Introduction to statistical inference 2

Lauri Mehtätalo
University of Eastern Finland
School of Computing

May 28, 2018

II

Contents

Chapter 1

Recap from "Introduction to statistical inference 1"

1.1 Random variable

- Random variable is a function from sample space $\mathcal{S}$ of an experiment to sample space of the random variable $\mathcal{X}$, which is set of real numbers.

$$X : \mathcal{S} \to \mathcal{X}$$

- The sample space is a set of all possible values random variable can get.
- $\mathcal{X}$ can be
  - an interval of real axis (continuous random variable).

$$\mathcal{X} = [0, 10),\ \mathcal{X} = [0, 10],\ \mathcal{X} = (0, 10)$$

  - An uncountable set of integers (discrete random variable)

$$\mathcal{X} = \{0, 1, 2, \ldots\}$$

  - A countable set of integers or real numbers (discrete random variable)

$$\mathcal{X} = \{0, 1\},\ \mathcal{X} = \{0, 0.5, 1\},\ \mathcal{X} = \{0, 1, \ldots, 10\}$$

- Probabilities associated with each value of X are defined by the cumulative distribution function (cdf for short).

$$F_X(x) = P(X \leq x), \text{ where } -\infty < x < \infty$$

Note: $F_X(x)$ is a step function if X is discrete.

Note: $F_X(x)$ is a continuous function if X is continuous.

- $F_X(x)$ or $F(x)$ is cdf, if

  1) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$

  2) $F(x)$ is non-decreasing

  3) $F(x)$ is right-continuous

- Cdf is useful in calculation of any probabilities; for example

$$P(a < X \leq b) = F(b) - F(a)$$

Note: Be careful with $<$ and $\leq$ when working with discreate random variables.

- The probability density function (pdf for short) is defined for continuous random variable as

$$f_X(x) = F'(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}, \ -\infty < x < \infty$$

  and

$$\int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = F_X(x)$$

- The probability mass function (pmf for short) is defined for discrete random variables as

$$f_X(x) = P(X = x)$$
$$F_X(x) = \sum_{k=1}^{x} f_X(k)$$

1.2 Transformations of random variable

- Consider a monotonic function $g : \mathcal{X} \to \mathcal{Y}$
- $Y = g(X)$ is also a random variable; function g is called an transformation (muunnos).
- If $g(x)$ is a increasing function of x, then

$$F_Y(y) = F_X(g^{-1}(y))$$

- If $g(x)$ is a decreasing function of x, then

$$F_Y(y) = 1 - F_X(g^{-1}(y))$$

- The pdf of continuous Y is

$$f_Y(y) = F_Y'(y)$$

1.3 Expected values

$$E(X) = \mu_X = \begin{cases} \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x & \text{if } X \text{ is continuous} \\ \sum_{x\in\mathcal{X}} x f(x) & \text{if } X \text{ is discrete} \end{cases}$$

$$E(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x & \text{if } X \text{ is continuous} \\ \sum_{x\in\mathcal{X}} g(x) f(x) & \text{if } X \text{ is discrete} \end{cases}$$

1.4 Variance

$$\begin{aligned} \sigma_X^2 = Var(X) &= E(X-\mu_X)^2 \\ &= E(X^2 - 2X\mu_X + \mu_X^2) \\ &= E(X^2) - E(2X\mu_X) + E(\mu_X^2) \\ &= E(X^2) - \underbrace{2\mu_X \underbrace{E(X)}_{\mu_X}}_{2\mu_X^2} + E(\mu_X^2) \\ &= E(X^2) - \mu_X^2 \end{aligned}$$

$$sd(X) = \sqrt{Var(X)} = \sigma_X$$

1.5 Bivariate random variables

- For two discrete random variables, the joint pmf is defined as

$$f_{X,Y}(x,y) = P(X = x, Y = y)$$

- For two continuous random variables, we define the joint pdf $f_{X,Y}(x,y)$ as

$$P((X,Y) \in A) = \iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

- The expected value for transformation $g(X,Y) : \mathbb{R}^2 \to \mathbb{R}$ (for example, $g(X,Y) = XY$ or $g(X,Y) = \frac{X}{Y}$) is

$$E(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

if (X,Y) is continuous, and

$$E(g(X,Y)) = \sum_{x,y\in\,\mathbb{R}^2} g(x,y) f(x,y)$$

if (X, Y) is discrete.

- The marginal pmf / pdf for X are

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) \quad \text{(pmf)}$$
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y \quad \text{(pdf)}$$

  and correspondingly for Y.

- The conditional pmf / pdf are both defined as

$$f(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad \text{(for both discrete and continuous random variables)}$$

  and correspondingly for $x \mid y$.

1.6 Independence

- Random variables are said to be independent ($X \perp\!\!\!\perp Y$) if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

- For independent random variables, conditional distribution $y \mid x$ is

$$f(y \mid x) = f_y(y)$$

  regardless of the value of x.

1.7 Covariance

- Covariance measures the linear association between two random variables X and Y

$$\begin{aligned} cov(X, Y) &= E((X - \mu_X)(Y - \mu_Y)) \\ &= E(XY) - \mu_X \mu_Y \\ cov(X, Y) &= cov(Y, X) \\ corr(X, Y) &= \rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \qquad \text{where } -1 \leq \rho_{XY} \leq 1 \end{aligned}$$

Note: $cov(X, X) = Var(X)$

Note: If $X \perp\!\!\!\perp Y$, then $cov(X, Y) = 0$, but if $cov(X, Y) = 0$, it does not mean X and Y are necessarily independent.

1.8 Random vectors

- Random vectors generalize the bivariate random variables to a n-variate case.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \text{where } X_i,\ i = 1, 2, \ldots, n \text{ are scalar random variables}$$

- If $\boldsymbol{x}$ is a continuous random vector, then

$$P(\boldsymbol{X} \in A) = \int_A \cdots \int f(\boldsymbol{x}) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n \quad f : \mathbb{R}^2 \to \mathbb{R}$$

  where $f(\boldsymbol{x})$ is a joint pdf.

- If $\boldsymbol{x}$ is a discrete random vector, then

$$P(\boldsymbol{X} \in A) = \sum \ldots \sum f(\boldsymbol{x})$$

  where $f(\boldsymbol{x})$ is a joint pmf.

- Let $g(\boldsymbol{x})$ be a transformation $g : \mathbb{R}^2 \to \mathbb{R}$. Then, the expected value is

$$E(g(\boldsymbol{X})) = \begin{cases} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n & \text{if } \boldsymbol{X} \text{ is continuous} \\ \sum \ldots \sum_{\boldsymbol{X} \in \mathbb{R}^2} g(\boldsymbol{x}) f(\boldsymbol{x}) & \text{if } \boldsymbol{X} \text{ is discrete} \end{cases}$$

- Let us partition the n-variate random vector $\boldsymbol{X}$ as follows:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$$

  where $\boldsymbol{X}_1$ has lenght k and $\boldsymbol{X}_2$ has length $n - k$.

- The joint pdf of $\boldsymbol{X}$ can be written as

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

- The marginal density of $\boldsymbol{X}_1$ is

$$\begin{aligned} f(\boldsymbol{x}_1) &= \int_{\mathbb{R}^{n-k}} f(\boldsymbol{x}_1, \boldsymbol{x}_2) \, \mathrm{d}\boldsymbol{x}_2 && \text{if } \boldsymbol{X} \text{ is continuous} \\ f(\boldsymbol{x}_1) &= \sum_{\mathbb{R}^{n-k}} f(\boldsymbol{x}_1, \boldsymbol{x}_2) && \text{if } \boldsymbol{X} \text{ is discrete} \end{aligned}$$

  where $f(\boldsymbol{x}_1)$ is a k-variate pdf/pmf.

- The conditional pdf/pmf for $\boldsymbol{X}_2 \mid \boldsymbol{X}_1$ is

$$f(\boldsymbol{X}_2 \mid \boldsymbol{X}_1) = \frac{f(\boldsymbol{X}_1, \boldsymbol{X}_2)}{f(\boldsymbol{X}_2)}$$

- Expected value of random vectors is defined as vector

$$E(\boldsymbol{x}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{bmatrix}_{n\times 1} = \begin{bmatrix} E(\boldsymbol{X}_1) \\ E(\boldsymbol{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$$

- The variance of a random vector is $n \times n$ symmetric matrix called variance-covariance matrix.

$$Var(\boldsymbol{x})_{n\times n} = \begin{bmatrix} Var(X_1) & cov(X_1, X_2) & cov(X_1, X_3) & \dots & cov(X_1, X_n) \\ cov(X_2, X_1) & Var(X_2) & cov(X_2, X_3) & \dots & cov(X_2, X_n) \\ cov(X_3, X_1) & cov(X_3, X_2) & Var(X_3) & \dots & cov(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & cov(X_n, X_3) & \dots & Var(X_n) \end{bmatrix}$$

$$= \begin{bmatrix} Var(\boldsymbol{x}_1)_{k\times k} & cov(\boldsymbol{x}_1, \boldsymbol{x}_2')_{k\times(n-k)} \\ cov(\boldsymbol{x}_2, \boldsymbol{x}_1')_{(n-k)\times k} & Var(\boldsymbol{x}_2)_{(n-k)\times(n-k)} \end{bmatrix}$$

- The correlation matrix is defined as

$$corr(\boldsymbol{x})_{n\times n} = \begin{bmatrix} 1 & corr(X_1, X_2) & corr(X_1, X_3) & \dots & corr(X_1, X_n) \\ corr(X_2, X_1) & 1 & corr(X_2, X_3) & \dots & corr(X_2, X_n) \\ corr(X_3, X_1) & corr(X_3, X_2) & 1 & \dots & corr(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ corr(X_n, X_1) & corr(X_n, X_2) & corr(X_n, X_3) & \dots & 1 \end{bmatrix}$$

- If $\boldsymbol{X}$ has a n-variate normal distribution, then, with

$$E(\boldsymbol{x}) = \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad Var(\boldsymbol{x})_{n\times n} = \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2 \end{bmatrix}$$

then $\boldsymbol{X}_1 \mid \boldsymbol{X}_2$ has a k-variate normal distribution with

$$E(\boldsymbol{X}_1 \mid \boldsymbol{X}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2)$$

$$Var(\boldsymbol{X}_1 \mid \boldsymbol{X}_2) = \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_{21}$$

Note: $Var(\boldsymbol{X}_1 \mid \boldsymbol{X}_2) \leq Var(\boldsymbol{X}_1)$

1.9 Computing using expected values and variances

Let a, b and c be constants, and let X, Y and Z be (scalar) random variables. The following rules hold regardless of the distribution of random variables X, Y and Z.

