TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

Prof. Dr. Björn Sprungk
Faculty of Mathematics and Computer Science
Institute of Stochastics

Mathematics of machine learning

Appendix: Probability theory

Winter term 2024/25

What is randomness?

Many processes in nature and everyday life are random:

- Flipping a coin
- Rolling a dice
- Lifespan of a lamp or electronic device
- Payback of loans by bank customers
- Returns of investmens
- Taking a cat picture from the internet

What is randomness?

Many processes in nature and everyday life are random:

- Flipping a coin
- Rolling a dice
- Lifespan of a lamp or electronic device
- Payback of loans by bank customers
- Returns of investmens
- Taking a cat picture from the internet

To describe such situations we consider a basic (and abstract) concept:

Random Experiment

A random experiment is a well-defined and repeatable experiment with a known set of possible outcomes, but beforehand unknown actual outcome.

Mathematical description of randomness

- The sample space Ω is the set of all possible outcomes ω of a random experiment, e.g.,
  - Flipping a coin: $\Omega = \{\text{“heads”}, \text{“tails”}\}$
  - Rolling a die: $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$
  - Measuring the lifetime of a device: $\Omega = [0, \infty)$

Mathematical description of randomness

- The sample space Ω is the set of all possible outcomes ω of a random experiment, e.g.,
  - Flipping a coin: $\Omega = \{\text{“heads”}, \text{“tails”}\}$
  - Rolling a die: $\Omega = \{\epsdice{1}, \epsdice{2}, \epsdice{3}, \epsdice{4}, \epsdice{5}, \epsdice{6}\}$
  - Measuring the lifetime of a device: $\Omega = [0, \infty)$
- An event A is a subset possible outcomes $A \subseteq \Omega$. We say the event A occurs if the actual outcome of the random experiment belongs to A: $\omega \in A$
  - Rolling a dice: $A =$“number of dots is even”$= \{\epsdice{2}, \epsdice{4}, \epsdice{6}\}$
  - Lifetime of a device: $B =$“lifetime exceeds 5”$= (5, \infty)$

Mathematical description of randomness

- The sample space Ω is the set of all possible outcomes ω of a random experiment, e.g.,
  - Flipping a coin: $\Omega = \{\text{“heads”}, \text{“tails”}\}$
  - Rolling a die: $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$
  - Measuring the lifetime of a device: $\Omega = [0, \infty)$
- An event A is a subset possible outcomes $A \subseteq \Omega$. We say the event A occurs if the actual outcome of the random experiment belongs to A: $\omega \in A$
  - Rolling a dice: $A =$“number of dots is even”$= \{$⚁, ⚃, ⚅$\}$
  - Lifetime of a device: $B =$“lifetime exceeds 5”$= (5, \infty)$
- Given an event A the complementary event A^c is the set of all possible outcomes not belonging A:

$$A^c = \Omega \setminus A = \{\omega \in \Omega \colon \omega \notin A\}$$

e.g., Rolling a dice: $A =$“number of dots is even”, $\quad A^c = \{$⚀, ⚂, ⚄$\}$

Combining events

We use set-theoretic notation to work with combinations of events A, B:

- $A \cup B$: Union of A and B , i.e., A **or** B happens
- $A \cap B$: Intersection of A and B , i.e., A **and** B happens
- $A \subseteq B$: A is a subset of B, i.e., B includes A

Example: Consider rolling a dice as well as the events

$$A = \text{“number of dots is even”} = \{\text{⚁}, \text{⚃}, \text{⚅}\},$$
$$B = \text{“number of dots greater four”} = \{\text{⚄}, \text{⚅}\}.$$

Then

$$A \cup B = \text{“number of dots is even or greater than four”} = \{\text{⚁}, \text{⚃}, \text{⚄}, \text{⚅}\},$$
$$A \cap B = \text{“number of dots is even and greater than four”} = \{\text{⚅}\}.$$

Combining multiple events

For an (infinite) sequence $A_1,\ A_2,\ldots$ of events of a sample space Ω we define

- their union $\bigcup_{n=1}^{\infty} A_n$ as the event which consists of all outcomes that are in A_n for at least one value of $n = 1, 2, \ldots$

Combining multiple events

For an (infinite) sequence $A_1,\ A_2, \ldots$ of events of a sample space Ω we define

- their union $\bigcup_{n=1}^{\infty} A_n$ as the event which consists of all outcomes that are in A_n for at least one value of $n = 1, 2, \ldots$
- their intersection $\bigcap_{n=1}^{\infty} A_n$, as the event consisting of those outcomes that are in all of the events $A_n,\ n = 1, 2, \ldots$

Combining multiple events

For an (infinite) sequence $A_1,\ A_2,\dots$ of events of a sample space Ω we define

- their union $\bigcup_{n=1}^{\infty} A_n$ as the event which consists of all outcomes that are in A_n for at least one value of $n = 1, 2, \dots$
- their intersection $\bigcap_{n=1}^{\infty} A_n$, as the event consisting of those outcomes that are in all of the events $A_n,\ n = 1, 2, \dots$

