \theta^2} \right)$$

is the so-called **Fisher information**.

If we have a sample of size n , we interpret this as one n -dimensional random variable, and so we get:

$$\text{Var}_\theta(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)}$$

If $\hat{\theta}$ is an unbiased estimator and T a sufficient statistic, then $\tilde{\theta} = \mathbb{E}_\theta(\hat{\theta} | T)$ is also an unbiased estimator and has a variance which is not greater than the one of $\hat{\theta}$. This means, that if we look for an efficient estimator, we only need to consider functions of T .

Finally, if T is an sufficient statistic and has the property

$$\mathbb{E}_\theta(f(T)) = 0 \quad \forall \theta \in \Theta \quad \Rightarrow \quad f \equiv 0$$

then if $\hat{\theta} = g(T)$ is unbiased, this estimator $\hat{\theta}$ is efficient.

2.2 Confidence Intervals

$[A(X_1, \dots, X_n), B(X_1, \dots, X_n)]$ is called a **confidence interval** with **coverage probability** γ if for all θ

$$\mathbb{P}_\theta(A \leq \theta \leq B) \geq \gamma$$

If X_1, X_2, \dots are i.i.d. normal random variables with mean μ and variance σ^2 , then

- (1) \bar{X}_n has a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.
- (2) $\frac{n-1}{\sigma^2} \cdot S_n^2$ has χ^2 -distribution with $n - 1$ degrees of freedom.

(3) \bar{X}_n and S_n^2 are independent.

(4) $\frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}}$ has a t-distribution with $n - 1$ degrees of freedom.

With the help of this theorem one can obtain confidence intervals for the normal distribution:

(1) confidence interval for μ (σ^2 known):

$$\left[\bar{X}_n - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\sigma^2}{n}}, \bar{X}_n + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\sigma^2}{n}} \right]$$

(2) confidence interval for μ (σ^2 unknown):

$$\left[\bar{X}_n - t_{n-1, \frac{1+\gamma}{2}} \cdot \sqrt{\frac{S_n^2}{n}}, \bar{X}_n + t_{n-1, \frac{1+\gamma}{2}} \cdot \sqrt{\frac{S_n^2}{n}} \right]$$

(3) confidence interval for σ^2 (μ known):

$$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, \frac{1+\gamma}{2}}^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, \frac{1-\gamma}{2}}^2} \right]$$

(4) confidence interval for σ^2 (μ unknown):

$$\left[\frac{(n-1) \cdot S_n^2}{\chi_{n, \frac{1+\gamma}{2}}^2}, \frac{(n-1) \cdot S_n^2}{\chi_{n, \frac{1-\gamma}{2}}^2} \right]$$

When looking at proportions, that is to say

$$\mathbb{P}(X = 1) = \theta, \quad \mathbb{P}(X = 0) = 1 - \theta \quad \text{for a } \theta \in (0, 1),$$

one can use the fact that \bar{X}_n has an approximate normal distribution with mean θ and variance $\frac{\theta(1-\theta)}{n}$. This leads to an approximate confidence interval

$$\left[\bar{X}_n - z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\theta(1-\theta)}{n}}, \bar{X}_n + z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\theta(1-\theta)}{n}} \right]$$

but, unfortunately, we do not know the exact value of θ . So we could replace it by its estimator \bar{X}_n , or solve the equation

$$\bar{X}_n \pm z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\theta(1-\theta)}{n}} = \theta$$

with respect to θ and use the two solutions as the limits of our confidence interval.

3 Testing

A **hypothesis** is any subset H of the set of all possible probability distributions. In a parametric model, one speaks of a **parametric hypothesis**.

If the hypothesis contains only one distribution, it is called a **simple hypothesis**, otherwise it is a **composite hypothesis**.

In case of parametric hypotheses, one can distinguish **one-sided** (e.g. $\theta < \theta_0$ or $\theta > \theta_0$) and **two-sided hypotheses** (e.g. $\theta \neq \theta_0$).

