Introduction to statistical inference 2

Lauri Mehtätalo
University of Eastern Finland
School of Computing

April 22, 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

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

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

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

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, fullfils 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).