$E(c) = c$	
$E(cX) = cE(X)$	
$E(X+Y) = E(X) + E(Y)$	
$E(X+c) = E(X) + c$	
$E(XY) = E(X)E(Y)$	Only if $X \perp\!\!\!\perp Y$.
$E(g(X)) = g(E(X))$	Only in some special cases, like when $g(X)$ is a linear transformation.

$Var(X+Y) = Var(X) + Var(Y) + 2cov(X,Y)$	
$Var(X-Y) = Var(X) + Var(Y) - 2cov(X,Y)$	
$Var(aX) = a^2 \cdot Var(X)$	
$cov(X, Y+Z) = cov(X,Y) + cov(X,Z)$	
$cov(aX, bY) = ab \cdot cov(XY)$	
$Var(X+a) = Var(X)$	
$cov(X+a, Y+b) = cov(X,Y)$	
$E(X) = E_Y\left[E_{X\|Y}(X \mid Y)\right]$	
$Var(X) = E_Y\left[Var_{X\|Y}(X \mid Y)\right] + Var_Y\left[Var_{X\|Y}(X \mid Y)\right]$	

Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be fixed vectors and $\boldsymbol{X}$ and $\boldsymbol{Y}$ random vectors so that the dimensions in the equations match.

$E(\boldsymbol{a}'\boldsymbol{X}) = \boldsymbol{a}'E(\boldsymbol{X})$	
$Var(\boldsymbol{a}'\boldsymbol{X}) = \boldsymbol{a}'Var(\boldsymbol{X})\boldsymbol{a}$	Compare to $Var(aX) = a^2 \cdot Var(X) = a \cdot Var(X) \cdot a$
$cov(\boldsymbol{a}'\boldsymbol{X}, \boldsymbol{b}'\boldsymbol{Y}) = \boldsymbol{a}'cov(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{b}$	

Note: These equations need to be remembered by hearth!

Chapter 2

Random samples

Definition 2.1. Random variables $X_1, .., X_n$ are called random sample of size n from population $f(x)$, if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf/pmf if each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called independent and identically distributed random variables (i.i.d.) with pdf/pmf $f(x)$.

Note: Sample $X_1, \ldots, X_n$ can also be denoted by $\boldsymbol{X}$, where $\boldsymbol{X} = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}^T$

- If follows from the mutual independence, of $X_1, \ldots, X_n$ that the joint pdf or pmf of $\boldsymbol{X}$ is

$$f(x_1, \ldots, x_2) = f(x_1)f(x_2)\ldots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

- All univariate marginal distributions $f(x_i)$ are the same by definition 2.1.

Example 2.1. Let $X_1, \ldots, X_n$ be a random sample from $Exponential(\beta)$ population. X_i specifies the time until failure for n identical cellphones. The exponential pdf is

$$f(x_i) = \frac{1}{\beta} e^{-x_i/\beta}$$

The joint pdf pf the sample is

$$f(x_1, x_2, \ldots, x_n \mid \beta) = \prod_{i=1}^{n} \tfrac{1}{\beta} e^{-x_i/\beta} = \tfrac{1}{\beta^n} e^{\frac{1}{\beta}\sum x_i} \quad \text{Recall: } a^b a^c = a^{b+c}$$

What is the probability that all n cellphones last more than 2 years?

$$
\begin{aligned}
P(X_1 > 2, \ldots, X_n > 2) &= \int_2^\infty \cdots \int_2^\infty f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&= \int_2^\infty \cdots \int_2^\infty \textstyle\prod_{i=1} \frac{1}{\beta} e^{-x_i/\beta}\,\mathrm{d}x_1 \ldots \mathrm{d}x_n \\
&= \int_2^\infty \cdots \int_2^\infty \textstyle\prod_{i=2} \frac{1}{\beta} e^{-x_i/\beta} \underbrace{\int_2^\infty \frac{1}{\beta} e^{-x_1/\beta}\,\mathrm{d}x_1}_{\substack{\text{Integral over the exponential pdf} \\ \int_2^\infty f(x_1)\,\mathrm{d}x_1 = 1 - F(2) \\ = 1 - (1 - e^{-\frac{1}{\beta}2}) = e^{\frac{-2}{\beta}}}} \mathrm{d}x_2 \ldots \mathrm{d}x_n \\
&= e^{\frac{-2}{\beta}} \underbrace{\int_2^\infty \cdots \int_2^\infty \prod_{i=2} \frac{1}{\beta} e^{-x_i/\beta}\,\mathrm{d}x_2 \ldots \mathrm{d}x_n}_{\substack{\text{Identical to original integral except that} \\ \text{there are only n-1 terms to be integrated}}} = \ldots \\
&= e^{-2/\beta} \cdot e^{-2/\beta} \int_2^\infty \cdots \int_2^\infty \textstyle\prod_{i=3} \frac{1}{\beta} e^{-x_i/\beta}\,\mathrm{d}x_3 \ldots \mathrm{d}x_n \\
&= (e^{-2/\beta})^n = e^{-2n/\beta}
\end{aligned}
$$

A much more simpler solution: notice that $P(X_i > 2) = 1 - F(2) = e^{-1/\beta}$. Because X_i's are independent, then also events $(X_i > 2)$ are independent event, and so

$$P(\text{All } X_i > 2) = \prod_{i=1}^{n} P(X_i > 2) = (e^{-2/\beta})^n = e^{-2n/\beta}$$

Illustration See `ExponentialSample.R` for the case where $n = 2$ and $\beta = 3$. Notice that R uses parametrization $\lambda = \frac{1}{\beta}$ for the exponential distribution, so in R, exponential pdf is $f(x \mid \lambda) = \lambda e^{-\lambda x}$.

Note: Definition 2.1 assumes that $X_1, \ldots, X_n$ are independent. In practice, we often do analysis on dependent data. For that use, we could define term "dependent random sample". Mathematical treatment of dependent sample requires more specific description of dependence structure, e.g. through spatial or temporal autocorrelation models, or explicit models for grouped data.

Dependent data are modeled by assuming that random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are vectors of appropriate length, and there are n independent realizations of them in the data. Quite often $n = 1$ and each replicate to $\boldsymbol{X}_1$, which is a rather long vector.

Example 2.2. Let $\boldsymbol{X} = \begin{bmatrix} X(u_1) & X(u_2) & X(u_3) \end{bmatrix}^T$ include random variables X at locations u_1, u_2 and u_3. Assume that $\boldsymbol{X}_1$ is normally distributed and the correlation $\rho_{ij} = corr(X(u_i), X(u_j))$ depends only on the spatial distance $s_{ij} = \|u_i - u_j\|$ between locations u_i and u_j.

The marginal means and variances of $X(u_i)$ are $E(X(u_i)) = \mu$ for all i and $Var(X(u_i)) = \sigma^2$ for all i. Consider case where $u_1 = (1, 0)$, $u_2 = (0, 0)$, $u_3 = (0, 2)$ and $\rho_{ij} = e^{-s_{ij}/2}$. Correlations and covariances related to different random variables can be seen on table 2.1.

Table 2.1: Correlations between the random variable pairs of example 2.2

Pair	s_{ij}	ρ_{ij}	$cov(X(u_i), X(u_j))$
1,2	1	0.61	$0.61\sigma^2$
1,3	$\sqrt{5}$	0.33	$0.33\sigma^2$
2,3	2	0.37	$0.37\sigma^2$

Let

$$\boldsymbol{\mu} = \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & 0.61 & 0.33 \\ 0.61 & 1 & 0.37 \\ 0.33 & 0.37 & 1 \end{bmatrix}$$

The joint pdf of $\boldsymbol{X}$ is 3 variate normal distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$

$$f(x_1 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^3}} |\boldsymbol{\Sigma}|^{-1/2} e^{\frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_1 - \boldsymbol{\mu})}$$

Note: In random sample, we assume that $X_1, \ldots, X_n$ are identically distributed. In the case of random vectors, whis is specified by saying that the marginal distributions of the elements of $\boldsymbol{X}_i$ are identical.

Note: We can also have independent replicates of random vectors, e.g. if we have grouped data where the groups are independent replicates from the process that generates the groups, and the observations within the groups are dependent. This is related to mixed-effects models.

Note: Definition 2.1 specifies sampling from an infinite population (or population model): the sampling procedure does not change the population.

The population may also be finite, like numbers in hat. In that case, the sampling can be made with replacement: whenever a number has been drawn, the value is recorded but the number is put back to the hat. This procedure, which is useful e.g. in Bootstrapping, fulfils the conditions of definition 2.1.

If we do the sampling without replacement (which often makes much sense), then the conditions of definition 2.1 are not fulfilled: each draw changes the population by removing one unit from the finite population. This leads to so called design-based inference, which is covered in the literature of sampling theory.

The difference to definition 2.1 is important especially if the sample size n is large compared to the size of population, but may be unimportant if $N >> n$. In this course, we are considering only sampling according to definition 2.1.

Random samples can be summarized to a well defined summary called statistic.

Definition 2.2. Let $\boldsymbol{X} = X_1, \ldots, X_n$ be a random sample of size n from population $f(x)$ and let $T(\boldsymbol{X})$ be a real-valued or vector-valued function, whose domain includes

the sample space of $\boldsymbol{X}$. The random variable $Y = T(\boldsymbol{X})$ is called a statistic (statistiikka, tunnusluku). The probability distribution of Y is called the sampling distribution of Y (otantajakauma).

Examples of statistics

Sample minimum	$min(\boldsymbol{X})$
Sample maximum	$max(\boldsymbol{X})$
Sample median	$median(\boldsymbol{X})$

Also other like sample mean, variance, standard deviation, quartiles etc.

Example 2.3. Let $\boldsymbol{X}$ be a random sample of size 5 from $Exponential(2)$ population. R script `statistics.R` illustrates the distributions of $min(\boldsymbol{X})$, $max(\boldsymbol{X})$, $median(\boldsymbol{X})$ and $mean(\boldsymbol{X})$ through simulation. Script repeates the sampling from Exponential distribution $M = 10000$ times and illustrates the distribution of the above mentioned sample statistic when $n = 5$.

Definition 2.3. The sample mean is the statistic defined by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Definition 2.4. The sample variance is the statistic defined by

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

- The sample standard deviation is the statistic defined by $S = \sqrt{S^2}$.
- The observed values of these random variables are denoted by $\bar{X}$, s^2 and s.

