Machine Learning: Recap Statistics

Dozenten: Prof. Dr. M. O. Franz, Prof. Dr. O. Dürr

HTWG Konstanz, WS 2026/27

<– WS 2026/27 Ganz gut gelaufen, die Dinge wie unbaisedness und variance an der Tafel, erklärt. Auch das mit der Z-Trafo wurde an der Tafel erklärt. –>

Remark: Much of the material have been developed together with Beate Sick (ZHAW/Uni Zurich).



AI use

The elephant in the room

Photo: Steve Jurvetson, “The elephant in the room at Arsenale”, CC BY 2.0, via Wikimedia Commons

AI is making huge progress: example mathematics

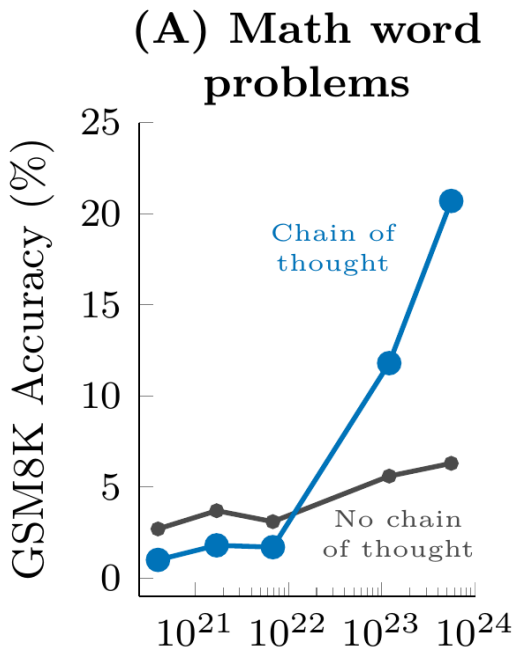
> "Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?" — typical test question 2021 (GSM8K)

2021	2023	2024	2025	Sept. 2026
word problems like the one above	*the same word problems*	*Mathematical Olympiad (IMO)*	*Mathematical Olympiad (IMO)*	*open research problem*
GPT-3: only about **one in three** correct (34 %)	GPT-4: **92 %** correct	AlphaProof: **silver** level	Gemini Deep Think & OpenAI: **gold**, 5 of 6 problems	**Navier–Stokes Millennium Problem solved** (OpenAI): 165 pages, formally verified in Lean

In five years: from "Natalia's clips" to the Millennium Problem.

Sources: Cobbe et al. 2021 (GSM8K, GPT-3); OpenAI GPT-4 Technical Report 2023; Google DeepMind 2024, 2025; OpenAI, "On the Navier–Stokes Millennium Prize Problem", 8.9.2026; Quanta Magazine, 8.9.2026

Side note: emergence – or mirage?



Model size: compute during training (FLOPs, log scale)

Emergence (Wei et al. 2022): on GSM8K, small models stay close to 0 %. From about 10^{23} FLOPs, performance suddenly jumps. On the question of emergence see also (Schaeffer et al. 2023).

For comparison today: GPT-4 (2023) ≈ $2 \cdot 10^{25}$ FLOPs, largest known runs 2025/26 ≈ $5 \cdot 10^{26}$ – about 20 and 500 times as much as at the right edge of the figure. Trend: ×4–5 per year.

Figure: Wei et al. (2022), Emergent Abilities of Large Language Models, TMLR, Fig. 3A (model: LaMDA), CC BY 4.0

Live demo: one exercise, two ways

PLACEHOLDER – exercise sheet 1, task 2

Way 1: Claude chat

- Copy the task from the exercise sheet into the chat
- Result: a solution – but not always with the methods from the lecture

Way 2: Claude Code + lecture PDFs

- Exercise sheet **and** lecture slides as context
- Result: a solution that fits the lecture, including Python code

```
Prompt (way 2): Solve the tasks in the attached exercise sheet. Use only methods from the attached lecture slides. Write Python code and give each result as a comment directly next to the calculation.
```

We are all in the same boat

How papers could be written today: AI takes the method from one paper and applies it to another problem.

