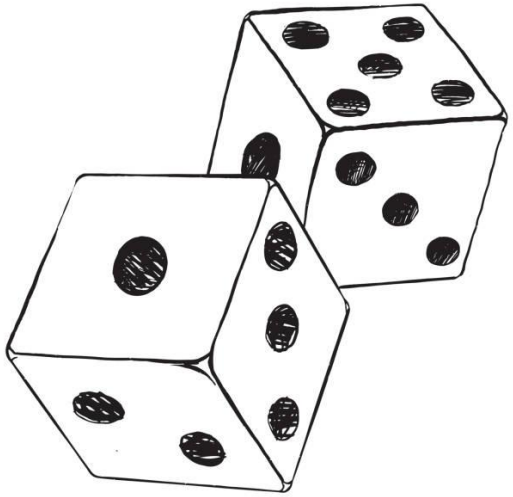
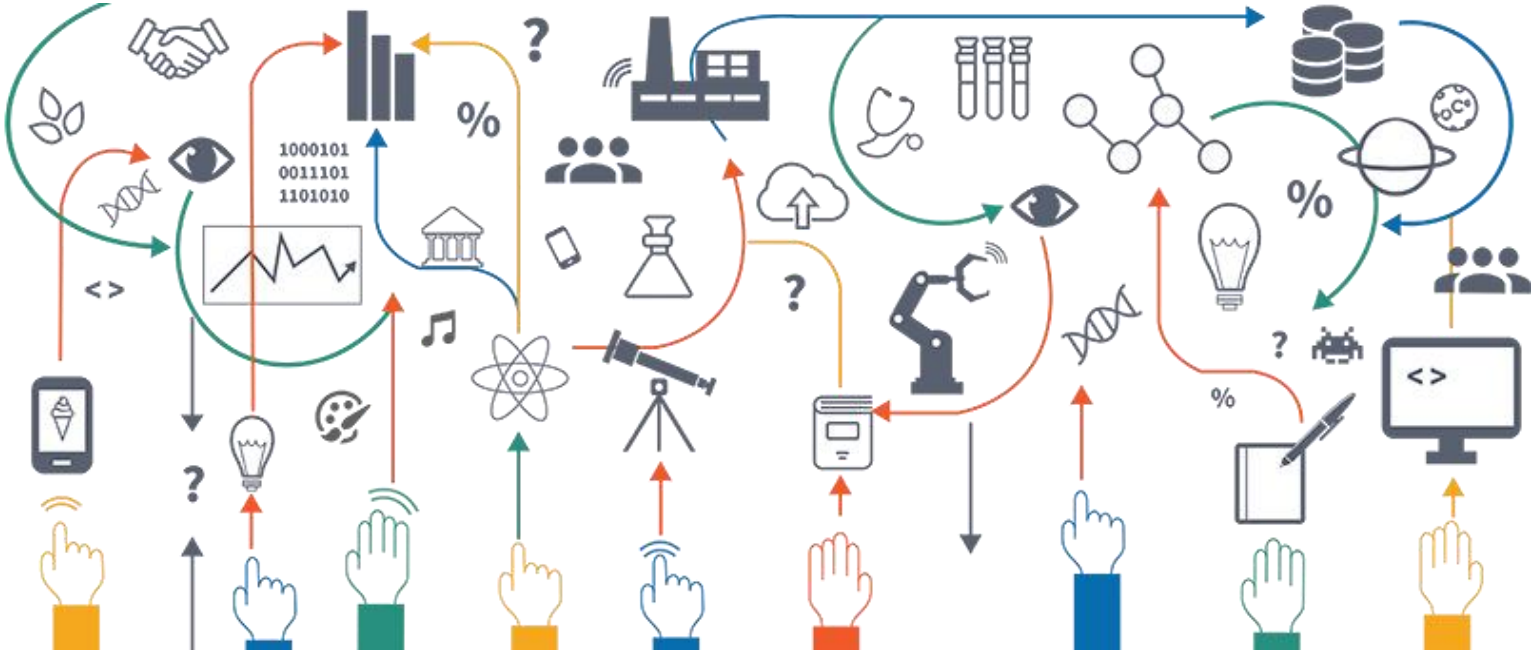
1 — Foundations of Probability Theory

Mathematics of Data Science



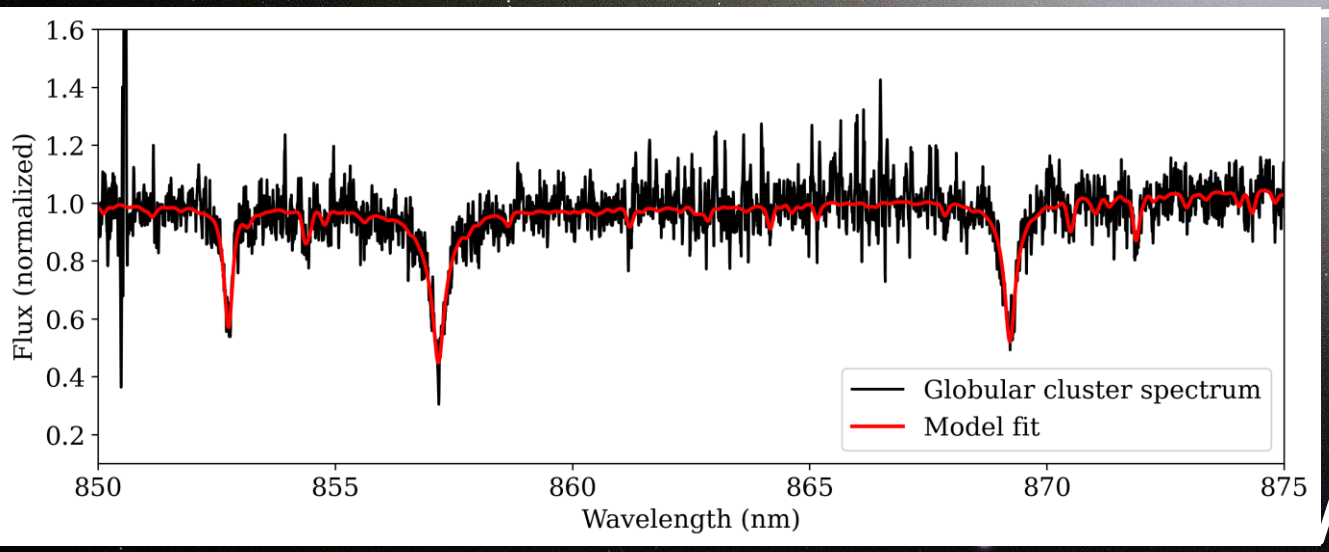
Lecturer: Dominik Dold



University of Vienna, WiSe 2025

Master's programme in Data Science

The world, and thus data, is inherently probabilistic



Credit: Katja Fahrion, Wikipedia (Hubble

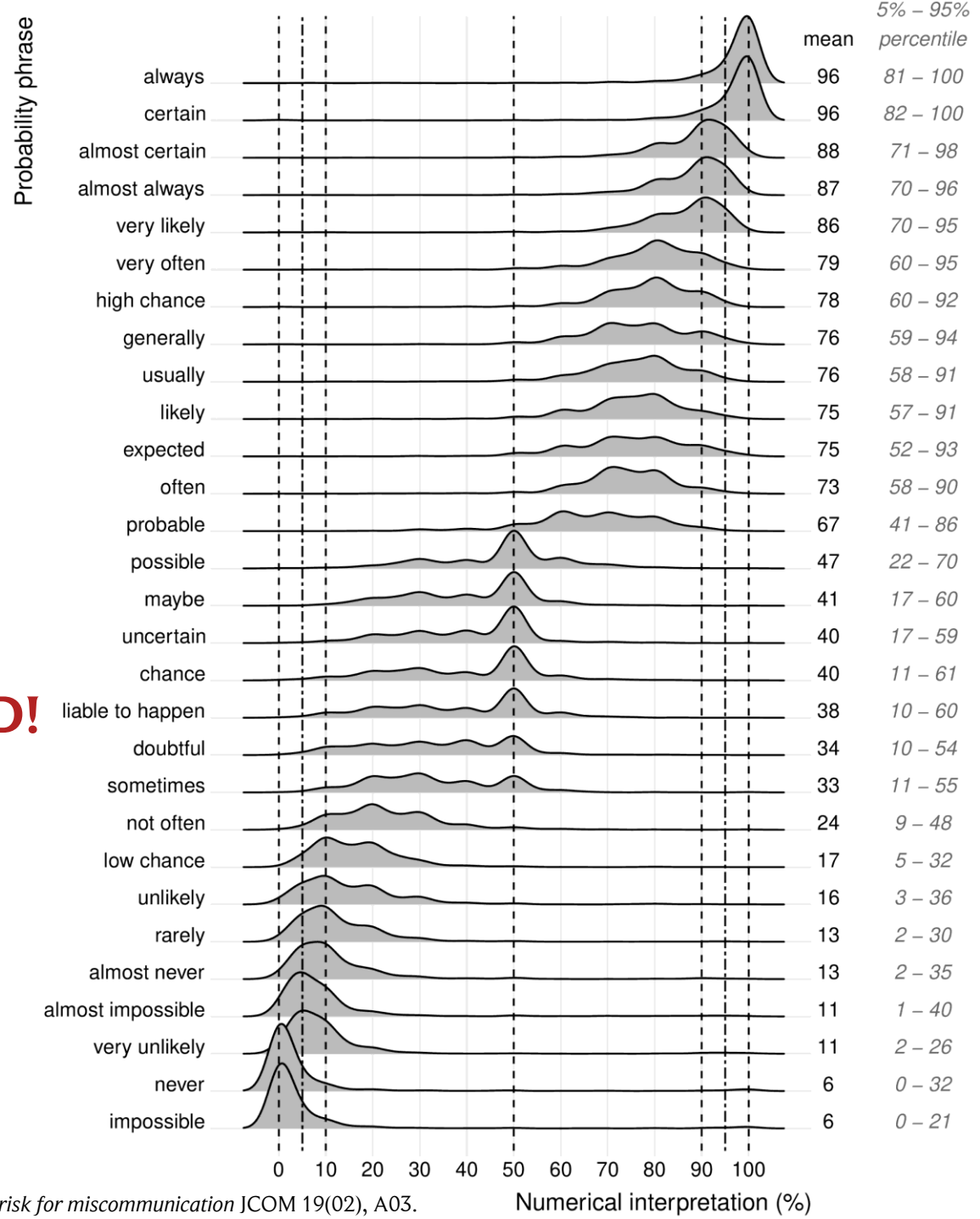
Content

- Sets
- Kolmogorov axioms
- Bayes Theorem
- Random variables
- Chebyshev Inequality
- Law of Large Numbers
- Transforming random variables
- Moment-generating function
- Central limit theorem
- Chernoff Bounds and Hoeffding's Inequality
- Shannon Entropy

Colloquial interpretation of probability phrases features a wide spread

↓

Developing an intuition about probability is HARD!