Theorem 2.1. Let $X_1, \ldots, X_n$ be any numbers. Then

a) $min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$

b) $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \underbrace{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}_{\substack{\text{Compare to} \\ Var(\boldsymbol{X})=E(\boldsymbol{X}^2)-(E(\boldsymbol{X}))^2}}$

Lemma 2.1. Let $X_1, \ldots, X_n$ be a random sample form a population and the $g(X)$ be a function such that $E(g(X_i))$ and $Var(g(X_i))$ exist. Then

$$E\left[\sum_{i=1}^{n} g(X_i)\right] = n \cdot E(g(X_i))$$

and

$$Var(\sum_{i=1}^{n} g(X_i)) = n \cdot Var(g(X_i))$$

Proof

$$E(\sum g(X_i)) = \sum E(g(X_i)) = *$$

Because X_i's are identically distributed

$$* = nE(g(X_i))$$

$$\begin{aligned} Var(\sum g(X_i)) &= E\left[\sum g(X_i) - E(\sum g(X_i))\right]^2 \\ &= E\left\{\sum \left[\underbrace{g(X_i) - E(g(X_i))}_{t_i}\right]\right\}^2 = * \end{aligned}$$

With t_i, we can get following

$$\begin{aligned} &E(\sum t_i \sum t_i) \\ &= E(\sum t_i^2 + \textstyle\sum\sum_{i \neq j} t_i t_j) \\ &= \sum E(t_i^2) + \textstyle\sum\sum_{i \neq j} E(t_i t_j) \end{aligned}$$

which we can expand back to original equation

$$* = \underbrace{\sum E(\left[g(X_i) - E(g(X_i))\right]^2)}_{\sum E(t_i^2)} + \underbrace{\textstyle\sum\sum_{i \neq j} E\left[(g(X_i) - E(g(X_i)))(g(X_j) - E(g(X_j)))\right]}_{\sum\sum_{i \neq j} E(t_i t_j)}$$

$$\begin{aligned} &= \sum Var(g(X_i)) + \textstyle\sum\sum_{i \neq j} cov(g(X_i), g(X_j)) \\ &= \underbrace{n \cdot Var(X_1)}_{\substack{\text{Because } X_i\text{'s} \\ \text{identically distributed}}} + n(n-1) \cdot \underbrace{0}_{\substack{\text{Because } X_i\text{'s} \\ \text{are independent}}} = n \cdot Var\left[g(X_1)\right] \end{aligned}$$

Note: The proof of $E(\sum g(X_i))$ did not utilize the assumption of independence. Therefore it is valid also for dependent samples. The poof of $Var(\sum g(X_i))$ used the assumption of independence. Therefore it is not valid for dependent samples.

- For dependent samples, we get

$$Var(\sum g(X_i)) = \sum_{i=1}^{n} \sum_{j=1}^{n} cov\left[g(X_i), g(X_j)\right]$$

- Proof left as an exercise (recall, that $Var(X_i) = cov(X_i, X_i)$).

$$Var(\boldsymbol{x}) = \begin{bmatrix} Var(X_1) & cov(X_1, X_2) & \dots & cov(X_1, X_n) \\ cov(X_2, X_1) & Var(X_2) & \dots & cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & \dots & Var(X_n) \end{bmatrix}$$

$Var(\sum X_i) = \sum\sum cov(X_i, X_j)$

- That is, the sum of all elements of matrix $Var(\boldsymbol{X})$.

Note: This extends the rule

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2 \cdot cov(X_1, X_2)$$

to general sum $\sum X_i$.

Theorem 2.2. Let $X_1, \dots, X_n$ be a random sample form a population with mean μ and variance $\sigma^2 < \infty$. Then

a) $E(\bar{X}) = \mu$

b) $Var(\bar{X}) = \frac{\sigma^2}{n} \quad E(\bar{X} - \mu)^2$

c) $E(S^2) = \sigma^2$

- Proof for a and b are familiar from ISI1 course. Proof of c applier the theorem 2.1 and is left as an exercise.

Note: Because $E(\bar{X}) = \mu$, the statistic $\bar{X}$ is said to be an unbiased estimator (harhaton estimaattori) of population mean μ. Also S^2 is an unbiased estimator of population variance σ^2.

Example 2.4. Illustrate these results through simulation. Intuitively, mean of an observed value of a statistic over a large number of samples should be close to expected value of the statistic in question. Therefore, (a), (b) and (c) of theorem 2.2 can be demonstrated by simulating M samples of a fixed size n and exploring how the means of the sample values of $\bar{x}$, $(\bar{x} - \mu)^2$ and s^2 behave as $M \to \infty$.

R-script `demonstratebias.R` implements this by assuming that the population has the $Uniform(0, 10)$ distribution and $n = 10$.

2.1 Distribution of sum

Theorem 2.3. (Convolution formula)

If X and Y are independent, continuous random variables with pdf's $f_X(x)$ and $f_Y(y)$, then the pdf of the sum $Z = x + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z-w)\,\mathrm{d}w$$

Proof: See Casella & Berger, p 215-216.

Illustration: See example 1.20 of `notes.pdf`

Note: To compute the complete distribution of a sum $Z = \sum_{i=1}^{n}$, where X_i's do not need to be identically distributed, but they are independent, we need to apply theorem 2.3 iteratively. E.g. to find the distribution of $X_1 + X_2 + X_3$, you may first find the distribution of $Z = X_1 + X_2$ using theorem 2.3 and thereafter use theorem 2.3 again to find the pdf of $Z + X_3$. This is not trivial in very general case, but there are easier ways to do this, e.g. when X_i's are identically distributed and independent (i.i.d).

However, recall that in any case, the moments are easy

$$E(\sum X_i) = \sum E(X_i)$$
$$Var(\sum X_i) = \sum Var(X_i) \quad \text{(If } X_i\text{'s are independent)}$$

2.2 Sampling from Normally distributed population

Theorem 2.4. Let $\boldsymbol{X}_{n\times 1}$ be random sample of size n from a $N(\mu, \sigma^2)$ distribution and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Then

a) $\bar{X}$ and S^2 are independent random variables.

b) $\bar{X}$ has $N(\mu, \frac{\sigma^2}{n})$

c) $(n-1)S^2/\sigma^2$ has chi-squared distribution with parameter $n-1$. Parameter $n-1$ is commonly called the "degrees of freedom".

Note: The chi-squared distribution with p degrees of freedom has the pdf

$$f_X(x) = \frac{1}{\underbrace{\Gamma(p/2)}_{\text{Gamma -function}}} X^{p/2-1} e^{-x/2}$$

- Gamma -function: $\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x}\,\mathrm{d}x$

Lemma 2.2. We use notation χ_p^2 for a chi-squared random variables with p degrees of freedom

- If Z is a $N(0,1)$ random variable, then $Z^2 \sim \chi_1^2$, that is, the square of a standard normal random variable is a chi-squared random variable.

- If $X_1, \ldots, X_n$ are independent and $X_i \sim \chi^2_{p_i}$ then $\sum X_i \sim \chi^2_{\sum p_i}$. That is, independent chi-squared random variables add to a chi-squared random variable, and the degrees of freedom also add.

Example 2.5. Illustrate the result of theorem 2.4 using R-script. In file `normalchi.R`, we implement the following:

Repeat M times

1. Generate random variables $X_i \sim N(0,1)$, where $n = 1, \ldots, 9$
2. Compute $Z = \sum x_i^2$ for each sample
3. Plot the histogram of the M obtained values of Z and compare to the distribution χ^2_n.

Note: If $\boldsymbol{X}_n$ is a random sample from $N(\mu, \sigma^2)$ population, then the random variable

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

where $\mu = E(\bar{X})$ and $\sigma/\sqrt{n} = Var(\bar{X})$.

- This transformation could be used to make inference on μ using the observed $\bar{x}$, if σ^2 is known.
- However, that is not the case, and therefore we use $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ which can also be written as
$$\frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}$$
where $(\bar{X} - \mu)/(\sigma/\sqrt{n}) \sim N(0,1)$ and $\sqrt{S^2/\sigma^2} \sim \chi^2_{n-1}$.

Theorem 2.5. Let $\boldsymbol{X}$ be a random sample from $N(\mu, \sigma^2)$. The random variable

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

has Student's t-distribution with parameter (n-1) "degrees of freedom".

In general, a random variable T has Student's t-distribution with p degrees of freedom ($T \sim t_p$) if T has the pdf

$$f_T(t) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})} \frac{1}{(p\pi)^{1/2}} \frac{1}{(1 + t^2/p)^{(p+1)/2}}$$

Proof: See Casella & Berger, p.223-224.

Illustration: File `normalchi.R`

- We use samples of sizes $3, 4, \ldots, 100$ and $M = 10000$.

- Plot empirical histogram of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$
- Compare to t_{n-1} distribution and to $N(0, \sigma^2)$.
- $\mu = 5$ and $\sigma^2 = 3^2$.
- Because $S^2 \to \sigma^2$ as n increases, the difference between $N(0,1)$ and t_{n-1} becomes meaningless as $n \to \infty$.

Note: t_p has only p moments. Especially $E(T_p) = 0$ if $p > 1$.

- A third important derived distribution under sampling from a normal population is the Snedecor's/Fisher's F-distribution.
- It is the theoretical distribution of the ratio of variances.

Example 2.6. Let $\boldsymbol{X}$ be a random sample from $N(\mu_X, \sigma_X^2)$ and let $\boldsymbol{Y}$ be a random sample from $N(\mu_Y, \sigma_Y^2)$. If we were interested in comparing the variability in these two populations, a quality of interest would be σ_X^2/σ_Y^2. Information about σ_X^2/σ_Y^2 is contained in s_X^2/s_Y^2; ratio of the sample variances.

The F-distribution allows this comparison by giving te distribution for random variable

$$\frac{s_X^2/s_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$$

where ratios s_X^2/σ_X^2 and s_Y^2/σ_Y^2 are independent, scaled χ^2-distributed random variables.

Definition 2.5. Let $\boldsymbol{X}_{n\times 1}$ be a random sample from a $N(\mu_X, \sigma_X^2)$ population and let $\boldsymbol{Y}_{m\times 1}$ be a random sample from an independent $N(\mu_Y, \sigma_X^2)$ population. The random variable

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$$

has the F-distribution with parameters (numerator and denominator degrees of freedom) $n-1$ and $m-1$. Equivalently, random variable F has the F-distribution with p and q degrees of freedom ($F \sim F_{p,q}$) if F has the pdf

$$f_F(x) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} \left(\frac{p}{q}\right)^{p/2} \frac{x^{p/2-1}}{\left[1+\frac{p}{q}x\right]^{(p+q)/2}}$$

Note:

a) If $X \sim F_{p,q}$, then $1/X \sim F_{q,p}$

b) If $X \sim t_q$, then $X^2 \sim \chi_1^2$ (recall that if $X \sim N(0,1)$, then $X^2 \sim \chi_1^2$)

2.3 Convergence

- What happens if $n \to \infty$.

