Statistical Methods
(summer term 2024)

Probability and probability distributions

(based on original lectures by Prof. Dr. N. Christlieb and Dr. Hans-G. Ludwig)

Dr Yiannis Tsapras

ZAH – Heidelberg

Overview

- Concept of probability

- A bit of probability calculus

- Common probability distributions

- Law of large numbers

Let us try to define some terms

- **random experiment**: a mechanism that produces a definite outcome that cannot be predicted with certainty

- **sample space**:The collection of all possible outcomes or results of a **random experiment**
  e.g. tossing a coin 3 times giving head or tails. The sample space here is

  $$\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$$

- **event**: A subset of the sample space
  e.g. the event A of obtaining two heads (the outcome observed) is

  $$A = \{hht, hth, thh\}$$

- **probability measure**: a mapping from the power set of Ω to the interval [0,1]. i.e. each subset of Ω has an associated probability and their sum is 1

- **random variable**: a function that maps a set of events to associated probabilities
  e.g. the random variable X defined as number of heads in 3 tosses of a coin

Setting the probability

- **Frequentist** approach: The probability of an event A, denoted as $\mathrm{P}(A)$, is
  - defined as the long-run relative frequency of the event A occurring in repeated trials of a random experiment (actually tossing coins).
  - relies on the notion that probabilities are objective and can be determined through repeated experimentation and observation.

- **Bayesian** approach: $\mathrm{P}(A)$ is interpreted as a degree of belief in the occurrence of the event, given prior knowledge or evidence
  - This contrasts with the frequentist interpretation, where probability is viewed as the long-run relative frequency of the event occurring
  - Remark: $\mathrm{P}(A)$ is called the prior probability (distribution) of the event A, and it is updated every time when new data comes in

Problems of probability definitions

- **Frequentist** approach:
  - Implicitly assumes that all elements of the sample space are equally likely (*Laplace's principle of indifference*). Is this always true?
  - Relies on a random experiment that is carried out a large number of times and under the same conditions. Is this feasible?

- **Bayesian** approach:
  - "Belief" is subjective (but quantifiable). Probabilities therefore represent degrees of belief rather than long-run frequencies.
  - Prior probabilities (beliefs) can be updated based on new evidence.
  - In most cases, the results do not critically depend on the particular assumptions on the prior probability.

- **Resolution(?) through Kolmogorov Axioms:** These provide a consistent logical foundation for probability theory that is applicable to both frequentist and Bayesian interpretations.

Kolmogorov axioms

- **Non-negativity:** For any event A, the probability is positive: $0 \leq \mathrm{P}(A) \leq 1$

- **Normalization:** The probability that at least one of the events in the entire sample space S will occur: $\mathrm{P}(S) = 1$

- **Additivity:** For any two mutually exclusive events, A and B: $\mathrm{P}(A \,\mathrm{or}\, B) \equiv \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

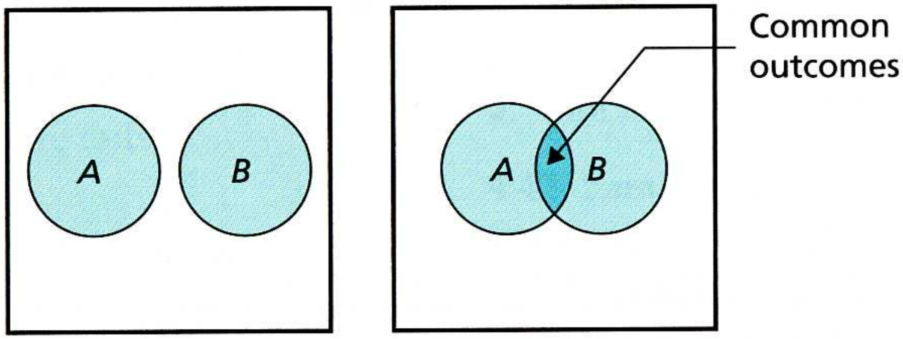


Illustration via sets

Example for throwing dice:

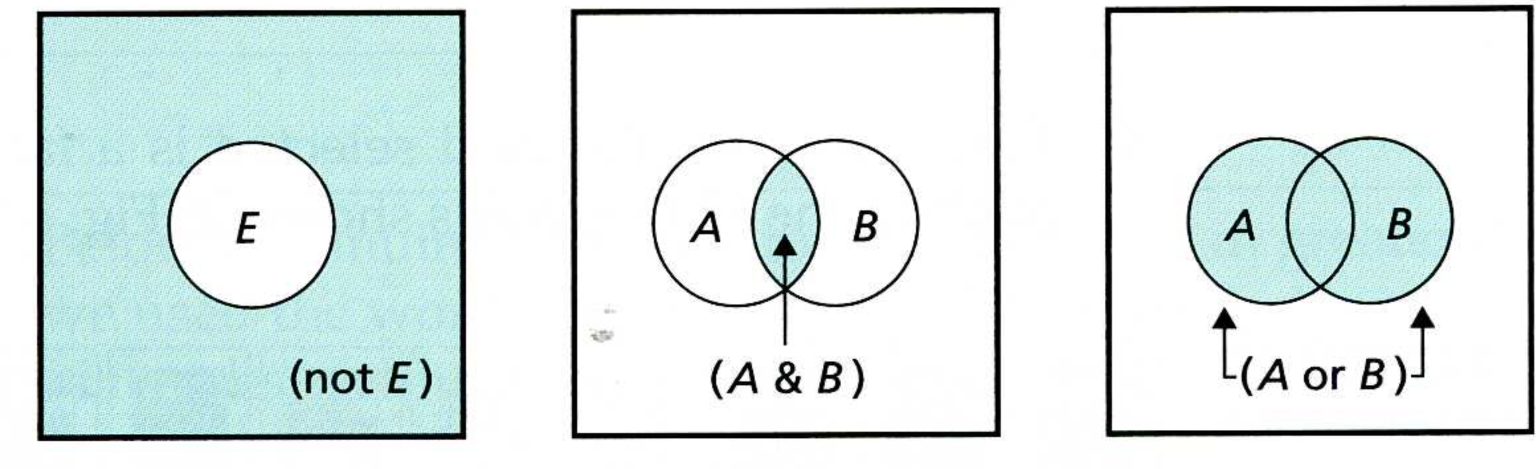
- Change of specific outcome (e.g. 1): $1/6$

