# Probability Theory

Maximilian Soelch

Technische Universität München

# Probability Theory

Probability Theory is the study of uncertainty. Uncertainty is all around us.

Mathematical probability theory is based on *measure theory*—we do not work at this level.

Slides are mostly based on *Review of Probability Theory* by Arian Meleki and Tom Do.

# Probability Theory

The basic problem that we study in probability theory:

Given a data generating process,
what are the properties of the outcomes?

The basic problem of statistics (or better statistical *inference*) is the *inverse* of probability theory:

Given the outcomes,
what can we say about the process that generated the data?

Statistics uses the formal language of probability theory.

# Basic Elements of Probability

**Sample space** $\Omega$:

The set of all outcomes of a random experiment.

e.g. rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
e.g. rolling a die twice: $\Omega' = \Omega \times \Omega = \{(1,1), (1,2), \ldots, (6,6)\}$

**Set of events** $\mathcal{F}$ (event space):

A set whose elements $A \in \mathcal{F}$ (*events*) are subsets of $\Omega$.
$\mathcal{F}$ ($\sigma$-field) must satisfy

- $\emptyset \in \mathcal{F}$,
- $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$,
- $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$.

e.g. "die outcome is even" $\rightsquigarrow$ event $A = \{\omega \in \Omega : \omega \text{ even}\} = \{2, 4, 6\}$

e.g. (smallest) $\sigma$-field that contains $A$: $\mathcal{F} = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}\}$

# Basic Elements of Probability ctd.

**Probability measure** $\mathrm{P} : \mathcal{F} \to [0, 1]$

with *Axioms of Probability*:

- $\mathrm{P}(A) \geq 0$ for all $A \in \mathcal{F}$,
- $\mathrm{P}(\Omega) = 1$,
- If $A_1, A_2, \ldots$ are disjoint events $\left(A_i \cap A_j = \emptyset, i \neq j\right)$ then

$$\mathrm{P}(\cup_i A_i) = \sum_i \mathrm{P}(A_i).$$

e.g. for rolling a die $\mathrm{P}(A) = \frac{|A|}{|\Omega|}$

**The triple ($\Omega$, $\mathcal{F}$, $\mathrm{P}$) is called a probability space.**

# Important Properties

The three axioms from the previous slide suffice to show:

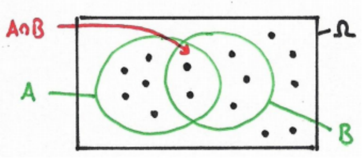
- If $A \subseteq B \Rightarrow \mathrm{P}(A) \leq \mathrm{P}(B)$
- $\mathrm{P}(A \cap B)\,(\equiv \mathrm{P}(A, B)) \leq \min(\mathrm{P}(A)\,, \mathrm{P}(B))$
- $\mathrm{P}(A \cup B) \leq \mathrm{P}(A) + \mathrm{P}(B)$
- $\mathrm{P}(\Omega \setminus A) = 1 - \mathrm{P}(A)$
- If $A_1, A_2, \ldots, A_k$ are sets of *disjoint* events such that $\cup_i A_i = \Omega$ then $\sum_k \mathrm{P}(A_k) = 1$ (*Law of total probability*).

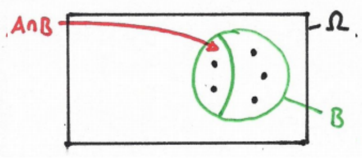
# Conditional Probability

Let $A, B \subseteq \Omega$ be two events with $\mathrm{P}(B) \neq 0$, then:

$$\mathrm{P}(A \mid B) := \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

$\mathrm{P}(A \mid B)$ is the probability of $A$ conditioned on $B$ and represents the probability of $A$, if it is known that $B$ was observed.