Theorem 2.6. Markov inequality

Let X be a random variable such that $P(X \geq 0) = 1$ (i.e. X gets only positive values). Then, for every number $t > 0$

$$P(x > t) \leq \frac{E(X)}{t}$$

Proof: Consider only case where X is discrete random variable.

$$\begin{aligned} E(X) &= \textstyle\sum_X xf(x) \\ &= \textstyle\sum_{x<t} xf(x) + \sum_{x\geq t} xf(x) \end{aligned}$$

Because $X \geq 0$, both terms are positive

$$\begin{aligned} E(X) &\geq \textstyle\sum_{X\geq t} xf(x) \geq \sum_{X\geq t} tf(x) = t \underbrace{\sum_{X\geq t} f(x)}_{P(X\geq t)} \\ &= t \cdot P(X \geq t) \end{aligned}$$

$$\begin{aligned} E(X) &\geq t \cdot P(X \geq t) || : t \ \ (> 0) \\ \tfrac{E(X)}{t} &\geq P(X \geq t) \end{aligned}$$

Example 2.7. Let X be a non-negative random variable with $E(X) = 1$

$$P(X \geq 100) \leq \frac{E(X)}{100} = 0.01$$

Theorem 2.7. Chebyshev's inequality

Let X be a random variable such that $Var(X)$ exists. Then for every number $t > 0$

$$P(\big|X - E(X)\big| \geq t) \leq \frac{\sigma^2}{t^2}$$

Proof: Let $Y = (X - E(X))^2$. Now $P(Y \geq 0) = 1$ and $E(Y) = Var(X)$

$$P(\big|X - E(X)\big| \geq t) = P(Y \geq t^2) \leq \frac{E(Y)}{t^2}$$

$$P(\big|X - E(X)\big| \geq t) \leq \frac{Var(X)}{t^2}$$

Example 2.8. If $Var(X) = \sigma^2$ and we select $t = 3\sigma$

$$P(\big|X - E(X)\big| \geq 3\sigma) \leq \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9}$$

Definition 2.6. A sequence of random variables $X_1, X_2, \ldots$ converges in probability (konvergoi todennäköisyysmielessä) to a random variable X if for every $\epsilon > 0$

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) = 0$$

or equivalently

$$\lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1$$

A commonly used notation for this kind of convergence is $X_n \xrightarrow{P} X$. The sequence $X_1, \ldots, X_n$ is not usually an i.i.d. sample but e.g. a statistic based on a sample of size n. Also, X is often a fixed constant, as it is in the following theorem.

Theorem 2.8. (The weak law of large numbers (Heikko suurten lukujen laki), WLLN)

Let $X_1, X_2, \ldots$ be i.i.d. random variables with expected value $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for every $\epsilon > 0$

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$$

meaning sample mean convergences in probability to population mean ($\bar{X}_n \xrightarrow{P} \mu$). We will have the sample mans of an i.i.d. sample arbitrarily close to μ if just n is high enough.

Proof: Use Chebyshev's inequality for the complement event:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\overbrace{Var(\bar{X}_n)}^{=\sigma^2/n}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

$$P(|\bar{X}_n - \mu| < \epsilon) = 1 - P(|\bar{X}_n - \mu| \geq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

where $\lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$ and $\lim_{n\to\infty} 1 - \frac{\sigma^2}{n\epsilon^2} = 1$.

Note The property that the same sample quantity approaches a fixed constant as $n \to \infty$ is called consistency (konsistenssi).

Note WLLN was already used in R-script `demonstrateBias.R`

Note WLLN also justifies the use of a histogram from a large number of replicates as an approximation of the true density as $n \to \infty$. This is because every class of the histogram gives $P(c_1 \leq x < c_1) = F(c_2) - F(c_1)$. Weather x belongs to a class $[c_1, c_2[$ is a Bernoulli(p) distributed random variable with

$$p = E(Y) = F(c_1) - F(c_2)$$

Note Another way of specifying the law of large numbers is the strong law of strong numbers (SLLN).

$$P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1.$$

$\bar{X}$ converges to μ with probability 1, meaning almost sure convergence.

Definition 2.7. A sequence of random variables converges in distribution to random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points of x where $F_X(x)$ is a continuous cdf. Notation used for convergence in distribution is $X_n \xrightarrow{d} X$.

Example 2.9. (Maximum of uniforms)

If $X_1, X_2, \ldots$ are i.i.d. Uniform(0,1) distributed random variables and let $X_{(n)} = \max_{1\le i\le n} X_i$. Let us explore if and to where $X_{(n)}$ converges in distribution.

When $n \to \infty$, $X_{(n)} \to 1$ and as $X_{(n)} < 1$, we have for any $\epsilon > 0$

$$P(|X_{(n)} - 1| \ge \epsilon) = P(1 - X_{(n)} \ge \epsilon) = P(X_{(n)} \le 1 - \epsilon)$$

Because sample is independent

$$\begin{aligned} P(X_{(n)} < 1-\epsilon) &= P(\text{All } X_i \le 1-\epsilon) \\ &= \left[\underbrace{P(X \le 1-\epsilon}_{F_X(1-\epsilon)}\right]^n && X \sim Unif(0,1) \\ &= (1-\epsilon)^n && \lim_{n\to\infty}(1-\epsilon)^n = 0 \\ &\to X_{(n)} \xrightarrow{P} 1 \end{aligned}$$

Let us take $\epsilon = \frac{t}{n}$ to rewrite $P(X_{(n)} \le 1 - \epsilon)$ as

$$P(X_{(n)} \le 1 - \frac{t}{n}) = (1 - \frac{t}{n})^n$$

for which $\lim_{n\to\infty}(1 - \frac{t}{n})^n = e^{-t}$.

$$\begin{aligned} & X_{(n)} \le 1 - \tfrac{t}{n} \\ & X_{(n)} - 1 \le -\tfrac{t}{n} && | \cdot (-1) \\ & 1 - X_{(n)} \le \tfrac{t}{n} && | \cdot n \\ & n(1 - X_{(n)}) \le t \\ & P(n(1 - X_{(n)}) \le t) = e^{-t} \\ & P(n(1 - X_{(n)}) \ge t) = \underbrace{1 - e^{-t}}_{\text{Exponential(1) cdf}} \end{aligned}$$

Illustration: See R-script `MaxOfUniforms.R`

How fast does the convergence occur, i.e. how large does the n need to be to have $n(1 - X_{(n)}) \sim Exponential(1)$?

We generate $M = 10000$ samples of each of the sample sizes $n = 2, 3, 4, 5, 10, 20$. For each sample, find $X_{(n)}$ and compute $y = n(1 - X_(n))$. Plot the histogram if y and compare to Exponential(1) pdf.

The approximation looks quite nice already, when $n \geq 10$. But notice the behaviour in the tails. Especially $n(1 - X_{(n)})$ is bounded to interval $[0, n]$.

Theorem 2.9. (The central limit theorem (CTL) (Keskeinen raja-arvolause))

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite $E(X) = \mu$ and finite variance $Var(X) = \sigma^2 > 0$. Let $\bar{X}_n = \sum_{i=1}^n X_i$. Let $G_{Z_n}(z)$ denote the cdf of

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\bar{X}_n - \overbrace{\mu}^{E(\bar{X}_n)}}{\underbrace{\sigma/\sqrt{n}}_{sd(\bar{X}_n)}}$$

Then $\lim_{n\to\infty} G_{Z_n} z = \underbrace{\int_{-\infty}^{z} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}_{N(0,1)\text{ pdf}} \mathrm{d}y}_{N(0,1)\text{ cdf}}$ That is: Z_n has limited standard normal distribution, meaning $Z_n \xrightarrow{d} N(0, 1)$. Equivalently, $\bar{X} \xrightarrow{d} N(\mu, \sigma^2/n)$. **Notes:**

- We assume only finite μ and σ^2 and end up to normlity.
- The rate of convergence is affected by the original distribution of X_i
  - The closer the distribution of X is the Normal curve, the faster the convergence is.
  - The rate needs to be evaluated case by case.
- Provides an all purpose approximation og the distribution of sums of i.i.d. random variables.

Proof: See Casella & Berger pages 237-238.

Illustration: See R-script `unifmean.R`. Script takes $M = 10000$ samples using each of the following sample sizes $n = 1, 2, 3, 4, 5, 6$ from a Uniform(0,10) population, and demonstrates the distribution of $\bar{X}$ using a histogram and compares it to $N(\mu, \sigma^2/n)$ distribution. Expected value and variance of Uniform(0,10) population are

$$\mu = \frac{b-a}{2} = \frac{10}{2} = 5$$

$$\sigma^2 = \frac{1}{12}(b-a)^2 = \frac{(10-0)^2}{12} = \frac{100}{12} = \frac{25}{3}$$

The visual evaluation of histogram gives an impression that the approximation is good already when $n > 4$. However, we are usually interested in the behaviour in the tails,

which needs more careful evaluation, e.g. compute the 0.95^{th} quartile of the Normal distribution. Check which proportion of the simulated samples had the mean above this quartile. The proportion should be close to 0.05 if the approximation is good.

Theorem 2.10. (Slutsky) If $X_n \xrightarrow{d} X$ and if $Y_n \xrightarrow{P} a$, where a is a constant, then

a) $Y_n X_n \xrightarrow{d} aX$

b) $X_n + Y_n \xrightarrow{d} X + a$

Note Proof for this is not shown.

Example 2.10. (Normal approximation with estimated variance)

The central limit theorem said that under mild conditions

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

However, σ^2 is often unknown, but can be unbiasedly and consistently estimated by $s_n^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$. It can also be shown that (see Casella & Berger example 5.5.3 and exercise 3.32)

$$\frac{\sigma}{s_n} \xrightarrow{P} 1$$

Note

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} = \underbrace{\frac{\sigma^2}{s_n}}_{\frac{\sigma}{s_n} \xrightarrow{P} 1} \underbrace{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma^2}}_{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)} \xrightarrow{d} N(0,1)$$

So even tough the variance is estimated, $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ converges to a normal distribution. However, intuitively, the convergence should be more slow than when σ is known.

Chapter 3

Estimation

In estimation, we want to use sample $X_1, \ldots, X_n$ to make inference on an unknown population parameter θ. It means that we want to find a statistic $T(\boldsymbol{X})$ that somewhat optimally captures the information of the sample on the parameter of interest, θ.