- Chance for each of the sides coming up: $1/6$

- Chance of rolling a 1 or 2 (these are exclusive events!): $\mathrm{P}(1 \cup 2) = 2/6 = 1/6 + 1/6$

More on probability calculus

- Probability of complementary event: $\mathrm{P}(\text{not}\, E) = \mathrm{P}\big(\overline{E}\big) = 1 - \mathrm{P}(E)$
  - Note: The probability of the complement of event E can be also written in other ways: $\mathrm{P}(E^c) = 1 - \mathrm{P}(E)$ or $\mathrm{P}(\neg E) = 1 - \mathrm{P}(E)$
  - Both expressions represent the probability that event E does not occur.



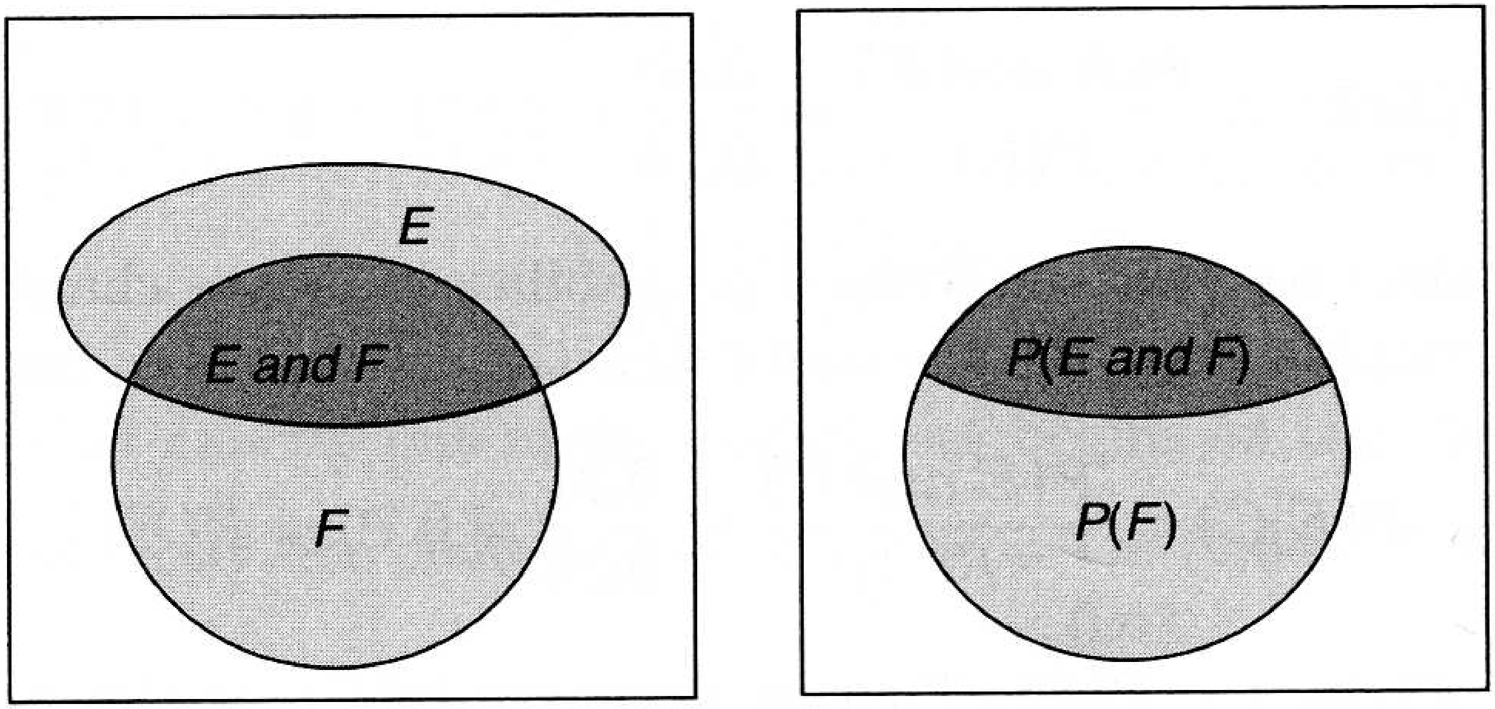
- If $\{E_i\}$ is a set of mutually exclusive and exhaustive events (i.e., includes *all* elementary events), then

$$\sum_{i=1}^{n} \mathrm{P}(E_i) = 1$$

Conditional probability

- Let E and F be two events with $\mathrm{P}(F) > 0$. Then the *conditional probability* of E *given* that F has occurred is

$$\mathrm{P}(E \,|\, F) = \frac{\mathrm{P}(E \,\mathrm{and}\, F)}{\mathrm{P}(F)} \equiv \frac{\mathrm{P}(E \cap F)}{\mathrm{P}(F)}$$



Example:

- Suppose we have a deck of 52 playing cards. We want to find the probability of drawing an Ace given that we have already drawn a card from the deck and it is a Spade (♠) (which we saw by looking only at the tip of the card).