Example: If $A_1 = \{1,2,3\},\ A_2 = \{2,3,4\},\ A_3 = \{3,4,5\}$, then

$$\bigcup_{n=1}^{3} A_n = \{1,2,3,4,5\} \quad \text{and} \quad \bigcap_{n=1}^{3} A_n = \{3\}.$$

Combining multiple events

For an (infinite) sequence $A_1,\ A_2, \ldots$ of events of a sample space Ω we define

- their union $\bigcup_{n=1}^{\infty} A_n$ as the event which consists of all outcomes that are in A_n for at least one value of $n = 1, 2, \ldots$
- their intersection $\bigcap_{n=1}^{\infty} A_n$, as the event consisting of those outcomes that are in all of the events $A_n,\ n = 1, 2, \ldots$

Example: If $A_1 = \{1, 2, 3\},\ A_2 = \{2, 3, 4\},\ A_3 = \{3, 4, 5\}$, then

$$\bigcup_{n=1}^{3} A_n = \{1, 2, 3, 4, 5\} \quad \text{and} \quad \bigcap_{n=1}^{3} A_n = \{3\}.$$

Meaning:

- union $\bigcup_{n=1}^{\infty} A_n$: One of the events A_n happens
- union $\bigcap_{n=1}^{\infty} A_n$: All of the events A_n happen

Probability of an event

Axioms of probability

A probability measure $\mathbb{P}$ assigns to each event A of a sample space Ω a real number $\mathbb{P}(A)$ called the probability of the event A such that the following three conditions are satisfied

Probability of an event

Axioms of probability

A probability measure $\mathbb{P}$ assigns to each event A of a sample space Ω a real number $\mathbb{P}(A)$ called the probability of the event A such that the following three conditions are satisfied

1. $0 \leq \mathbb{P}(A) \leq 1$

Probability of an event

Axioms of probability

A probability measure $\mathbb{P}$ assigns to each event A of a sample space Ω a real number $\mathbb{P}(A)$ called the probability of the event A such that the following three conditions are satisfied

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(\Omega) = 1$

Probability of an event

Axioms of probability

A probability measure $\mathbb{P}$ assigns to each event A of a sample space Ω a real number $\mathbb{P}(A)$ called the probability of the event A such that the following three conditions are satisfied

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(\Omega) = 1$
3. For any sequence of pairwise mutually exclusive events $A_1, A_2, \ldots$ (i.e. $A_i \cap A_j = \varnothing$ for each $i \neq j$):

Probability of an event

Axioms of probability

A probability measure $\mathbb{P}$ assigns to each event A of a sample space Ω a real number $\mathbb{P}(A)$ called the probability of the event A such that the following three conditions are satisfied

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(\Omega) = 1$
3. For any sequence of pairwise mutually exclusive events $A_1, A_2, \ldots$ (i.e. $A_i \cap A_j = \varnothing$ for each $i \neq j$):

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Probability of an event

Axioms of probability

A probability measure $\mathbb{P}$ assigns to each event A of a sample space Ω a real number $\mathbb{P}(A)$ called the probability of the event A such that the following three conditions are satisfied

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(\Omega) = 1$
3. For any sequence of pairwise mutually exclusive events $A_1, A_2, \ldots$ (i.e. $A_i \cap A_j = \varnothing$ for each $i \neq j$):

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Frequentist interpretation: If we repeat a random experiment over and over again, then the proportion of time that event A occurs will converge to the value $\mathbb{P}(A)$.

Arithmetics of probability

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\varnothing) = 0$ (impossible event)
- $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ (monotonicity)
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ (additivity)
- $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ (Union bound)

Arithmetics of probability

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\varnothing) = 0$ (impossible event)
- $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ (monotonicity)
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ (additivity)
- $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ (Union bound)

Example: For a fair die we have $\mathbb{P}(\text{⚀}) = \mathbb{P}(\text{⚁}) = \ldots = \mathbb{P}(\text{⚅}) = \frac{1}{6}$. Consider now
$A =$"number of dots is even"$= \{\text{⚁}, \text{⚃}, \text{⚅}\}$, $B =$"number of dots is greater four"$= \{\text{⚄}, \text{⚅}\}$. Then:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{2}{3}$$

Independence

Two events are called independent if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Random variables

- Often we are interested in the outcome of a certain *quantity* related to a random experiment, e.g.,
  - Rolling two dice: "sum of dots", i.e., with realizations in $\{2, 3, \ldots, 12\}$
  - Flipping a coin: "number of tries until first time heads" with realizations in $\{1, 2, 3, ...\}$
  - Production of steel bars: "Young's modulus" with realizations in $(0, \infty)$

Random variables

- Often we are interested in the outcome of a certain *quantity* related to a random experiment, e.g.,
  - Rolling two dice: "sum of dots", i.e., with realizations in $\{2, 3, \dots, 12\}$
  - Flipping a coin: "number of tries until first time heads" with realizations in $\{1, 2, 3, ...\}$
  - Production of steel bars: "Young's modulus" with realizations in $(0, \infty)$

- These are examples of (real-valued) random variables

Random variables

- Often we are interested in the outcome of a certain *quantity* related to a random experiment, e.g.,
  - Rolling two dice: "sum of dots", i.e., with realizations in $\{2, 3, \ldots, 12\}$
  - Flipping a coin: "number of tries until first time heads" with realizations in $\{1, 2, 3, ...\}$
  - Production of steel bars: "Young's modulus" with realizations in $(0, \infty)$

- These are examples of (real-valued) random variables

- Mathematically, a random variable X is a mapping

$$X \colon \Omega \to \mathbb{R}$$

i.e., a random variable X assigns to each outcome of a random experiment a real number.

