Statistical inference (MVE155/MSG200)

Introduction

Statistical inference

Lectures: Aila Särkkä
Tuesday 13:15-15:00
Friday 13:15-15:00

Exercises: TonyJohansson
Monday 13-15-15:00
Wednesday 13:15-15:00

Exam: Tuesday, March 14, 14:00-18:00

Course literature:

- Compendium "Statistical inference" by Serik Sagitov
- Additional textbook: Mathematical statistics and data analysis, 3rd edition (2nd edition is also OK), by John Rice (Cremona).

Interested in taking notes?

CHALMERS

FUNKA IS LOOKING FOR A NOTE-TAKER

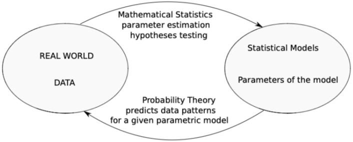
- You help other students who would otherwise find it difficult to take up their studies in a good way.
- You also help yourself as your own notes make you perform better and remember longer.
- Through the work, you do you will learn to plan and structure your notes, which you will benefit from even after studying.
- You also get paid for your work as a note-taker.

https://student.portal.chalmers.se/en/chalmersstudies/disabilities/notetakingsupport/Pages/for-those-who-want-to-become-a-note-taker.aspx

Introduction

Statistical analysis consists of

- collecting data
- organising and summarising data
- analysing and interpreting data (inference).



In this course, we will talk about

- parameter estimation and hypothesis testing based on properly collected, relatively small data sets.
- basic principles of experimental design, such as randomisation, blocking, and replication, are recalled.

List of course topics

- Parametric models (different distribution)
- Random sampling (simple random sampling, stratified sampling)
- Parameter estimation (method of moments, maximum likelihood)
- Hypothesis testing (likelihood ratio test)
- Bayesian inference
- Summarising data (QQ-plots, skewness and kurtosis)
- Comparing two samples (means, proportions, paired data)
- Analysis of variance
- Categorical data analysis (χ^2-test)
- Multiple regression

Some definitions and notations

We denote random variables by capital letters, X, Y, Z, ..., and their values/realizations by the corresponding small letters x, y, z...

Recall that the expected value (mean) of a random variable X is defined as

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i \, \mathrm{P}(X = x_i) = \sum_{i=1}^{\infty} x_i p_i$$

if X is a discrete with the probability mass function $p_i = \mathrm{P}(X = x_i)$, $i = 1, \ldots$, and as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \, f(x) \, dx$$

if X is a continuous with the density function f. The mean of X is often denoted by μ, i.e.

$$\mathbb{E}(X) = \mu.$$

Some definitions and notations

The variance of X is defined as

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}((X - \mu)^2) = \mathbb{E}(X^2) - \mu^2,$$

where $X - \mu$ is called the deviation from the mean. The square root of the variance is called the standard deviation.

The variance of X is often denoted by σ^2, i.e. $\mathrm{Var}(X) = \sigma^2$, and the standard deviation by σ.

Sometimes, a standardised version of X,

$$Z = \frac{X - \mu}{\sigma},$$

called a z-score, is used. It has mean 0 and variance 1.

Some definitions and notations

Covariance between two random variables X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) = \mathbb{E}(XY) - \mu_X \mu_Y$$

and correlation as

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where μ_X and μ_Y, and σ_X and σ_Y, are the means and standard deviations of X and Y, respectively.

Note that $-1 \leq \rho \leq +1$ and ρ is -1 or $+1$ if X is a linear function of Y.

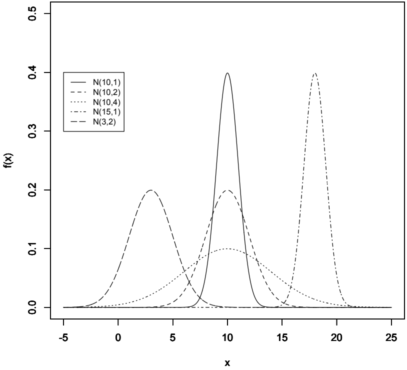
If X and Y are independent, $\rho = 0$, but not necessarily the other way around.