Willems, S., Albers, C. and Smeets, I. (2020). *Variability in the interpretation of probability phrases used in Dutch news articles — a risk for miscommunication* JCOM 19(02), A03.

Towards probabilities: set theory

Bertrand Russel (1872 - 1970)

Definition of a set Q:

$$Q = \{x \mid x \text{ has required property}\}$$

Or: x has required property $\leftrightarrow x \in Q$

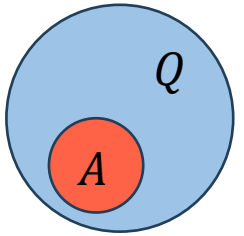
- Sets are very intuitive: collection of elements (with certain property)
- Example: the set of all Skywalkers $Q = \{$  $\}$
- Set with no elements: empty set {}
- Beware: this simple notion of sets can lead to problems
  Logical contradiction by Russel: R = **the set of sets** that are not elements of themselves

If $R \notin R$, it follows that $R \in R$, which is a contradiction!

Set operations

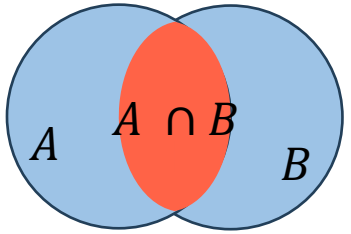
Set operations always reduce to logic operations.

- **Subset** $A \subset Q: x \in A \rightarrow x \in Q$ (but not necessarily the opposite!)



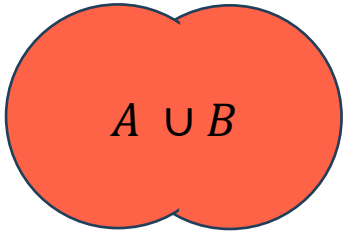
- **Intersection:** $A \cap B =$ set of elements that are both in A and B

$$\{1,2,3\} \cap \{1,3\} = \{1, 3\}$$



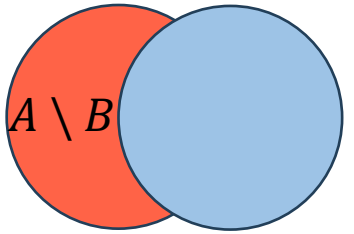
- **Union:** $A \cup B =$ set of elements that are in A or B

$$\{1,2,3\} \cup \{2,4\} = \{1,2,3,4\}$$

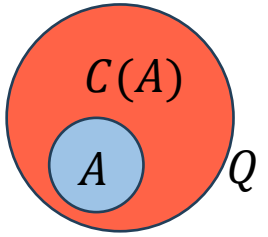


- **Difference:** $A \setminus B =$ set of elements that are only in A, not in B

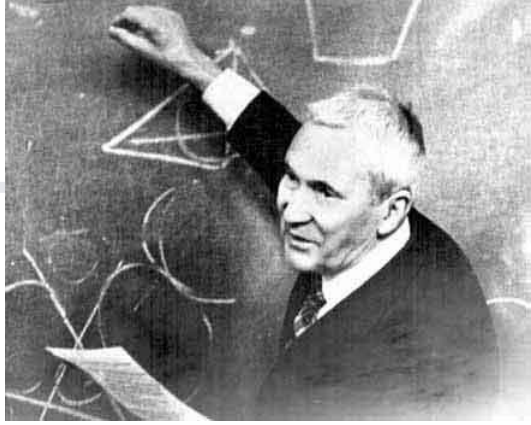
$$\{1,2,3\} \setminus \{1,3\} = \{2\}$$



- **Complement:** for $A \subset Q, C(A) =$ set of all elements in Q that are not in A



Probability: Kolmogorov axioms


Andrey Kolmogorov (1903 - 1987)

Some simple definitions to set up the stage for our random experiment. A random experiment is defined **using sets and set operations**:

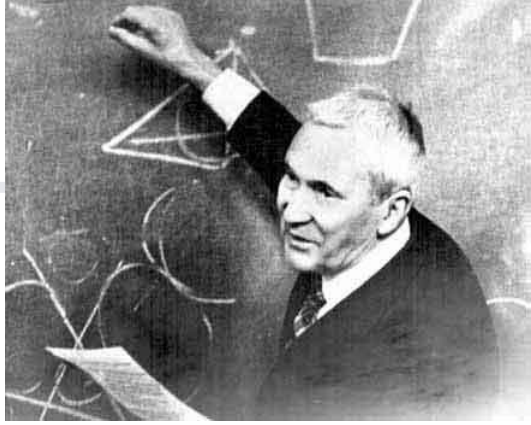
- **Outcome space** Ω = set of all possible outcomes
  Examples: 6-sided die: $\{1,2,3,4,5,6\}$
  2 toin cosses: $\{HH, HT, TH, TT\}$
- **Event** E = subset of Ω that might or might not happen
  Examples: Roll an even number $\{2,4,6\}$
  Get exactly one tail: $\{HT, TH\}$

Then we define a **probability measure** P such that (Kolmogorov axioms):

1. $P(E) \geq 0$ for all events E
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ for all events $A \cap B = \{\}$

$P(E)$ = probability of event E
(Ω, P) = probability space

Some intuition behind the axioms



Andrey Kolmogorov (1903 - 1987)

1. $P(E) \geq 0$ for all events E
   Each event has a probability of 0 (doesn't happen) or higher.
2. $P(\Omega) = 1$
   Normalization of all events. The prob. of ***any event happening*** *is 1.*
3. $P(A \cup B) = P(A) + P(B)$ for all events $A \cap B = \{\}$
   Probabilities of mutually exclusive events add up.

   Example: throwing a 1 ***or*** *a 2,* $A = \{1\}, B = \{2\}$

   $$P(\{1,2\}) = P(\{1\}) + P(\{2\})$$

   Similarly, we have $P(\{1,2,3,4,5,6\}) = \sum_{i=1}^{6} P(\{i\}) = 1$

Exercise time

Prove the following relations using the Kolmogorov axioms:

1. $P\big(C(A)\big) + P(A) = 1$, in particular $P(\{\}) = 0$
2. $A \subset B \rightarrow P(A) \leq P(B)$
3. $P(A \setminus B) = P(A) - P(A \cap B)$, with $A \setminus B = A \ \cap C(B)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Some reminders:

Kolmogorov axioms

1. $P(E) \geq 0$ for all events E
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ for all events $A \cap B = \{\}$

Set operations

Subset $A \subset Q$: $x \in A \rightarrow x \in Q$ (but not necessarily the opposite!)

Intersection: $A \ \cap B =$ set of elements that are both in A and B

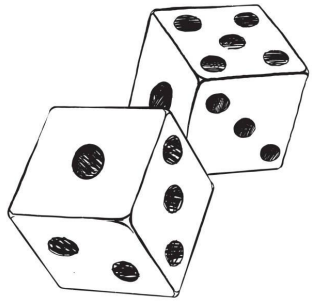
Union: $A \ \cup B =$ set of elements that are in A or B

Difference: $A \setminus B =$ set of elements that are only in A, not in B

Complement: for $A \subset Q, C(A) =$ set of all elements in Q that are not in A

Example: rolling a die

Outcomes $\Omega = \{1,2,3,4,5,6\}$



We assume a rigged die:

$$P(\{1\}) = P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = 0.15, \quad P(\{6\}) = 0.25$$

What's the probability of rolling either an even number **or** either 3 or 6?

Events: $A = \{2,4,6\}, \ B = \{3,6\}$

Probabilities: $P(A) = P(\{2\}) + P(\{4\}) + P(\{6\}) = 0.55, \quad P(B) = 0.4$

$$P(A \cap B) = P(\{6\}) = 0.25$$

Result: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7$

Laplacian probability

Pierre-Simon Laplace (1749-1827)

- Note that we can choose P however we want as long as the axioms hold!
- **Special case:** every elementary outcome has the same probability!
  For example: rolling a fair die, $P(\{1\}) = \cdots = P(\{6\}) = 1/6$

In this case, we can use **Laplace's rule** to calculate probabilities:

$$P(B) = \frac{|B|}{|\Omega|}$$

where $|B|$ is the number of elements in B.

Thus, calculating probabilities is equal to counting events!

Example: rolling an even number $P(\{2,4,6\}) = \frac{|\{2,4,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = \frac{1}{2}$

But **ONLY** applicable if all elementary outcomes have the same probability! E.g., lottery, rolling fair dice, shuffling cards, …

Conditional probabilities

Conditional probability: probability of event A under the condition that event B_i happened.

$$P(A \mid B_i) = \frac{|A \cap B_i|}{|\mathrm{B_i}|} = \frac{P(A \cap B_i)}{P(B_i)}$$

Note that generally $P(A|B_i) \neq P(B_i|A)$ *For example:* $P(\text{woman} \mid \text{pregnant}) = 1$, but $P(\text{pregnant} \mid \text{woman}) < 1$

We denote by $P(A \cap B_i) = P(A, B_i)$ the probability that both A and B_i happen.