Random variables

- Often we are interested in the outcome of a certain *quantity* related to a random experiment, e.g.,
  - Rolling two dice: "sum of dots", i.e., with realizations in $\{2, 3, \dots, 12\}$
  - Flipping a coin: "number of tries until first time heads" with realizations in $\{1, 2, 3, ...\}$
  - Production of steel bars: "Young's modulus" with realizations in $(0, \infty)$

- These are examples of (real-valued) random variables

- Mathematically, a random variable X is a mapping

$$X : \Omega \to \mathbb{R}$$

  i.e., a random variable X assigns to each outcome of a random experiment a real number.

- An actual outcome $x = X(\omega) \in \mathbb{R}$, $\omega \in \Omega$, of a random variable X is called a realization of X.

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values
- Countable means that all values can be indexed by (a subset of) the positive integers.

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values
- Countable means that all values can be indexed by (a subset of) the positive integers.
- Thus, also countably infinite number of outcomes is possible

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values
- Countable means that all values can be indexed by (a subset of) the positive integers.
- Thus, also countably infinite number of outcomes is possible

Continous random variables:

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values
- Countable means that all values can be indexed by (a subset of) the positive integers.
- Thus, also countably infinite number of outcomes is possible

Continous random variables:

- Can take a continuum of values

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values
- Countable means that all values can be indexed by (a subset of) the positive integers.
- Thus, also countably infinite number of outcomes is possible

Continous random variables:

- Can take a continuum of values
- Uncountably infinite number of outcomes

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values
- Countable means that all values can be indexed by (a subset of) the positive integers.
- Thus, also countably infinite number of outcomes is possible

Continous random variables:

- Can take a continuum of values
- Uncountably infinite number of outcomes
- E. g. lifetime of a car which may be any value within $(0, \infty)$

Discrete and continuous random variables

Discrete random variables:

- Can take only a countable number of values
- Countable means that all values can be indexed by (a subset of) the positive integers.
- Thus, also countably infinite number of outcomes is possible

Continous random variables:

- Can take a continuum of values
- Uncountably infinite number of outcomes
- E. g. lifetime of a car which may be any value within $(0, \infty)$

In general, mixtures of the two may occur (not considered here)

Working with random variables

- In the following we consider probabilities of events related to a random variable X such as

$$\mathbb{P}(X = x) \;=\; \mathbb{P}(\{\omega \in \Omega \colon X(\omega) = x\})$$

Working with random variables

- In the following we consider probabilities of events related to a random variable X such as

$$\mathbb{P}(X = x) \;=\; \mathbb{P}(\{\omega \in \Omega \colon X(\omega) = x\})$$

- **Example:** Rolling two fair dice we consider X as the "sum of dots". Now here $\omega = [\epsdice[black]{4}, \epsdice{6}]$ denotes the outcome of the random experiment with first dice being black (showing 4) and second dice being (showing 6). Then

$$\begin{aligned}
\mathbb{P}(X = 4) &= \mathbb{P}(\{\, [\epsdice[black]{1}, \epsdice{3}],\ [\epsdice[black]{2}, \epsdice{2}],\ [\epsdice[black]{3}, \epsdice{1}] \,\}) \\
&= \mathbb{P}(\ [\epsdice[black]{1}, \epsdice{3}]\ ) + \mathbb{P}(\ [\epsdice[black]{2}, \epsdice{2}]\ ) + \mathbb{P}(\ [\epsdice[black]{3}, \epsdice{1}]\ ) \\
&= \mathbb{P}(\epsdice[black]{1}) \cdot \mathbb{P}(\epsdice{3}) + \mathbb{P}(\epsdice[black]{2}) \cdot \mathbb{P}(\epsdice{2}) + \mathbb{P}(\epsdice[black]{3}) \cdot \mathbb{P}(\epsdice{1}) \\
&= \frac{1}{12}
\end{aligned}$$

Can you explain this calculation?

Describing random variables

- **Discrete:** For a discrete random variable X we have the probability mass function (pmf)

$$p_X(x) := \mathbb{P}(X = x)$$

Describing random variables

- **Discrete:** For a discrete random variable X we have the probability mass function (pmf)

$$p_X(x) := \mathbb{P}(X = x)$$

- **Continuous:** For a continuous random variable X we have the probability density function (pdf) $f_X\colon \mathbb{R} \to [0, \infty)$ defined via

$$\mathbb{P}\left(X \in [a, b]\right) = \int_a^b f_X(x)\,\mathrm{d}x, \qquad a \leq b$$

  e.g., $\mathbb{P}(X = x) = 0$

Describing random variables

- **Discrete:** For a discrete random variable X we have the probability mass function (pmf)