Some distributions: normal distribution

- Normal distribution plays a central role in probability theory and statistics.
- If a random variable X is normally distributed, $X \sim N(\mu, \sigma)$, it has the density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$.



Some distributions: normal distribution

- If $X \sim N(\mu, \sigma)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

  (standard normal distribution).

- Measurement error (random noise) is often modelled by a normal $N(0, \sigma)$ variable, i.e.

$$Y = \mu + \sigma Z,$$

  where $Z \sim N(0, 1)$. The standard deviation is called the size of the noise.

Law of large numbers and central limit teorem

Let $X_1, X_2, ..., X_n$ be independent and identically distributed (iid) random variables with mean μ and variance σ^2.

The law of large numbers states that $\bar{X} \to \mu$ as $n \to \infty$.

According to the central limit theorem, the mean

$$\bar{X} \approx N(\mu, \sigma/\sqrt{n})$$

if the sample size n is large enough.

Mixtures of normal distributions

Example: We have particles of two different sizes in a liquide. Diameter distribution of one of them, X_1, is $N(\mu_1, \sigma_1)$ (with the density function f_1) and of the other, X_2, $N(\mu_2, \sigma_2)$ (with f_2). The distribution of the diameter is then a mixture of the two normal distributions,

$$f(y) = w_1 f_1(y) + w_2 f_2(y),$$

where the particle (its diameter) comes from the distribution f_1 with probability w_1 and from the distribution f_2 with probability w_2, and $w_1 + w_2 = 1$.

Mixtures of normal distributions

In general, if we have k different components (particle sizes), $X_1, X_2, ..., X_k$, each having a a normal distribution $N(\mu_i, \sigma_i)$, $i = 1, ..., k$, the variable Y from the mixture distribution

$$f(y) = w_1 f_1(y) + ... + w_k f_k(y),$$

and comes from the ith normal distribution with probability w_i, $i = i, ..., k$, where $w_1 + w_2 + ... + w_k = 1$.

The mean $\mathbb{E}(Y) = \mu$ and variance $\mathrm{Var}(Y) = \sigma^2$ become

$$\mu = \sum_{i=i}^{k} w_i \mu_i \text{ and } \sigma^2 = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2 + \sum_{i=1}^{k} w_i \sigma_i^2.$$

Note that the variance has two parts, between and within components.

t-distribution

Let $X_1, X_2, ..., X_n$ be iid variables from $N(\mu, \sigma)$. Then, $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ and

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

What is the distribution if σ is unknown and needs to be estimated?

t-distribution

Recall that $\bar{X}$ is an unbiased estimator for μ (with variance σ^2/n) and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

an unbiased estimator for σ^2 (with variance $\frac{\sigma^4}{n}(\mathbb{E}((\frac{X-\mu}{\sigma})^4) - \frac{n-3}{n-1})$.

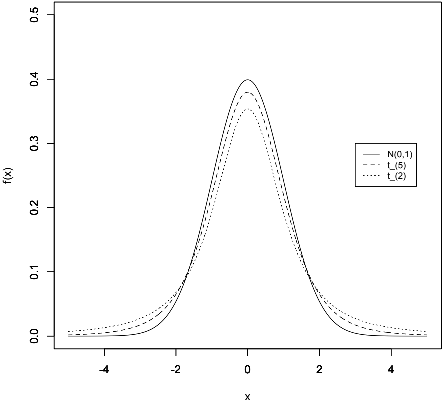
Then,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has the t-distribution with $n-1$ degrees of freedom, t_{n-1}.

t-distribution

As the number of degrees of freedom increases, the t-distribution approaches $N(0,1)$-distribution.



t-distribution

The density function of the t-distribution with $k \geq 0$ degrees of freedom is

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\,\Gamma(\frac{k}{2})}(1 + \frac{x^2}{k})^{-\frac{k+1}{2}}, \quad -\infty < x < \infty,$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$

and

$$\Gamma(k) = (k-1)! \text{ for } k = 1, 2, ...$$

The mean of the t-distribution is always zero and the variance depends on k (can be infinite or undefined).

Gamma and exponential distributions

Positive valued continuous distributions.