This is a **chain** of events: $P(A, B_i) = P(A \mid B_i) \cdot P(B_i)$

First, B_i has to happen

If B_i happened, how likely is it that A happened?

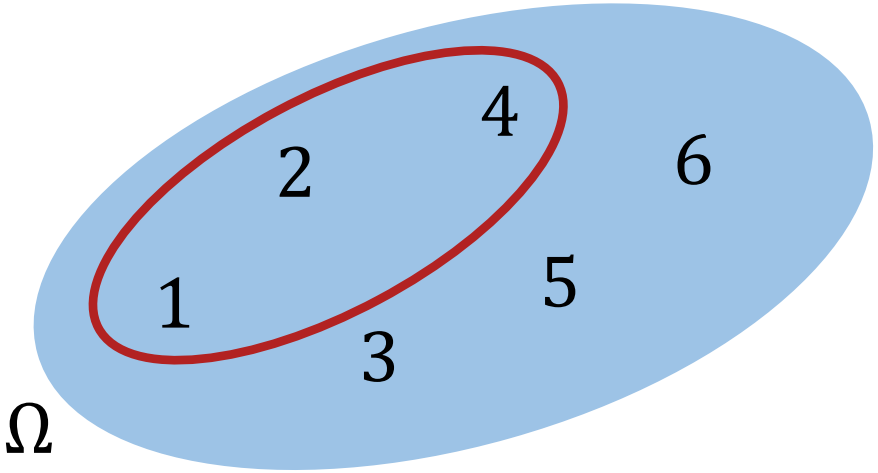
Conditional probabilities: visualization

Our events:

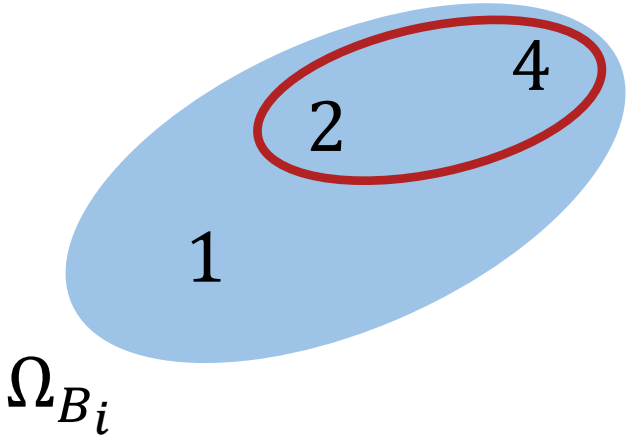
Rolling a power of 2: $B_i = \{1,2,4\}$

Rolling an even number: $A = \{2,4,6\}$

First, B_i has to happen



If B_i happened, how likely is it that A happened?



$$P(B_i) = \frac{1}{2}$$

$$P(A \mid B_i) = \frac{2}{3}$$

$$P(A, B_i) = \frac{1}{3}$$

See also: https://setosa.io/ev/conditional-probability/

Bayes theorem

Thomas Bayes (1701-1761)

Bayes law: follows directly from the definition of conditional probability.

Used to swap condition! $$P(B_i \mid A) = \frac{P(A \mid B_i)\, P(B_i)}{P(A)}$$

NEVER forget this relation!

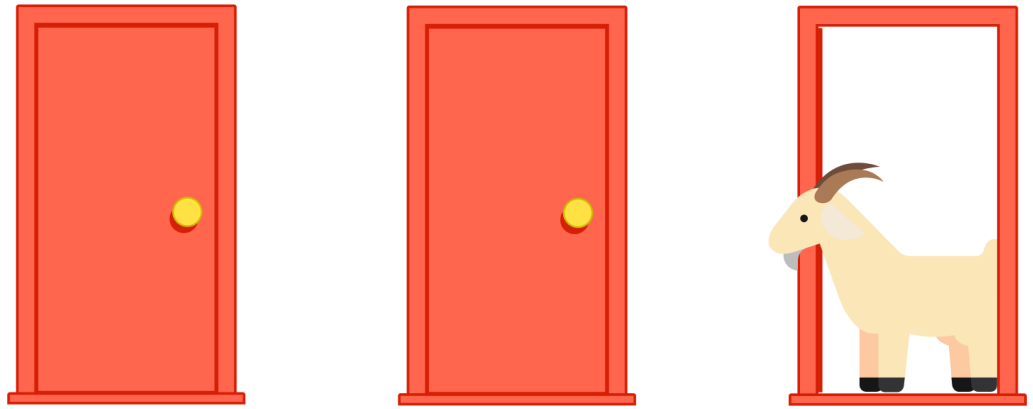
Always take into account all possible events! $$P(A) = \sum_i P(A \mid B_i) P(B_i), \quad \sum_i P(B_i) = 1$$

Basically: account for all events B_i that could have led to event A.

Also ensures normalization: $$\sum_i P(B_i \mid A) = \frac{\sum_i P(A \mid B_i)\, P(B_i)}{P(A)} = 1$$

Example: the Monty Hall problem ("*Ziegenproblem*")

You are in a gaming show and have to select one of three doors. Two have goats (no win), and one a big prize. After making your decision, the moderator opens one of the remaining doors (but not the one you chose). They give you the chance to swap doors. What are you doing?



A B C

Initially, we have to guess…

$$P(\text{Prize@}A) = P(\text{Prize@}B) = P(\text{Prize@}C) = \frac{1}{3}$$

Assume we chose *A* and *B* was opened by the moderator

$$P(\text{Open } B \mid \text{Prize@}A) = 1/2 \qquad P(\text{Open } B \mid \text{Prize@}C) = 1 \qquad P(\text{Open } B \mid \text{Prize@}B) = 0$$

$$P(\text{Prize@}C \mid \text{Open } B) = \frac{P(\text{Open } B \mid \text{Prize@}C)P(\text{Prize@}C)}{\sum_{i\in\{A,BC\}} P(\text{Open } B \mid \text{Prize@}i)P(\text{Prize@}i)} = \frac{\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3} + 0\cdot\frac{1}{3} + 1\cdot\frac{1}{3}} = \frac{2}{3}$$

Image credit: https://brilliant.org/wiki/monty-hall-problem/

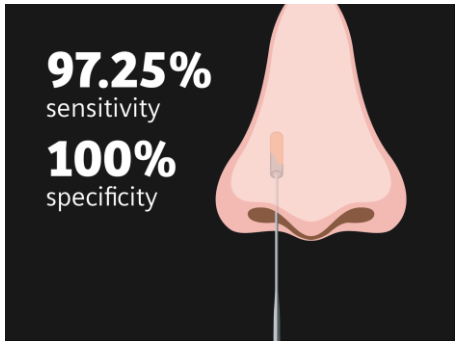
Another example

Sensitivity (true positive rate)**:**

$p(\text{tested positive} \mid \text{sick})$

Specificity (true negative rate)**:**

$p(\text{tested negative} \mid \text{not sick})$



Question: If your test is positive, does this mean your chances of being sick is 97.25%?

NO! Apply Bayes theorem: $p(\text{sick} \mid \text{tested positive}) = \underbrace{\frac{p(\text{tested positive} \mid \text{sick})}{p(\text{tested positive})}}_{\text{Obtained from manual / lab tests}} \cdot \underbrace{p(\text{sick})}_{\text{How common is the disease in your area?}}$

Note: $p(A \mid B)$ does not mean A was caused by B.
It only measures correlation (as seen by Bayes theorem: we can switch the order)!

Random variables

A **random variable** X is a mapping from random events to numbers

$$X\colon\ \Omega \to\ \mathbb{R}\ (\text{or } \mathbb{R}^N, \mathbb{C}, \mathbb{C}^N, \dots)$$

I.e., a random variable assigns a value to random events!

Examples:

- Rolling a die is a random experiment, the number that comes up a random variable.
- Drawing a lottery ticket is a random experiment, the money won a random variable.
- Measuring the temperature is a random experiment, the obtained value a random variable.

Turning the tables: We can (and usually do) phrase everything in terms of random variables!

$p(x) = \mathrm{P}(\{\omega \in \Omega\colon\quad X(\omega) = x\})$ is called the **probability distribution** of x**.**

Getting probabilities for random variables

Random variable:
values x returned by a random experiment.

Probability distribution:
a distribution $p(x)$ over values that x can take. Probabilities are obtained via integration.

In case of discrete random variables:
We have only a finite (countable) number of values $\{x_0, x_1, \dots\}$.
Then $p(x_i)$ is the probability of observing the value x_i in a random experiment.

In case of continuous random variables:
E.g., an $\mathbb{R}$-valued random variable. In this case, only integral statements make sense.

$$P(a \leq x \leq b) = \int_a^b p(x)\mathrm{d}x$$ is the probability for observing a value in the interval $[a, b]$.

I.e., $p(x)$ is the distribution, and by integrating we get probabilities!

Characterizing probability distributions

Expectation value: $\mathrm{E}(x) = \int x \cdot p(x)\,\mathrm{d}x$ $\quad$ ($\mathrm{E}(x) = \sum_i x_i p(x_i)$ in discrete case)

If we repeat the random experiment many times, which value do we get on average?