MY PROMPT (JUNE)

> "… Couldn't we use a method similar to the one used in the causal forest paper for our method? …"



AI (Fable)

works for about 6 h

RESULT

New method finished.

It works.

But: as long as I have not understood it, I cannot publish it.

(I am still working on it.)

How do you stay relevant in times of ever better AI?

Ideas:



Cartoon: taz, 28 February

What remains: human judgement

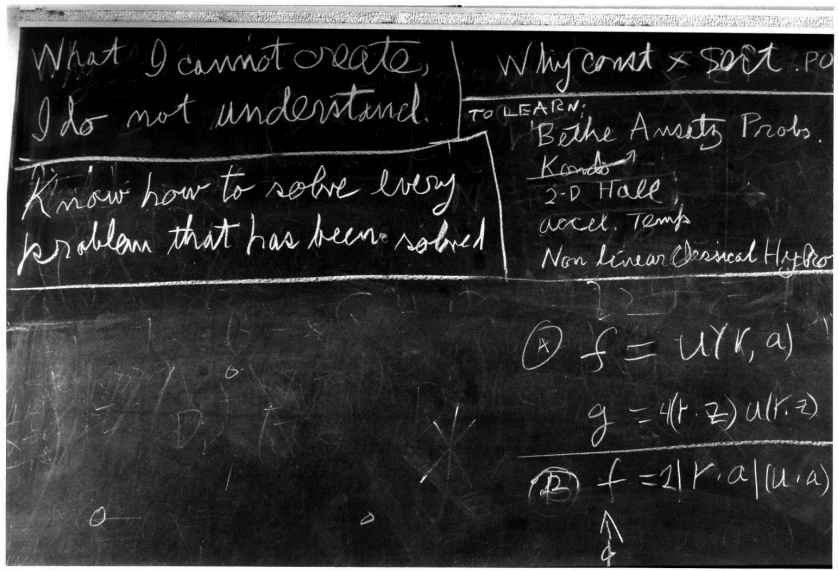
Before you use an AI result:

- Does the answer fit the question?
- Have errors happened? Has something been hallucinated?
- Is the solution humanly justifiable?

If you don’t know the method, you can only believe the AI.

In machine learning just as in research.

How do you gain understanding?



Richard Feynman's blackboard on the day of his death, 15 February 1988.

Caltech Archives, Image 1.10-29

"What I cannot create, I do not understand."

Richard Feynman (1918–1988), Nobel Prize in Physics 1965

Understanding means: being able to do it yourself.

What does *create* mean today?

What research says

Generation effect

What you **produce yourself**, you remember better than what you only read.

(Generation Effect) Slamecka & Graf 1978, replicated many times

Does it have to be by hand? Nobody knows – the studies on this are not really robust.

AI when practising maths (PNAS 2025)

Field experiment, almost 1000 school students, three groups:

	While practising	Exam without AI
without AI	–	–
standard ChatGPT	**+48 %**	**−17 %**
tutor mode (hints instead of solutions)	**+127 %**	no difference

Bastani et al. 2025, compared with the group without AI

It is not *whether* you use AI that decides – it is *how*.

Learning is like strength training: if you let a forklift lift the weights in the gym, you will not get stronger.

How to use AI for learning

Don’t train every muscle – but the right ones. No forklift in the gym, but gladly on the building site. The exercises here are training.

1. **Tutor instead of solution machine.** AI should ask questions and give hints, not deliver the solution. Short, fast interaction.
2. **Think for yourself.** First try it yourself, then AI. You must have the feeling: *I did that.*
3. **Check instead of believe.** Do the plot and the mean fit the question? Are the confidence intervals plausible? Are the assumptions met?
4. **The test:** Can you explain it effortlessly to a fellow student without AI? *Sprezzatura:* effort when learning – ease when explaining.

Stay curious: Try out what happens when you change something – with AI this costs almost nothing.