- The probability of drawing an Ace given that we have already drawn a Spade is given by:

$$P(\text{Ace}|\text{Spade}) = \frac{P(\text{Ace} \cap \text{Spade})}{P(\text{Spade})}$$

- $P(\text{Ace} \cap \text{Spade})$ is the probability of drawing the Ace of Spades: $\frac{1}{52}$

- $P(\text{Spade})$ is the probability of drawing any Spade: $\frac{13}{52}$

- Therefore

$$P(\text{Ace}|\text{Spade}) = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}$$

Corollary: joint probability

- **Special case:** If events A and B are $independent$ (i.e. do not affect each others probability of happening), then: $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$

  For example in successive dice rolls, the score of one roll does not affect subsequent rolls

  In such cases the events can be considered independent, but we cannot assume this is always true

- In conditional probabilities the order of the events generally does not matter

$$\mathrm{P}(A \,|\, B) = \frac{\mathrm{P}(A \,\mathrm{and}\, B)}{\mathrm{P}(B)} \equiv \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

$$\iff$$

$$\mathrm{P}(A \,\mathrm{and}\, B) \equiv \mathrm{P}(A \cap B) \equiv \mathrm{P}(A, B) = \mathrm{P}(A \,|\, B)\,\mathrm{P}(B) = \mathrm{P}(B \,|\, A)\,\mathrm{P}(A)$$

- Has to be like this since the logical “and” is commutative. Note, the writing of the “and” with comma or using the intersection symbol ($\cap$) from standard Set theory.

Law of total probability

- Since the event B either happens or not, we can write

$$\mathrm{P}(A) = \mathrm{P}(A \,\mathrm{and}\, B) + \mathrm{P}\big(A \,\mathrm{and}\, \overline{B}\big) = \mathrm{P}(A \,|\, B)\,\mathrm{P}(B) + \mathrm{P}\big(A \,|\, \overline{B}\big)\,\mathrm{P}\big(\overline{B}\big)$$

- This can be generalized:
  If $\{B_i\}$ is the set of all possible (mutually exclusive) events then

$$\mathrm{P}(A) = \sum_{i=1}^{n} \mathrm{P}(A \cap B_i) = \sum_{i=1}^{n} \mathrm{P}(A \,|\, B_i)\,\mathrm{P}(B_i)$$

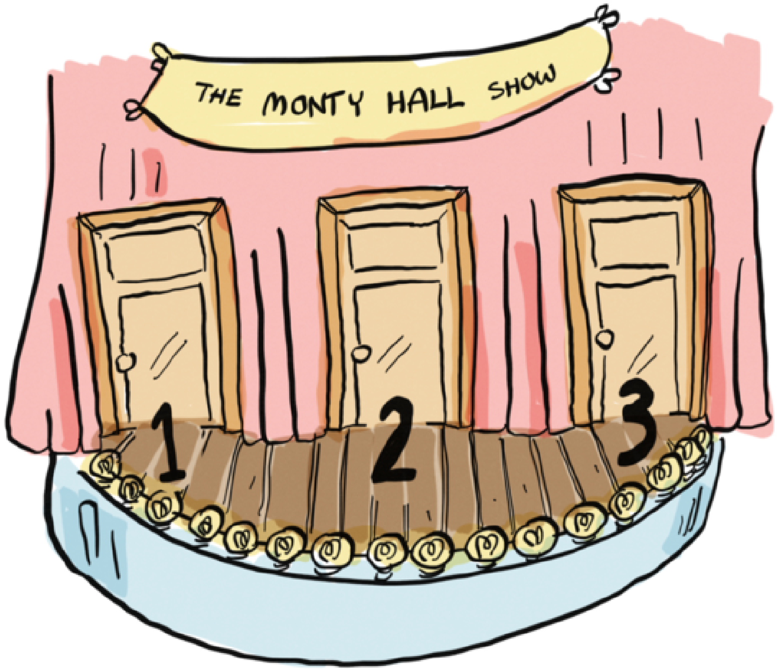
This relatio is called the **law of total probability**.

$\rightarrow$Blackboard

The Monty Hall Problem

- Consider the following scenario: You are a contestant in a game show and you are presented with three closed doors, behind only one of which is a prize (the other two have nothing behind them). You are asked to pick a door. Then, the host opens one of the OTHER two doors, which has nothing behind it. You are offered a choice of staying with the door you originally picked or switching to the other remaining door. Whichever door you choose will be your final choice.

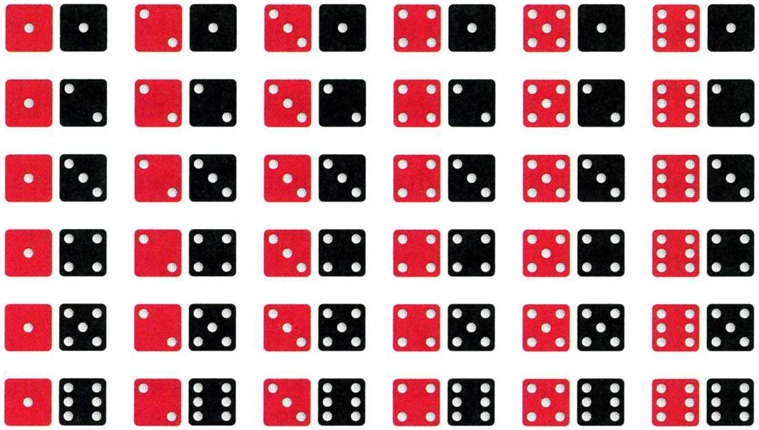
  Should you stay with the door you chose or switch? Does it even matter?