⟶ *measures the "offset" of* $p(x)$

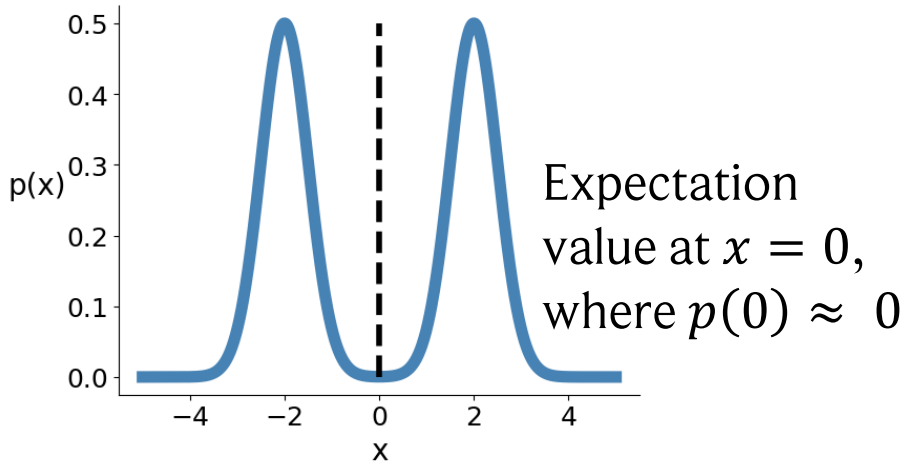
Very important: Expectation value is a weighted average, and **NOT** necessarily the most likely value of x!

Variance: $\mathrm{Var}(x) = \mathrm{E}\left(\left(X - \mathrm{E}(X)\right)^2\right)$

On average, how far away is the observed value from the expectation value?

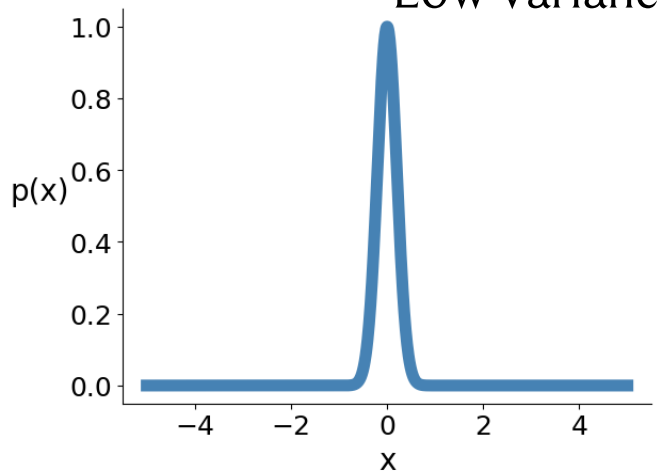
⟶ *measures the "width" of* $p(x)$

Note: distribution can be skewed!

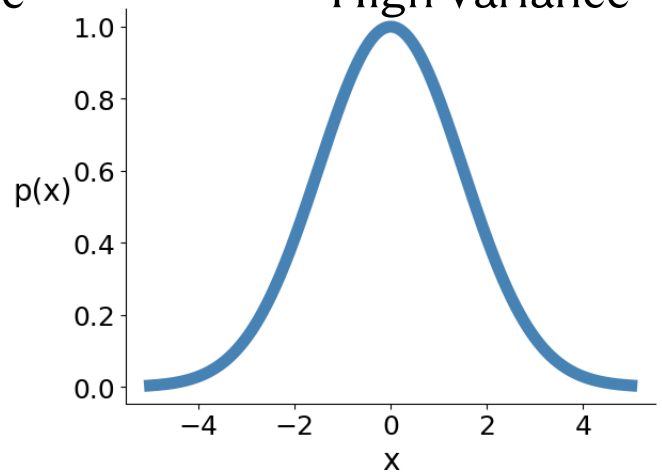


Expectation value at $x = 0$, where $p(0) \approx 0$

Low variance



High variance



Characterizing probability distributions

Cumulative distribution: $P(a) = P(x \leq a) = \int_{-\infty}^{a} p(x)\mathrm{d}x$

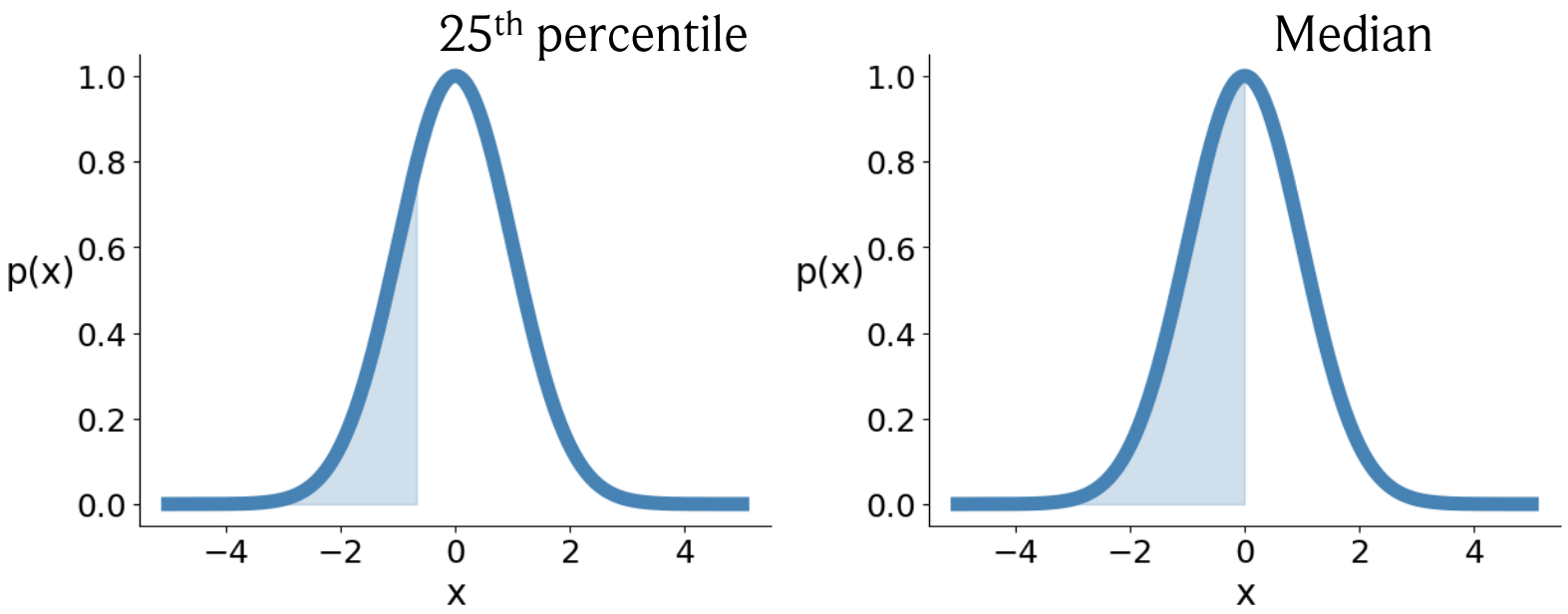
Probability of observing a value smaller or equal to a.

k$^{\text{th}}$ percentile: a_k such that $P(a_k) = k\%$

Good for measuring spread of asymmetric distributions! E.g., using 25^{th} *and* 75^{th} *percentile.*

Median: 50$^{\text{th}}$ percentile

Alternative to the expectation value. Less prone to outliers than the expectation value!



Exercise time

Show that:

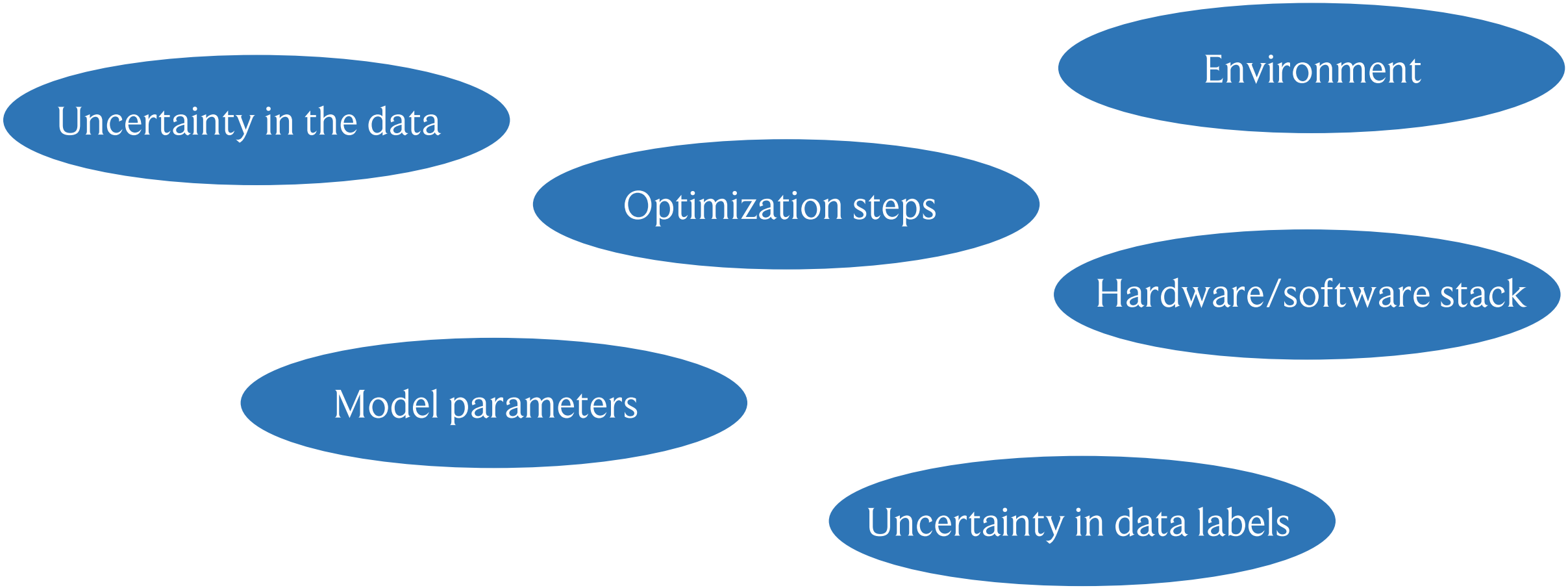
1. $\mathrm{E}(\lambda x) = \lambda E(x)$
2. $\mathrm{E}(x + y) = \mathrm{E}(x) + \mathrm{E}(y)$
3. If x and y are independent, then $\mathrm{E}(x \cdot y) = \mathrm{E}(x)\mathrm{E}(y)$