Based on a sample, we decide in favour of a so-called **null hypothesis** H_0 or against it. This can be described by a **(non-randomized) test** which is a function φ from \mathbb{R}^n to $\{0, 1\}$. If $\varphi(X_1, \dots, X_n) = 0$, we accept H_0 , otherwise we reject it.

A **randomized test** is a function $\varphi : \mathbb{R}^n \rightarrow [0, 1]$ where $\varphi(X_1, \dots, X_n)$ is the probability that we reject H_0 .

An **error of the first kind** occurs, if we reject H_0 although it is true, whereas an **error of the second kind** occurs, if we accept H_0 even though it is wrong. A test φ is said to have **level of significance** α , if the probability of a first kind error is not greater than α , which means

$$\mathbb{E}_\theta(\varphi) \leq \alpha \quad \forall \varphi \in H_0$$

A **best test** of level α is a test of level α with the smallest probability of a second kind error.

3.1 The Neyman-Pearson and the Likelihood Ratio Test

If $H_0 = \{P_0\}$ and $H_1 = \{P_1\}$, the best (randomized) test of level α is the **Neyman-Pearson test**:

$$\varphi(x) = \begin{cases} 1 & \text{if } f_1(x) > k \cdot f_0(x) \\ c & \text{if } f_1(x) = k \cdot f_0(x) \\ 0 & \text{if } f_1(x) < k \cdot f_0(x) \end{cases}$$

where f_i ($i = 0, 1$) denotes the Radon-Nikodym derivative of P_i with respect to a common dominating measure, and $k \in [0, \infty]$ and $c \in [0, 1]$ are calculated from the equation

$$E_0(\varphi) = \alpha$$

The **likelihood ratio test** is given by

$$\varphi(X_1, \dots, X_n) = \begin{cases} 1 & l < \lambda \\ c & l = \lambda \\ 0 & l > \lambda \end{cases}$$

where the **likelihood ratio statistic** l is defined by

$$l := \frac{\sup_{\theta \in H_0} L(X_1, \dots, X_n, \theta)}{\sup_{\theta \in \Theta} L(X_1, \dots, X_n, \theta)}$$

and $\lambda \in [0, \infty]$ and $c \in [0, 1]$ are calculated as above.

3.2 Unbiased Tests

A test of level α is called **unbiased**, if $\mathbb{E}_\theta(\varphi) \geq \alpha \quad \forall \theta \in H_1$. Such a test can be constructed similarly to the Neyman-Pearson test:

$$\varphi(X_1, \dots, X_n) = \begin{cases} 1 & l(\theta_1) > k \cdot l(\theta_0) + \tilde{k} \cdot l'(\theta_0) \\ c & l(\theta_1) = k \cdot l(\theta_0) + \tilde{k} \cdot l'(\theta_0) \\ 0 & l(\theta_1) < k \cdot l(\theta_0) + \tilde{k} \cdot l'(\theta_0) \end{cases}$$

where $l(\theta)$ denotes the likelihood function $L(X_1, \dots, X_n, \theta)$, and $k, \tilde{k} \in [0, \infty]$ and $c \in [0, 1]$ are determined by the equations

$$\mathbb{E}_{\theta_0}(\varphi) = \alpha \quad \text{and} \quad \frac{\partial}{\partial \theta} (\mathbb{E}_\theta(\varphi))|_{\theta=\theta_0} = 0$$

3.3 Special tests for the normal distribution

3.3.1 Tests for μ

(1) σ^2 known

- $H_0 : \mu \leq \mu_0 / H_1 : \mu > \mu_0$: reject if $\bar{X}_n > \mu_0 + z_{1-\alpha} \cdot \sqrt{\frac{\sigma^2}{n}}$
- $H_0 : \mu \geq \mu_0 / H_1 : \mu < \mu_0$: reject if $\bar{X}_n > \mu_0 - z_{1-\alpha} \cdot \sqrt{\frac{\sigma^2}{n}}$
- $H_0 : \mu = \mu_0 / H_1 : \mu \neq \mu_0$: reject if $|\bar{X}_n - \mu_0| > z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}}$