# Multiplication law

Let $A_1, \ldots, A_n$ be events with $\mathrm{P}(A_1 \cap \ldots \cap A_n) \neq 0$. Then:

$$\mathrm{P}(A_1 \cap \ldots \cap A_n) = \prod_{i=1}^{n} \mathrm{P}\left( A_i \mid \bigcap_{j<i} A_j \right)$$

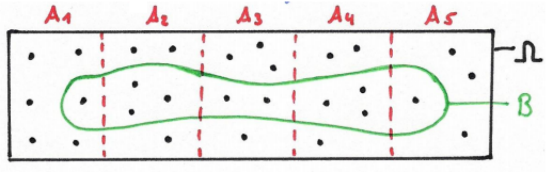
$$= \mathrm{P}(A_1) \cdot \mathrm{P}(A_2 \mid A_1) \cdot \mathrm{P}(A_3 \mid A_1 \cap A_2) \cdot \ldots \cdot \mathrm{P}(A_n \mid A_1 \cap \ldots \cap A_{n-1}).$$

# Law of total probability (revisited)

Let $B$ be an event and $\Phi$ a partition of $\Omega$ with $\mathrm{P}(A) > 0$ for all $A \in \Phi$. Then:

$$\mathrm{P}(B) = \sum_{A \in \Phi} \mathrm{P}(B \cap A) = \sum_{A \in \Phi} \mathrm{P}(A) \cdot \mathrm{P}(B \mid A).$$

Graphical representation for a 5-partition $\Phi = \{A_1, \ldots, A_5\}$ of $\Omega$:



# Bayes' rule

Let $A$ and $B$ be two events with $\mathrm{P}(A), \mathrm{P}(B) \neq 0$. Then:

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A) \cdot \mathrm{P}(A)}{\mathrm{P}(B)}.$$

Bayes' rule applies the multiplication rule twice to set $\mathrm{P}(A \mid B)$ and $\mathrm{P}(B \mid A)$ in relation:

$$\mathrm{P}(B \mid A) \cdot \mathrm{P}(A) = \underbrace{\mathrm{P}(A \cap B)}_{=\mathrm{P}(A,B)} = \mathrm{P}(A \mid B) \cdot \mathrm{P}(B).$$

Bayes' rule is always used if the conditional probability $\mathrm{P}(A \mid B)$ is easy to calculate or given, but the conditional probability $\mathrm{P}(B \mid A)$ is searched for.

# Independence

Two events $A$, $B$ are called *independent* if and only if

$$\mathrm{P}(A, B) = \mathrm{P}(A)\,\mathrm{P}(B)\,,$$

or equivalently $\mathrm{P}(A \mid B) = \mathrm{P}(A)$. What does that mean in words?

Two events $A$, $B$ are called *conditionally independent* given a third event C if and only if

$$\mathrm{P}(A, B \mid C) = \mathrm{P}(A \mid C)\mathrm{P}(B \mid C).$$

# Random variables

We are usually only interested in some aspects of a random experiment.

Random variable $X : \Omega \to \mathbb{R}$ (actually not every function is allowed ...).

A random variable is usually just denoted by an upper case letter $X$ (instead of $X(\omega)$). The value a random variable may take is denoted by the corresponding lower-case letter.

For a discrete random variable

$$\mathrm{P}(X = x) := \mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) \,.$$

For a continuous random variable

$$\mathrm{P}(a \leq X \leq b) := \mathrm{P}(\{\omega \in \Omega : a \leq X(\omega) \leq b\}) \,.$$

Note the usage of $\mathrm{P}$ here.

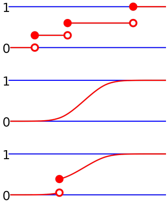
# Cumulative distribution function – CDF

A probability measure $\mathrm{P}$ is specified by a *cumulative distribution function (CDF)*, a function $F_X : \mathbb{R} \to [0,1]$:

$$F_X(x) \equiv \mathrm{P}(X \leq x)\,.$$

Properties:

- $0 \leq F_X(x) \leq 1$
- $\lim_{x\to-\infty} F_X(x) = 0$
- $\lim_{x\to\infty} F_X(x) = 1$
- $x \leq y \to F_X(x) \leq F_X(y)$




Let $X$ have CDF $F_X$ and $Y$ have CDF $F_Y$. If $F_X(x) = F_Y(x)$ for all $x$, then $\mathrm{P}(X \in A) = \mathrm{P}(Y \in A)$ for all (measurable) $A$.
We call $X$ and $Y$ identically distributed (or equal in distribution).