4. $\mathrm{Var}(x) = \mathrm{E}(x^2) - \mathrm{E}(x)^2$
5. $\mathrm{Var}(\lambda x) = \lambda^2 \mathrm{Var}(x)$
6. $\mathrm{Var}(x + y) = \mathrm{Var}(x) + \mathrm{Var}(y) + 2\,[\mathrm{E}(\mathrm{x} \cdot y) - \mathrm{E}(\mathrm{x})\mathrm{E}(\mathrm{y})]$
   and if x and y are independent, then $\mathrm{Var}(x + y) = \mathrm{Var}(x) + \mathrm{Var}(y)$

7. **Bonus:** assume you toss two dice. The outcome of the first die is X, the outcome of the second is Y. What is $\mathrm{E}(X/Y)$, and what is $\mathrm{E}(Y/X)$? Anything weird here?

Time to reflect a bit…

Question: Where do random variables appear in data science?



Useful bound: Chebyshev inequality

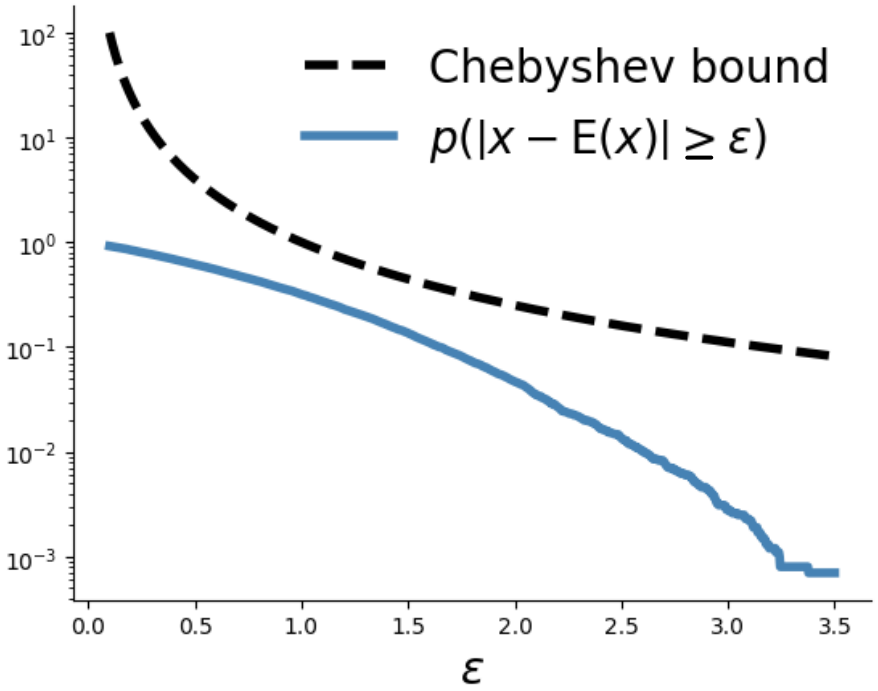
Chebyshev inequality: Assume a random variable X with finite expectation value $\mathrm{E}(X)$ and finite variance $\mathrm{Var}(X)$. Then the Chebyshev inequality states that

$$P(\,|X\, - \mathrm{E}(X)| \geq \epsilon) \leq \frac{\mathrm{Var}(x)}{\epsilon^2} \quad \forall \epsilon > 0$$

Proof: see handwritten notes

In other words: the Chebyshev inequality

- *quantifies* how often values that deviate by more than ϵ from the expectation value appear
- *relates* this to the variance of the probability distribution of X



A direct consequence: law of large numbers (LLN)

Law of large numbers: Assume random variables $x_1, x_2, \ldots, x_n$ with equal expectation value and finite variance $\mathrm{Var}(x_i) \leq M < \infty$. Then the law of large numbers states:

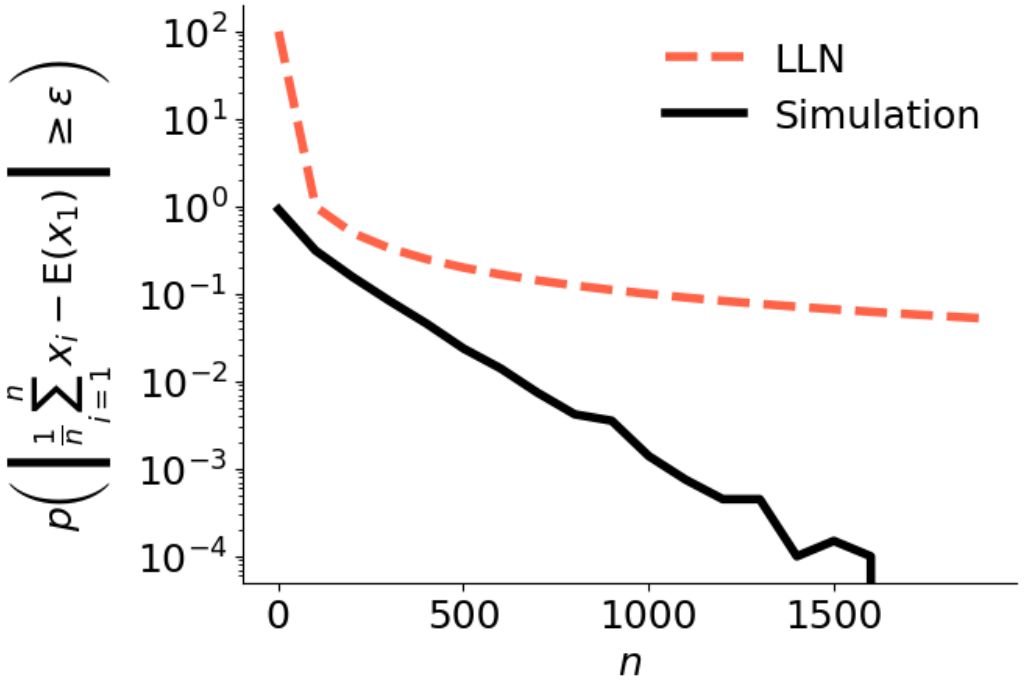
$$P\left(\left|\frac{1}{n}\sum_i x_i - \mathrm{E}(x_1)\right| \geq \epsilon\right) \leq \frac{M}{n\epsilon^2} \quad \forall \epsilon > 0$$

Proof: see handwritten notes

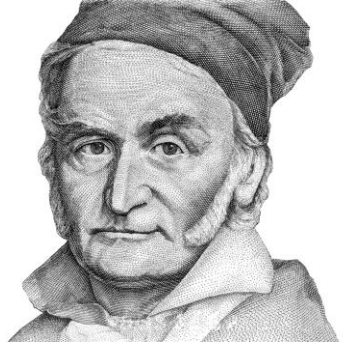
In other words: the law of large numbers

- Shows that the arithmetic mean $\frac{1}{n}\sum_i x_i$ *converges to the expectation value* for large n.
- Set the probability $\delta = \frac{M}{n\epsilon^2}$.

  Then the error ϵ is given by $\epsilon = \sqrt{\frac{M}{n\delta}}$, i.e., $\epsilon \propto \frac{1}{\sqrt{n}}$

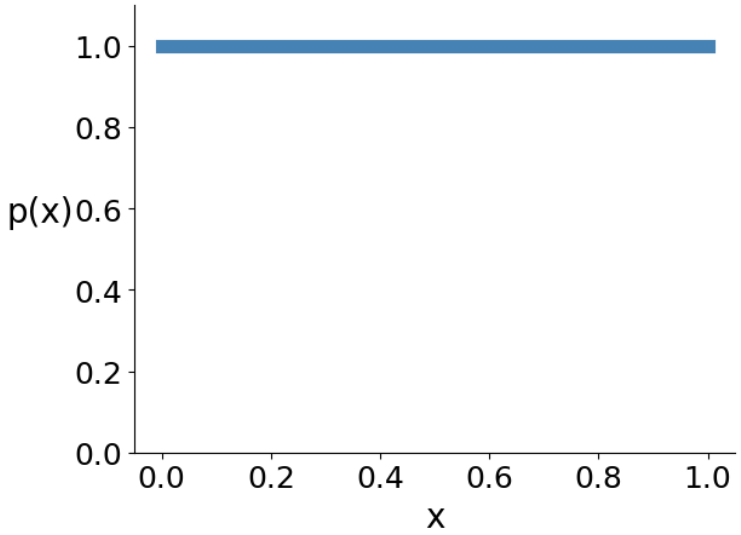


Some useful random distributions



Carl. F Gauß (1777 - 1855)

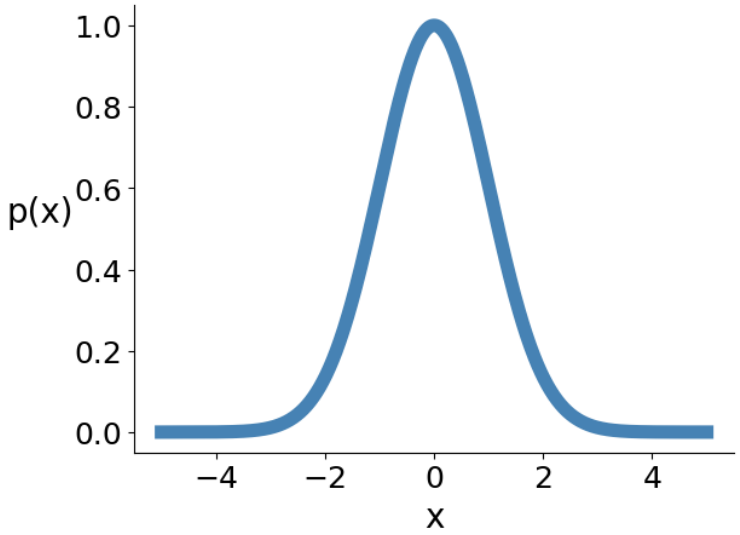
Uniform distribution in [0,1]:



$$p(x) = \Theta(x)\Theta(1-x)$$

$$\text{with} \quad \Theta(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

Normal distribution:



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

μ: mean
σ^2: variance

Incredibly fundamental, appears super often!
So don't forget this one!