(2) σ^2 unknown

- $H_0 : \mu \leq \mu_0 / H_1 : \mu > \mu_0$: reject if $\bar{X}_n > \mu_0 + t_{n-1;1-\alpha} \cdot \sqrt{\frac{S_n^2}{n}}$
- $H_0 : \mu \geq \mu_0 / H_1 : \mu < \mu_0$: reject if $\bar{X}_n > \mu_0 - t_{n-1;1-\alpha} \cdot \sqrt{\frac{S_n^2}{n}}$
- $H_0 : \mu = \mu_0 / H_1 : \mu \neq \mu_0$: reject if $|\bar{X}_n - \mu_0| > t_{n-1;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S_n^2}{n}}$

3.3.2 Tests for σ^2

(1) μ known

- $H_0 : \sigma^2 \leq \sigma_0^2 / H_1 : \sigma^2 > \sigma_0^2$: reject if $\sum (X_i - \mu)^2 > \sigma_0^2 \cdot \chi_{n;1-\alpha}$
- $H_0 : \sigma^2 \geq \sigma_0^2 / H_1 : \sigma^2 < \sigma_0^2$: reject if $\sum (X_i - \mu)^2 < \sigma_0^2 \cdot \chi_{n;\alpha}$
- $H_0 : \sigma^2 = \sigma_0^2 / H_1 : \sigma^2 \neq \sigma_0^2$: reject if $\sum (X_i - \mu)^2 < \sigma_0^2 \cdot \chi_{n;\frac{\alpha}{2}}$
or if $\sum (X_i - \mu)^2 > \sigma_0^2 \cdot \chi_{n;1-\frac{\alpha}{2}}$

(2) μ unknown

- $H_0 : \sigma^2 \leq \sigma_0^2 / H_1 : \sigma^2 > \sigma_0^2$: reject if $S_n^2 > \frac{\sigma_0^2}{n-1} \cdot \chi_{n-1;1-\alpha}$

- $H_0 : \sigma^2 \geq \sigma_0^2 / H_1 : \sigma^2 < \sigma_0^2$: reject if $S_n^2 < \frac{\sigma_0^2}{n-1} \cdot \chi_{n-1;\alpha}$
- $H_0 : \sigma^2 = \sigma_0^2 / H_1 : \sigma^2 \neq \sigma_0^2$: reject if $S_n^2 < \frac{\sigma_0^2}{n-1} \cdot \chi_{n-1;\frac{\alpha}{2}}$
or if $S_n^2 > \frac{\sigma_0^2}{n-1} \cdot \chi_{n-1;1-\frac{\alpha}{2}}$

3.3.3 Two-Sample Tests

(1) Tests for equality of means:

If σ_1^2 and σ_2^2 are known, reject H_0 if

$$|\bar{X}_n - \bar{Y}_m| > z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

and if the variance are not known, but are at least equal, reject H_0 if

$$|\bar{X}_n - \bar{Y}_m| > t_{n+m-2;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{m+n}{mn(m+n-2)} \cdot ((n-1)S_X^2 + (m-1)S_Y^2)}$$

where S_X^2 and S_Y^2 denote the sample variances of the samples (X_1, \dots, X_n) and (Y_1, \dots, Y_m) respectively.

(2) Test for equality of the variances:

Reject H_0 if

$$\frac{S_X^2}{S_Y^2} > F_{n-1,m-1;1-\frac{\alpha}{2}} \quad \text{or} \quad \frac{S_X^2}{S_Y^2} < F_{n-1,m-1;\frac{\alpha}{2}}$$

4 Analysis of Variance

4.1 The Fisher-Cochran Theorem

Let X_1, \dots, X_n be independent standard normal random variables and

$$Y_i := \sum_{j=1}^n \alpha_{i,j} \cdot X_j \quad (i = 1, \dots, k)$$

where $\alpha_{i,j}$ (for $i \in \{1, \dots, k\}$ and $j \in \{1, \dots, n\}$) are real numbers. Then

$$S := \sum_{i=1}^k Y_i^2$$

is called a **sum of squares**.

The maximal number of linearly independent random variables among the Y 's is called the number of **degrees of freedom** of a sum of squares. This number can also be obtained by subtracting the number of (linearly independent) linear relations among the Y 's from k .