Learning prompt to copy

```
I am working on the task in the attached exercise sheet. Do not give me a solution. Ask me questions
and give me only hints until I find it out myself. Use only methods from the attached lecture slides.
When I show you my Python code, check it, but do not correct it for me – tell me where I should look.
```

Tools

- **NotebookLM** (Google): works only with your sources, e.g. the lecture slides
- **Ask Alma**: HTWG's AI, free and privacy-friendly, via Moodle
- **Learning modes** in ChatGPT, Claude and Gemini: they ask back instead of giving the solution

Rules for this lecture

- **You do not have to use AI.** Learning without AI is completely fine.
- **There is only one really dumb thing: using AI the wrong way.** If you let AI give you the solutions while practising, you do worse in the exam than without AI.
- Questions about the material: ask AI if a concept is unclear to you. Be critical of the answers (“hallucination”).
- Code / exercises: you may use AI. Work iteratively (in several steps).
  - **Ownership: you are the author of the code, you are responsible and you must understand the code!**
- Exam: you have no access to AI in the exam.

Those who understand stay relevant.

Literature

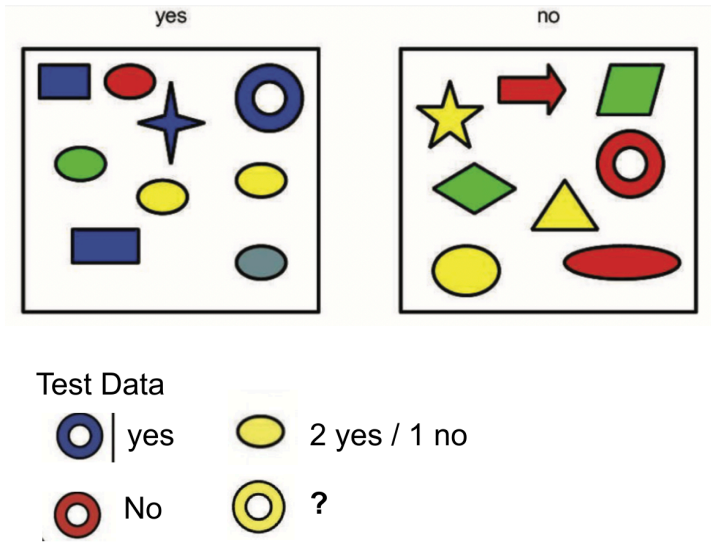
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö. & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *PNAS*, 122, e2422633122.
- Slamecka, N. J. & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), 592–604.
- Caltech Archives (1988). *Richard Feynman's blackboard at time of his death.* Image 1.10-29. digital.archives.caltech.edu/collections/Images/1.10-29
- Cobbe, K. et al. (2021). Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 (GSM8K).
- Wei, J. et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 08/2022. arXiv:2206.07682.
- Schaeffer, R., Miranda, B. & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? *NeurIPS 2023*.
- OpenAI (2023). *GPT-4 Technical Report.* arXiv:2303.08774.
- Google DeepMind (2024). AI achieves silver-medal standard solving International Mathematical Olympiad problems. Blog, 25.07.2024.
- Google DeepMind (2025). Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. Blog, 21.07.2025.
- OpenAI (2026). *On the Navier–Stokes Millennium Prize Problem.* openai.com/index/navier-stokes-solution, 08.09.2026.
- Quanta Magazine (2026). AI Has Solved One of Math's $1 Million Millennium Prize Problems. 08.09.2026.

Overview

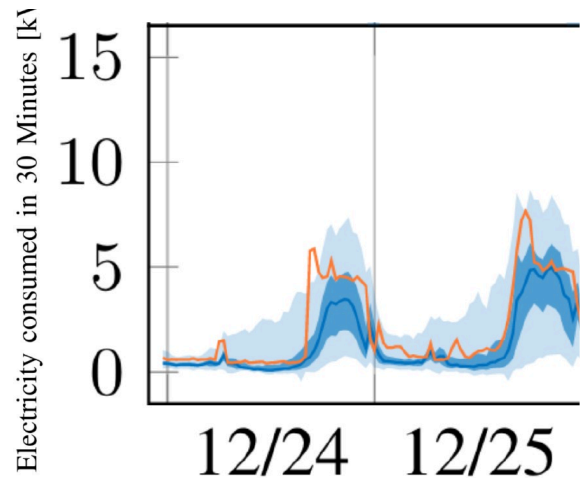
- Recap statistics
  - Why Probability and Statistic is useful
  - Sampling
  - Parameter Estimation
  - Quality of parameter estimation

