Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fourth lecture, November 4, 2024

Expected value and covariance

Example

If a random variable X takes finitely many values $x_1, \ldots, x_n$ with equal probability, it is natural to define the *average* of X as the arithmetic average $\frac{1}{n} \sum_{i=1}^{n} x_i$.

More generally, if X takes the value x_i with probability p_i, then it is natural to define the average of X as the weighted average $\sum_{i=1}^{n} p_i x_i$, i.e., values x_i which are more likely to be realized are assigned a larger weight and *vice versa* for values x_i which are less likely to occur.

The *expected value* of a random variable is used to formalize the notion of "mean" or "average" of a real-valued random variable X.

Definition (Expected value of a discrete, real-valued RV)

Let X be a discrete, real-valued random variable with target space $E \subset \mathbb{R}$ and PMF p_X. The expected value (also called mean) of X is

$$\mathbb{E}[X] = \sum_{x \in E} x \, p_X(x). \tag{1}$$

Definition (Expected value of a continuous, real-valued RV)

Let X be a continuous, real-valued random variable with PDF f_X. The expected value (also called mean) of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, f_X(x) \, \mathrm{d}x. \tag{2}$$

A random variable X is called integrable if

- X is a discrete, real-valued random variable and the series (1) is absolutely convergent.
- X is a continuous, real-valued random variable and the integral (2) is absolutely convergent.

Example

The expected value of X can be interpreted as the value that X will take on average. If we observe realizations $x_1, \ldots, x_n$ of X, then for large n, the empirical mean should be close to $\mathbb{E}[X]$:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \approx \mathbb{E}[X].$$

Example

Assume that X is deterministic, i.e., there exists $x \in \mathbb{R}$ such that $X = x$ almost surely†. Then $\mathbb{E}[X] = x$.

Example

Let X be a discrete random variable with a finite target space $E \subset \mathbb{R}$. Suppose that X is uniformly distributed in E. Then

$$\mathbb{E}[X] = \frac{1}{|E|}\sum_{x \in E} x,$$

so the expected value of X coincides with the algebraic average of the values $x \in E$.

†The term "almost surely", abbreviated "a.s.", means that the probability of this outcome is 1.

Example

Let $a < b$ and assume that $X \sim \mathcal{U}(a,b)$. Then $f_X(x) = \frac{\mathbf{1}_{(a,b)}(x)}{b-a}$, and

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{\mathbf{1}_{(a,b)}(x)}{b-a} \, \mathrm{d}x = \int_a^b \frac{x}{b-a} \, \mathrm{d}x = \frac{a+b}{2}.$$

Example

Let $\mu \in \mathbb{R}$ and $\sigma > 0$ and consider $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2} \, \mathrm{d}x.$$

Performing the change of variables $y = x - \mu$, we obtain

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} (y+\mu) \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}y^2} \, \mathrm{d}y \\
&= \frac{1}{2\pi\sigma^2} \underbrace{\int_{-\infty}^{\infty} y\mathrm{e}^{-\frac{1}{2\sigma^2}y^2} \, \mathrm{d}y}_{= 0 \text{ as an odd function of } y} + \mu \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2\sigma^2}y^2} \, \mathrm{d}y}_{= 1 \text{ (PDF integrates to 1 over } \mathbb{R})} \\
&= \mu.
\end{aligned}$$

This justifies calling the parameter μ the mean of the Gaussian RV X.

In many cases, one is interested in the expected value of some derived quantity of the random variable X. The following result makes this simple.

Theorem (Law of the unconscious statistician)

- *If X is a discrete random variable with PMF p_X and $g\colon E \to \mathbb{R}$, then*

$$\mathbb{E}[g(X)] = \sum_{x \in E} g(x) p_X(x).$$

- *If X is a continuous RV with PDF f_X and $g\colon \mathbb{R} \to \mathbb{R}$ continuous,*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

- *If X is a continuous $\mathbb{R}^k$-valued RV with PDF f_X and $g\colon \mathbb{R}^k \to \mathbb{R}^k$ continuous,*

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^k} g(x) f_X(x)\,\mathrm{d}x.$$

In other words, it is enough to know the distribution of X in order to be able to compute $\mathbb{E}[g(X)]$ for any continuous function g. It is not necessary to solve the distribution of $g(X)$.

Example

A stick of length 1 is broken into two pieces at a uniformly random point between 0 and 1. Let Y denote the length of the larger piece and we wish to know $\mathbb{E}[Y]$.