Transforming random variables

Suppose you know the distribution $p(x)$ of a random variable x.

Question: can you write down the distribution of a function $y = f(x)$?

Yes: we require that probabilities are conserved locally. So, we substitute in the integral!

$$\int p(x)\mathrm{d}x = \int p\big(x(y)\big)\frac{\mathrm{d}x}{\mathrm{d}y}\,\mathrm{d}y = \int q(y)\mathrm{d}y \quad \text{with} \quad \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}\,f^{-1}(y)$$

probability density of interval

f rescales interval length

Thus, we have: $q(y) = p\big(y(x)\big)\dfrac{\mathrm{d}}{\mathrm{d}y}\,f^{-1}(y)$

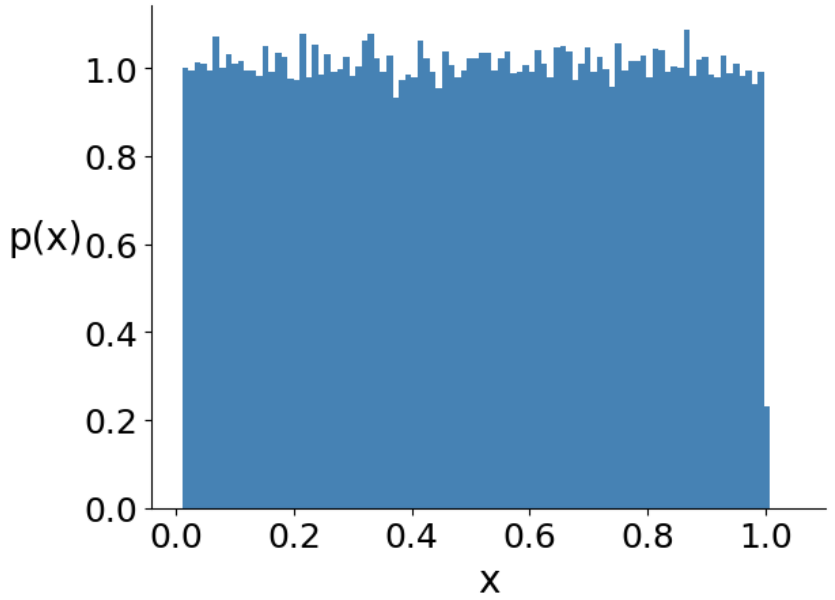
Examples

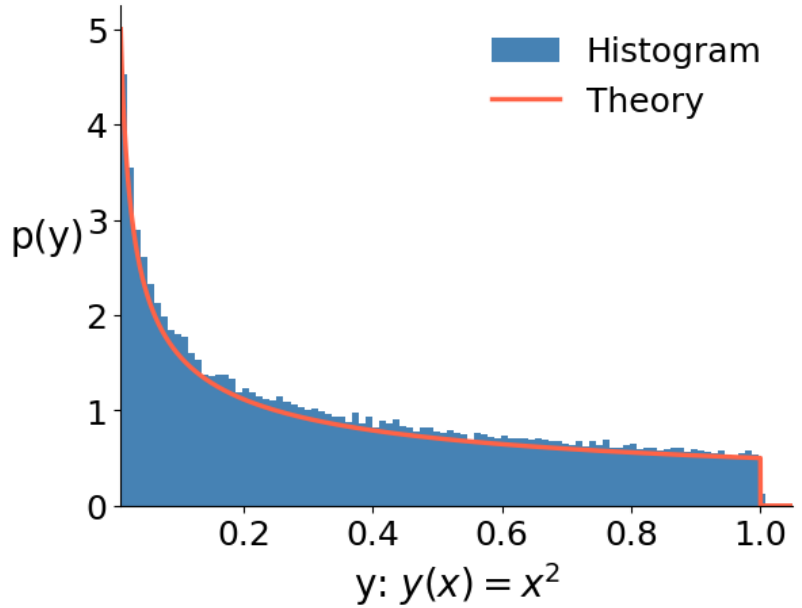
Assume: $y = f(x) = x^2$

Then: $f^{-1}(y) = \sqrt{y}$ and $\frac{d}{dy} f^{-1}(y) = \frac{1}{2\sqrt{y}}$

Thus: $q(y) = p(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}}$

Example: $p(x)$ is the uniform distribution





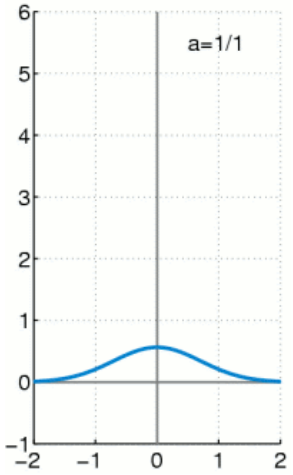
There is a way to construct transformations by hand: "inversion sampling" (not covered in this lecture)!

Adding (and subtracting) random variables

Question: assume we add two random variables x and y, which both follow two different random distributions $p(x)$ and $p(y)$. What is the random distribution p(z) of $z = x + y$?

A helpful tool for these types of questions: the **Dirac-Delta distribution** δ

$$\int_a^b f(x)\delta(x-c)\mathrm{d}x = \begin{cases} f(\mathrm{c}) \text{ if } \mathrm{c} \in [a,b] \\ 0 \text{ otherwise} \end{cases}$$



Idea:
for a value z, sum up the probabilities of all combinations of values of x and y that sum up to z

$$p(z) = \int \mathrm{d}x \int \mathrm{d}y\, p(x)p(y)\delta(z-x-y) = \int \mathrm{d}x\, p(x)p(z-x)$$

Thus, the sum distribution is the convolution of the two individual distributions!

Multiplying and dividing random variables

Question: assume we multiply two random variables x and y, which both follow two different random distributions $p(x)$ and $p(y)$. What is the random distribution p(z) of $z = xy$?

Same idea as for adding random variables!

$$p(z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \, p(x)p(y)\delta(z - xy) = \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{|x|} p(x)p(z/x)$$

Question: What is the random distribution p(z) of $z = y/x$?

$$p(z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \, p(x)p(y)\delta(z - y/x) = \int_{-\infty}^{\infty} \mathrm{d}x \, |x| \, p(x)p(z \cdot x)$$

Exercise time

Assume two random variables x, y ($x, y \in [0,1]$) with independent uniform distributions. Calculate:

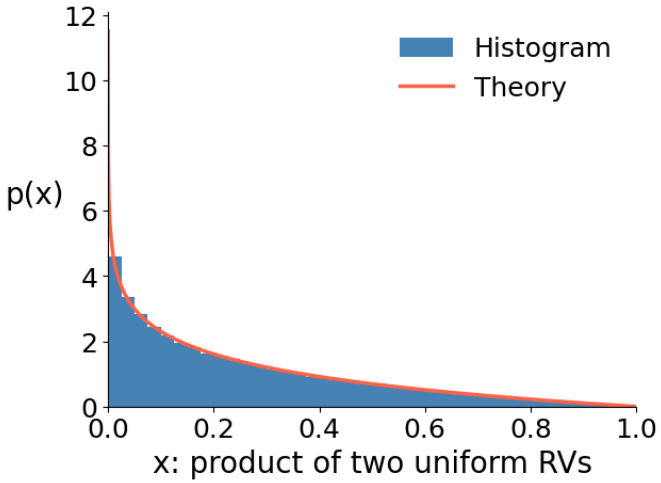
1. The distribution p(z) of $z = x \cdot y$
2. The distribution p(z) of $z = x/y$ using $\delta(ax) = \frac{1}{|a|}\delta(x)$

Some reminders:

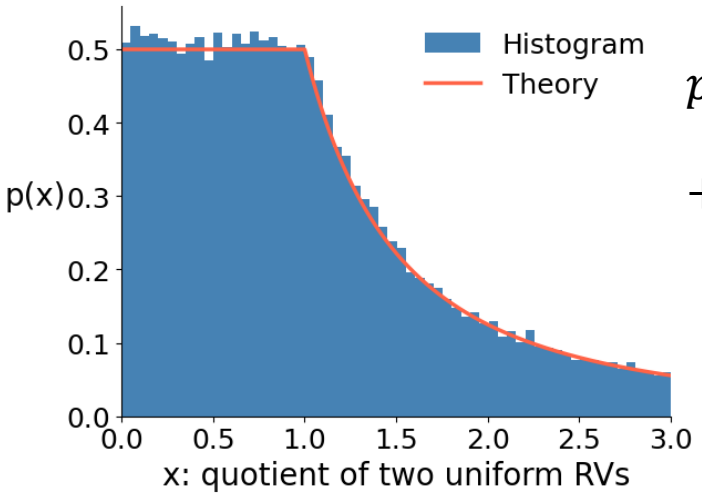
1. Uniform distribution in [0,1]: $p(x) = \Theta(x)\Theta(1-x)$ with $\Theta(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ otherwise} \end{cases}$
2. $p(z) = \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{|x|} p(x)p(z/x)$ for $z = x \cdot y$
3. $p(z) = \int_{-\infty}^{\infty} \mathrm{d}x\, |x|\, p(x)p(z \cdot x)$ for $z = x/y$

