Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald

(2) Random variables

Random variables

- Definition and notation
- Cumulative distribution functions
- Probability mass and density functions

Random variables

- **Definition and notation**
- Cumulative distribution functions
- Probability mass and density functions

Random variables and distributions

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathcal{X}$ be a function.
- Let $\mathcal{S}$ be a σ-algebra on $\mathcal{X}$.
- For every $S \in \mathcal{S}$ let the *preimage* of S be

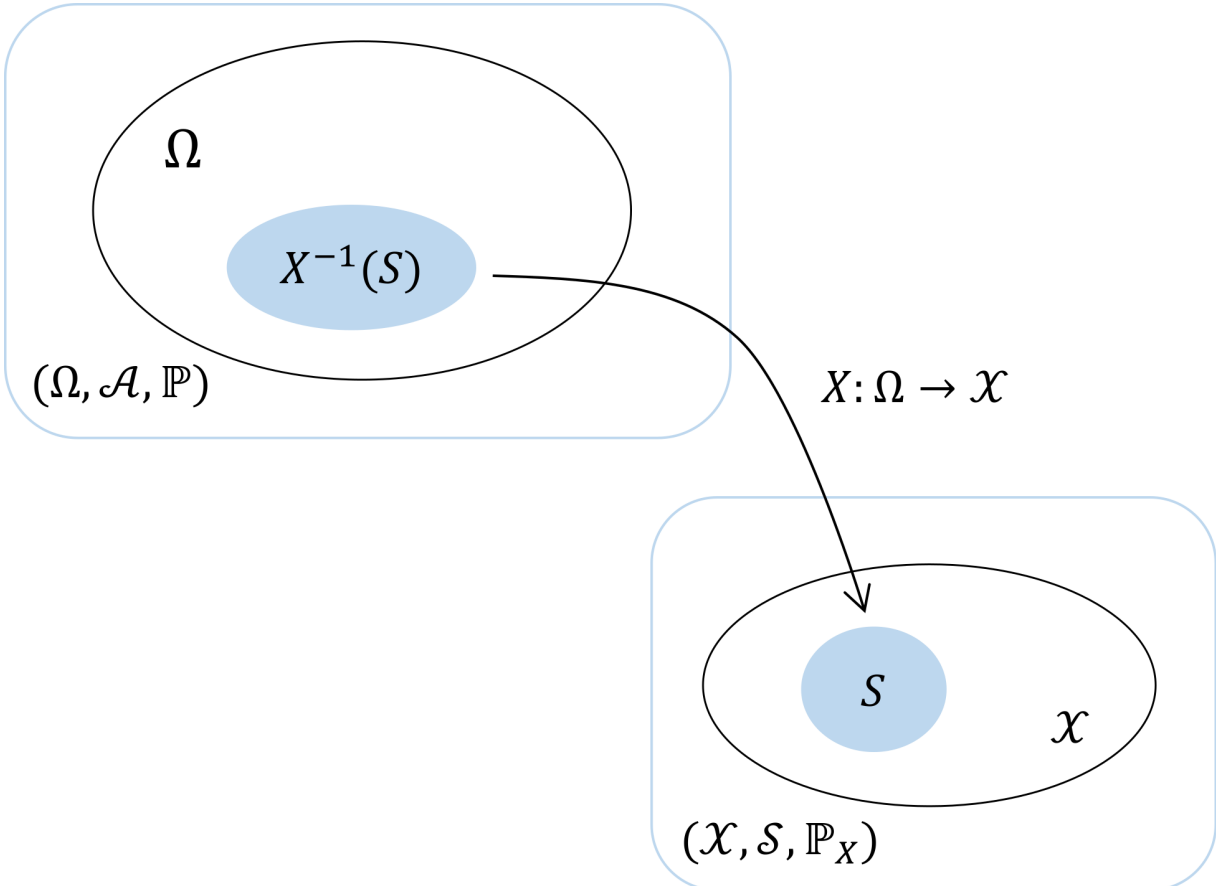
$$X^{-1}(S) := \{\omega \in \Omega | X(\omega) \in S\}. \tag{1}$$