# Probability density function—PDF

For some continuous random variables, the CDF $F_X(x)$ is continuous on $\mathbb{R}$. The *probability density function* is then defined as the piecewise derivative

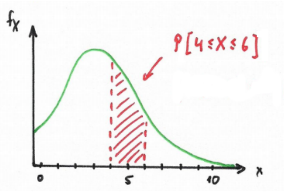
$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x},$$

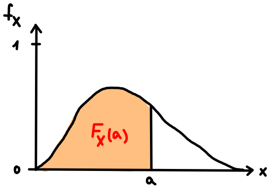
and $X$ is called continuous.

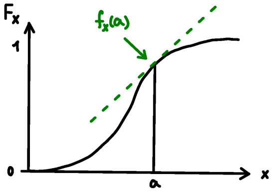
$$\mathrm{P}(x \le X \le x + \Delta x) \approx f_X(x) \cdot \Delta x.$$

Properties:

- $f_X(x) \ge 0$
- $\int_{x \in A} f_X(x)\,\mathrm{d}x = \mathrm{P}(X \in A)$
- $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$



# Probability mass function—PMF

$X$ takes on only a *countable* set of possible values (*discrete* random variable).

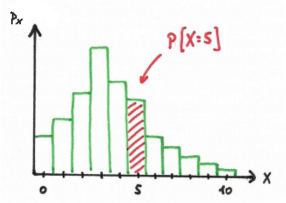
A probability mass function $p_X : \Omega \to [0, 1]$ is a simple way to *represent* the probability measure associated with $X$:

$$p_X(x) = \mathrm{P}(X = x)$$

(Note: We use the probability measure $\mathrm{P}$ on the random variable $X$)

Properties:

- $0 \leq p_X(x) \leq 1$
- $\sum_x p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = \mathrm{P}(X \in A)$



# Transformation of Random Variables

Given a (continuous) random variable $X$ and a strictly monotonic (increasing *or* decreasing) function $s$, what can we say about $Y = s(X)$?

$$f_Y(y) = f_X\big(t(y)\big)\,|t'(y)|,$$

where $t$ is the inverse of $s$.

# Expectation

For any measurable function $g : \mathbb{R} \to \mathbb{R}$, we define the *expected value*:

$$\mathrm{E}[g(X)] = \sum_x g(x) p_X(x) \qquad \text{discrete}$$

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

Special case: $\mathrm{E}[X]$, i.e., $g(x) = x$, is called the *mean* of $X$.

Properties:

- $\mathrm{E}[a] = a$ for any *constant* $a \in \mathbb{R}$
- $\mathrm{E}[af(X)] = a\mathrm{E}[f(X)]$ for any constant $a \in \mathbb{R}$
- $\mathrm{E}[f(X) + g(X)] = \mathrm{E}[f(X)] + \mathrm{E}[g(X)]$

For any $A \in \mathbb{R}$: $\mathrm{E}[\mathbb{I}_A(X)] = P(X \in A)$

# Variance and Standard Deviation

*Variance* measures the concentration of a random variable's distribution around its mean.

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

Properties:

- $\mathrm{Var}(a) = 0$ for any constant $a \in \mathbb{R}$.
- $\mathrm{Var}(af(X)) = a^2\,\mathrm{Var}(f(X))$ for any constant $a \in \mathbb{R}$.

$\sigma(X) = \sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of $X$.

# Entropy

The *Shannon entropy* or just *entropy* of a discrete random variable $X$ is

$$H[X] \equiv -\sum_x \mathrm{P}(X = x) \log \mathrm{P}(X = x) = -\mathrm{E}[\log \mathrm{P}(X)].$$

Given two probability mass fuctions $p_1$ and $p_2$, the *Kullback-Leibler divergence* (or *relative entropy*) between $p_1$ and $p_2$ is

$$\mathrm{KL}(p_1 \parallel p_2) \equiv -\sum_x p_1(x) \log \frac{p_2(x)}{p_1(x)}$$

Note that the KL divergence is not symmetric.