Examples

Uniform distributions

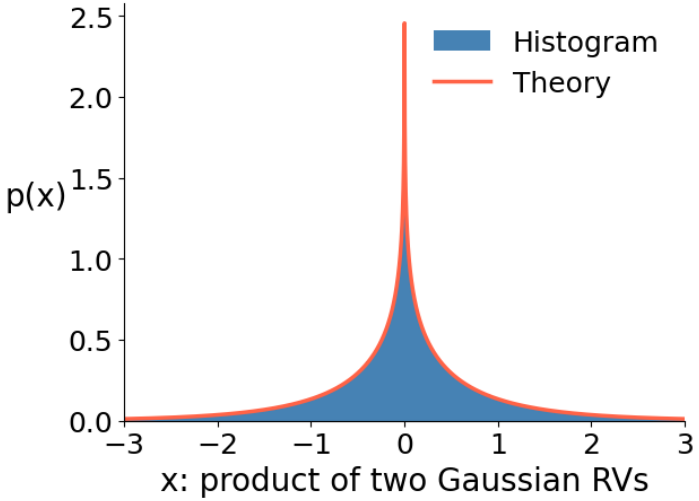


$$p(x) = \ln\left(\frac{1}{x}\right)\Theta(x)$$



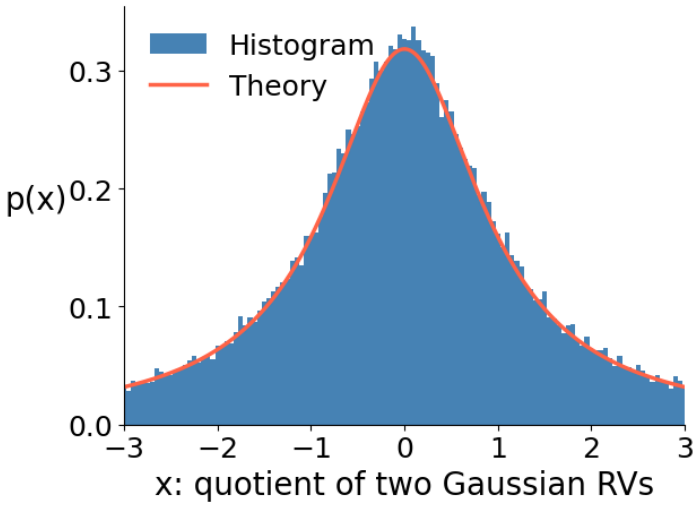
$$p(x) = \frac{1}{2}\Theta(x)\Theta(1-\mathrm{x}) + \frac{1}{2\mathrm{x}^2}\Theta(x-1)$$

Normal distributions



$$p(x) = \frac{K_0(|x|)}{\pi}$$

K_0: K-form Bessel function



$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

Cauchy distribution

NO finite mean and variance. Chebyshev & law of large numbers do not apply!

Moment-generating function

Given a probability distribution $p(x)$, its corresponding moment-generating function is:

$$\varphi_x(t) = \mathrm{E}(e^{tx}) = \int_{-\infty}^{\infty} \mathrm{d}x\, e^{tx} p(x)$$

Laplace transform of $p(x)$!

Why **moment-generating** function?

The name becomes clear when plugging in the Taylor expansion of $e^{tx} = \sum_{i=0}^{\infty} \frac{(tx)^i}{i!}$

$$\varphi_x(t) = \int_{-\infty}^{\infty} \mathrm{d}x\, e^{tx} p(x) = \sum_{i=0}^{\infty} \frac{1}{i!} \int_{-\infty}^{\infty} dx\, (tx)^i p(x) = \sum_{i=0}^{\infty} \frac{t^i}{i!} E(x^i)$$

Moments extractable by differentiation!

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} \varphi_x(t) \Big|_{t=0} = \mathrm{E}(x^n)$$

Moment-generating function of a Gaussian

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \longrightarrow \varphi_x(t) = e^{t\mu+\frac{1}{2}t^2\sigma^2}$$

Proof: see handwritten notes

Unique property of normal distributions:

Define the cumulants of a probability distribution as $\kappa_n = \frac{\mathrm{d}^n}{\mathrm{d}t^n} \ln\varphi_x(t)\Big|_{t=0}$

$$\text{thus } \ln\varphi_x(t) = \sum_{i=1}^{\infty} \frac{\kappa_n(x)}{n!}\, t^n$$

A probability distribution is Gaussian if only **the first two cumulants are non-zero!**

$$\kappa_1 = \mu \qquad \kappa_2 = \sigma^2$$

Application 1: Central limit theorem (never forget!)

Central limit theorem: Assume n independent random variables x_i with finite expectation value $\mathrm{E}(x_i) \leq A < \infty$ and finite variance $\mathrm{Var}(x_i) \leq B < \infty$. Let $z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i$.

For $n \to \infty$, z is Gaussian distributed with mean $\mu = \frac{1}{\sqrt{n}} \sum_i \mathrm{E}(x_i)$, variance $\sigma^2 = \frac{1}{n} \sum_i \mathrm{Var}(x_i)$.

Proof: see handwritten notes

THIS is the reason the normal distribution is everywhere!

BUT: Only true in the limit and with rescaling ($\frac{1}{\sqrt{n}}$).

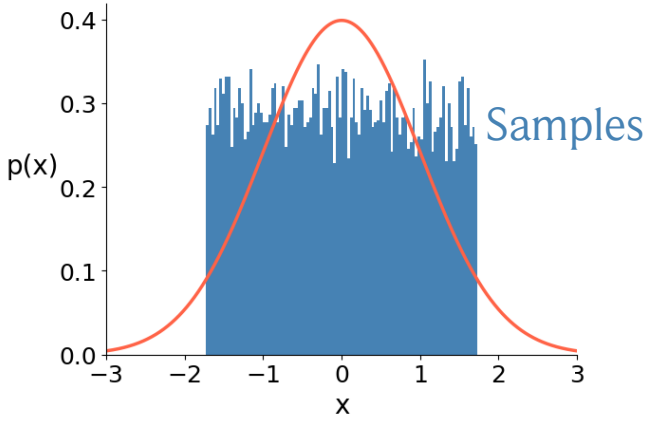
In practice: often already works well for small values of n.

AND: Only true for random variables with finite variance. For example: The sum of two random variables that follow a Cauchy distribution also follows a Cauchy distribution…

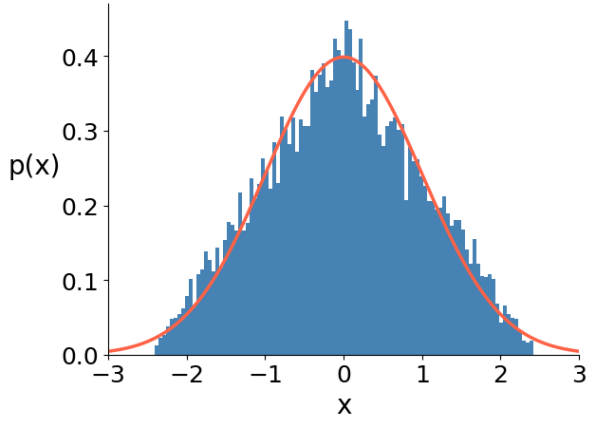
Examples

Summing up uniform distributed random variables (samples shifted and rescaled)

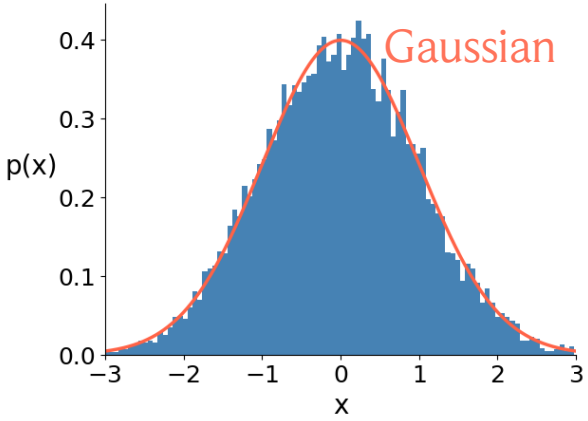
$n = 1$



$n = 2$

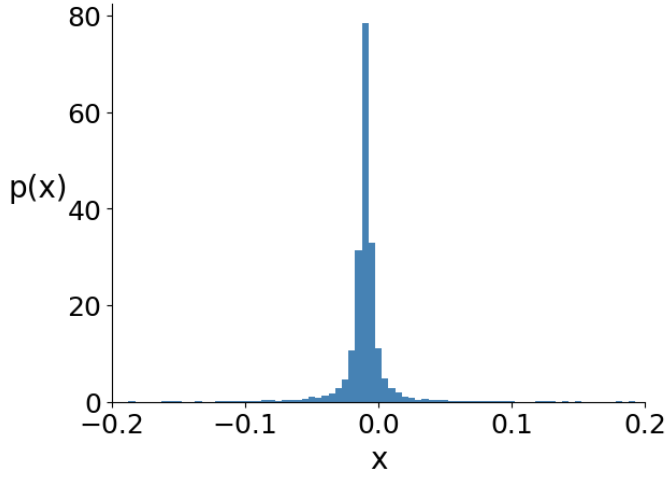


$n = 4$

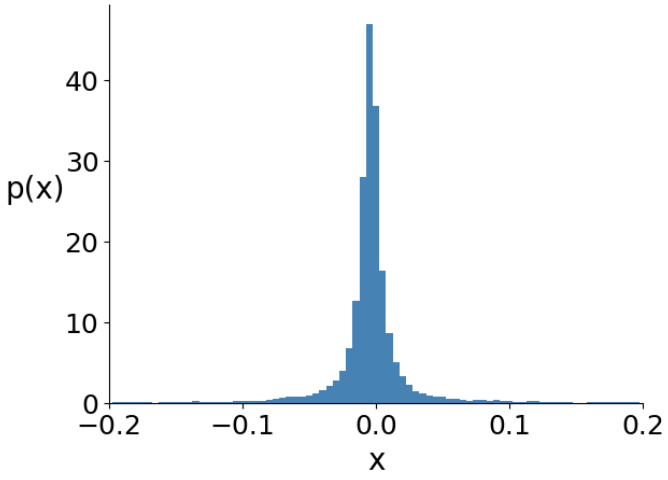


Sum of random variables following a Cauchy distribution (obviously not Gaussian)