Most commonly. θ is the population mean, but it can also be related to population variance,covariance, median, maximum ect. Very often the population parameters, especially the mean, is not a scalar number, but it's function of some known characteristics of sampled units i, and therefore θ includes parameters of that function.

Example 3.1. Simple linear regression

Consider case of simple linear regression, where X is a fixed known predictor, and Y is a random variable we want to predict with values of X. The model can be of following for

$$Y_i = \mu(X_i) + e_i$$

Where $\mu(X_i)$ is a fixed mean of Y

$$\mu_i = \mu(X_i) = \beta_0 + \beta_1 X_i$$

and e_i is the random error related to value of Y_i

$$e_i \sim N(0, \sigma^2)$$

This would mean that distribution of Y_i would be

$$Y_i \sim N(\mu(x_i), \sigma^2)$$

In this case, we are interested in estimating the the regression coefficients

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Example 3.2. Lets consider a random variable

$$Y_i = \begin{cases} 1 & \text{if tree } i \text{ is dead} \\ 0 & \text{if tree } i \text{ is alive} \end{cases}$$

Parameter of interest might be the probability of a tree being dead.

$$Y_i \sim Bernoulli(p),\ \theta = p$$

If we also know the age of the tree, and it is justified to assume that older trees are more commonly dead than younger trees, then

$$Y_i \sim Bernoulli(p_i)$$

$$p_i = \beta_0 + \beta_1 \text{Age}_i,\ \ \rightarrow \boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Because $\beta_0 + \beta_1 \text{Age}_i$ is not restricted to $[0, 1]$, a better model uses some function for $p(\text{Age}_i)$. Most commonly we use

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Age}_i$$

Which is a logit transformation $[0, 1] \rightarrow \mathbb{R}$. This provides us with logistic regression.
Note: See also example 1.2 and 1.3 in `notes.pdf`

3.1 Two principles

A sufficient statistic (tyhjentävä statistiikka) for a population parameter θ is such statistic $T(\boldsymbol{X})$, that in certain sense it captures all the information about θ contained in the sample.

3.2 The sufficient principle

If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then any inference about θ should depend on $\boldsymbol{X}$ only through the value of $T(\boldsymbol{X})$.

Definition 3.1. A statistic $T(\boldsymbol{X})$ is sufficient statistic for θ if the conditional distribution of the sample given $T(\boldsymbol{X})[\boldsymbol{X}|T(\boldsymbol{X})]$ does not depend on θ.

Example 3.3. Let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$ population, where σ^2 is known. Now the sample mean $\bar{X}$ is a sufficient statistic for μ.

That is, if we have the complete data and you know only the sample mean, we both still have all information on μ the sample includes. However, if a third person knows

only $min(\boldsymbol{X})$, $max(\boldsymbol{X})$ and $median(\boldsymbol{X})$, he still does not have much information on μ as you have when you know $\bar{X}$.

If also σ^2 is unknown, then $\boldsymbol{T}(\boldsymbol{X}) = \begin{bmatrix} \bar{X} \\ s^2 \end{bmatrix}$ is sufficient for $\boldsymbol{\theta} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$

For more formal discussion on the topic, see Casella & Berger section 6.2.

3.3 The likelihood function

Definition 3.2. Let $f(\boldsymbol{x}, \boldsymbol{\theta})$ denote the joint pdf or pmf of the sample $\boldsymbol{X} = X_1, \ldots, X_n$. Then Given that $\boldsymbol{X} = \boldsymbol{x}$ has been observed, the function of θ defined by

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

is called the likelihood function (uskottavuusfunktio).

Note We do not assume $X_1, \ldots, X_n$ to be independent.

- If $\boldsymbol{X}$ is a discrete random vector, then

$$L(\boldsymbol{\theta}|\boldsymbol{X}) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$$

- If we compare the likelihood at two parameter points $\theta = \theta_1$ and $\theta = \theta_2$, and find out that

$$P_{\theta_1}(\boldsymbol{X} = \boldsymbol{x}) = L(\theta_1|\boldsymbol{x}) > L(\theta_2|\boldsymbol{x}) = P_{\theta_2}(\boldsymbol{X} = \boldsymbol{x})$$

then the sample we observed is more likely to have occurred when $\theta = \theta_1$ than when $\theta = \theta_2$.

$\rightarrow$ θ_1 is more plausible (uskottava) value for θ than θ_2.

- If X is a continuous, real valued random variable, then for a small value ϵ

$$P_\theta(X - \epsilon < X < X + \epsilon) \approx 2\epsilon f(X, \theta) = 2\epsilon L(\theta|X)$$

therefore

$$LR = \frac{P_{\theta_1}(X - \epsilon < X < X + \epsilon)}{P_{\theta_2}(X - \epsilon < X < X + \epsilon)} \approx \frac{L(\theta_1|X)}{L(\theta_2|X)}$$

$\rightarrow$ If the $LR > 1$, then the sample was more likely to occur when $\theta = \theta_1$, and if $LR < 1$, then the sample was more likely to occur when $\theta = \theta_2$.

Example 3.4. Let X have negative binomial distribution. The Negative binomial distribution specifies the pmf for the number of successes before a fixed number of a failures (r) in a series of Bernoulli trials with success probability p. The pmf

$$P(X = x|r, p) = \binom{r + x - 1}{x} p^r (1 - p)^x, \text{ where } x = 0, 1, \ldots \text{ and } 0 \leq p \leq 1$$

Consider an experiment where $r = 3$ and it was observed that $X = 2$. The likelihood as a function of $\theta = p$

$$L(p|2) = P_p(X = 2) = \binom{4}{2} p^3(1-p)^2$$

$\rightarrow$ The likelihood is a 5$^{\text{th}}$ order polynomial with respect to p.

In general, if $X = x$ has been observed,

$$L(p|x) = P_p(X = x) = \binom{3+x-1}{2} p^3(1-p)^x$$

in which case the likelihood is a polynomial of order $x + 3$.

Example 3.5. Let $\boldsymbol{X}_{3\times 1}$ have 3-variate normal distribution so that all components have a common, unknown mean of μ and a known, common variance $\sigma^2 = 3.69$. In addition, $cov(X_1, X_2) = 2.25$ and $cov(X_1, X_3) = cov(X_2, X_3) = 0$.

This kind of situation could stem from a grouped structure of the data: X_1 and X_2 might belong to the same group (e.g. school class) and X_3 may originate from another group. That is

$$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } \mu = \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 3.69 & 2.25 & 0 \\ 2.25 & 3.69 & 0 \\ 0 & 0 & 3.69 \end{bmatrix}$$

The likelihood of this model for observation $\boldsymbol{x} = \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix}$ has been illustrated in R-script `MVNlikelihood.R`.

Note The likelihood function is not a pdf as a function of θ. Therefore we say, that θ_1 is more plausible than θ_2 (not more probable). Note also, that θ is thought to be fixed, therefore it makes no sense to specify a pdf for it. (In Bayesian statistics, θ is thought to be random.)

3.4 The likelihood principle

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two observed sample points such that $L(\theta|\boldsymbol{x})$ is proportional to $L(\theta|\boldsymbol{y})$, that is there exists a constant $C(\boldsymbol{x}, \boldsymbol{y})$ such that

$$L(\theta|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y}) L(\theta|\boldsymbol{y}) \text{ for all } \theta$$

then the conclusions drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ should be identical. **Note** $C(\boldsymbol{x}, \boldsymbol{y})$ may be different for different $(\boldsymbol{x}, \boldsymbol{y})$ pairs, but does not dependent on θ.

3.5 Point estimation

Definition 3.3. A point estimator is any function $W(X_1, \ldots, X_n)$. That is, any statistic is a point estimator.

- We now first consider the methods to find estimators and there after consider criteria to evaluate, whether the estimators are good or bad.

Note Estimator is a function of the sample $X_1, \ldots, X_n$ and estimate is the numerical value the estimator got after the sample was actually taken.

→ Estimator is a random variable, and estimate is its realized value in the sample.

- Often, there is a natural, intuitive candidate fro a point estimator, e.g. $\bar{X}$ is a intuitive choice for a point estimator of μ.

3.6 Estimating the population mean using least squares

Based on theorem 2.1, the ordinary least squares estimator of parameter μ using observed sample values $y_1, \ldots, y_n$ is found by

$$min_\theta \sum_{i=1}^{n} (y_i - \theta)^2$$

The solution is $\theta = \bar{y}$ The approach also generalized to cases where $E(Y_i)$ is a function of a fixed variable x_i. Consider a model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where Y_i is a random variable and x_i is another fixed observable variable. We can β_0 such that $E(\epsilon_i) = 0$ and

$$E(Y_i) = E(\underbrace{\beta_0 + \beta_1 x_i}_{\text{fixed}} + \underbrace{\epsilon_i}_{E(\epsilon_i)=0}) = \beta_0 + \beta_1 x_i = E(Y_i|x_i) = \mu(x_i, \boldsymbol{\beta})$$

Parameters $\boldsymbol{\beta}$ are "second order" parameters. They quantify the (assumed) linear relationship between x_i and $E(Y_i)$. Using the idea of theorem 2.1, $\boldsymbol{\beta}$ can be estimated using

$$min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

The solutions are

$$\widehat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \text{ and } \widehat{\beta_0} = \bar{y} - \beta_1 \bar{x}$$

- A more general formulation results by defining

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

  and then the model is $\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$

- The least squares estimator of $\boldsymbol{\beta}$ is found by minimizing

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta})$$

- The solution is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

- This is a general solution for the multiple regression case as well.

The least squares method is optimal and intuitive way to find the estimators for population mean. However, it does not provide tools to estimate other parameters of distribution. For that purpose, we will next present three other approaches for parameter estimation in general case:

- The method of moments
- Maximum likelihood
- Bayesian approach

3.7 The method of moments

- This method dates back to late 1800's.

Let $X_1, \ldots, X_n$ be a sample from population with pmf or pdf $f(x|\theta_1, \ldots, \theta_k)$. The method of moments (momenttimenetelmä), moment matching, finds estimates of $\theta_1, \ldots, \theta_k$ by equating the first k sample moments to the corresponding population moments.