The **Fisher-Cochran theorem** states the following: Let S be a sum of squares which has a χ^2 -distribution with f degrees of freedom and S_i ($i = 1, \dots, k$) a sum of squares with f_i degrees of freedom such that

$$S = S_1 + \dots + S_k$$

Then, S_1, \dots, S_k are independent and S_i ($i = 1, \dots, k$) has a χ^2 -distribution with f_i degrees of freedom if and only if

$$f = f_1 + \dots + f_k$$

4.2 One-Way Analysis of Variance

With the help of the Fisher-Cochran theorem, one can construct a test for the equality of the means of k random samples $(X_{i1}, \dots, X_{in_i})$, where X_{ij} has a normal distribution with mean μ_i and variance σ^2 . We reject our null hypothesis if

$$\frac{(N - k) \cdot \sum_{i=1}^k n_i \cdot (\bar{X}_i - \bar{X})^2}{(k - 1) \cdot \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2} > F_{k-1, n-k; 1-\alpha}$$

where

$$N := \sum_{i=1}^k n_i$$

and

$$\begin{aligned} \bar{X}_i &:= \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} X_{ij} \\ \bar{X} &:= \frac{1}{N} \cdot \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij} \end{aligned}$$

are the sample means of the i -th and the combined sample, respectively.

If we want to have a confidence interval for μ_i we may use

$$\left[\bar{X}_i - t_{n-k, \frac{1+\gamma}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2}{n_i}}, \bar{X}_i + t_{n-k, \frac{1+\gamma}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2}{n_i}} \right]$$

where

$$\hat{\sigma}^2 := \frac{1}{n - k} \cdot \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

is used as an estimator for σ^2 (because it is more accurate than S_i^2).

Sometimes it is also useful to have a confidence interval for $\mu_i - \mu_j$ which is given by

$$\left[\bar{X}_i - \bar{X}_j - t_{n-k, \frac{1+\gamma}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2 \cdot (n_i + n_j)}{n_i \cdot n_j}}, \bar{X}_i - \bar{X}_j + t_{n-k, \frac{1+\gamma}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2 \cdot (n_i + n_j)}{n_i \cdot n_j}} \right]$$

4.3 Two-Way Analysis of Variance

In this case, we may observe for example the length, weight, etc. of a product which is manufactured in two different stages: first by one of the machines A_1, \dots, A_n , then by one of the machines B_1, \dots, B_k . Then, we would like to know whether all machines of type A have the same influence on the final length, weight, etc.

If we assume that we have one product (with size X_{ij}) for each possible combination of machines A_i and B_j then we reject H_0 if

$$\frac{(k-1) \cdot k \cdot \sum_{i=1}^n (X_{i.} - X_{..})^2}{\sum_{i=1}^n \sum_{j=1}^k (X_{ij} - X_{i.} - X_{.j} + X_{..})^2} > F_{n-1, (n-1)(k-1); 1-\alpha}$$

where

$$\begin{aligned} X_{i.} &:= \frac{1}{k} \cdot \sum_{j=1}^k X_{ij} & (i = 1, \dots, n) \\ X_{.j} &:= \frac{1}{n} \cdot \sum_{i=1}^n X_{ij} & (j = 1, \dots, k) \\ X_{..} &:= \frac{1}{k \cdot n} \cdot \sum_{i=1}^n \sum_{j=1}^k X_{ij} \end{aligned}$$

5 Linear Regression

5.1 Simple Linear Regression

In this chapter, we will assume that two random variables x and Y are related through the linear equation

$$Y = a \cdot x + b$$

where a and b are (unknown) real numbers. When measuring Y for some fixed x , however, measurement errors have to be taken into account. Thus, we rather have the relation

$$Y = a \cdot x + b + e$$

where e is some (random) error, which should have expectation 0 and variance σ^2 . Furthermore, the errors of two different observations should be uncorrelated.