Example: law of total probability

Two dice A and B are tossed simultaneously. What is the probability that A=1?

The probability for each of the possible combinations is 1/36. Therefore ...



$$
\begin{aligned}
\mathrm{P}(A=1) &= \sum_{i=1}^{n} \mathrm{P}(A=1 \,|\, B_i)\,\mathrm{P}(B_i) \\
&= \sum_{i=1}^{n} \mathrm{P}(A=1 \text{ and } B_i) \\
&= \mathrm{P}(A=1 \text{ and } B=1) + \mathrm{P}(A=1 \text{ and } B=2) + \ldots \\
&= 6 \times \frac{1}{36} = \frac{1}{6}
\end{aligned}
$$

In this example the answer was easy to arrive at from the outset. However, there are more complex situations where it is not as obvious!

Independent events

- Two events A and B are *independent* if

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

  i.e. the probability of event A is independent of the occurrence of event B (and vice versa)

- As we saw before, two events A and B are independent if and only if

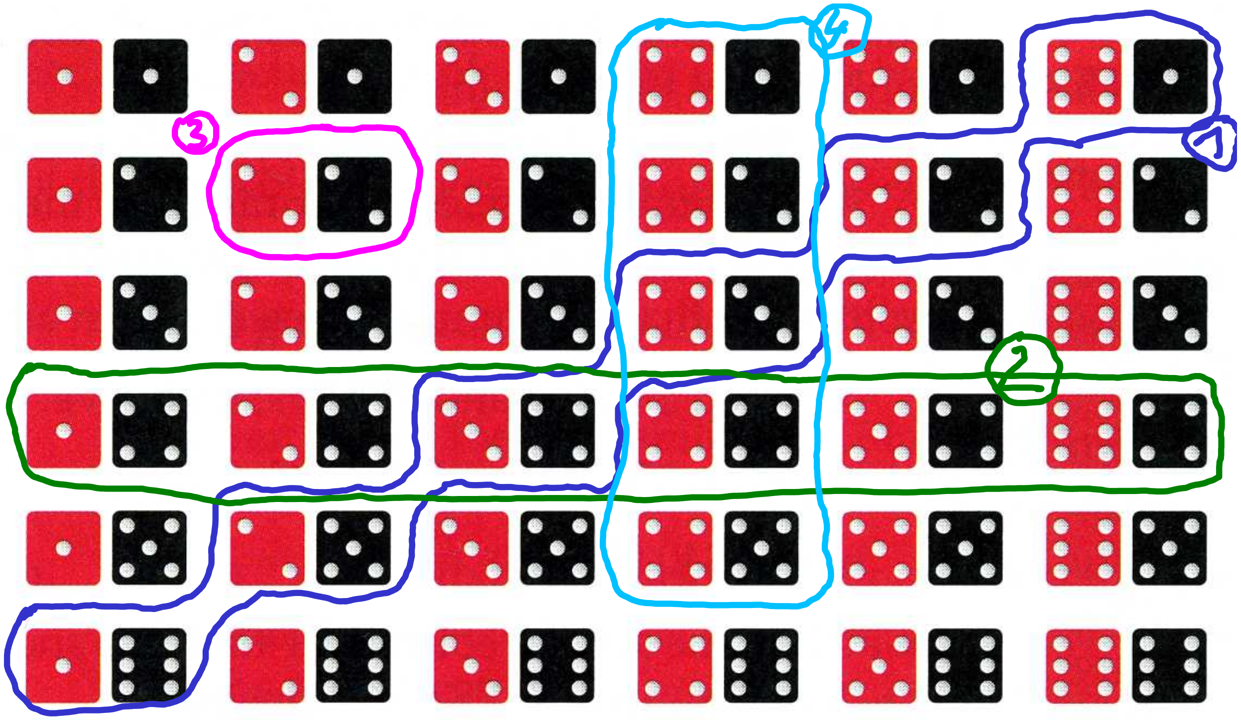
  $$\mathrm{P}(A \,\mathrm{and}\, B) \equiv \mathrm{P}(A \cap B) = \mathrm{P}(A)\ \mathrm{P}(B)$$

- In the case of independent events, we can simply multiply their probabilities to get the probability of their intersection, which makes things easier.

Example: independent events

Sample space of tossing zwo dice considered with four different events. Note, similarity with Venn diagram.

- event 1: sum is 7
- event 2: black die shows 4
- event 3: double of 2s
- event 4: red die shows 4, black die 1 to 5



- This means: $P(1) = 6/36$, $P(2) = 6/36$, $P(3) = 1/36$, $P(4) = 5/36$
- Are events 3 and 4 independent?
- Are events 1 and 2 independent? →Blackboard

Bayes' theorem, or law of inverse probability

Joint probability can be arranged as

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

This expression is one of the most important expressions in probability theory. It looks mostly harmless but becomes far more interesting if we substitute 'data' for A and 'model/hypothesis' for B:

$$\mathrm{P}(\mathrm{model} \,|\, \mathrm{data}) = \frac{\mathrm{P}(\mathrm{data} \,|\, \mathrm{model})\ \mathrm{P}(\mathrm{model})}{\mathrm{P}(\mathrm{data})} = \frac{\mathrm{P}(\mathrm{data} \,|\, \mathrm{model})\ \mathrm{P}(\mathrm{model})}{\sum_{i=1}^{n} \mathrm{P}(\mathrm{data} \,|\, \mathrm{model}_i)\,\mathrm{P}(\mathrm{model}_i)}$$