$$p_X(x) := \mathbb{P}(X = x)$$

- **Continuous:** For a continuous random variable X we have the probability density function (pdf) $f_X : \mathbb{R} \to [0, \infty)$ defined via

$$\mathbb{P}\left(X \in [a, b]\right) = \int_a^b f_X(x)\,\mathrm{d}x, \qquad a \leq b$$

e.g., $\mathbb{P}(X = x) = 0$

- **Both:** For any random variable we have the cumulative distribution function (cdf)

$$F_X(x) := \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}$$

Describing random variables

- We have for the two cases

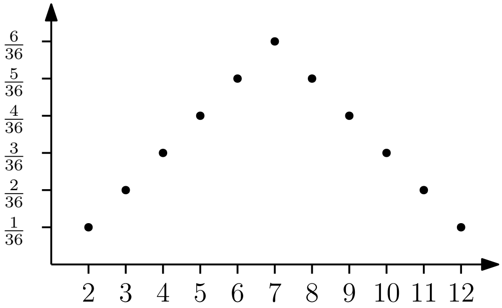
$$F_X(x) = \sum_{y \leq x} p_X(y) \qquad \text{or} \qquad F_X(x) = \int_{-\infty}^{x} f_X(y)\, \mathrm{d}y$$

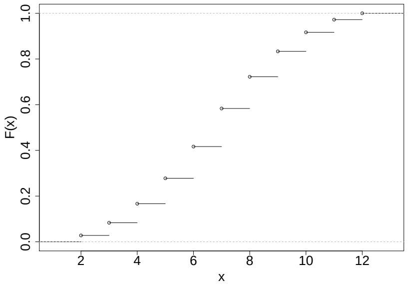
Describing random variables

- We have for the two cases

$$F_X(x) = \sum_{y \le x} p_X(y) \qquad \text{or} \qquad F_X(x) = \int_{-\infty}^{x} f_X(y)\, \mathrm{d}y$$

- **Example:** pmf and cdf for X being the sum of two fair dice:





Moments of random variables

- The cdf or pmf/pdf completely describes the random behaviour of a random variable

Moments of random variables

- The cdf or pmf/pdf completely describes the random behaviour of a random variable

- Nonetheless, we are often interested in average characteristics of random variables.

- The mean or expectation of a random variable describes its "average value":

$$\mathbb{E}[X] = \sum_x x \; p_X(x) \qquad \text{or} \qquad \mathbb{E}[X] = \int_{-\infty}^{\infty} x \; f_X(x) \, \mathrm{d}x$$

Moments of random variables

- The cdf or pmf/pdf completely describes the random behaviour of a random variable

- Nonetheless, we are often interested in average characteristics of random variables.

- The mean or expectation of a random variable describes its "average value":

$$\mathbb{E}[X] = \sum_x x \; p_X(x) \qquad \text{or} \qquad \mathbb{E}[X] = \int_{-\infty}^{\infty} x \; f_X(x) \, \mathrm{d}x$$

- The variance and standard deviation of a random variable describes its "spread around the mean"

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right], \qquad \mathbb{S}\mathbf{td}[X] = \sqrt{\mathbb{V}[X]}$$

Calculating with moments

Given real-valued random variables X and Y and $a \in \mathbb{R}$ we have

- For the mean

$$\mathbb{E}[X \pm Y] = \mathbb{E}[X] \pm \mathbb{E}[Y], \qquad \mathbb{E}[aX] = a\mathbb{E}[X]$$

  and for the variance

$$\mathbb{V}[X \pm Y] = \mathbb{V}[X] + \mathbb{V}[Y] \pm 2\mathbb{C}\mathbf{ov}[X, Y], \qquad \mathbb{V}[aX] = a^2\mathbb{V}[X]$$

  where $\mathbb{C}\mathbf{ov}[X, Y]$ denotes the covariance of X and Y (see later)

- Moreover, for functions $g \colon \mathbb{R} \to \mathbb{R}$ we have for the random variable $g(X)$

$$\mathbb{E}[g(X)] = \sum_x g(x)\ p_X(x) \qquad \text{or} \qquad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\ f_X(x)\ \mathrm{d}x$$

- For events $A \subseteq \Omega$ we have for the random variable $\mathbf{1}_A$ (indicator function of set A)

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

Some common discrete and continuous distributions

Uniform distribution $\mathrm{U}[n]$, $n \in \mathbb{N}$

Models, e.g., rolling a fair dice with n faces.

It is defined by

$$X \sim \mathrm{U}[n] \quad \iff \quad \mathbb{P}(X=1) = \mathbb{P}(X=2) = \cdots = \mathbb{P}(X=n) = \frac{1}{n}$$

We have

$$\mathbb{E}[X] = \frac{n+1}{2}, \qquad \mathbb{V}[X] = \frac{n^2-1}{12}.$$

Bernoulli distribution $\mathrm{B}(p)$, $0 \leq p \leq 1$

Models binary outcomes of a random experiment, e.g., flipping a coin, surviving an accident or producing flawless components.