Under these circumstances, we can find estimators for a and b simply by minimizing

$$\sum_{i=1}^n (Y_i - a \cdot x_i - b)^2$$

which leads us to

$$\hat{a} = \frac{\sum x_i \cdot (Y_i - \bar{Y})}{\sum x_i \cdot (x_i - \bar{x})} = \frac{\sum x_i \cdot Y_i - \bar{Y} \cdot \sum x_i}{\sum x_i^2 - \frac{1}{n} \cdot (\sum x_i)^2} = \frac{\sum (x_i - \bar{x}) \cdot (Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{b} = \bar{Y} - \hat{a} \cdot \bar{x}$$

For σ^2 , we arrive at

$$\hat{\sigma}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (Y_i - \hat{a} \cdot x_i - \hat{b})^2$$

and for $\hat{Y}(x)$ we get

$$\hat{Y}(x) = \hat{a} \cdot x + \hat{b}$$

If we assume that e is standard normal distributed, then we can even calculate confidence intervals

$$\left[\hat{a} - t \cdot \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}, \hat{a} + t \cdot \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}} \right]$$

for a

$$\left[\hat{b} - t \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)}, \hat{b} + t \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)} \right]$$

for b and

$$\left[\hat{Y}(x) - t \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}, \hat{Y}(x) + t \cdot \sqrt{\hat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} \right]$$

for $Y(x) = a \cdot x + b$, where t stands for $t_{n-2; \frac{1+\gamma}{2}}$.

Sometimes, one also wants a confidence interval for $Y(x) = a \cdot x + b + e$, which is called a **prediction interval** and is given by

$$\left[\hat{Y}(x) - t \cdot \sqrt{\hat{\sigma}^2 \cdot \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}, \hat{Y}(x) + t \cdot \sqrt{\hat{\sigma}^2 \cdot \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} \right]$$

5.2 Multiple Linear Regression

In this case, we consider functions of the form

$$Y = a_1 \cdot x_1 + \dots + a_k \cdot x_k$$

where we would like to estimate a_1, \dots, a_k with the help of n observations $(Y_i, x_{i1}, \dots, x_{ik})$ ($i = 1, \dots, n$).

If we write

$$\begin{aligned} Y &:= (Y_1 \ \cdots \ Y_k)^T \\ a &:= (a_1 \ \cdots \ a_k)^T \\ X &:= \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} \end{aligned}$$

we finally arrive at

$$\begin{aligned} \hat{a} &= (X^T X)^{-1} \cdot X^T Y \\ \hat{Y}(x) &= x^T \cdot \hat{a} \\ \hat{\sigma}^2 &= \frac{1}{n-k} \cdot (Y^T Y - Y^T X \cdot (X^T X)^{-1} \cdot X^T Y) \end{aligned}$$

which leads us (again for standard normal distributed errors e) to the confidence intervals

$$\left[\hat{a}_i - t \cdot \sqrt{\hat{\sigma}^2 \cdot (X^T X)^{-1}_{ii}}, \hat{a}_i + t \cdot \sqrt{\hat{\sigma}^2 \cdot (X^T X)^{-1}_{ii}} \right]$$

for a_i and

$$\left[\hat{Y}(x) - t \cdot \sqrt{\hat{\sigma}^2 \cdot x^T (X^T X)^{-1} x}, \hat{Y}(x) + t \cdot \sqrt{\hat{\sigma}^2 \cdot x^T (X^T X)^{-1} x} \right]$$

for $Y(x)$, as well as to the prediction interval

$$\left[\hat{Y}(x) - t \cdot \sqrt{\hat{\sigma}^2 \left(1 + x^T (X^T X)^{-1} x \right)}, \hat{Y}(x) + t \cdot \sqrt{\hat{\sigma}^2 \left(1 + x^T (X^T X)^{-1} x \right)} \right]$$

for $Y(x)$, where t stands for $t_{n-k; \frac{1+\gamma}{2}}$.

5.3 Other Functional Relations

Sometimes, one wants to use a nonlinear function, such as a polynomial or an exponential function. The former can be done with multiple linear regression by letting $x_{ji} := x_i^j$, whereas for the latter one has to change the setting a little bit: By taking the logarithm on both sides of the equation $Y = a \cdot e^{bx}$ one arrives at $\ln Y = b \cdot x + \ln a$, and one can apply simple linear regression. However, one has to be aware of the fact that in this case, the error term is not an additive term, but is added in the exponent.