Let $X \sim \mathcal{U}(0,1)$ denote the position of the breaking point. Then $Y = \max(X, 1-X)$. By the law of the unconscious statistician, we obtain

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \max(x, 1-x)\, \mathbf{1}_{(0,1)}(x)\, \mathrm{d}x = \int_0^1 \max(x, 1-x)\, \mathrm{d}x$$

$$= \int_0^{1/2} (1-x)\, \mathrm{d}x + \int_{1/2}^1 x\, \mathrm{d}x = \frac{1}{2} - \frac{1}{8} + \frac{1}{2} - \frac{1}{8} = \frac{3}{4}.$$

Example (Moments)

An important class of maps g are given by $g(x) = x^k$. Then

$$\mathbb{E}[X^k] = \begin{cases} \sum_{x\in E} x^k p_X(x) \text{ if } X \text{ is a discrete RV with target space } E \subset \mathbb{R} \\ \int_{-\infty}^{\infty} x^k f_X(x)\, \mathrm{d}x \text{ if } X \text{ is a continuous, real-valued RV} \end{cases}$$

is the k^{th} moment of X. (If $\mathbb{E}[|X|^k] = \infty$, the moment is said not to exist.) If this expression is finite for $k = 2$, then X is called square-integrable.

Example

Let $a < b$ and assume that $X \sim \mathcal{U}(a, b)$. Then

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \frac{\mathbf{1}_{(a,b)}(x)}{b-a} \, \mathrm{d}x = \int_a^b x^2 \frac{1}{b-a} \, \mathrm{d}x$$
$$= \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}.$$

The probability of an event A of a probability space $(\Omega, \mathbb{P})$ can be written as the expected value of the indicator function for set A.

Proposition

Let $(\Omega, \mathbb{P})$ be a probability space and let $A \subset \Omega$ be an event. Define the random variable $\mathbf{1}_A \colon \Omega \to \mathbb{R}$,

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Then

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).$$

Proof. Since $X = \mathbf{1}_A$ is a discrete random variable taking values in $E = \{0, 1\}$, its PMF satisfies

$$p_X(0) = \mathbb{P}(A^{\complement}) = 1 - \mathbb{P}(A), \quad p_X(1) = \mathbb{P}(A).$$

Hence

$$\mathbb{E}[X] = 0 \cdot p_X(0) + 1 \cdot p_X(1) = \mathbb{P}(A). \quad \square$$

Properties of the expected value

Proposition

Let X be a real-valued random variable and $a, b \in \mathbb{R}$. Then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Proof. For continuous random variables: $\mathbb{E}[aX + b] = \int_{\mathbb{R}} (ax + b) f_X(x) \, \mathrm{d}x$ $= a \underbrace{\int_{\mathbb{R}} x f_X(x) \, \mathrm{d}x}_{=\mathbb{E}[X]} + b \underbrace{\int_{\mathbb{R}} f_X(x) \, \mathrm{d}x}_{=1}$. The proof is similar for discrete RVs. □

Theorem

1. *If $X \geq 0$ almost surely, then $\mathbb{E}[X] \geq 0$. (Similarly, if $X \leq 0$ almost surely, then $\mathbb{E}[X] \leq 0$.)*
2. *If $X_1, \dots, X_n$ are real-valued random variables and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, then*
$$\mathbb{E}\left[\sum_{i=1}^{n} \alpha_i X_i\right] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i].$$
3. *If $X \leq Y$ almost surely, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.*

Finally, the expected value of a product of independent random variables is the product of the expected values.

Theorem

Let $X_1, \ldots, X_n$ be independent real-valued random variables. Then

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i].$$

Variance

Definition

Let X be a real-valued random variable with mean $\mu = \mathbb{E}[X]$. The variance of X is defined as

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

Note that this quantity is well-defined provided that $\mathbb{E}[X^2] < \infty$. The standard deviation of X is defined as

$$\sigma_X = \sqrt{\mathrm{Var}(X)}.$$

- Note that $\mathrm{Var}(X) = \sum_{x\in E}(x-\mu)^2 p_X(x)$ if X is a discrete random variable with PMF p_X, and $\mathrm{Var}(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f_X(x)\,\mathrm{d}x$ is X if a continuous random variable with PDF f_X.
- The variance $\mathrm{Var}(X)$ is always *nonnegative*. While $\mathbb{E}[X]$ represents the *average value* of X, $\mathrm{Var}(X)$ quantifies how far realizations of X can spread away from this average value.

Theorem (Variance translation)

Let $\mu = \mathbb{E}[X]$ denote the mean of random variable X. Then

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

Proof.

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] &= \mathbb{E}[X^2] - 2\mu\underbrace{\mathbb{E}[X]}_{=\mu} + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2. \quad \square\end{aligned}$$

Remark. If the random variable X satisfies $\mathbb{E}[X] = 0$, then we say that X is centered. In this case, we simply have $\mathrm{Var}(X) = \mathbb{E}[X^2]$.