- We can derive the probability of the model (being the correct one) given the data. $\Longrightarrow$ **Bayesian inference**
- The individual terms have names: $\mathrm{P}(B)$ is called the *prior*, $\mathrm{P}(A \,|\, B)$ is the *likelihood*, $\mathrm{P}(B \,|\, A)$ is the *posterior* and $\mathrm{P}(A)$ the *evidence*
- The term *inverse probability* is motivated by the fact that the conditional probabilities of A and B are interchanged →blackboard

Example: The patient screening test

Consider the following scenario:
A serious disease affects about 1% of population. You are worried that you might be infected so you go to see a doctor to test whether you have it. The doctor performs a screening test that has 95% sensitivity –that is, 95% of people who have the disease test positive (true positive rate) – and 99% specificity – that is 99% of the healthy people test negative (true negative rate).

Question: If you test positive, what are the chances that you have the disease?

Using Bayes' Theorem for the patient screening test

- Let D be the event of having the disease and T the event of testing positive

- Only 1% of the population have the disease: $\mathrm{P}(D) = 0.01$ and $\mathrm{P}(\neg D) = 0.99$

- Sensitivity: $\mathrm{P}(T|D) = 0.95$ and Specificity: $\mathrm{P}(\neg T|\neg D) = 0.99$, so $\mathrm{P}(T|\neg D) = 0.01$

- **Bayes' Theorem**: $P(D|T) = \frac{P(T|D) \cdot P(D)}{P(T)}$

- **Calculate** $P(T) = P(T|D) \cdot P(D) + P(T|\neg D) \cdot P(\neg D) = 0.0194$

- **Calculate** $P(D|T) = \frac{0.95 \times 0.01}{0.0194} \approx 0.49$
  - Despite the test's high sensitivity and specificity, the probability that you have the disease given a positive result is approximately 49% (about half the people who test positive do not have the disease)
  - This is due to the low prevalence of the disease in the general population
  - What if you take another test a few days later and test positive again?

Tested positive twice? Bad news

- Let T_1 and T_2 be the events of testing positive on the first and second tests, respectively. We want to evaluate $P(D|T_1 \cap T_2)$

- $\mathrm{P}(T_1|D) = \mathrm{P}(T_2|D) = 0.95$ and $\mathrm{P}(T_1|\neg D) = \mathrm{P}(T_2|\neg D) = 0.01$

- We already calculated $P(D|T_1) \approx 0.49$

- The probability we test positive a second time after we already tested positive the first time is $\mathrm{P}(T_2|T_1) = P(T_2|D) \cdot P(D|T_1) + P(T_2|\neg D) \cdot P(\neg D|T_1)$
  - The probability that we don't have the disease if we tested positive the first time is $\mathrm{P}(\neg D|T_1) = 1 - P(D|T_1) = 1 - 0.49 = 0.51$
  - Substituting back we get $\mathrm{P}(T_2|T_1) \approx 0.47$

- $P(D|T_1 \cap T_2) = \frac{P(T_2|D) \cdot P(D|T_1)}{P(T_2|T_1)} = \frac{0.95 \times 0.49}{0.47} \approx 0.99$
  - If you test positive twice, the probability that you have the disease is approximately 99%!

Univariate probability distributions

- A univariate probability distribution describes the probability of outcomes for a single random variable

- These distributions can be $\textit{discrete}$ or $\textit{continuous}$, depending on the nature of the random variable

- A $\textit{discrete}$ probability distribution is applicable to $\textit{discrete}$ random variables, i.e. variables that take a countable number of distinct values

- The **probability mass function** (PMF), $f(x) = \mathrm{P}(X = k)$, gives the probability that a $\textit{discrete}$ random variable X is exactly equal to some value k. Here, k is a value from the range of values available to X (e.g. $1, 2, 3, ...$)

Examples of $\textit{discrete}$ probability distributions:

- **Binomial Distribution**: $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

- **Poisson Distribution**: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Binomial distribution

- Represents the number of successes in a fixed number of independent Bernoulli trials (e.g., number of heads in 10 coin flips). [A Bernoulli trial is a random experiment with exactly two possible outcomes.]

$$f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- $P(X = k)$ represents the probability of obtaining exactly k successes out of n trials.
- The term $\binom{n}{k}$, read 'n choose k', is called the **binomial coefficient** and represents the number of ways to choose k successes from n trials: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- p^k is the probability of having k successes, where p just represents the probability of having a success in each trial
- $(1-p)^{n-k}$ is the probability of having $n-k$ failures, where $(1-p)$ represents the probability of failure in each trial
- has an expectation value (mean) $\mathrm{E}[k] = np$
- has variance $\mathrm{Var}[k] = np(1-p)$

→Notebook

Poisson distribution

- Represents the number of events occurring in a fixed interval of time or space, given a constant mean rate of occurrence (e.g., number of emails received in an hour).

$$f(k) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- $P(X = k)$ represents the probability of observing exactly k events in a fixed interval
- The term on the right, $\frac{\lambda^k e^{-\lambda}}{k!}$, is the Poisson probability formula:
  - ⋆ λ^k is the rate of occurrence raised to the power of k, where λ is the average number of events per interval
  - ⋆ $e^{-\lambda}$ is the exponential decay factor, accounting for the probability that fewer events occur as λ increases
  - ⋆ $k!$ is the number of ways to arrange k events
- has an expectation value (mean) $\mathrm{E}[k] = \lambda$
- has variance $\mathrm{Var}[k] = \lambda$