# Bernoulli distribution

A Bernoulli-distributed random variable $X \sim \mathrm{Ber}(\mu), \mu \in [0, 1]$ models the outcome of an experiment. It is positive with a probability of $\mu$ and negative with a probability of $1 - \mu$.

$$p_X(x) = \begin{cases} \mu, & \text{if } x = 1 \\ 1 - \mu, & \text{if } x = 0 \\ 0 & \text{else} \end{cases}$$

For calculations the following equation is more useful:

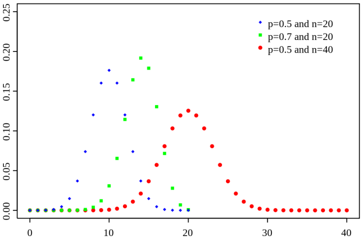
$$\mathrm{Ber}(x \mid \mu) = \mu^x \cdot (1 - \mu)^{1-x}$$

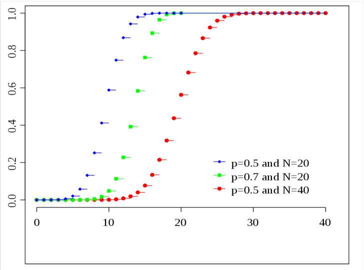
# Binomial distribution

A Binomial random variable
$X \sim \text{Bin}(N, \mu),\ \ N \geq 1,\ \mu \in [0, 1]$
shows the number of successes by performing $N$ trials, where each trial is independent from the others. The success probability is $\mu$.

For $x \in \{0, 1, \ldots, N\}$:

$$\text{Bin}(x \mid N, \mu) = \binom{N}{x} \cdot \mu^x \cdot (1-\mu)^{N-x}$$



# Poisson distribution

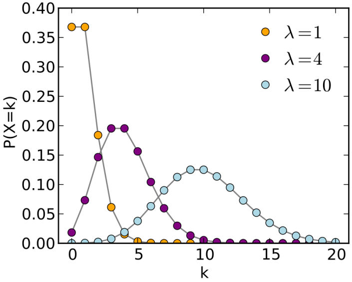
A Binomial random variable with large $N$ and small $\mu$ can be approximated by a Poisson random variable $X \sim \mathrm{Poi}(\lambda)$.

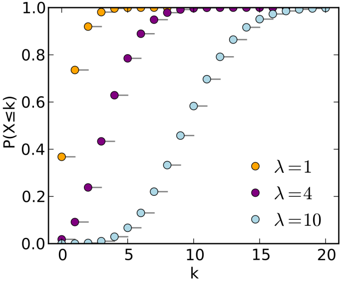
For $\lambda = N\mu$ and as $N \to \infty$:

$$X \sim \mathrm{Bin}(N, \mu) \to X \sim \mathrm{Poi}(\lambda)$$

For $x \in \mathbb{N}_0$:

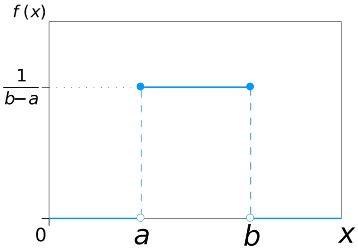
$$\mathrm{Poi}(x \mid \lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$



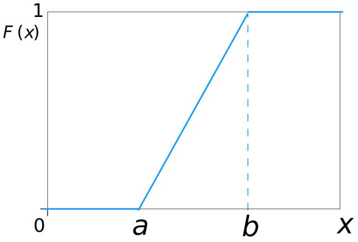
# Uniform distribution

A uniformly distributed random variable $X \sim \mathrm{U(a, b)}$, $a, b \in \mathbb{R}$, $a < b$, takes any value on the interval $[a, b]$ with equal probability.




For $x \in [a, b]$:

$$\mathrm{U}(x \mid a, b) = \frac{1}{b - a}$$

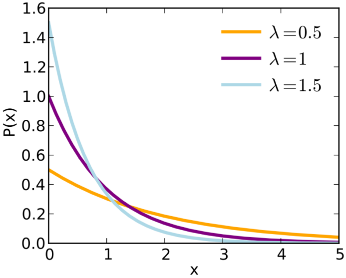


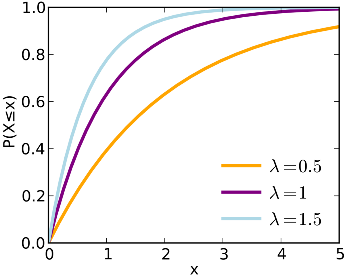
# Exponential distribution

An exponentially distributed random variable $X \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$ can be referred to as the latency until an event ("success") occurs the first time. $\lambda$ corresponds to the expected number of successes in one unit of time.