Why Stats: Need for probabilistic prediction

Classification
Training Data



Regression
Next Day Power Consumption
Data until 30 minutes before midnight

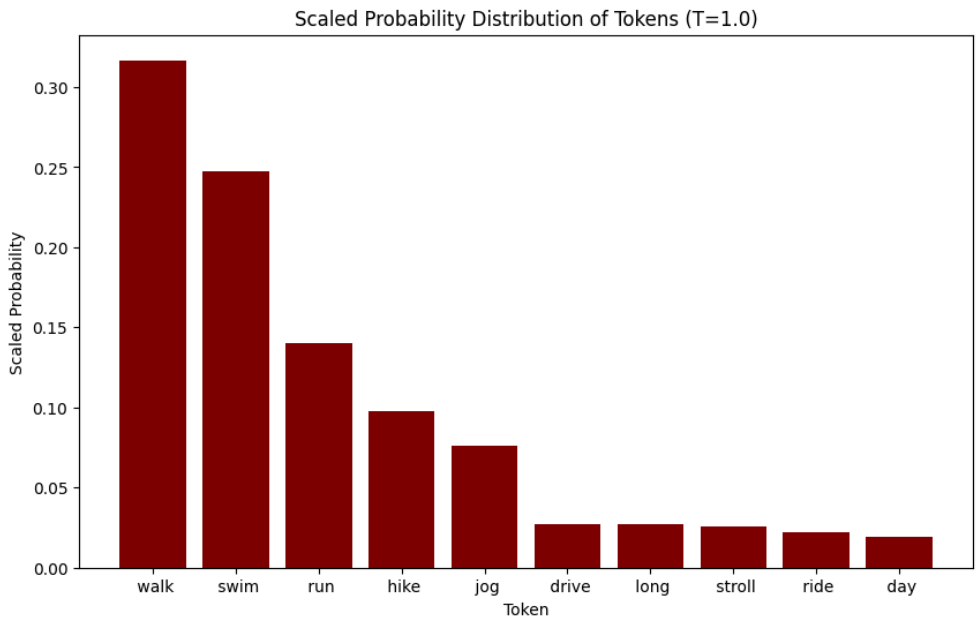


In many ML applications probabilistic forecasts are needed.

Left: Adapted from Murphy 2012. Right: https://arxiv.org/abs/2204.13939

LLM: Probabilities for next word

Prompt
The weather is really nice today. I'm thinking about going for a …



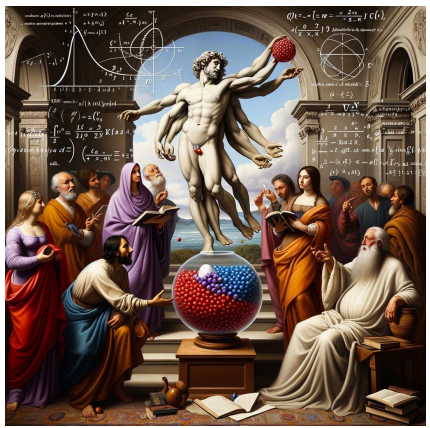
https://github.com/oduerr/llm_playground/tree/main/hugging_face/sampling_from_LLM

Short Recap of Statistics*

Eat your own dogfood: this section is accompanied by the ipython notebook
https://github.com/ioskn/mldl_htwg/blob/master/intro/intro.ipynb

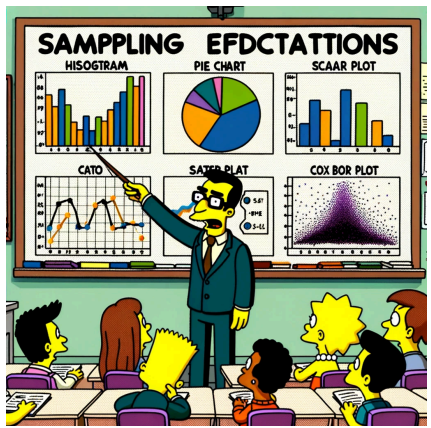
Models vs. Data

Models („probability land“)