It is defined by

$$X \sim \mathrm{B}(p) \quad \Longleftrightarrow \quad \mathbb{P}(X = 0) = 1 - p, \quad \mathbb{P}(X = 1) = p$$

We have

$$\mathbb{E}[X] = p, \qquad \mathbb{V}[X] = p(1 - p).$$

Binomial distribution $\mathrm{Bin}(n, p)$, $0 \leq p \leq 1$, $n \in \mathbb{N}$

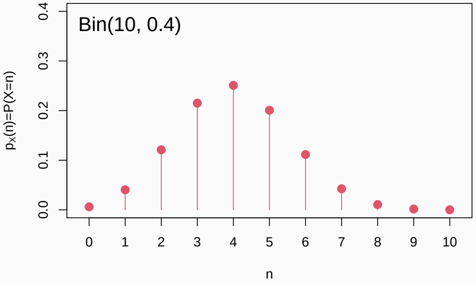
Models the number of successes in a sequence of n independent Bernoulli experiments, each having success probability p.

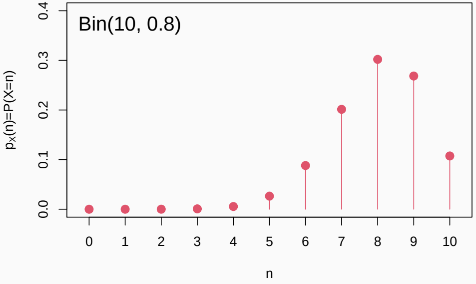
It is denotes by $X \sim \text{Bin}(n, p)$ and defined by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

We have

$$\mathbb{E}[X] = np, \qquad \mathbb{V}[X] = np(1-p).$$





Uniform distribution $\mathrm{U}[a,b]$, $a,b \in \mathbb{R}$

Models "pure" randomness within the interval $[a,b]$, i.e., there is no preference which of those values might be the actual outcome.

It is defined by

$$X \sim \mathrm{U}[a,b] \quad \Longleftrightarrow \quad f_X(x) = \frac{1}{b-a}\,\mathbf{1}_{[a,b]}(x)$$

We have

$$\mathbb{E}[X] = \frac{b+a}{2}, \qquad \mathbb{V}[X] = \frac{(b-a)^2}{12}.$$

Normal or Gaussian distribution $\mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$

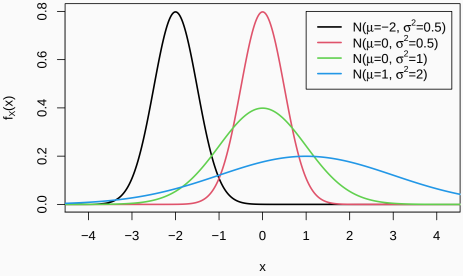
Models symmetric deviation from mean value. Appears quite often in nature, i.e., measurement errors, logarithm of conductivities, exchange rates or blood pressure.

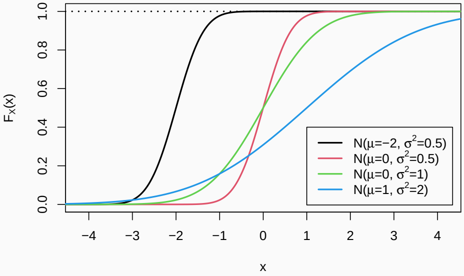
Carl F. Gauss
(1777 – 1855)

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad \Longleftrightarrow \quad f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

We have

$$\mathbb{E}[X] = \mu, \qquad \mathbb{V}[X] = \sigma^2.$$





Multivariate Distributions

- We consider now pairs (X, Y) and, more general, tuples $(X_1, \ldots, X_n)$ of random variables.
- For two random variables X and Y we have the joint cumulative distribution function

$$F(x, y) = P(X \leq x, Y \leq y), \qquad x, y \in \mathbb{R}$$

- If X and Y are **discrete** we define their joint probability mass function by

$$p(x, y) = P(X = x, Y = y), \qquad F(x, y) = \sum_{s \leq x} \sum_{t \leq y} p(s, t)$$

- If X and Y are **continuous** we define their joint probability density function $f \colon \mathcal{R} \times \mathcal{R} \to [0, \infty)$ by $f(x, y)$, defined for all $x \in \mathcal{R}$ and $y \in \mathcal{R}$ such that for all real numbers $a_1 \leq b_1$, $a_2 \leq b_2$

$$\mathbb{P}(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y) \, \mathrm{d}y \, \mathrm{d}x, \qquad F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t) \, \mathrm{d}t \, \mathrm{d}s$$

Example (Rolling two fair dice)

- We roll two fair dice independently of each other.
- The probability of getting 3 on the red dice and 5 on the black dice is

$$p(3,5) = \mathbb{P}(X = 3, Y = 5) = \frac{1}{36}.$$

- The joint pmf is uniformly:

$$p(x,y) =$$

Example (Rolling two fair dice)

- We roll two fair dice independently of each other.
- The probability of getting 3 on the red dice and 5 on the black dice is

$$p(3,5) = \mathbb{P}(X = 3, Y = 5) = \frac{1}{36}.$$

- The joint pmf is uniformly:

$$p(x,y) = \mathbb{P}(X = x, Y = y) =$$

Example (Rolling two fair dice)

- We roll two fair dice independently of each other.
- The probability of getting 3 on the red dice and 5 on the black dice is