For $x \in \mathbb{R}_0^+$:

$$\mathrm{Exp}(x \mid \lambda) = \lambda \cdot e^{-\lambda x}$$



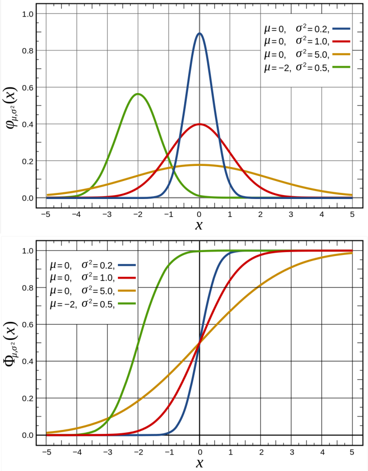



# Normal/Gaussian distribution

A Normal or Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2),\ \mu, \sigma \in \mathbb{R}$, $\sigma > 0$ has approximately the same distribution as the sum of many independently, arbitrarily but identically distributed random variables.

For $x \in \mathbb{R}$:

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

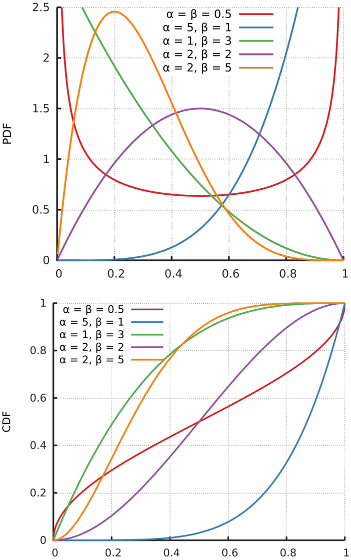


# Beta distribution

Random variables $X \sim \mathrm{Beta}(a, b)$, $a, b > 0$, following a Beta distribution can often be seen as the success probability for a binary event.

For $x \in [0, 1]$:

$$\mathrm{Beta}(x \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}$$



# Gamma distribution

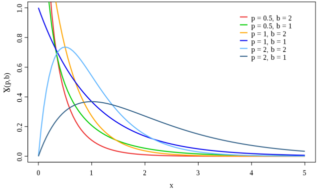
Random variables $X \sim \mathrm{Gamma}(a, b)$ following a Gamma distribution are governed by the parameters $a, b > 0$ . For $x > 0$:

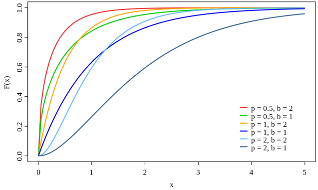
$$\mathrm{Gamma}(x \mid a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx}$$

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\, \mathrm{d}t$$

$$\Gamma(n+1) = n! \text{ for } n \in \mathbb{N}_0^+$$

The Gamma distribution is the conjugate prior for the precision (inverse variance) of a univariate Gauss distribution.