Define

$$\begin{aligned} m_1 &= \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i & \mu_1' &= E(X^1) \\ m_2 &= \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^2 & \mu_2' &= E(X^2) \\ &\vdots & &\vdots \\ m_k &= \tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^k & \mu_k' &= E(X^k) \end{aligned}$$

- The population moments will typically be functions of $\boldsymbol{\theta}$, $\mu_j'(\theta_1, \ldots, \theta_k)$

- The moment estimator $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ of $(\theta_1, \ldots, \theta_k)$ is found by solving the following system of equations for $(\theta_1, \ldots, \theta_k)$

$$\begin{cases} m_1 = \mu_1'(\boldsymbol{\theta}) \\ m_2 = \mu_2'(\boldsymbol{\theta}) \\ \vdots \\ m_k = \mu_k'(\boldsymbol{\theta}) \end{cases}$$

Example 3.6. Let $X_1, \ldots, X_n$ be i.i.d. with $X_i \sim N(\mu, \sigma^2)$, and both μ and σ^2 are unknown.

$$\begin{array}{ll} m_1 = \bar{X} & m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \\ \mu_1' = \mu & \mu_2' = E(X^2) = \mu^2 + \sigma^2 \end{array}$$

Recall $Var(X) = E(X^2) - (E(X))^2 \to E(X^2) = Var(X) + (E(X))^2 = \sigma^2 + \mu^2$
Following system of equation needs to be solved for μ and σ^2

$$\begin{cases} \bar{X} = \mu \\ \frac{1}{n}\sum X^2 = \mu^2 + \sigma^2 \end{cases}$$

$$\begin{cases} \bar{X} - \mu = 0 \\ \frac{1}{n}\sum X^2 - \mu^2 - \sigma^2 = 0 \end{cases}$$

We get

$$\tilde{\mu} = \bar{X}$$

$$\tilde{\sigma}^2 = \underbrace{\frac{1}{n}\sum (X_i - \bar{X})}_{\text{Biased estimator}} \neq \underbrace{\frac{1}{n-1}\sum (X_i - \bar{X})^2}_{\text{Unbiased estimator}}$$

But the bias vanishes as $n \to \infty$.

This example was easy, since the system of equations we needed to solve was linear.

Example 3.7. A numerical solution Let $X_1, \ldots, X_n$ be i.i.d. from the Weibull(a,b) distribution, which has the pdf

$$f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-(x/b)^a} \text{ where } 0 < x < \infty \text{ and } a, b > 0$$

where a is called te shape parameter and b is called the scale parameter.

The first two moments are

$$E(X) = \int_0^\infty f(x) x \, \mathrm{d}x = \ldots = b\Gamma\left(1 + \frac{1}{a}\right)$$

$$E(X^2) = \int_0^\infty f(x) x^2 \, \mathrm{d}x = \ldots = b^2\Gamma\left(1 + \frac{2}{a}\right)$$

where $\Gamma(z) = \int_0^\infty z^{u-1}e^{-z}\,\mathrm{d}z$ cannot be written in closed form, but can be well approximated using function gamma in R.

The moment-matching system of equations is

$$\begin{cases} \frac{1}{n}\sum x_i - b\Gamma\left(1+\frac{1}{a}\right) = 0 \\ \frac{1}{n}\sum X^2 - b^2\Gamma\left(1+\frac{2}{a}\right) = 0 \end{cases}$$

This is nonlinear with respect to a!

We solve it for a and b using the Gauss-Newton method, which is a multivariate generalization of the Newton-Raphson method. Example in R-script `WeibullMoment.R` generates first a random sample of size 30 from Weibull(2,10) population. Using the generalized sample, it finds the moment estimates of a, b numerically using R-function NRnum.

The observed sample moments are

$$m_1 = \bar{X} = 9.604,\ m_2 = \frac{1}{n}\sum x^2 = 113.22$$

We solve

$$\begin{cases} 9.604 - b\Gamma\left(1+\frac{1}{a}\right) = 0 \\ 113.22 - b^2\Gamma\left(1+\frac{2}{a}\right) = 0 \end{cases}$$

for (a, b) ($\boldsymbol{\theta} = \begin{bmatrix} a & b \end{bmatrix}^T$). We use $a' = 1$ and $b' = 9.60$ as starting values, and we get

$$\tilde{a} = 2.21\ \tilde{b} = 10.84$$

3.8 Newton-Raphson

Consider function $f(a) = 0$, which needs to be solved for a.

1. Start with some guess a', compute $f(a')$ and $f'(a')$ to approximate $f(a)$ by a linear function. Find a such that the approximator = 0.

2. Set the solution as a' and repeat step 1.

If the initial guess a' was good enough and f is continuous and differentiable, we will find the approximate solution within few iteration.

Gauss-Newton generalized this to vector valued $\boldsymbol{a}$ and $\boldsymbol{x}$. The algorithm has been implemented in R-function NRnum of package lmfor. The function evaluates also the required derivatives $\frac{\mathrm{d}f(\boldsymbol{x})}{\mathrm{d}\theta_i}$ numerically.

3.9 Maximum likelihood

Definition 3.2 defined the likelihood as

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

where

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix}$$

Note For i.i.d. data $X_1, \ldots, X_n$ from population $f(x|\theta)$, the likelihood simplifies to

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} f(\boldsymbol{x}|\theta)$$

Definition 3.4. For each sample point $\boldsymbol{x}$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{x})$ attains its maximum value as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ fixed. A maximum likelihood estimator of $\boldsymbol{\theta}$ based on sample $\boldsymbol{x}$ is this function $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$. The maximum likelihood estimate (ML estimate) is the numerical value of that function in the data point we have.

Note We use abbreviation MLE for both ML estimator and ML estimate.

- MLE is the parameter point for which the observed sample is most likely.
- The MLE also fulfills some optimality criteria, which will be discussed later.
- How to find MLE?
  - If $L(\boldsymbol{\theta}|\boldsymbol{x})$ is differentiable with respect to all θ_i, then the candidates for MLE are the values that solve

$$\frac{\mathrm{d}}{\mathrm{d}\theta_i} L(\boldsymbol{\theta}|\boldsymbol{x}) = 0,\ i = 1, \ldots, k$$

  - These together with the boundaries of the parameter space provide candidates of MLE.
  - Of the candidates the one that gives the maximum value of $L(\boldsymbol{\theta}|\boldsymbol{x})$ is the MLE.

Example 3.8. Example ML1

Let $X_1, \ldots, X_n$ be i.i.d. from $N(\theta, 1)$, meaning only the population mean θ is unknown.

$$\begin{aligned} L(\theta, \boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}(x_i-\theta)^2} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \\ \frac{\mathrm{d}}{\mathrm{d}\theta} L(\theta|\boldsymbol{x}) &= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} (-\tfrac{1}{2}) \sum_{i=1}^{n}(-2(x_i-\theta)) \\ &= L(\theta|\boldsymbol{x}) \sum_{i=1}^{n}(x_i-\theta) \end{aligned}$$

$\frac{\mathrm{d}}{\mathrm{d}\theta} L(\theta|\boldsymbol{x}) = 0$ if

$$\begin{array}{lll} \text{either} & L(\theta|\boldsymbol{x}) = 0 & \text{A} \\ \text{or} & \sum_{i=1}^{n}(x_i-\theta) = 0 & \text{B} \end{array}$$

But if $L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) > 0$

$\rightarrow$ There is no solution for A.

B gives

$$\begin{aligned} &\sum (x_i - \theta) = 0 \\ &\sum x_i - \sum \theta = 0 \\ &\sum x_i - n\theta = 0 <=> \hat{\theta} = \frac{\sum x_i}{n} = \bar{x} \end{aligned}$$

So the sample mean, which is the least squares and moment estimator for the population mean is also the ML estimator for the population mean.

To confirm that $\bar{X}$ is the MLE, we still ned to make sure that

1) $\bar{x}$ is the maximum, not minimum

2) The boundaries of the support of θ do not provide higher value of θ

We can conclude that

1) This is maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} L(\theta|\boldsymbol{x})_{\theta=\bar{x}} < 0$$

2) We can notice that

$$\lim_{\theta\to\infty} L(\theta|\boldsymbol{x}) = \lim_{\theta\to-\infty} L(\theta|\boldsymbol{x}) = 0$$

and $L(\bar{x}|\boldsymbol{x}) > 0$.

- In most cases, the solution of the estimation problem becomes easier if the natural logarithm of the likelihood is used instead of likelihood.

- The maximum of $L(\boldsymbol{\theta}|\boldsymbol{x})$ and $ln(L(\boldsymbol{\theta}|\boldsymbol{x}))$ coincide (are at the same point $\boldsymbol{\theta}$) because ln-function is increasing on $(0, \infty)$.

Example 3.9. Example ML2 (Bernoulli MLE)

Let $X_1, \ldots, X_n$ $Bernoulli(p)$, i.i.d.

$$L(p|\boldsymbol{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = (*)$$

Recall that

$$\text{PMF} : f(x_i|p) = \begin{cases} p & \text{if } x_i = 1 \text{ (success)} \\ 1-p & \text{if } x_i = 0 \text{ (failure)} \end{cases}$$

$$(*) = p^{\sum x_i}(1-p)^{\sum(1-x_i)} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

The log-likelihood is

$$ln(L(p|\boldsymbol{x})) = ln(p^{\sum x_i}(1-p)^{n-\sum x_i}) = \sum x_i ln(p) + (n - \sum x_i) ln(1-p)$$

$$\frac{\mathrm{d}}{\mathrm{d}p} ln(L) = \sum x_i \frac{1}{p} + (n - \sum x_i)\frac{-1}{1-p} = \ldots = \frac{\sum x_i - np}{p(1-p)}$$

$$\begin{aligned} &\tfrac{\mathrm{d}}{\mathrm{d}p} ln(L) = 0 \\ &\tfrac{\sum x_i - np}{p(1-p)} = 0 \qquad || \cdot p(1-p) \neq 0 \\ &\sum x_i - np = 0 \\ &p = \tfrac{\sum x_i}{n} = \bar{x} \end{aligned}$$

Check that this is a global maximum and $ln(L(\bar{x})) > ln(L(0))$ and $ln(L(\bar{x})) > ln(L(1))$ to confirm that the solution is also the MLE.

Example 3.10. Example ML3 (MLE in dependent sample)

Recall the grouped data example 3.5, where $Y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume that Y has been observed to be $\boldsymbol{y} = \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix}$ $(\bar{y} = 5)$. Assume (unrealistically) that $\boldsymbol{\Sigma}$ is known to be

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3.69 & 2.25 & 0 \\ 2.25 & 3.69 & 0 \\ 0 & 0 & 3.69 \end{bmatrix}$$

We want to estimate parameter θ of the unknown vector $\boldsymbol{\mu} = \begin{bmatrix} \theta \\ \theta \\ \theta \end{bmatrix}$ (i.e. we assume that all three components of $\boldsymbol{\mu}$ have the common mean of θ). The likelihood is directly

the pdf of the multivariate normal distribution and the log-likelihood is the natural logarithm of that.