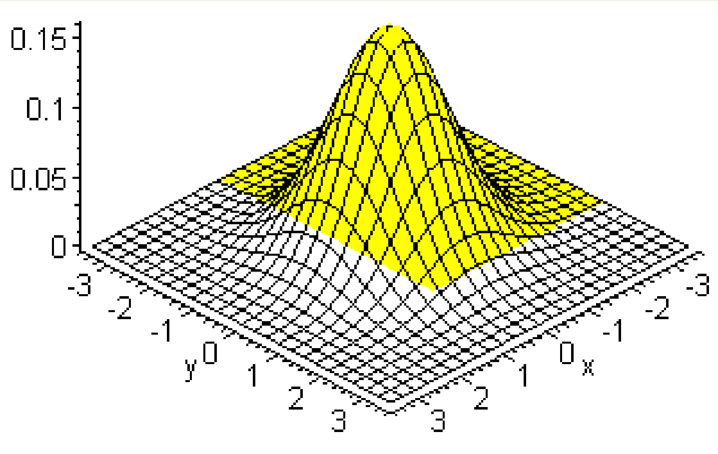
$$p(3,5) = \mathbb{P}(X = 3, Y = 5) = \frac{1}{36}.$$

- The joint pmf is uniformly:

$$p(x,y) = \mathbb{P}(X = x, Y = y) = \begin{cases} \frac{1}{36} & \text{for } x = 1,2,\ldots,6 \\ & \text{and } y = 1,2,\ldots,6 \\ 0 & \text{otherwise} \end{cases}$$

Example (Joint density function)

The probability of simultaneously $-4 < X \leq 1$ and $-4 < Y \leq 2$ is the yellow **volume** under the joint density.



Marginals

- Describe the randomness of X or Y given a joint distribution of (X, Y)
- **Discrete**: The marginal mass function of X is given by:

$$p_X(x) = \mathbb{P}\{X = x\} = \sum_y p(x, y)$$

- **Continuous**: The marginal density function of X is given by:

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

- **Both:** The marginal distribution function of X is given by:

$$F_X(x) = \mathbb{P}(X \leq x), \qquad F_X(x) = F(x, \infty)$$

Conditioning

Example

- When rolling a fair dice twice the sample space is $\Omega = \{(1,1), (1,2), \dots, (6,5), (6,6)\}$.
- Each of the 36 possible outcomes in Ω is equally likely to occur and hence has probability $\frac{1}{36}$.

Conditioning

Example

- When rolling a fair dice twice the sample space is $\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$.
- Each of the 36 possible outcomes in Ω is equally likely to occur and hence has probability $\frac{1}{36}$.
- Suppose now that the first dice shows a **4**. What is the probability that the sum of the two dice is equal to **6**?

Conditioning

Example

- When rolling a fair dice twice the sample space is $\Omega = \{(1,1),(1,2),\dots,(6,5),(6,6)\}$.
- Each of the 36 possible outcomes in Ω is equally likely to occur and hence has probability $\frac{1}{36}$.
- Suppose now that the first dice shows a **4**. What is the probability that the sum of the two dice is equal to **6**?
- This is a conditional probability, as it is conditional upon the occurrence of a prior event (**4** on the roll of the first dice).

Conditioning

Example

- When rolling a fair dice twice the sample space is $\Omega = \{(1,1),(1,2),\dots,(6,5),(6,6)\}$.
- Each of the 36 possible outcomes in Ω is equally likely to occur and hence has probability $\frac{1}{36}$.
- Suppose now that the first dice shows a **4**. What is the probability that the sum of the two dice is equal to **6**?
- This is a conditional probability, as it is conditional upon the occurrence of a prior event (**4** on the roll of the first dice).
- There are now 6 equally likely allowed outcomes: $(4,1)$, $(4,2)$, $(4,3)$, $(4,4)$, $(4,5)$, $(4,6)$.

Conditioning

Example

- When rolling a fair dice twice the sample space is $\Omega = \{(1,1),(1,2),\dots,(6,5),(6,6)\}$.
- Each of the 36 possible outcomes in Ω is equally likely to occur and hence has probability $\frac{1}{36}$.
- Suppose now that the first dice shows a **4**. What is the probability that the sum of the two dice is equal to **6**?
- This is a conditional probability, as it is conditional upon the occurrence of a prior event (**4** on the roll of the first dice).
- There are now 6 equally likely allowed outcomes: $(4,1)$, $(4,2)$, $(4,3)$, $(4,4)$, $(4,5)$, $(4,6)$.
- Only one of them, $(4,2)$ gives a sum equal to **6**. Thus, the answer is $\frac{1}{6}$.
- This is different from the unconditional probability of getting a sum of 6 which is $\frac{5}{36}$!

Definition

Let A end B be two events in a sample space Ω with $\mathbb{P}(B) > 0$. The conditional probability of A given B is given by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

Definition

Let A end B be two events in a sample space Ω with $\mathbb{P}(B) > 0$. The conditional probability of A given B is given by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

- In a way conditioning changes the sample space from Ω to B. Hence, the normalisation by $\mathbb{P}(B)$.