# Overview: probability distributions

| Distribution | Notation | Param. | Co-dom. | PMF / PDF | Mean | Variance |
|---|---|---|---|---|---|---|
| Bernoulli* | $\mathrm{Ber}(\mu)$ | $\mu \in [0,1]$ | $x \in \{0,1\}$ | $\mu^x(1-\mu)^{1-x}$ | $\mu$ | $\mu(1-\mu)$ |
| Binomial* | $\mathrm{Bin}(N,\mu)$ | $N \geq 1, \mu \in [0,1]$ | $x \in \{0,1,\dots,N\}$ | $\binom{N}{x}\mu^x(1-\mu)^{N-x}$ | $N\mu$ | $N\mu(1-\mu)$ |
| Poisson* | $\mathrm{Poi}(\lambda)$ | $\lambda > 0$ | $x \in \mathbb{N}_0^+$ | $\frac{e^{-\lambda}\lambda^x}{x!}$ | $\lambda$ | $\lambda$ |
| Uniform | $\mathrm{U}(a,b)$ | $a,b \in \mathbb{R}, a < b$ | $x \in [a,b]$ | $\frac{1}{b-a}$ | $\frac{a+b}{2}$ | $\frac{1}{12}(b-a)^2$ |
| Exponential | $\mathrm{Exp}(\lambda)$ | $\lambda > 0$ | $x \in \mathbb{R}_0^+$ | $\lambda e^{-\lambda x}$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| Normal/Gauss | $\mathcal{N}(\mu,\sigma^2)$ | $\mu \in \mathbb{R}, \sigma > 0$ | $x \in \mathbb{R}$ | $\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ | $\mu$ | $\sigma^2$ |
| Beta | $\mathrm{Beta}(a,b)$ | $a,b > 0$ | $x \in [0,1]$ | $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ | $\frac{a}{a+b}$ | $\frac{ab}{(a+b)^2(a+b+1)}$ |
| Gamma | $\mathrm{Gamma}(a,b)$ | $a,b > 0$ | $x \in \mathbb{R}_0^+$ | $\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$ | $\frac{a}{b}$ | $\frac{a}{b^2}$ |

*Discrete distributions

With the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$, with the property that $\Gamma(n+1) = n!$ for $n \in \mathbb{N}_0^+$.

# Two random variables—*Bivariate* case

Two random variables $X$ and $Y$ can interact. We need to consider them simultaneously for statistical analysis. To this end, we introduce the *joint cumulative distribution function* of $X$ and $Y$:

$$F_{XY}(x, y) = \mathrm{P}(X \leq x,\, Y \leq y)\,.$$

$F_X(x)$ and $F_Y(y)$ are the *marginal cumulative distribution function* of $F_{XY}(x, y)$.

Properties:

- $0 \leq F_{XY}(x, y) \leq 1$
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$
- $F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$

# Two *continuous* random variables

Most properties can be defined analogously to the univariate case.
*Joint probability density function*:

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}.$$

Properties:

- $f_{XY}(x, y) \geq 0$
- $\int \int_A f_{XY}(x, y)\, dxdy = \mathrm{P}((X, Y) \in A)$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\, dxdy = 1$

If we remove the effect of one of the random variables, we yield the *marginal probability density function* or *marginal density* for short:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, \mathrm{d}y.$$

# Relations between $f_{X,Y}$, $f_X$, $f_Y$, $F_{X,Y}$, $F_X$ and $F_Y$

$$\begin{tikzcd}[column sep=huge, row sep=huge]
f_X(x) \arrow[d, bend right, "\int_{-\infty}^{x} f_X(u)\,du"] & f_{X,Y}(x,y) \arrow[l, "\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy"'] \arrow[r, "\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx"] \arrow[d, bend right, "\int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv"'] & f_Y(y) \arrow[d, bend right, "\int_{-\infty}^{y} f_Y(v)\,dv"'] \\
F_X(x) \arrow[u, bend right, "\frac{d}{dx} F_X(x)"'] & F_{X,Y}(x,y) \arrow[l, "\lim_{y \to \infty} F_{X,Y}(x,y)"] \arrow[r, "\lim_{x \to \infty} F_{X,Y}(x,y)"'] \arrow[u, bend right, "\frac{d}{dy}\frac{d}{dx} F_{X,Y}(x,y)"'] & F_Y(y) \arrow[u, bend right, "\frac{d}{dy} F_Y(y)"']
\end{tikzcd}$$

# Two *discrete* random variables

*Joint probability mass function*:

$$p_{XY}(x, y) = \mathrm{P}(X = x, Y = y)\,.$$

Properties:

- $0 \leq p_{XY}(x, y) \leq 1$
- $\sum_x \sum_y p_{XY}(x, y) = 1$

In order to get the *marginal probability mass function* $p_X(x)$, we need to *sum out* all possible $y$ (*marginalization*):

$$p_X(x) = \sum_y p_{XY}(x, y).$$

# Conditional distributions/Bayes' rule

| | discrete | continuous |
|---|---|---|
| Definition | $p_{Y\mid X}(y \mid x) = \frac{p_{XY}(x,y)}{p_X(x)}$ | $f_{Y\mid X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}$ |
| Bayes' rule | $p_{Y\mid X}(y \mid x) = \frac{p_{X\mid Y}(x\mid y)p_Y(y)}{p_X(x)}$ | $f_{Y\mid X}(y \mid x) = \frac{f_{X\mid Y}(x\mid y)f_Y(y)}{f_X(x)}$ |
| Probabilites | $p_{Y\mid X}(y \mid x) = \mathrm{P}(Y = y \mid X = x)$ | $\mathrm{P}(Y \in A \mid X = x) = \int_A f_{Y\mid X}(y \mid x)\,\mathrm{d}y$ |

# Independence

Two random variables $X$, $Y$ are *independent* if $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all values $x$ and $y$.

Equivalently:

- $p_{XY}(x, y) = p_X(x)\, p_Y(y)$
- $p_{Y|X}(y \mid x) = p_Y(y)$
- $f_{XY}(x, y) = f_X(x)f_Y(y)$
- $f_{Y|X}(y \mid x) = f_Y(y)$

# Independent and identically distributed—i.i.d.

If two random variables $X$ and $Y$ are called *identically distributed* it means that the following holds:

$$f_X(x) = f_Y(x),$$
$$F_X(x) = F_Y(x).$$

As a consequence (among many others):

$$\mathrm{E}[X] = \mathrm{E}[Y],$$
$$\mathrm{Var}(X) = \mathrm{Var}(Y).$$

It does not mean that $X = Y$ is true! $X$ and $Y$ following the same distribution does not imply that they always provide the same values!
If $X$ and $Y$ are also independent, we call them *independent and identically distributed* (i.i.d.).

# Expectation and covariance

Given two random variables $X$, $Y$ and $g : \mathbb{R}^2 \to \mathbb{R}$.

- $\mathrm{E}[g(X, Y)] := \sum_x \sum_y g(x, y) p_{XY}(x, y)$.
- $\mathrm{E}[g(X, Y)] := \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) \, \mathrm{d}x \, \mathrm{d}y$.

*Covariance*

- $\mathrm{Cov}(X, Y) := \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$.
- When $\mathrm{Cov}(X, Y) = 0$, $X$ and $Y$ are *uncorrelated*.
- *Pearson* correlation coefficient $\rho(X, Y)$:

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1, 1].$$

- $\mathrm{E}[f(X, Y) + g(X, Y)] = \mathrm{E}[f(X, Y)] + \mathrm{E}[g(X, Y)]$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.
- If $X$ and $Y$ are independent, then $\mathrm{Cov}(X, Y) = 0$.
- If $X$ and $Y$ are independent, then $\mathrm{E}[f(X)g(Y)] = \mathrm{E}[f(X)]\mathrm{E}[g(Y)]$.

# Multiple random variables—Random vectors

Generalize previous ideas to more than two random variables. Putting all these random variables together in one vector $\boldsymbol{X}$, a *random vector* ($\boldsymbol{X} : \Omega \to \mathbb{R}^n$). The notions of joint CDF and PDF apply equivalently, e.g.

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \mathrm{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)\,.$$

Expectation of a continuous random vector for $g : \mathbb{R}^n \to \mathbb{R}$:

$$\mathrm{E}[g(\boldsymbol{X})] = \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n) f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \ldots \mathrm{d}x_n\,.$$

If $g : \mathbb{R}^n \to \mathbb{R}^m$ then the expected value of $g$ is the *element-wise* values of the output vector:

$$\mathrm{E}[g(\boldsymbol{X})] = \begin{bmatrix} \mathrm{E}[g_1(\boldsymbol{X})] \\ \mathrm{E}[g_2(\boldsymbol{X})] \\ \vdots \\ \mathrm{E}[g_m(\boldsymbol{X})] \end{bmatrix}.$$

