Introduction to statistical inference 2

Lauri Mehtätalo
University of Eastern Finland
School of Computing

April 19, 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

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

- 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

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

  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

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

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.

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

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

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

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

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

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

  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

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

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