R-script `MVNMLE.R` does the estimation. Define the likelihood as we did previously in the R-script `MVNLikelihood.R`. Find the value of θ that maximizes the likelihood.

1. Simple solution (for demonstration). Compute $L(\theta)$ for values $\theta = (0, 0.001, 0.002, \ldots, 10)$ and select the one that provides the highest value. This direct search or grid search algorithm have the estimate $\hat{\theta} = 4.831 < \bar{y} = 5$, because the independent $Y_3 = 4$ gets more weight in estimation than the two mutually dependent elements $\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$. Therefore $\hat{\theta}$ is shrunken towards $Y_1 = 4$ compared to $\bar{y} = 5$.

2. We solved this also using function mle of R-package stats4.

1. Define a R-function that evaluates the negative log-likelihood $-ln(L)$; call it e.g. nll.
2. Specify some good starting values of the parameters in a named list.
3. Make a call to mle: `mle(nll, start=list(theta=5))` $\rightarrow$ gives $\hat{\theta} = 4.831081$.

Note To get more stable and faster solution, you could also provide analytical derivative of the function you used

Note mle gives only the candidate of MLE, so you are responsible for checking that it really is the MLE.

- If $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is multidimensional, then we need to maximize the $ln(L)$ with respect to all components θ_i simultaneously.
- If the log-likelihood is differentiable with respect to all θ_i, we hust need to find the practical derivatives of $ln(L)$ (or L) with respect to all θ_i and set them to 0

  $\rightarrow$ We get system of k equations.

Example 3.11. Example ML4 Let $X_1, X_2, \ldots, X_n$ be i.i.d. form $N(\theta, \sigma^2)$ population, and let both θ and σ^2 be unknown. Now

$$L(\theta, \sigma^2|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)}^{n/2} e^{\frac{1}{2}\sum(x_i-\theta)^2/\sigma^2}$$

$$ln(L(\theta, \sigma^2|\boldsymbol{x})) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2}\frac{\sum(x_i-\theta)^2}{\sigma^2}$$

The ML equations becomes

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}\theta}ln(L(\theta, \sigma^2|\boldsymbol{x})) = \frac{1}{\sigma^2}\sum(x_i - \theta) \\ \frac{\mathrm{d}}{\mathrm{d}\sigma^2}ln(L(\theta, \sigma^2|\boldsymbol{x})) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i - \theta)^2 \end{cases}$$

Setting these to 0 and solving for $\boldsymbol{\theta} = \begin{bmatrix} \theta \\ \sigma^2 \end{bmatrix}$ gives

$$\hat{\theta} = \bar{x} \text{ and } \sigma^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$$

These are the candidates for the MLE, but it can be shown that they are also the MLE's. See Casella & Berger p.322 for more formal argumentation.

The so called invariance property of MLE's is highly useful in making restrictions to the parameters. Suppose that the distribution is indexed using θ, but we want to estimated a function of it, $\tau(\theta)$.

Theorem 3.1. If $\hat{\theta}$ is the MLE of θ, and $\tau(\theta)$ is any function of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.

Proof: See Casella & Berger, p.320.

Example 3.12. Example ML5

Recall the Weibull population from the example 3.7 for which we estimated the parameters a and b using the method of moments. Because a and b are both positive, it is convenient to reparametrize the pdf in terms of $ln(a)$ and $ln(b)$.

Weibull MLE: $\boldsymbol{x} = (14.7, \ldots, 17.56)$

The MLE's of $\boldsymbol{\theta} = (a, b)$ is

Table 3.1: The MLE's of $\boldsymbol{\theta} = (a, b)$.

	In the original scale	$exp(\widehat{ln(\theta)})$
$\hat{a}$	2.24	2.24
$\hat{b}$	10.88	10.88

In R-script `WeibullMLEInvariance.R` $A = ln(a)$ and $B = ln(b)$.

$$\widehat{ln(a)} = \hat{A} = 0.81 \quad exp(\hat{A}) = 2.243$$
$$\widehat{ln(b)} = \hat{B} = 2.89 \quad exp(\hat{B}) = 10.876$$

Example 3.13. Example ML6 (MLE in dependent data)

Recall exercise 2.2 where we had spatial, multivariate normal data. Consider observations of Y_i in the following 2-dimensional grid, $i = 1, \ldots, 25$. $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_25 \end{bmatrix}$ is generated by following model

$Y \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{25} \end{bmatrix}$, $\mu_i = 0$ for all i, and $Var(Y_i) = \sigma_i^2 = 2^2$ for all i

$$corr(Y_i, Y_j) = exp^{-\phi s_i j}, \text{ where } \phi = \frac{1}{2}$$

and $s_i j$ is the euclidian distance between the location i and j, $s_i j = ||u_i - u_j||$.

Let us first generate the data and illustrate is graphically; see R-script `MLESpatial.R`. Next, given the simulated data, let us estimate the parameters μ (common mean for all i), σ^2 (common variance for all i) and ϕ (the parameter of correlation function).

Because $\sigma^2 > 0$ and $\phi > 0$, we parametrize the likelihood function in terms of $ln(\sigma^2)$ and $ln(\phi)$, and define the negative log-likelihood in R-function nll. The script provides following resulting estimates.

$\hat{\mu} = -1.87$	True: 0
$\widehat{ln(\phi)} = -2.03 \rightarrow \hat{\phi} = 0.13$	True: 0.5
$\widehat{ln(\sigma^2)} = 2.35 \rightarrow \hat{\sigma}^2 = 10.47$	True: 4

3.10 Bayes estimators

In the Bayesian approach, the fundamental difference to the previous methods is to think the parameters $\boldsymbol{\theta}$ as random variable (vector), not as fixed quantity.

- The variability of $\boldsymbol{\theta}$ in the population is described by some distribution called the prior distribution.
  - It is the belief of the experimenter about $\boldsymbol{\theta}$, and is formulated before the data are seen.
  - This prior is updated with the data $\boldsymbol{x}$ to get the posterior distribution of $\boldsymbol{\theta}$
- There are two main reasons to use Bayesian approach:
  1. We may be willing to combine the information we already have on θ with the new information we get from the data.
  2. We do not actually believe on the prior, but we use the Bayesian approach because it provides tools for estimation in the situation where e.g. ML is very tricky.
     - Especially, the Bayesian computational methods provide estimation tools for such cases.
- Let $\pi(\boldsymbol{\theta})$ be the prior pdf and let $f(\boldsymbol{x}|\boldsymbol{\theta})$ be the sampling pdf of the data $\boldsymbol{x}$.

– The posterior distribution is the conditional distribution of $\boldsymbol{\theta}$ given the sample $\boldsymbol{x}$

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})}$$

where $m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\theta$

Note θ means both the random variable and it's realized value here.

Example 3.14. Bayes 1 (Binomial Bayes)

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p). Then

$$\sum X_i \sim Binomial(n, p)$$

Assume that prior of the p is the $Beta(\alpha, \beta)$ distribution with some values of α and β, which are set before the data are seen. The joint pdf of (y, p) is

$$f(y,p) = \underbrace{f(y|p)}_{Binomial}\ \underbrace{\pi(p)}_{Beta} = \left[\binom{n}{y} p^y (1-p)^{n-p}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]$$

$$= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

The marginal pdf of Y is

$$f(y) = \int_0^1 f(y,p)\,\mathrm{d}p = \binom{n}{p}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

which is called the beta-binomial pdf. The posterior distribution is

$$f(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

which is the Beta($y+\alpha$,$n-y+\beta$).

A natural point estimator of p is the expected value $E(p|\boldsymbol{x})$ of posterior distribution:

$$\hat{p}_\beta = \frac{y-\alpha}{\alpha+\beta+n}$$

It is a weighted mean of the prior mean $\frac{\alpha}{\alpha+\beta}$ and the ML-estimator $\frac{y}{n}$:

$$\frac{n}{\alpha+\beta+n}\frac{y}{n} + \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} = w\frac{y}{n} + (1-w)\frac{\alpha}{\alpha+\beta},\ \text{where } w = \frac{n}{\alpha+\beta+n} \in [0,1]$$

Note The Beta distribution is the conjugate prior for the Binomial distribution, because the posterior distribution belongs to the same family. The use of conjugate priors leads to mathematically beautiful results, but are not otherwise justified by any theory.

- In the conditional distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\theta)}{m(\boldsymbol{x})}$$

the denominator $m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\theta$ is hard to compute in general case.

  - The modern computational methods, so called Markov chain Monte Carlo (MCMC) methods allow simulation of observations from $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ without the need to evaluate $m(\boldsymbol{x})$.

- The most simple MCMC method is the Metropolis algorithm. (Denote $f(x|\theta)\pi(\theta)$ by $a(\theta)$)

  1. Choose an arbitrary starting point θ_0 form Θ (the sample space of θ). Choose an arbitrary symmetric pdf (the proposal density) $g(\theta_{t+1}|\theta_t)$, which suggests the new value of θ (θ_{t+1}) given the previous value θ_t. Usually we use $N(\theta_t|\sigma^2)$ as the proposal density.

  2. At each iteration t

     2.1 Generate a candidate of θ: θ' using $g(\theta_{t+1}|\theta_t)$.

     2.2 Compute the acceptance ratio

     $$\alpha = \frac{a(\theta')}{a(\theta_t)}$$

     2.3 If $\alpha < 1$, then accept θ' as θ_{t+1} with probability $alpha$.

     Repeat until a desired sample size has been obtained.

Example 3.15. Bayes 2

Consider a sample $\boldsymbol{x}' = \begin{bmatrix} 9.76 & 9.08 & 16.23 & 10.28 & 10.51 \end{bmatrix}$. The aim is to estimate the population mean θ. We use restrictive prior for it:

$$\pi(\theta) \sim Uniform(7, 13)$$

In R-script `Bayes.R` there is function postProfile that implements function $a(\theta)$. $N(\theta_t, 1)$ is used as $g(\theta_{t+1}|\theta_t)$. We generate a sample size of $n = 10000$ and use the posterior sample mean as the point estimate of θ.

Notes

- MCMC methods require only that $\pi(\theta)f(\boldsymbol{x}|\theta)$ can be evaluated.

- The sample generated from the posterior is not independent: the sequence $\theta_1, \ldots, \theta_t, \ldots, \theta_N$ has autocorrelation so that successive values are dependent.

- Commonly used MCMC methods are the Metropolis-Hastings-algorithm and the Gibbs sampler.