# Independence of more than two random variables

The random variables $X_1, \ldots, X_n$ are independent if for all subsets $I = \{i_1, \ldots, i_k\} \subset \{1, \ldots, N\}$ and all $(x_{i_1}, \ldots, x_{i_k})$

$$f_{X_{i_1},\ldots,X_{i_k}}(x_{i_1}, \ldots, x_{i_k}) = f_{X_{i_1}}(x_{i_1}) \cdot \ldots \cdot f_{X_{i_k}}(x_{i_k}),$$

or equivalently

$$F_{X_{i_1},\ldots,X_{i_k}}(x_{i_1}, \ldots, x_{i_k}) = F_{X_{i_1}}(x_{i_1}) \cdot \ldots \cdot F_{X_{i_k}}(x_{i_k}),$$

hold.

If there exists a combination of values so that the equations above do not hold then the random variables are not independent.
For better distinction, this notion of independence is sometimes called *mutual independence*.

# Covariance matrix

For a random vector $\boldsymbol{X} : \Omega \to \mathbb{R}^n$, the *covariance matrix* $\boldsymbol{\Sigma}$ is the $n \times n$ square *symmetric*, *positive definite* matrix whose entries are

$$\boldsymbol{\Sigma}_{ij} = \operatorname{Cov}(X_i, X_j).$$

$$\boldsymbol{\Sigma} = \mathrm{E}\big[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^T\big] = \mathrm{E}\Big[\boldsymbol{X}\boldsymbol{X}^T\Big] - \mathrm{E}[\boldsymbol{X}]\mathrm{E}[\boldsymbol{X}]^T$$

# Multinomial distribution

The multivariate version of the Binomial is called a *Multinomial*, $\boldsymbol{X} \sim \mathrm{Multinomial}(N, \boldsymbol{\mu})$. We have $k \geq 1$ mutually exclusive events with a success probability of $\mu_k$ (such that $\sum_{i=1}^{k} \mu_k = 1$).
We draw $N$ times independently.

$$p_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k) = \binom{N}{x_1 \;\; x_2 \ldots x_k} \mu_1^{x_1} \mu_2^{x_2} \ldots \mu_k^{x_k}.$$

where

$$\begin{aligned}
&\sum_k x_k = N, \\
&\binom{N}{x_1 \;\; x_2 \ldots x_k} = \frac{N!}{x_1! \; x_2! \; \ldots x_k!}, \\
&\mathrm{E}[\boldsymbol{X}] = (N\mu_1, N\mu_2, \ldots, N\mu_k), \\
&\mathrm{Var}(X_i) = N\mu_i(1 - \mu_i), \\
&\mathrm{Cov}(X_i, X_j) = -N\mu_i\mu_j.
\end{aligned}$$

Example: An urn with $n$ balls of $k \geq 1$ different labels, drawn $N \geq 1$ times with replacement and probabilities $\mu_k = \frac{\#k}{n}$.
The marginal distribution of $X_i$ is $\mathrm{Bin}(n, \mu_i)$.

# Multivariate Gaussian

The multivariate version of the Gaussian $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is very similar to the univariate, except that it allows for dependencies between the individual components. $\boldsymbol{\mu} \in \mathbb{R}^k$ is the mean vector, the positive definite, symmetric $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ the covariance matrix

$$f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right),$$

where

$$\begin{aligned} \mathrm{E}[\boldsymbol{X}] &= \boldsymbol{\mu}, \\ \mathrm{Var}(X_i) &= \boldsymbol{\Sigma}_{ii}, \\ \mathrm{Cov}(X_i, X_j) &= \boldsymbol{\Sigma}_{ij}. \end{aligned}$$

The marginal distribution of $X_i$ is $\mathcal{N}(\mu_i, \boldsymbol{\Sigma}_{ii})$.

# Notation in the lecture

Consider

$$p_X(x), \quad x \in \mathbb{R} \qquad \text{vs} \qquad p_X(y), \quad y \in \mathbb{R}$$

$p_X(x), f_X(x), p_{XY}(x, y), f_{XY}(x, y)$ are written as $p(x)$ or $p(x, y)$.
Likewise $p_Y(y)$ is written as $p(y)$.