- If $X^{-1}(S) \in \mathcal{A}$ for all $S \in \mathcal{S}$, then X is called *measurable*.
- Let $X : \Omega \to \mathcal{X}$ be measurable. All $S \in \mathcal{S}$ get allocated the probability

$$\mathbb{P}_X : \mathcal{S} \to [0,1], S \mapsto \mathbb{P}_X(S) := \mathbb{P}\left(X^{-1}(S)\right) = \mathbb{P}\left(\{\omega \in \Omega | X(\omega) \in S\}\right) \tag{2}$$

- X is called a *random variable* and $\mathbb{P}_X$ is called the *distribution* of X.
- $(\mathcal{X}, \mathcal{S}, \mathbb{P}_X)$ is a probability space.
- With $\mathcal{X} = \mathbb{R}$ and $\mathcal{S} = \mathcal{B}$ the probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ takes center stage.

Random variables and distributions



$$\mathbb{P}\left(X^{-1}(S)\right) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \in S\}) =: \mathbb{P}_X(S)$$

Definition (Random variable)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote a probability space. A (real-valued) random variable is a mapping

$$X : \Omega \to \mathbb{R}, \omega \mapsto X(\omega), \tag{3}$$

with the *measurability property*

$$\{\omega \in \Omega | X(\omega) \in S)\} \in \mathcal{A} \text{ for all } S \in \mathcal{S}. \tag{4}$$

Remarks

- Random variables are neither "random" nor "variables".
- Intuitively, $\omega \in \Omega$ gets randomly selected according to $\mathbb{P}$ and $X(\omega)$ realized.
- The distributions (probability measures) of random variables are central.

Random variables and distributions

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ and $(\mathcal{X}, \mathcal{S}, \mathbb{P}_X)$ denote probability spaces for $X : \Omega \to \mathcal{X}$.
- The following notations for events $A \in \mathcal{A}$ w.r.t. X are conventional:

$$\{X \in S\} := \{\omega \in \Omega | X(\omega) \in S\}, S \subset \mathcal{X}$$
$$\{X = x\} := \{\omega \in \Omega | X(\omega) = x\}, x \in \mathcal{X}$$
$$\{X \leq x\} := \{\omega \in \Omega | X(\omega) \leq x\}, x \in \mathcal{X}$$
$$\{X < x\} := \{\omega \in \Omega | X(\omega) < x\}, x \in \mathcal{X}$$

- These conventions entail the following conventions for distributions:

$$\mathbb{P}_X(X \in S) = \mathbb{P}(\{X \in S\}) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \in S\}), S \subset \mathcal{X}$$
$$\mathbb{P}_X(X \leq x) = \mathbb{P}(\{X \leq x\}) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\}), x \in \mathcal{X}$$

- Often, the random variable subscript in distribution symbols is omitted:

$$\mathbb{P}(X \in S) = \mathbb{P}_X(X \in S), S \subset \mathcal{X}$$
$$\mathbb{P}(X \leq x) = \mathbb{P}_X(X \leq S), x \in \mathcal{X}$$

- Distributions can be defined using *cumulative distribution functions*, *probability mass functions*, and *probability density functions*.

Random variables

- Definition and notation
- **Cumulative distribution functions**
- Probability mass and density functions

Definition (Cumulative distribution function)

The cumulative distribution function (CDF) of a random variable X is defined as

$$P : \mathbb{R} \to [0, 1], x \mapsto P(x) := \mathbb{P}(X \leq x). \tag{5}$$

Remarks

- CDFs can be used to define distributions.
- CDFs exist for both discrete and continuous random variables.

Example (Cumulative distribution function)

Consider a random variable with outcome space $\mathcal{X} = \{0, 1, 2\}$ and distribution defined by

$$\mathbb{P}(X = 0) = \frac{1}{4}, \quad \mathbb{P}(X = 1) = \frac{1}{2}, \quad \mathbb{P}(X = 2) = \frac{1}{4} \tag{6}$$

Then its distribution function is given by

$$P : \mathbb{R} \to [0, 1], x \mapsto P(x) := \begin{cases} 0 & x < 0, \\ \frac{1}{4} & 0 \leq x < 1, \\ \frac{3}{4} & 1 \leq x < 1, \\ 1 & x \geq 2. \end{cases} \tag{7}$$

Remarks

- P is right-continuous.
- P is defined for all $x \in \mathbb{R}$, while $X \in \{0, 1, 2\}$.