The density function of Gamma distribution, $\text{Gam}(\alpha, \lambda)$, is

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \quad x > 0,$$

where $\alpha > 0$ is a shape parameter and $\lambda > 0$ is the (inverse) scale parameter.

Exponential distribution is a special case of Gamma distribution, namely

$$\text{Gam}(1, \lambda) = \text{Exp}(\lambda).$$

Also, if $X_1, X_2, ..., X_k$ are independent $\text{Exp}(\lambda)$-variables, then

$$X_1 + X_2 + ... + X_k \sim \text{Gam}(k, \lambda), \quad k = 1, 2, ...$$

The mean of $\text{Gam}(\alpha, \lambda)$ is α/λ and the variance α/λ^2.

χ^2 distribution

χ^2-distribution can be defined by using $N(0,1)$-distribution: Let $Z_1, Z_2, ..., Z_n$ be independent $N(0,1)$-distributed random variables. Then

$$Z_1^2 + Z_2^2 + ... + Z_n^2 \sim \chi_n^2.$$

Also, for independent $N(\mu, \sigma)$-distributed random variables $X_1, X_2, ..., X_n$

$$\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + ... + (X_n - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2.$$

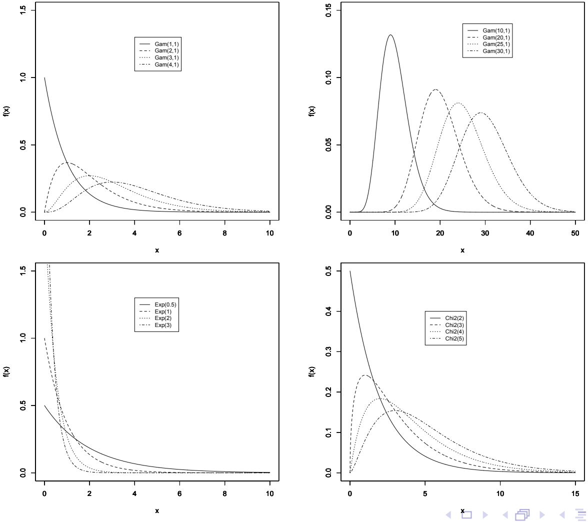
χ^2-distribution is also a special case of Gamma distribution, namely

$$\mathrm{Gam}(\frac{k}{2}, \frac{1}{2}) = \chi^2(k)$$

(k is a positive integer).

Examples of gamma, exponential, and χ^2 distributions

Gamma (top), exponential (bottom left), and χ^2:



Note on t-distribution

t-distribution is defined as a ratio of two independent random variables: a $N(0,1)$-distributed random variable Z and a square root of a χ^2-distributed random variable V divided by the number of its degrees of freedom df, i.e.

$$\frac{Z}{\sqrt{V/df}} \sim t_{df}.$$

Let $X \sim N(\mu, \sigma)$ and S^2 the sample variance based on a sample of size n. Then,

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S} = \frac{\sqrt{n}\,(\bar{X}-\mu)/\sigma}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} \sim t_{n-1}$$

since $\sqrt{n}(\bar{X}-\mu)/\sigma \sim N(0,1)$ and $V = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Bernoulli distribution

We flip a coin and define X to be a stochastic variable that gets the value 1 if the result is "heads" and value 0 if the result is "tails". Let the probability of "heads" be p.

Then, X is Bernoulli distributed with parameter $p \in [0, 1]$ with

$$\mathrm{P}(X = 1) = p \text{ and } \mathrm{P}(X = 0) = 1 - p.$$

Binomial distribution

Let us flip the coin n times and denote the (stochastic) number of "heads" by Y. Then, Y is binomially distributed, $\mathrm{Bin}(n, p)$. For $Y \sim \mathrm{Bin}(n, p)$,

$$\mathrm{P}(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, ..., n.$$

Y is a sum of n independent random variables from Bernoulli distribution with parameter p. Therefore, Bernoulli distributed random variable is $\mathrm{Bin}(1, p)$.