Example

Let $a < b$ and suppose that $X \sim \mathcal{U}(a, b)$. We have already computed that

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \mathbb{E}[X^2] = \frac{a^2+ab+b^2}{3}.$$

Therefore

$$\mathrm{Var}(X) = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12},$$

and the standard deviation $\sigma_X = \frac{b-a}{2\sqrt{3}}$. Hence, the larger the interval $[a, b]$ for the uniform distribution, the larger the standard deviation.

Example

Let $\mu \in \mathbb{R}$ and $\sigma > 0$ and suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x-\mu)^2} \, \mathrm{d}x.$$

Carrying out the change of variables $y = \frac{x-\mu}{\sigma}$, where $\mathrm{d}x = \sigma \, \mathrm{d}y$, we get

$$\mathrm{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \mathrm{e}^{-\frac{1}{2}y^2} \, \mathrm{d}y.$$

Since

$$\int_{-\infty}^{\infty} y^2 \mathrm{e}^{-\frac{1}{2}y^2} \, \mathrm{d}y = \sqrt{2\pi} \tag{3}$$

(see the following slide for an argument), we conclude that

$$\mathrm{Var}(X) = \sigma^2.$$

This justifies calling the parameter σ^2 the variance of the Gaussian RV X.

Intermezzo – computing the value of the integral (3)

Let $a > 0$ be a parameter and consider the following *parametric* integral:

$$I(a) := \int_{-\infty}^{\infty} y^2 \mathrm{e}^{-\frac{1}{2}ay^2}\,\mathrm{d}y = -2\int_{-\infty}^{\infty} \frac{\partial}{\partial a}\mathrm{e}^{-\frac{1}{2}ay^2}\,\mathrm{d}y$$
$$\overset{(*)}{=} -2\frac{\mathrm{d}}{\mathrm{d}a}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}ay^2}\,\mathrm{d}y.$$

Applying $\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2\sigma^2}x^2}\,\mathrm{d}x = 1 \;\Leftrightarrow\; \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2\sigma^2}x^2}\,\mathrm{d}x = \sqrt{2\pi}\sigma$ with $\sigma = \frac{1}{\sqrt{a}}$ yields

$$I(a) = -2\frac{\mathrm{d}}{\mathrm{d}a}\frac{\sqrt{2\pi}}{\sqrt{a}} = \frac{\sqrt{2\pi}}{a^{3/2}}.$$

The value of the integral (3) corresponds to $I(1) = \sqrt{2\pi}$.

This technique is known as the "Leibniz integral rule", or "Feynman's differentiation under the integral sign". The difficult part is verifying that the order of integration and differentiation can be switched in $(*)$. This is allowed, e.g., when the integrand $f(a, y)$ is continuously differentiable.

Theorem

1. *If X is a real-valued random variable and $a, b \in \mathbb{R}$, then*

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

2. *If $X_1, \ldots, X_n$ are* independent *real-valued random variables and $a_1, \ldots, a_n \in \mathbb{R}$, then*

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i).$$

Covariance and correlation

Definition

Let X and Y be two real-valued random variables with means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Then the covariance of X and Y is

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

If $\sigma_X^2 = \mathrm{Var}(X)$ and $\sigma_Y^2 = \mathrm{Var}(Y)$ are the variances, then the correlation of X and Y is

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

Remark. The correlation always satisfies

$$-1 \leq \rho_{X,Y} \leq 1$$

as a consequence of the *Cauchy–Schwarz inequality*.

Theorem

Let X and Y be two real-valued random variables with means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Then

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

Proof.

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
&= \mathbb{E}[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\
&= \mathbb{E}[XY] - \mu_Y \underbrace{\mathbb{E}[X]}_{=\mu_X} - \mu_X \underbrace{\mathbb{E}[Y]}_{=\mu_Y} + \mu_X \mu_Y \\
&= \mathbb{E}[XY] - \mu_X \mu_Y. \quad \square
\end{aligned}$$

The random variables X and Y are said to be uncorrelated if $\mathrm{Cov}(X, Y) = 0$.

Theorem

If X and Y are independent, then X and Y are uncorrelated.

Proof. Since $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for independent X and Y, there holds

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X \mu_Y = \underbrace{\mathbb{E}[X]}_{=\mu_X}\underbrace{\mathbb{E}[Y]}_{=\mu_Y} - \mu_X \mu_Y = 0. \quad \square$$

Note that, in general, X, Y are uncorrelated $\not\Rightarrow$ X, Y are independent! (However, this converse statement does hold for *jointly Gaussian distributions* – we will formulate a special case of this in a moment.)