3.11 Methods for evaluating estimators

Definition 3.5. The mean squared error (MSE) of an estimator W of parameter θ is the function of θ defined by

$$E_\theta(W-\theta)^2$$

- In general, any increasing function of the aboslute distancec $|W-\theta|$ is a reasonable measure for the quality of W (e.g. mean absolute error $E_\theta(W-\theta)$), but MSE is 1) mathematically factorable and 2) has the interpretation

$$E_\theta(W-\theta)^2 = Var_\theta(W) + (E_\theta(W-\theta))^2$$

$$MSE = variance + bias^2$$

Definition 3.6. The bias of the point estimator W is the difference

$$Bias_\theta(W) = E_\theta(W) - \theta$$

- Estimator W is said to be unbiased if $Bias_\theta(W) = 0$ for all θ
- MSE as a criterion for an estimator leads to estimators that have small bias and variance.

Example 3.16. (Normal MSE)

Let $X_1, \ldots, X_n$ be i.i.d. N(μ, σ^2). The statistics $\bar{x}$ and s^2 are unbiased estimators of μ and σ^2 since

$$E(\bar{x}) - \mu = \mu - \mu = 0$$

$$E(s^2) = \sigma^2 \to E(s^2) - \sigma^2 = 0$$

Note These results are true even if the population is not normally distributed.

The MSE of these estimators are

$$(*)\ MSE(\bar{X}) = E(\bar{X}-\mu)^2 = Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$(**)\ MSE(s^2) = E(s^2-\sigma^2)^2 = Var(s^2) = \ldots = \frac{2\sigma^4}{n-1}$$

Note that (*) did not require normality, but (**) did. An alternative estimator for σ^2 is the MLE/moment estimator

$$\hat{\sigma}^2 = \frac{1}{n}(X_i - \bar{X})^2 = \frac{n-1}{n}s^2$$

$$E(\hat{\sigma}^2) = E\left(\frac{n-1}{n}s^2\right) = \frac{n-1}{n}E(s^2) = \frac{n-1}{n}\sigma^2$$

Now $E(\hat{\sigma}^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 \neq 0$, meaning $\hat{\sigma}^2$ is a unbiased estimator of σ^2.

The variance of $\hat{\sigma}^2$ is

$$Var(\hat{\sigma}^2) = Var(\frac{n-1}{n}s^2) = \left(\frac{n-1}{n}\right)^2 Var(s^2) = \frac{2(n-1)}{n^2}\sigma^4$$

The MSE of $\hat{\sigma}^2$ is

$$E(\hat{\sigma}^2 - \sigma^2)^2 = variance + bias^2 = \frac{2(n-1)}{n^2}\sigma^4 + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \frac{2n-1}{n^2}\sigma^2$$

$$\underbrace{E(\hat{\sigma}^2 - \sigma^2)^2}_{MSE(\hat{\sigma}^2)} = \frac{2n-1}{n^2}\sigma^2 < \frac{2\sigma^4}{n-1} = \underbrace{E(s^2 - \sigma^2)^2}_{MSE(s^2)}$$

$$\frac{2n-1}{n^2} = \frac{2-\frac{1}{n}}{n} < \frac{2}{n-1}$$

$\rightarrow$ Allowing bias in the estimator of σ^2 leads to gain in terms of MSE.

- The bias is negative $\rightarrow$ ML underestimates the variance on average.
- In general, MSE is justified choice to evaluate estimators of location parameters, but not the scale parameters.

It is often impossible to find a single estimator that is best in terms of MSE. In addition, the bias of an estimator is often seen as a severe problem, or inconvenience. Therefore, we often restrict consideration in the class of unbiased estimatros. In that class

$$E_\theta(E - \theta)^2 = Var_\theta(W)$$

Therefore, form among unbiased estimators, we should select the one with the lowest variance.

Definition 3.7. An estimator W^* is the best unbiased estimator (BUE) of $\tau(\theta)$, if it satisfies $E_\theta(W^*) = \tau(\theta)$ for all θ and for any other estimator W with $E_\theta(W) = \tau(\theta)$ we have

$$Var_\theta(\tau^*) \leq Var_\theta(W) \text{ for all } \theta$$

W^* is also called uniform minimum variance unbiased estimator of $\tau(\theta)$ (UMVUE).

Example 3.17. Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ) and let $\bar{X}$ and s^2 be the sample mean and sample variance respectively. Recall that in Poisson(λ) population $E(X) = Var(X) = \lambda$. Therefore, both $\bar{X}$ and s^2 are unbiased estimators of λ. Which one to select?

$$Var_\lambda(\bar{X}) = \frac{\lambda}{n}$$

$$Var_\lambda(s^2) = \ldots \text{ long calculation}$$

Conclution however will be, that $Var_\lambda(\bar{X}) \leq Var_\lambda(s^2)$ for all λ. So, from among these two candidates, $\bar{X}$ should be selected. But is $\bar{X}$ really the best?

We could construct infintely many unbiased estimators as weighted mean of $\bar{X}$ and s^2:

$$W_W = w\bar{X} + (1-w)s^2 \text{ where } w \in [0,1]$$

Could some selection of w lead to smaller variance than $Var(\bar{X}) = \frac{\lambda}{n}$? Or is there some other estimator that has $Var(w) = \frac{\lambda}{n}$?

Theorem 3.2. (Cramer-Rao Inequality)

Let $X_1, \ldots, X_n$ be a sample with pdf $f(\boldsymbol{X}|\theta)$ and let $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying

$$(*)\ \frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta(W(\boldsymbol{X})) = \int_{\mathcal{X}} \frac{\mathrm{d}}{\mathrm{d}\theta} W(\boldsymbol{x}) f(\boldsymbol{x}|\theta)\, \mathrm{d}\boldsymbol{x}$$

and

$$Var_\theta(\boldsymbol{X}) < \infty$$

, then

$$Var_\theta(W(\boldsymbol{X})) \geq \frac{\left[\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta(W(\boldsymbol{X}))\right]^2}{E\left\{\left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{x}|\theta))\right]^2\right\}}$$

Note If $W(\boldsymbol{X})$ is unbiased for θ, then

$$\left[\frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta(W(\boldsymbol{X}))\right]^2 = 1 \text{ and } Var_\theta(W(\boldsymbol{X})) \geq \frac{1}{E\left\{\left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{x}|\theta))\right]^2\right\}}$$

Note Integral in condition (*) can be replaced by sum to apply the theorem to discrete random variables.

Proof: Based on Cauchy-Schwarz inequality

$$cov(X,Y)^2 \leq Var(X)Var(Y)\ (1)$$

where $X = W(\boldsymbol{X})$ and $Y = \frac{\mathrm{d}}{\mathrm{d}\theta} ln[f(\boldsymbol{X}|\theta)]$.

First note that

$$(2)\ \frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta(W(\boldsymbol{X})) = \int_{\mathcal{X}} \frac{\mathrm{d}}{\mathrm{d}\theta} W(\boldsymbol{X}) f(\boldsymbol{X}|\theta)\, \mathrm{d}\boldsymbol{x} \text{ condition (*)}$$

We multiply this by $1 = \frac{f(\boldsymbol{X}|\theta)}{f(\boldsymbol{X}|\theta)}$ and utilize $\frac{\mathrm{d}ln(f(x))}{\mathrm{d}x} = \frac{1}{f(x)} f'(x) = \frac{\frac{\mathrm{d}}{\mathrm{d}x} f(x)}{f(x)}$ to get

$$= \int_{\mathcal{X}} W(\boldsymbol{X}) \frac{\frac{\mathrm{d}}{\mathrm{d}\theta} f(\boldsymbol{X}|\theta)}{f(\boldsymbol{X}|\theta)} f(\boldsymbol{X}|\theta)\, \mathrm{d}x$$

Which we can write as an expected value of transformation

$$= E_\theta \left[W_\theta(\boldsymbol{X}) \frac{\mathrm{d}ln(f(\boldsymbol{X}|\theta))}{\mathrm{d}\boldsymbol{X}} \right] \quad (3)$$

If we consider here $X = W_\theta(\boldsymbol{X})$ and $Y = \frac{\mathrm{d}ln(f(\boldsymbol{X}|\theta))}{\mathrm{d}\boldsymbol{X}}$, we have

$$E(Y) = E\left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{X}|\theta) \right]$$

and if we assume that $X = W_\theta(\boldsymbol{X}) = 1$, and utilize the equation (3), we get

$$= E\left[1 \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{X}|\theta)) \right]$$

we can write this equation then in the form of equation (2)

$$= \frac{\mathrm{d}}{\mathrm{d}\theta} E_\theta(X) = \frac{\mathrm{d}}{\mathrm{d}\theta} 1 = 0 \ (4)$$

Now, $E_\theta \left[W_\theta(\boldsymbol{X}) \frac{\mathrm{d}ln(f(\boldsymbol{X}|\theta))}{\mathrm{d}\boldsymbol{X}} \right] = cov(W(\boldsymbol{X}), \frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{X}|\theta))$. Also, because of (4), we also have

$$(5)\ Var_\theta \left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{X}|\theta)) \right] = \underbrace{E\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(X)) \right)^2 \right]}_{Var(Y)=E(Y^2)-\underbrace{\left[E(Y)\right]^2}_{=0 \text{ because of (4)}}=E(Y^2)}$$

Using (5) and (2) in (1) gives

$$Var_\theta(W(\boldsymbol{X})) \geq \frac{\left[cov\left(W(\boldsymbol{X}), \frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{X}|\theta)) \right) \right]^2}{Var\left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(\boldsymbol{X}|\theta)) \right]}$$

$$= \frac{\left[\frac{\mathrm{d}}{\mathrm{d}\theta} E(W(\boldsymbol{X})) \right]^2}{E\left\{ \left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(X)) \right]^2 \right\}}$$

Corollary 3.1. (Cramer-Rao i.i.d. case)

$$Var(W(\boldsymbol{X})) \geq \frac{\left[\frac{\mathrm{d}}{\mathrm{d}\theta} E(W(\boldsymbol{X})) \right]^2}{n \cdot E_\theta \left\{ \left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(X_i|\theta)) \right]^2 \right\}}$$

Note The quantity $E_\theta \left\{ \left[\frac{\mathrm{d}}{\mathrm{d}\theta} ln(f(X_i|\theta)) \right]^2 \right\}$ is called the information number or Fisher's information.

Note Any candidate estimator satisfying $E_\theta[W(\boldsymbol{X})] = \tau(\theta)$ and attaining the Cramer-Rao bound is the beast unbiased estimator of $\tau(\theta)$.