Definition

Let A end B be two events in a sample space Ω with $\mathbb{P}(B) > 0$. The conditional probability of A given B is given by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

- In a way conditioning changes the sample space from Ω to B. Hence, the normalisation by $\mathbb{P}(B)$.
- It is easy to see that $\quad \mathbb{P}(A) = \mathbb{P}(A \mid \Omega)$.

Definition

Let A end B be two events in a sample space Ω with $\mathbb{P}(B) > 0$. The conditional probability of A given B is given by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

- In a way conditioning changes the sample space from Ω to B. Hence, the normalisation by $\mathbb{P}(B)$.
- It is easy to see that $\mathbb{P}(A) = \mathbb{P}(A \mid \Omega)$.

Definition

Two events A and B of a sample space Ω are **independent** if the knowledge that B has occurred does not affect the probability that A occurs, i..e.

Definition

Let A end B be two events in a sample space Ω with $\mathbb{P}(B) > 0$. The conditional probability of A given B is given by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

- In a way conditioning changes the sample space from Ω to B. Hence, the normalisation by $\mathbb{P}(B)$.
- It is easy to see that $\mathbb{P}(A) = \mathbb{P}(A \mid \Omega)$.

Definition

Two events A and B of a sample space Ω are **independent** if the knowledge that B has occurred does not affect the probability that A occurs, i..e.

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

or, equivalently,

$$P(A \cap B) = P(A)\, P(B) \tag{2}$$

Otherwise, A and B are called **dependent**.

Important theorems

Let $\Omega = \bigcup_k B_k$ with pairwise exclusive B_k. Then

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k).$$

Law of total probability

Given A and $B_1, \ldots, B_k$ as above we have

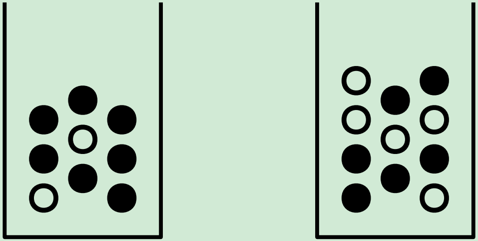
$$\mathbb{P}(A) = \mathbb{P}(A \mid B_1)\, \mathbb{P}(B_1) + \mathbb{P}(A \mid B_2)\, \mathbb{P}(B_2) + \cdots + \mathbb{P}(A \mid B_k)\, \mathbb{P}(B_k).$$

Bayes' rule

For any two events A, B of the sample space Ω with $\mathbb{P}(B) \neq 0$ we have

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

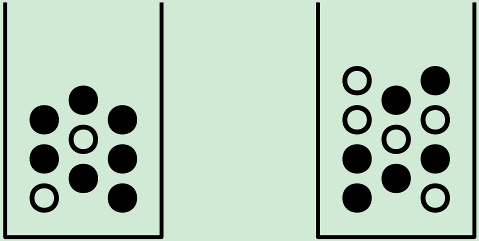
Example



We flip a fair coin.

- head: we draw ball from left urn
- tail: we draw ball from right urn

Example

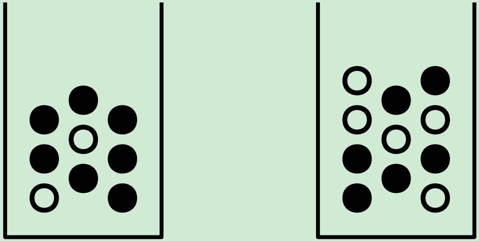


We flip a fair coin.

- head: we draw ball from left urn
- tail: we draw ball from right urn

What is the conditional probability that the outcome of the toss was "head" given that a white ball was drawn?

Example



We flip a fair coin.

- head: we draw ball from left urn
- tail: we draw ball from right urn

What is the conditional probability that the outcome of the toss was "head" given that a white ball was drawn?

Answer: $\frac{\frac{2}{9}}{\frac{67}{198}} \frac{1}{2} = \frac{22}{67}$

Independent random variables

We generalize the concept of independent events to random variables:

Definition

The random variables X and Y are independent if the following equivalent conditions hold for all x and y

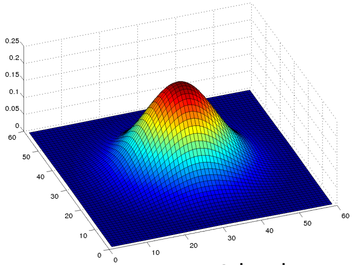
- $\mathbb{P}(X \leq x, Y \leq y) \;=\; \mathbb{P}(X \leq x)\, \mathbb{P}(Y \leq y)$.
- The joint distribution function of X and Y is separable

$$F(x, y) = F_X(x)\; F_Y(y).$$

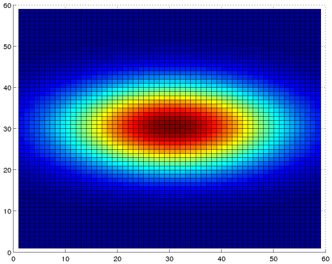
- The probability mass/density functions is separable