→Notebook

Univariate probability distributions

What about *continuous* random variables? These can take on an infinite number of possible values within a given range

- Let X represent a *continuous* random variable

The **probability density function** $f(x)$ gives the relative likelihood for X to take on a given value. The probability that X lies within the interval $[a, b]$ is

- $\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\, dx$. This is $= 1$ if the range covers the entire domain

- $f(x)$ is a single-valued non-negative number for all x

- f has as physical dimension the dimension of $1/x$
  e.g, probability per meter if x happens to be a length

In these lectures, *continuous* and *discrete* distributions typically are loosely referred to as *probability distribution*

Let us look at an example of a very important *continuous* probability distribution

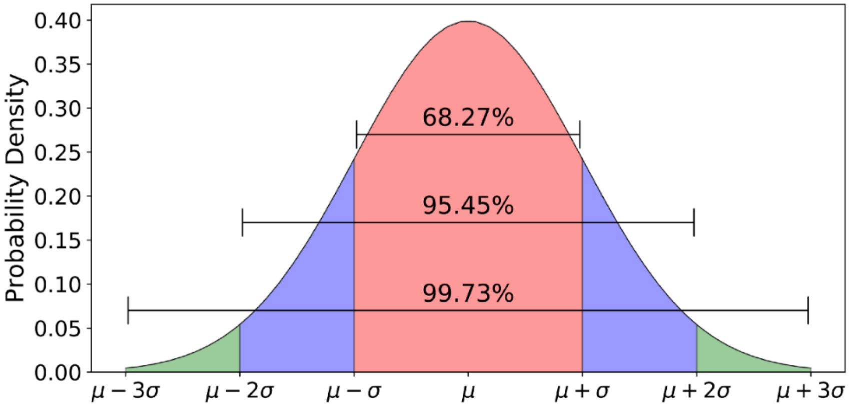
Normal (Gaussian) Distribution

- The Normal distribution, also known as the Gaussian distribution, is probably the most important *continuous* probability distribution because it approximates a wide variety of phenomena

- It is defined by its mean μ and standard deviation σ, which determine its centre and spread, respectively

- The probability density function (PDF) of the Normal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- Approximately 68, 95, 99.7% of the PDF lies within 1, 2, 3 σ of the mean, respectively

Cumulative distribution function (CDF)

The **Cumulative Distribution Function (CDF)** $F(x)$ of a random variable X is the probability that X will take a value less than or equal to x: $F(x) = \mathrm{P}(X \leq x)$

- For *discrete* distributions:

$$\mathrm{P}(X \leq x) = \sum_{x_i \leq x} \mathrm{P}(X = x_i)$$

- For *continuous* distributions:

$$\mathrm{P}(X \leq x) = \int\limits_{-\infty}^{x} f(t)dt$$

Note: t is a dummy variable of integration used to calculate the cumulative probability from $t = -\infty$ up to $t = x$.

Cumulative distribution function (CDF)

- The CDF is non-decreasing

- What is the difference between the CDF and the PDF?
  - The CDF is the probability that random variable has a value less than or equal to x: $F(x) \equiv \mathrm{P}(X \leq x)$
  - the PDF is the probability that a random variable X will take a value exactly equal to x: $f(x) \equiv \mathrm{P}(X = x)$.

- For *continuous* random variables, the PDF can be found by differentiating the CDF

- For *discrete* distributions, we can evaluate the PMF by using the fact that the PMF is the difference between consecutive values of the CDF
  - Specifically, the PMF at a point x_i can be obtained by subtracting the CDF value just before x_i from the CDF value at x_i: $P(X = x_i) = F(x_i) - F(x_{i-1})$

Example

- Consider a *discrete* random variable X with the following CDF:

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.2 & \text{if } 1 \leq x < 2 \\ 0.7 & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3 \end{cases}$$

- The corresponding PMF $P(X = x)$ is:

$$P(X = 1) = F(1) - F(0) = 0.2 - 0 = 0.2$$

$$P(X = 2) = F(2) - F(1) = 0.7 - 0.2 = 0.5$$
$$P(X = 3) = F(3) - F(2) = 1 - 0.7 = 0.3$$

→Notebook

Moments of a distribution

Moments are used to describe different characteristics of a probability distribution. In the following, integrals have to be taken over the domain of definition of the PDF

- Raw moments (calculated with respect to the origin:
  - The n-th raw moment is defined as: $\mu_n = \int x^n f(x)\, dx$
  - the first raw moment is the **mean** $\mu \equiv \mu_1 = \int x f(x)\, dx$
  - the mean is a scalar value, with the same physical dimension as x
- Central moments (calculated with respect to the mean)
  - The n-th central moment is defined as: $m_n \equiv \int (x - \mu)^n f(x)\, dx$
  - the second central moment is the **variance** $\sigma^2 \equiv m_2 = \int (x - \mu)^2 f(x)\, dx$
    - ⋆ measures the spread of the distribution around the mean
  - the third central moment is the **skewness** $\equiv m_3 / m_2^{(3/2)}$
    - ⋆ measures the asymmetry/lopsidedness of the distribution
  - the fourth central moment is the **kurtosis** $\equiv m_4 / m_2^2$
    - ⋆ measures the "peakedness" of the distribution

While the PDF (or CDF) encapsulates the complete information on a random variable, moments are often used to *summarize* a PDF. Note that many PDFs are fully characterized by a small set of moments

Quantiles (from inverse CDF)

Definition:

- A **quantile** is a value that divides the range of a probability distribution into continuous intervals with equal probabilities

- Quantiles provide a useful way to understand the distribution

- If $F(x)$ is the Cumulative Distribution Function (CDF) of a random variable X, the p-th quantile x_p is defined as:

  $$x_p = F^{-1}(p)$$

  where F^{-1} is the inverse of the CDF (the **quantile function**)

Types of Quantiles

- **Median:** The 0.5 quantile ($x_{0.5}$) divides the data into two equal parts
- **Quartiles:** Values that divide the data into four equal parts:
  - $x_{0.25}$: First quartile (Q1)
  - $x_{0.50}$: Second quartile (Q2), also the median
  - $x_{0.75}$: Third quartile (Q3)
- **Interquartile Range (IQR):** The difference between the third and first quartile (sometimes used as a measure for the width of a distribution):

$$\text{IQR} = x_{0.75} - x_{0.25}$$

- **Percentiles:** Values that divide the data into 100 equal parts
- **Deciles:** Values that divide the data into 10 equal parts

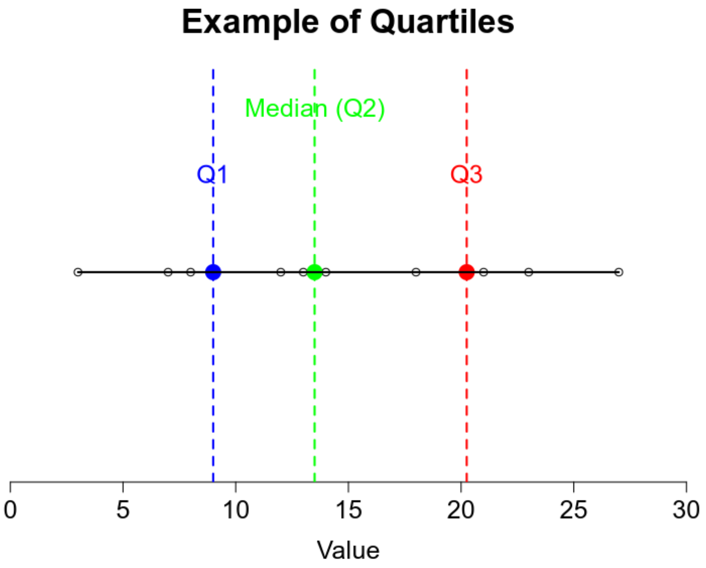
Example using Quartiles

Example Dataset:

$$\{3, 7, 8, 12, 13, 14, 18, 21, 23, 27\}$$

Calculate Quartiles:

- **Step 1:** Order the data from smallest to largest (already ordered)
- **Step 2:** Calculate the median $Q2 = \frac{13+14}{2} = 13.5$
- **Step 3:** Calculate the first quartile:
  - First half of the data: $\{3, 7, 8, 12, 13\}$
  - Q1 is the median of the first half: $Q1 = 8$
- **Step 4:** Calculate the third quartile:
  - Second half of the data: $\{14, 18, 21, 23, 27\}$
  - Q3 is the median of the second half: $Q3 = 21$



Most likely value (mode) vs. mean vs. median

- Mean: The arithmetic mean of the data

- Median: the middlemost value of the given ungrouped data if the data is arranged in ascending order

- Mode: the value that appears most often in the data

- Note: for an asymmetric probability distribution, the most likely value, mean and median are not identical

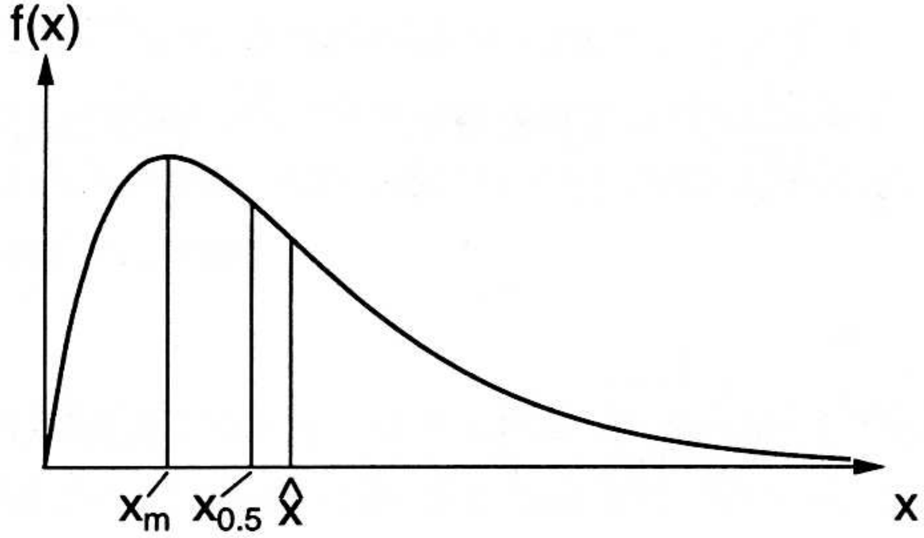


Bild 3.4: Wahrscheinlichster Wert x_m, Mittelwert $\widehat{x}$ und Median $x_{0.5}$ einer unsymmetrischen Verteilung.

Probability distributions in `R`

Replace _ in the list by . . .