Theorem

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y),$$
$$\mathrm{Var}(X-Y) = \mathrm{Var}(X) - 2\mathrm{Cov}(X, Y) + \mathrm{Var}(Y).$$

Joint random variables

Definition

Let $X = (X_1, \dots, X_d)$, $d \in \mathbb{N}$, be a joint random variable. We define the mean $\mu = (\mu_i)_{i=1}^d \in \mathbb{R}^d$ and the covariance matrix $C = (C_{i,j})_{i,j=1}^d \in \mathbb{R}^{d\times d}$ of X by

$$
\begin{aligned}
\mu_i &= \mathbb{E}[X_i] \quad \text{for } i = 1, \dots, d, \\
C_{i,j} &= \mathrm{Cov}(X_i, X_j) \quad \text{for } i, j = 1, \dots, d.
\end{aligned}
$$

Example

Let $X = (X_1, \dots, X_d)$ be a d-dimensional Gaussian random variable $X \sim \mathcal{N}(\mu, C)$, where $\mu = (\mu_i)_{i=1}^d \in \mathbb{R}^d$ and $C = (C_{i,j})_{i,j=1}^n \in \mathbb{R}^{d\times d}$ is a symmetric, positive definite matrix. Then

$$
\begin{aligned}
\mu_i &= \mathbb{E}[X_i] \quad \text{for } i = 1, \dots, d, \\
C_{i,j} &= \mathrm{Cov}(X_i, X_j) \quad \text{for } i, j = 1, \dots, d,
\end{aligned}
$$

meaning that μ is the mean of X and C is the covariance matrix of X.

Corollary (Independence of jointly Gaussian random variables)

Let $X = (X_1, \ldots, X_d) \sim \mathcal{N}(\mu, C)$ *for* $\mu = (\mu_j)_{j=1}^d \in \mathbb{R}^d$ *and symmetric, positive definite* $C = (C_{i,j})_{i,j=1}^d \in \mathbb{R}^{d\times d}$*. Then* $X_1, \ldots, X_d$ *are independent if and only if* C *is a diagonal matrix, i.e.,* $C_{i,j} = 0$ *whenever* $i \neq j$*.*

Proof. "$\Rightarrow$" If $X_1, \ldots, X_d$ are independent, then X_i and X_j are independent for all $i \neq j$. Independent random variables are uncorrelated, so the covariance

$$C_{i,j} = \mathrm{Cov}(X_i, X_j) = 0 \quad \text{whenever } i \neq j.$$

"$\Leftarrow$" Let $C = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. Then the marginal distribution of X_j is Gaussian, with PDF $f_{X_j}(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\mathrm{e}^{-\frac{1}{2\sigma_j^2}(x-\mu_j)^2}$. Hence,

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det C}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^{\mathrm{T}}C^{-1}(x-\mu)} = \prod_{j=1}^d \frac{1}{\sqrt{2\pi\sigma_j^2}}\mathrm{e}^{-\frac{1}{2\sigma_j^2}(x_j-\mu_j)^2},$$

i.e., $f_X(x) = f_{X_1}(x_1)\cdots f_{X_d}(x_d)$, meaning that $X_1, \ldots, X_d$ are independent. $\square$

Sample mean and sample variance

In practice, the random variables are not observed directly: we observe realizations, or a sample, thereof. It is useful to define notions of *sample mean* and *sample variance*, which are quantities that can be computed directly from the observed realizations.

Definition

Let $X_1, \dots, X_n$ be real-valued random variables[†]. The sample mean of is defined as the arithmetic average

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The sample variance is defined as

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

and the sample standard deviation is defined as $s_n = \sqrt{s_n^2}$.

Remark. Note that the sample mean $\overline{X}_n$ and the sample variance s_n^2 are themselves random variables. As we shall see, if $X_1, \dots, X_n$ are *independent and identically distributed* provided some integrability conditions are satisfied, then there holds for large n that

$$\overline{X}_n \approx \mathbb{E}[X_1] \quad \text{and} \quad s_n^2 \approx \mathrm{Var}(X_1).$$

[†] One may think of $X_1, \dots, X_n$ as representing a sample from some random variable X.

Sample covariance of vector-valued random variables

If $X_1, \ldots, X_n$ are vector-valued random variables taking values in $\mathbb{R}^d$, then their sample covariance matrix $\boldsymbol{Q} = (Q_{j,k})_{j,k=1}^n$ is defined as

$$Q_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{i,j} - \mu_j)(X_{i,k} - \mu_k), \quad j, k = 1, \ldots, d,$$

where $\mu = \overline{X} = (\overline{X}_1, \ldots, \overline{X}_d)$ is the mean.