6 The χ^2 -Family of Tests

In this section, we would like to cope with the following problem: Given a random sample, one wants to test whether this sample comes from a certain probability distribution.

So, let us assume that X can take k values with probabilities p_1, p_2, \dots, p_k , and we would like to test $H_0 : p_i = p_i^{(0)}$ ($i = 1, \dots, k$) against the alternative that at least one p_i is different from $p_i^{(0)}$. By calculating the likelihood ratio test and properly applying the Fisher-Cochran theorem, one arrives at the test statistic

$$\chi^2 = \sum_{i=1}^k \frac{\left(Y_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$$

(where Y_i is the number of occurrences of i among X_1, \dots, X_n), which has a χ^2 -distribution with $k - 1$ degrees of freedom.

Finally, we may reject H_0 if

$$\chi^2 = \sum_{i=1}^k \frac{\left(Y_i - np_i^{(0)}\right)^2}{np_i^{(0)}} > \chi_{k-1;1-\alpha}^2$$

However, this approximation is only valid for large n . As a rule of thumb, $np_i^{(0)}$ should be at least 5.

If we want to test $H_0 : p_i = p_i^{(0)}(\theta_1, \dots, \theta_s)$ ($i = 1, \dots, k$) then we may use the maximum likelihood estimators $\hat{\theta}_j$ for θ_j ($j = 1, \dots, s$). Our test statistic, however, has no longer $k - 1$ degrees of freedom, but only $k - 1 - s$. So, we reject H_0 if

$$\chi^2 = \sum_{i=1}^k \frac{\left(Y_i - np_i^{(0)}(\hat{\theta}_1, \dots, \hat{\theta}_s)\right)^2}{np_i^{(0)}(\hat{\theta}_1, \dots, \hat{\theta}_s)} > \chi_{k-1-s;1-\alpha}^2$$

6.1 The χ^2 -Goodness-of-Fit Test

For continuous distributions (e.g. normal distribution), we may estimate the parameters by the maximum likelihood method, construct classes C_1, \dots, C_k such that $n\mathbb{P}[X \in C_i] \geq 5$ and use the above test statistic to determine whether our data has this certain distribution. In other words, we reject our null hypothesis (that X has this very distribution with parameters $\hat{\theta}_1, \dots, \hat{\theta}_s$) if

$$\chi^2 = \sum_{i=1}^k \frac{\left(Y_i - np_i^{(0)}(\hat{\theta}_1, \dots, \hat{\theta}_s)\right)^2}{np_i^{(0)}(\hat{\theta}_1, \dots, \hat{\theta}_s)} > \chi_{k-1-s;1-\alpha}^2$$

where $Y_i := |\{j \mid X_j \in C_i\}|$.

6.2 The χ^2 -Test of Independence

Given a two-dimensional sample $((X_1, Y_1), \dots, (X_n, Y_n))$, where both X and Y have a discrete distribution such that

$$\begin{aligned}\mathbb{P}[X = j] &= p_j & (j = 1, 2, \dots, k) \\ \mathbb{P}[Y = l] &= q_l & (l = 1, 2, \dots, m)\end{aligned}$$

we would like to test whether X and Y are independent. This may be achieved by the test statistic

$$\sum_{j=1}^k \sum_{l=1}^m \frac{(Z_{jl} - Z_j \cdot Z_{\cdot l} / n)^2}{Z_j \cdot Z_{\cdot l} / n}$$

where

$$\begin{aligned}Z_{jl} &:= |\{i \mid X_i = j, Y_i = l\}| \\ Z_{\cdot l} &:= |\{i \mid Y_i = l\}| = \sum_{j=1}^k Z_{jl} \\ Z_j &:= |\{i \mid X_i = j\}| = \sum_{l=1}^m Z_{jl}\end{aligned}$$

So, we reject H_0 if

$$\sum_{j=1}^k \sum_{l=1}^m \frac{(Z_{jl} - Z_j \cdot Z_{\cdot l} / n)^2}{Z_j \cdot Z_{\cdot l} / n} > \chi_{(k-1)(m-1)}^2$$