For $y \sim \mathrm{Bin}(n, p)$,

$$\mu = np \text{ and } \sigma^2 = np(1-p).$$

Binomial distribution: approximation, continuity correction

Binomial distribution can be approximated by a normal distribution (due to the central limit theorem) when $np \geq 5$ and $n(1-p) \geq 5$:

$$\text{Bin}(n, p) \approx N(np, \sqrt{np(1-p)}).$$

If n is small, the approximation can be improved by using the so-called continuity correction: For $Y \sim$Bin(n, p) (and $y = 1, ..., n$),

$$\mathrm{P}(Y \leq y) = \mathrm{P}(Y < y + 1)$$

and therefore, y can be replaced by any number in the interval $[y, y+1)$, e.g. by $y + \frac{1}{2}$.

Binomial distribution: continuity correction

Then,

$$\mathrm{P}(Y \leq y) = \mathrm{P}(Y \leq y + \frac{1}{2}) \approx \Phi\left(\frac{y + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

and

$$\mathrm{P}(Y < y) = \mathrm{P}(Y \leq y - \frac{1}{2}) \approx \Phi\left(\frac{y - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right),$$

where Φ is the distribution function of the standard normal distribution $N(0,1)$.

Multinomial distribution

In binomial distribution, there are two possible outcomes like "heads" and tails" or "success" and "failure". If there are more than two outcomes (e.g. six sides of a dice), we have a multinomial distribution. Then, $(X_1, ..., X_r) \sim \text{Mn}(n; p_1, ..., p_r)$ and

$$\mathrm{P}(X_1 = x_1, ..., X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r},$$

where $x_i = 0, ..., n,\ i = 1, .., r$ and $(p_1, ..., p_r)$ is a vector of probabilities such that

$$p_1 + ... + p_r = 1.$$

Note that $\text{Bin}(n, p) = \text{Mn}(n; p, 1 - p)$ and that the marginal distributions of $X_i, ..., X_r$ are $\text{Bin}(n, p_i)$. Also, the different counts X_i and X_j, $i \neq j$, are negatively correlated.

Poisson distribution

Poisson distribution is used to describe the number of rear events, e.g. earthquakes, during a given time interval.

For a Poisson distributed random variable $X \sim \text{Pois}(\mu)$,

$$\mathrm{P}(X = x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, \dots.$$

The mean and variance are both equal to μ.

$\text{Pois}(\mu)$ can be obtained as a limit of $\text{Bin}(n, p)$ as $n \to \infty$, $p \to 0$, and $np \to \mu$ (and $\text{Bin}(n, p)$ can be approximated by $\text{Pois}(np)$).

Geometric distribution

We have a sequence of coin flips (Bernoulli trials) with probability p for "heads". The number of trials, X, needed until we get the first "heads" has a geometric distribution with parameter p, $p \sim \text{Geom}(p)$, with

$$\mathrm{P}(X = x) = (1 - p)^{x-1} p, \quad x = 1, 2, \ldots$$

The mean and the variance are $\frac{1}{p}$ and $\frac{1-p}{p^2}$, respectively.

Hypergeometric distribution

Let us have B black balls and $W = N - B$ white balls in a box (with N balls) and let us draw n balls from the box without replacement. Then, the number of black balls among the n balls, X, has the distribution

$$\mathrm{P}(X = x) = \frac{\left(\begin{array}{c} B \\ x \end{array}\right)\left(\begin{array}{c} W \\ n - x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}, \quad \max(0, n - W) \leq x \leq \min(n, B),$$

and $X \sim \mathrm{HG}(N, n, p)$-distributed, where p is the portion of black balls, i.e. $B = Np$.

The mean and variance of X are $\mu = np$ and $np(1 - p)\frac{N-n}{N-1}$, where $\frac{N-n}{N-1}$ is called the finite population correction.

Hypergeometric distribution

If n is much smaller than N, $\frac{N-n}{N-1}$ is close to 1 and $\mathrm{HG}(N, n, p) \approx \mathrm{Bin}(n, p)$.

Also, $\mathrm{HG}(N, n, p)$ can be approximated by normal distribution, namely

$$\mathrm{HG}(N, n, p) \approx N\left(np, \sqrt{np(1-p)\frac{N-n}{N-1}}\right),$$

which can be used when $np \geq 5$ and $n(p-1) \geq 5$. Note that the drawings are not independent since they are drawn without replacement.