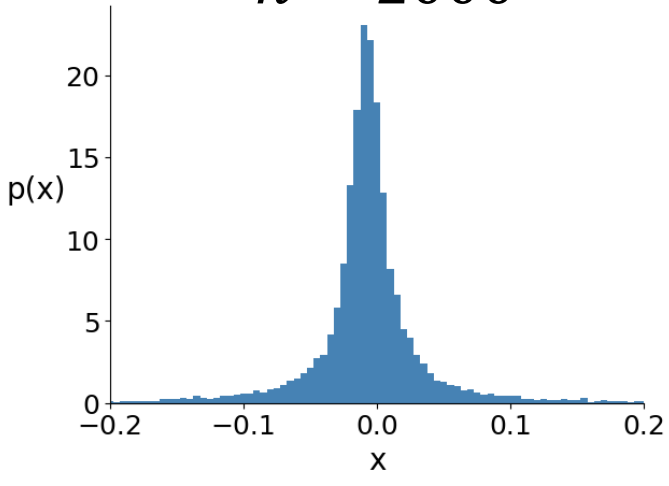
$n = 1$



$n = 2$



$n = 2000$



Application 2: improved bounds

Markov's inequality: Assume a positive random variable x. Then:

$$P(\, x \geq \epsilon) \leq \frac{\mathrm{E}(x)}{\epsilon} \quad \forall \epsilon > 0$$

Chernoff Bound: Assume a random variable x with density function $p(x)$. Further assume that $t > 0$. Then

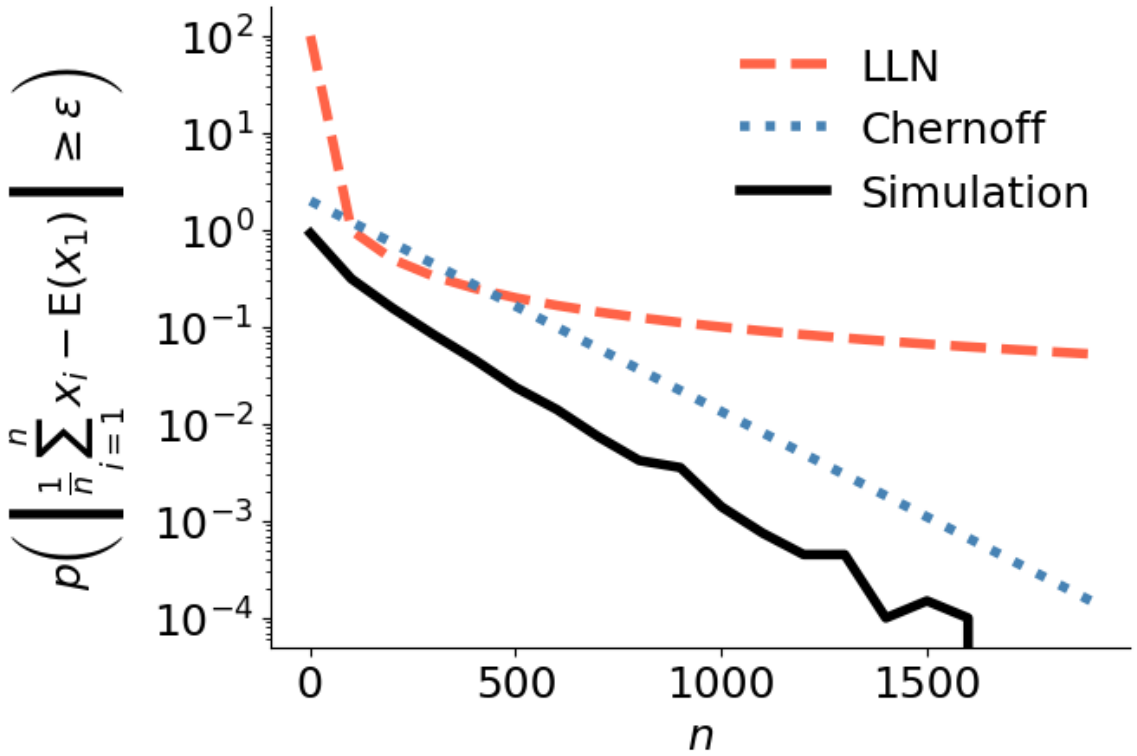
$$P(\, x \geq \epsilon) \leq \inf_t \left(e^{-t\epsilon} \varphi_x(t) \right) \quad \forall \epsilon > 0$$

Chernoff Bound for Gaussians: Assume random variables $x_1, x_2, \ldots, x_n$ with the same normal distribution. Then

$$P\left(\left| \frac{1}{n} \sum_i x_i \; - \mathrm{E}(x_1) \right| \geq \epsilon \right) \leq 2\, e^{-\frac{n\epsilon^2}{2\,\sigma^2}} \quad \forall \epsilon > 0$$

Proofs: see handwritten notes

Comparison with LLN



A fundamental theorem: Hoeffding's inequality

There are two more observations needed to get to a very general and powerful result:

1. The Chernoff Bound that we derived for Gaussians is **also** valid for any random variable with finite support, i.e., $x \in [a, b]$ is constrained to an interval.
2. **Popoviciu's inequality on variances:** Let x be a random variable with finite support x. Then $\mathrm{Var}(x) \leq \frac{(b-a)^2}{4}$.

Hoeffding's inequality: Assume independent random variables $x_1, x_2, \ldots, x_n$ from the same distribution and finite support $x \in [a, b]$. Then:

$$P\left(\left|\frac{1}{n}\sum_i x_i - \mathrm{E}(x_1)\right| \geq \epsilon\right) \leq 2\, e^{-\frac{2n\epsilon^2}{(b-a)^2}} \quad \forall \epsilon > 0$$

Example of applying Hoeffding's inequality

How many samples (e.g., training data samples!) from the random distribution $p(x)$ do we need such that with probability at least $1 - \delta$ the average does not deviate by more than ϵ from the expectation value?

Answer: Using Hoeffding's inequality, we set $\delta = 2\, e^{-\frac{2n\epsilon^2}{(b-a)^2}}$ from which we get:

$$\epsilon = \sqrt{\frac{(\ln(2) - \ln(\delta))(b-a)^2}{2n}} \qquad \text{and} \qquad n \geq \frac{(\ln(2) - \ln(\delta))(b-a)^2}{2\epsilon^2}$$

Information: measuring surprise of random variables

Assume we send a message. The message consists of n independent random variables $x_1, x_2, \dots, x_n$ that can take values ['A', 'B', 'C'] with probabilities $p(A), p(B), p(C)$. We concatenate the results to get a message.

Example message: A B A C

Likelihood of message: $p(A)p(B)p(A)p(C)$

Question: How much information does the message contain?
Or in other words: how many bits are required to encode/store the message?

First: intuitively, a message contains information if it allows us to *act*.

We require the following properties of an information measure I:

1. I continuously decreases with $p(\text{'}X\text{'})$ (*less probable events are more surprising*)
2. $I(p(\text{'}X\text{'}) = 1) = 0$ (*always the same signal is not surprising, i.e., has no information*)
3. $I\big(p(\text{'}X\text{'}) \cdot p(\text{'}Y\text{'})\big) = I(p(\text{'}X\text{'})) + I\big(p(\text{'}Y\text{'})\big)$ (*information is additive for indep. RVs*)

Solved by $I(p) = -\log_2(p)$ ← *log with basis 2 = bits!*

Image credit: shutterstock

Shannon entropy and data compression

Claude Shannon (1916 - 2001)

Shannon entropy $H(p)$: the expected information content of a random variable with probability distribution p.

$$H(p) = -\sum_i p_i \log_2 p_i \qquad (H(p) = -\int p(x) \log_2 p(x))$$

The Shannon entropy provides a lower limit of how messages composed of independent random variables can be compressed!

Shannon's source coding theorem: n independent random variables with distribution p cannot be compressed into less than $n\ H(p)$ bits without information loss.

Can we compress more in practice? **Yes!** 1) Utilize structure in data; 2) Compress with loss.

Kullback-Leibler divergence and thermodynamics

Rolf Landauer (1927 – 1999)

The Kullback-Leibler divergence D_{KL} measures how close we are to that optimum!

$$D_{KL}(p||q) = -\sum_i p_i \log_2 q_i - H(p) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

"Assume" probabilities q_i, but average is done with actual frequencies!

BUT it is not a metric! $D_{KL}(p||q) \neq D_{KL}(q||p)$

D_{KL} can be used to measure the similarity of probability distributions.

Curious observation:
thermodynamics defines the entropy of physical systems similarly!

$$S(p) = -k_B \sum_i p_i \ln p_i$$

Landauer's principle provides a link: the minimum heat W released when deleting one bit of information in a device operating at temperature T is

$$W = T\,\Delta S = k_B T \ln 2$$