Introduction to statistical inference 2

Lauri Mehtätalo
University of Eastern Finland
School of Computing

May 21, 2018

II

Contents

1 Recap from “Introduction to statistical inference 1” 1

1.1 Random variable 1
1.2 Transformations of random variable 2
1.3 Expected values 3
1.4 Variance 3
1.5 Bivariate random variables 3
1.6 Independence 4
1.7 Covariance 4
1.8 Random vectors 5
1.9 Computing using expected values and variances 7

2 Random samples 9

2.1 Distribution of sum 14
2.2 Sampling from Normally distributed population 15
2.3 Convergence 18

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

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

and correspondingly for Y.

- The conditional pmf / pdf are both defined as

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

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

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

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

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

  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

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

  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

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

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

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

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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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

$$\begin{aligned} * &= \underbrace{\sum E(\left[g(X_i) - E(g(X_i))\right]^2)}_{\sum E(t_i^2)} + \underbrace{\textstyle\sum\sum_{i \neq j} E\left[(g(X_i) - E(g(X_i)))(g(X_j) - E(g(X_j)))\right]}_{\sum\sum_{i \neq j} E(t_i t_j)} \\ &= \sum Var(g(X_i)) + \textstyle\sum\sum_{i \neq j} cov(g(X_i), g(X_j)) \\ &= \underbrace{n \cdot Var(X_1)}_{\text{Because } X_i\text{'s identically distributed}} + n(n-1) \cdot \underbrace{0}_{\text{Because } X_i\text{'s are independent}} = n \cdot Var\left[g(X_1)\right] \end{aligned}$$

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^2_p for a chi-squared random variables with p degrees of freedom

- If Z is a $N(0,1)$ random variable, then $Z^2 \sim \chi^2_1$, that is, the square of a standard normal random variable is a chi-squared random variable.

- If $X_1, \dots, X_n$ 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 $n = 1, \dots, 9$
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 $3, 4, \dots, 100$ 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 $t > 0$

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

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

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

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

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

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

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 $\epsilon > 0$

$$P(|X_{(n)} - 1| \geq \epsilon) = P(1 - X_{(n)} \geq \epsilon) = P(X_{(n)} \leq 1 - \epsilon)$$

Because sample is independent

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

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

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

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

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

Illustration: See 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 $[0, n]$.