Identity of CDFs

Let X have CDF P and let Y have CDF Q. If $P(x) = Q(x)$ for all x, then $\mathbb{P}(X \in S) = \mathbb{P}(Y \in S)$ for all events $S \in \mathcal{S}$.

Properties of CDFs

A function $P : \mathbb{R} \to [0, 1]$ is a CDF for some probability $\mathbb{P}$, if and only if P satisfies the following conditions

(1) P is *non-decreasing*: $x_1 < x_2$ implies that $P(x_1) \leq P(x_2)$.

(2) P is *normalized*: $\lim_{x \to -\infty} P(x) = 0$ and $\lim_{x \to \infty} P(x) = 1$.

(3) P is *right-continuous*: $P(x) = P(x^+)$ for all x, where $P(x^+) := \lim_{y \to x, y > x} P(y)$.

Random variables

- Definition and notation
- Cumulative distribution functions
- **Probability mass and density functions**

Definition (Probability mass functions, discrete random variables)

A random variable X is discrete, if it takes on countably many values in $\mathcal{X} := \{x_1, x_2, ...\}$. The probability mass function of X is defined as

$$p : \mathcal{X} \to [0, 1], x \mapsto p(x) := \mathbb{P}(X = x). \tag{8}$$

Remarks

- A set is countable, if it is finite or bijectively related to $\mathbb{N}$.
- A PMF is non-negative: $p(x) \geq 0$ for all $x \in \mathcal{X}$.
- A PMF is normalized: $\sum_i p(x) = 1$.
- The CDF of a PMF is $P(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} p(x_i)$.
- The CDF of a PMF is also referred to as a *cumulative mass function* (CMF).

Example (Bernoulli random variable)

Let X be a random variable with outcome set $\mathcal{X} = \{0, 1\}$ and probability mass function

$$p : \mathcal{X} \to [0, 1], x \mapsto p(x) := \mu^x (1 - \mu)^{1-x} \text{ for } \mu \in [0, 1]. \tag{9}$$

Then X is said to be distributed according to a *Bernoulli distribution* with parameter $\mu \in [0, 1]$, for which we write $X \sim \text{Bern}(\mu)$. We denote the probability mass function of a Bernoulli random variable by

$$\text{Bern}(x; \mu) := \mu^x (1 - \mu)^{1-x}. \tag{10}$$

Remarks

- A Bernoulli random variable can be used to model a single biased coin flip with outcomes "failure" 0 and "success" 1.
- μ is the probability for X to take the value 1,

$$\mathbb{P}(X = 1) = \mu^1 (1 - \mu)^{1-1} = \mu. \tag{11}$$

Definition (Probability density functions, continuous random variables)

A random variable X is continuous, if there exists a function

$$p : \mathbb{R} \to \mathbb{R}_{\geq 0}, x \mapsto p(x) \tag{12}$$

such that

- $p(x) \geq 0$ for all $x \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} p(x)dx = 1$,
- $\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\,dx$ for all $a, b \in \mathbb{R}, a \leq b$.

Remarks

- PDFs can take on values larger than 1 and $\mathbb{P}(X = a) = \int_a^a p(x)\,dx = 0$.
- Probabilities are obtained from PDFs by integration,
- (Probability) mass = (probability) density $\times$ (set) volume.
- The CDF of a PDF is $P(x) = \int_{-\infty}^{x} p(\xi)\,d\xi$, thus $p(x) = \frac{d}{dx}P(x)$.
- The CDF of a PDF is also referred to as *cumulative density function*.

Example (Gaussian random variable, standard normal variable)

Let X be a random variable with outcome set $\mathbb{R}$ and probability density function

$$p : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{13}$$

Then X is said to be distributed according to a *Gaussian distribution* with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, for which we write $X \sim N\left(\mu, \sigma^2\right)$. We abbreviate the PDF of a Gaussian random variable by

$$N\left(x; \mu, \sigma^2\right) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{14}$$

A Gaussian random variable with $\mu = 0$ and $\sigma^2 = 1$ is said to be distributed according to a *standard normal distribution* and is often referred to as a Z *variable*.