6.3 The χ^2 -Test of Homogeneity

Finally, we have k samples $(X_{i1}, \dots, X_{in_i})$ ($i = 1, \dots, k$), where each X_{ij} can take values from 1 to m . Our null hypothesis is that all samples have the same underlying distribution, which will be rejected if

$$\sum_{i=1}^k \sum_{l=1}^m \frac{(Y_{il} - n_i Z_l / n)^2}{n_i Z_l / n} > \chi_{(k-1)(m-1)}^2$$

where

$$\begin{aligned}Y_{il} &:= |\{j \mid X_{ij} = l\}| \\ Z_l &:= |\{(i, j) \mid X_{ij} = l\}| \\ n &:= \sum_{i=1}^k n_i\end{aligned}$$

7 The Kolmogorov-Smirnov Test

7.1 One-Sample Test

This test is another goodness-of-fit test: Given a sample (X_1, \dots, X_n) , we would like to test H_0 : distribution F / H_1 : distribution $\neq F$. By the Glivenko-Cantelli theorem, the empirical distribution function

$$F_n(x) := \frac{|\{i \leq n \mid X_i \leq x\}|}{n}$$

converges uniformly to the actual distribution function of X . So we choose

$$D_n := \|F_n - F\| = \sup_x |F_n(x) - F(x)|$$

as our test statistic, which should be small (under H_0).

One can prove that for continuous F , the null distribution of D_n does not depend on F , and D_n can be calculated as follows:

$$D_n = \max_{i=1}^n \left(\max \left(\left| F(X_{n:i}) - \frac{i}{n} \right|, \left| F(X_{n:i}) - \frac{i-1}{n} \right| \right) \right)$$

where $X_{n:1} < X_{n:2} < \dots < X_{n:n}$ denote the order statistics.

One can even calculate the asymptotic distribution function for large n : Let $\lambda_n := \sqrt{n} \cdot D_n$. Then, for $n \rightarrow \infty$,

$$\mathbb{P}[\lambda_n \leq x] \rightarrow K(x) := \begin{cases} 0 & \text{if } x \leq 0 \\ \sum_{k=-\infty}^{k=+\infty} (-1)^k \cdot e^{-2k^2x^2} & \text{if } x > 0 \end{cases}$$

7.2 Two-Sample Test

If we would like to compare the two samples (X_1, \dots, X_n) and (Y_1, \dots, Y_m) and would like to know whether the underlying distributions are the same, we may use the test statistic

$$D_{n,m} := \|F_n - G_m\|$$

where F_n and G_m resp. are the empirical distribution functions. Again, for continuous F , the null distribution of $D_{n,m}$ does not depend on F . $D_{n,m}$ may be calculated as follows:

$$D_{n,m} = \max \left(\max_{i \leq n} |F_n(X_{n:i}) - G_m(X_{n:i})|, \max_{i \leq m} |F_n(Y_{m:i}) - G_m(Y_{m:i})| \right)$$

Finally, $\lambda_n := \sqrt{\frac{nm}{n+m}} \cdot D_{n,m}$ has an asymptotic distribution function K .

Appendix

A Probability Distributions

The **uniform distribution** on $[a, b]$ has the density

$$f_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The **normal distribution** with mean μ and variance σ^2 has the density

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The **gamma distribution** with parameters (a, λ) has the density

$$f_{a,\lambda}(x) = \frac{\lambda^a}{\Gamma(a)} \cdot e^{-\lambda \cdot x} \cdot x^{a-1} \quad (x > 0)$$

A gamma distribution with parameters $(\frac{n}{2}, \frac{1}{2})$ is called **χ^2 -distribution** with n degrees of freedom, and a gamma distribution with parameters $(1, \lambda)$ is called **exponential distribution**.