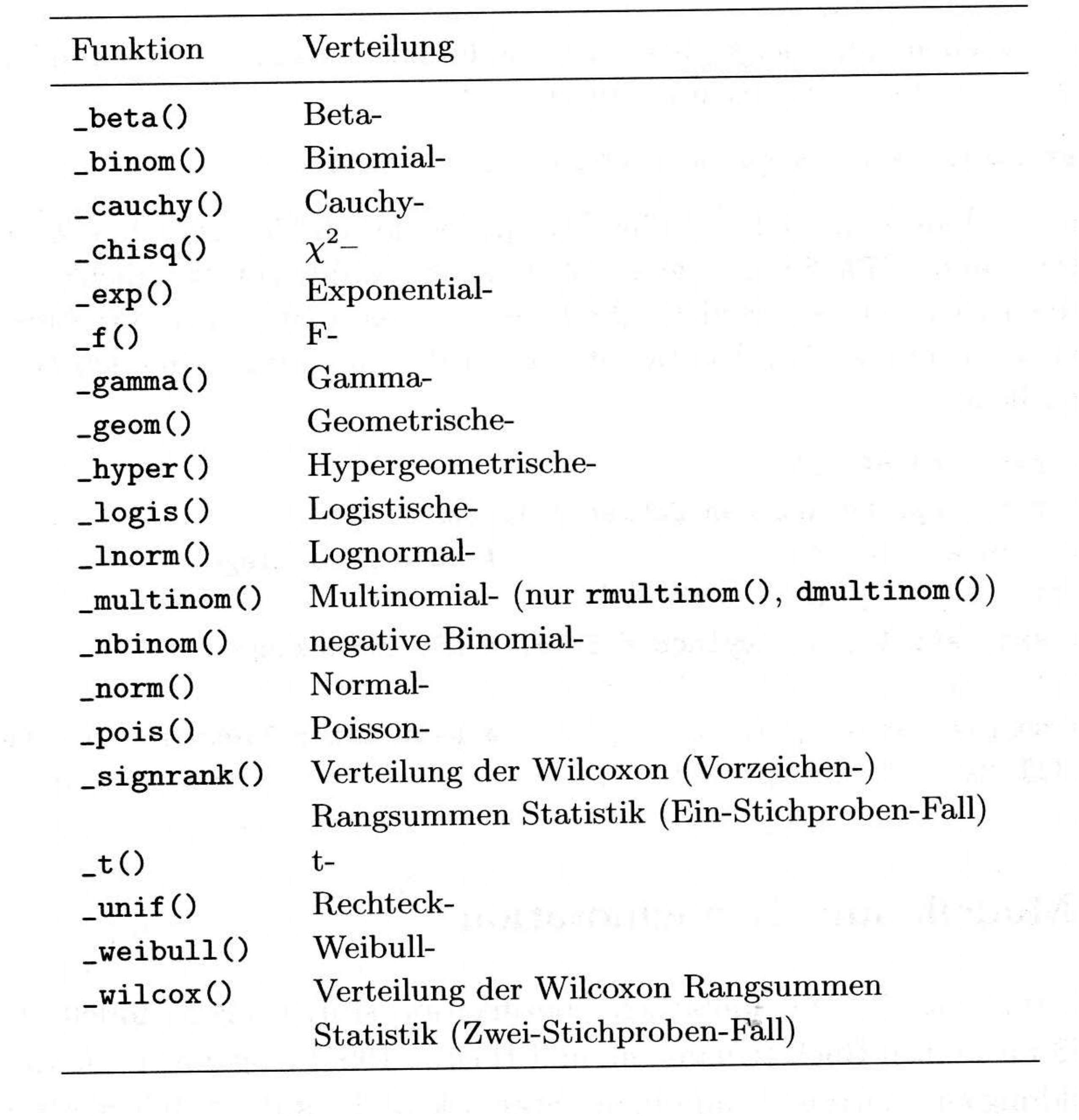
d: density of PDF

p: probability of CDF
$\int_{-\infty}^{x} \mathrm{PDF}(x')\, dx'$

q: quantile
→ inverse CDF

r: (random) generation of random numbers drawn from PDF

Example: `pnorm(-1)`
(by default, `pnorm` assumes a variance of one, and mean of zero)



Funktion	Verteilung
`_beta()`	Beta-
`_binom()`	Binomial-
`_cauchy()`	Cauchy-
`_chisq()`	χ^2–
`_exp()`	Exponential-
`_f()`	F-
`_gamma()`	Gamma-
`_geom()`	Geometrische-
`_hyper()`	Hypergeometrische-
`_logis()`	Logistische-
`_lnorm()`	Lognormal-
`_multinom()`	Multinomial- (nur `rmultinom()`, `dmultinom()`)
`_nbinom()`	negative Binomial-
`_norm()`	Normal-
`_pois()`	Poisson-
`_signrank()`	Verteilung der Wilcoxon (Vorzeichen-) Rangsummen Statistik (Ein-Stichproben-Fall)
`_t()`	t-
`_unif()`	Rechteck-
`_weibull()`	Weibull-
`_wilcox()`	Verteilung der Wilcoxon Rangsummen Statistik (Zwei-Stichproben-Fall)

Law of large numbers

- Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$. Let the sample mean be $\overline{X}_n \equiv 1/n \sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$ the **law of large numbers** states

$$\lim_{n\to\infty} \mathrm{P}\big(\big|\overline{X}_n - \mu > \epsilon\big|\big) \to 0\,.$$

- In words: As the sample size n increases, the probability that the sample mean $\overline{X}_n$ deviates more than ϵ from the true mean μ approaches zero

- More simply: For a large enough sample size, the sample mean $\overline{X}_n$ will be very close to the true mean of the distribution μ

- Important application: the expectation value of a random variable can be approximated by the arithmetic mean

- In practice, one has to average over $many$ events to get close to μ