$$f(x, y) = f_X(x)\; f_Y(y).$$

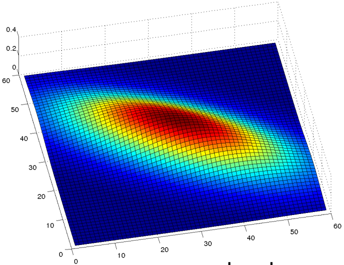
Example



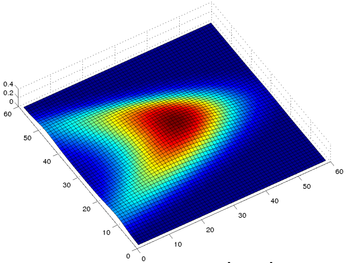
independent



independent



dependent



dependent

Conditional mass and density function

Definition

For two random variables X and Y with joint probability mass / density function $f(x,y)$ the conditional probability mass / density function of X given $Y=y$, is defined for all values of Y with $f_Y(y)>0$, by

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}.$$

Conditional mass and density function

Definition

For two random variables X and Y with joint probability mass / density function $f(x, y)$ the conditional probability mass / density function of X given $Y = y$, is defined for all values of Y with $f_Y(y) > 0$, by

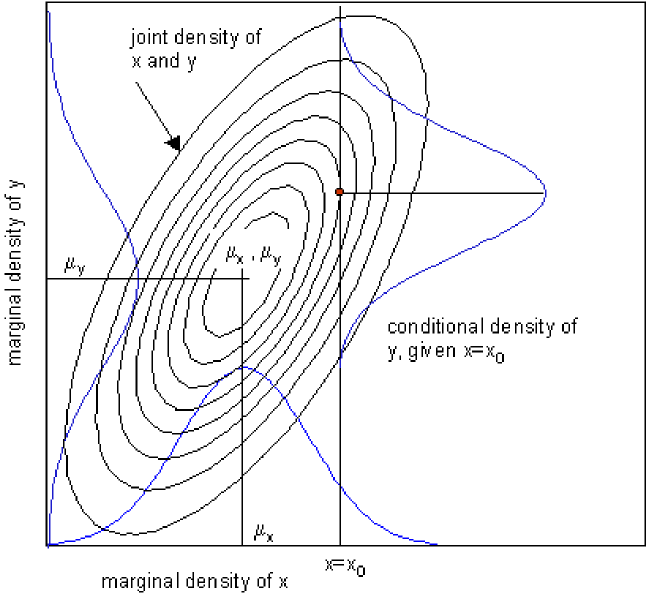
$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}.$$

- In general:

$$f_{Y|X}(y|x) \neq f_{X|Y}(x|y)$$

- If X and Y are independent, conditioning has no effect:

$$f_{Y|X}(y|x) = f_Y(y), \qquad f_{X|Y}(x|y) = f_X(x).$$


joint density of
x and y
μ_y
μ_x, μ_y
marginal density of y
conditional density of
y, given $x=x_0$
μ_x
$x=x_0$
marginal density of x

What is the average value we can expect for X, when $Y = y$?

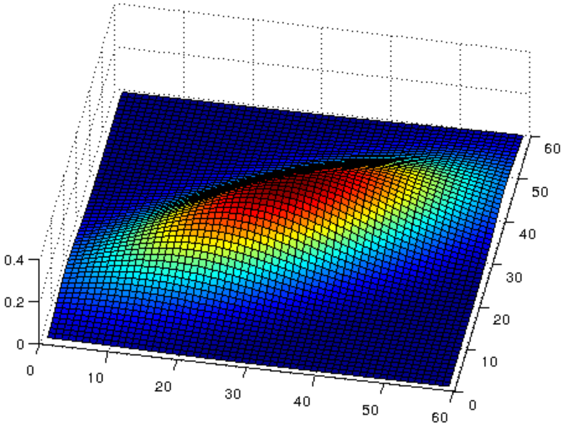
Definition

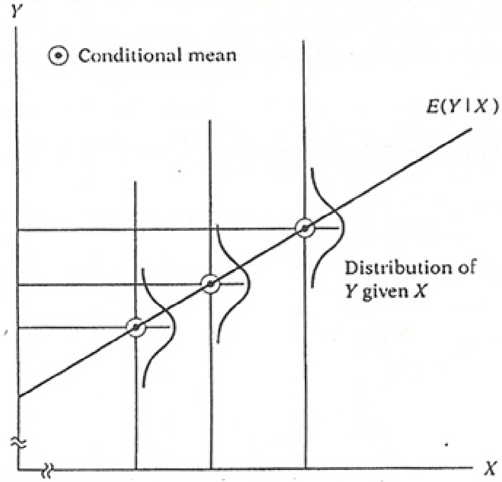
The conditional expectation of X, given that $Y = y$, is defined for all values of Y, such that $f_Y(y) > 0$, by

$$\mathbb{E}[X|Y = y] = \sum_x x\, f_{X|Y}(x|y) \qquad \text{(discrete)}$$

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\,\mathrm{d}x \qquad \text{(continuous)}$$

- It is also called conditional mean.
- The properties of $\mathbb{E}[X]$ also hold for $\mathbb{E}[X \mid Y = y]$ w.r.t. X.





- Taking the conditional expectation of $f_{Y|X}(y|x)$ for some x.
- This is not always a line!