If U and V are independent random variables with $U \sim N(0, 1)$ and $V \sim \chi_n^2$, then the distribution of $T = \frac{U}{\sqrt{V/n}}$ is called **t-distribution** with n degrees of freedom and has the density

$$f_n(x) = \frac{1}{\sqrt{\pi} \cdot n} \cdot \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

If U and V are independent random variables with $U \sim \chi_a^2$ and $V \sim \chi_b^2$, then the distribution of $T = \frac{U/a}{V/b}$ is called **F-distribution** with (a, b) degrees of freedom and has the density

$$f_{a,b}(x) = \frac{a}{b} \cdot \frac{\Gamma(\frac{a+b}{2})}{\Gamma(\frac{a}{2}) \cdot \Gamma(\frac{b}{2})} \cdot \frac{(\frac{a}{b}x)^{\frac{a}{2}-1}}{(1 + \frac{a}{b}x)^{\frac{a+b}{2}}} \quad \text{for } x > 0$$

B German Translations of Technical Terms

analysis of variance, one-way two-way	Varianzanalyse, einfache doppelte
confidence interval	Konfidenzintervall
coverage probability	Überdeckungswahrscheinlichkeit
degrees of freedom	Freiheitsgrade
distribution	Verteilung
distribution function, empirical	Verteilungsfunktion, empirische
estimator, consistent efficient unbiased	Schätzer, konsistenter effizienter erwartungstreuer, unverzerrter
exponential distribution	Exponentialverteilung
gamma distribution	Gammaverteilung
goodness-of-fit test	Anpassungstest
hypothesis, composite one-sided simple two-sided	Hypothese, zusammengesetzte einseitige einfache zweiseitige
level of significance	Signifikanzniveau
likelihood	Likelihood
likelihood ratio test	Likelihoodquotiententest
linear regression, multiple	lineare Regression, mehrfache
maximum likelihood estimator	Maximum Likelihood Schätzer
method of moments	Momentenmethode
normal distribution	Normalverteilung
null hypothesis	Nullhypothese
parameter	Parameter
parametric family	parametrische Familie
prediction interval	Vorhersageintervall
reject	ablehnen, verwerfen
rejection region	Verwerfungsbereich

sample	Stichprobe
sample mean	Stichprobenmittel
sample variance	Stichprobenvarianz
statistic,	Statistik,
sufficient	suffiziente
statistics,	Statistik,
non-parametric	nichtparametrische
parametric	parametrische
test,	Test, Signifikanztest,
randomized	randomisierter
unbiased	unverfälschter
uniform distribution	Gleichverteilung

Index

- analysis of variance, 7
 - one-way, 8
 - two-way, 9
- ANOVA, *see* analysis of variance
- χ^2 -distribution, 15
- χ^2 -test, 12
 - of homogeneity, 13
 - of independence, 13
- confidence interval, 3
- coverage probability, 3
- Cramér-Rao, 3
- degrees of freedom, 7
- empirical distribution function, 1
- error
 - of the first kind, 5
 - of the second kind, 5
- estimator, 2
 - asymptotically unbiased, 2
 - efficient, 2
 - strongly consistent, 2
 - unbiased, 2
 - weakly consistent, 2
- exponential distribution, 15
- F-distribution, 15
- Fisher information, 3
- Fisher-Cochran, 8
- gamma distribution, 15
- Glivenko-Cantelli, 1
- goodness-of-fit test, 12
- hypothesis, 5
 - composite, 5
 - one-sided, 5
 - parametric, 5
 - simple, 5
 - two-sided, 5
- Kolmogorov-Smirnov test, 14
- level of significance, 5
- likelihood method, 2
- likelihood ratio statistic, 5
- likelihood ratio test, 5
- likelihood-function, 2
- linear regression, 9
 - multiple, 11
- maximum likelihood estimator, 2
- method of moments, 2
- Neyman-Pearson test, 5
- normal distribution, 15
- null hypothesis, 5
- observation, 1
- parameters, 1
- parametric family, 2
- parametric statistics, 1
- prediction interval, 10
- random sample, 1
- sample mean, 1
- sample size, 1
- sample variance, 1
- statistic, 1
 - sufficient, 1
- statistics
 - non-parametric, 1
- sum of squares, 7
- t-distribution, 15
- test, 5
 - best, 5
 - randomized, 5
 - unbiased, 6
- uniform distribution, 15