Remarks

- The parameter μ specifies the location of highest probability density.
- The parameter σ^2 specifies the width of the distribution.
- The term $\frac{1}{\sqrt{2\pi\sigma^2}}$ is the normalization constant for $\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$.

Example (Uniform random variables)

Let X be a discrete random variable with a finite outcome set $\mathcal{X}$ and probability mass function

$$p : \mathcal{X} \to \mathbb{R}_{\geq 0}, x \mapsto p(x) := \frac{1}{|\mathcal{X}|}. \tag{15}$$

Then X is said to be distributed according to a *discrete uniform distribution*, for which we write $X \sim U(|\mathcal{X}|)$. We abbreviate the PMF of a discrete uniform random variable by

$$U(x; |\mathcal{X}|) := \frac{1}{|\mathcal{X}|}. \tag{16}$$

Similarly, let X be a continuous random variable with probability density function

$$p : \mathbb{R} \to \mathbb{R}_{>0}, x \mapsto p(x) := \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases} \tag{17}$$

Then X is said to be distributed according to a *continuous uniform distribution* with parameters a and b, for which we write $X \sim U(a, b)$. We abbreviate the PDF of a continuous uniform random variable by

$$U(x; a, b) := \frac{1}{b-a}. \tag{18}$$

Properties of cumulative density functions

- $\mathbb{P}(X > x) = 1 - P(x)$ (Exceedance distribution function)
- $\mathbb{P}(x < X \leq y) = P(y) - P(x)$ (Interval probability)
- With the properties of the Riemann integral, we have

$$\begin{aligned} P(y) - P(x) &= \mathbb{P}(x < X < y) = \mathbb{P}(x \leq X < y) \\ &= \mathbb{P}(x < X \leq y) = \mathbb{P}(x \leq X \leq y). \end{aligned} \tag{19}$$

Definition (Inverse cumulative distribution function)

Let X be a random variable with CDF P. Then the *inverse cumulative distribution function* or *quantile function* of X is defined as

$$P^{-1} : [0,1] \to \mathbb{R}, q \mapsto P^{-1}(q) := \inf\{x | P(x) > q\} \tag{20}$$

If P is invertible, i.e., strictly increasing and continuous, then $P^{-1}(q)$ is the unique real number x such that $P(x) = q$.

Remarks

- $P^{-1}(0.25)$ is called the *first quartile*.
- $P^{-1}(0.50)$ is called the median or *second quartile*.
- $P^{-1}(q)$ is also referred to as qth *percentile*.

Example (CDF and inverse CDF for Gaussian random variables)

Let X be a univariate Gaussian random variable with expectation parameter μ and variance parameter σ^2. Then, X has

- probability density function

$$p : \mathbb{R} \to \mathbb{R}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right),$$

- cumulative density function

$$P : \mathbb{R} \to [0,1], x \mapsto \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2\sigma^2}(\xi-\mu)^2\right) d\xi,$$

- and inverse cumulative density function

$$P^{-1} : [0,1] \to \mathbb{R}, q \mapsto P^{-1}(q) = \{x \in \mathbb{R} | P(x) = q\}.$$

Remark

- Let $\mu = 1, \sigma^2 = 1$. Then $p(2) = 0.24$, $P(2) = 0.84$, and $P^{-1}(0.84) = 2$.

Example (CDF and inverse CDF for standard normal variables)

Let Z be a standard normal variable. Then, Z has

- probability density function

$$\phi : \mathbb{R} \to \mathbb{R}, z \mapsto \phi(z) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right),$$

- cumulative density function

$$\Phi : \mathbb{R} \to [0,1], z \mapsto \Phi(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\xi^2\right) d\xi,$$

- inverse cumulative density function

$$\Phi^{-1} : [0,1] \mapsto \mathbb{R}, q \mapsto \Phi^{-1}(q) = \{z \in \mathbb{R} | \Phi(z) = q\}$$

Examples

- $\phi(1.645) = 0.102$, $\Phi(1.645) = 0.950$, $\Phi^{-1}(0.950) = \Phi^{-1}(1 - 0.050) = 1.640$.
- $\phi(1.960) = 0.058$, $\Phi(1.960) = 0.975$, $\Phi^{-1}(0.975) = \Phi^{-1}(1 - \frac{0.050}{2}) = 1.960$.