- Urn Models
- Gaussian distribution (and others)
- Linear Regression
- …
- Neural Networks (incl. ChatGPT)

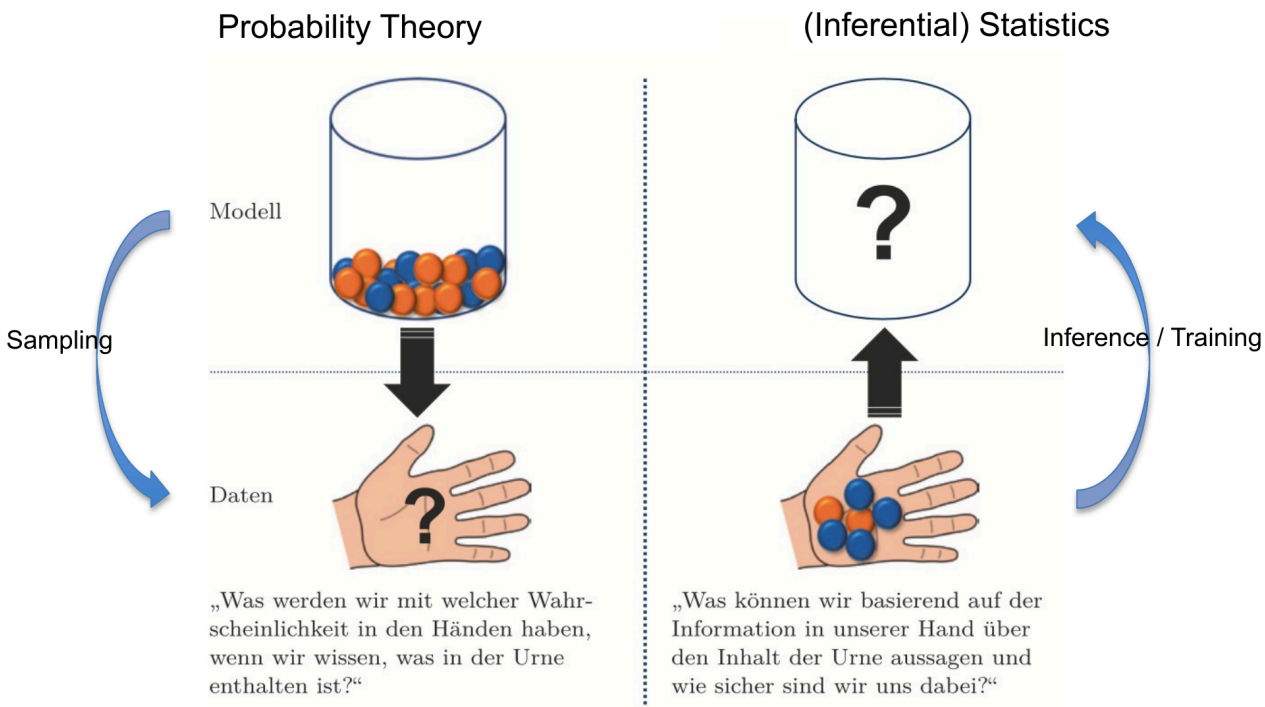
Data



Many ML and DL models are probabilistic models. They give probability how data is generated.

Generated with ChatGPT / Dalle-3

Probability vs Statistics (Inferential)

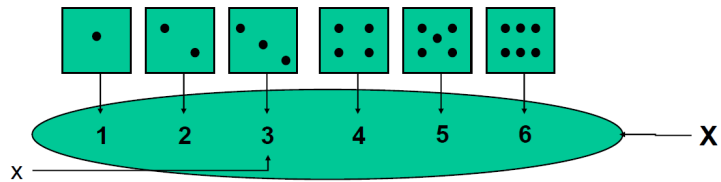


Aus Lukas Meier W‘keit und Statistik, adapted

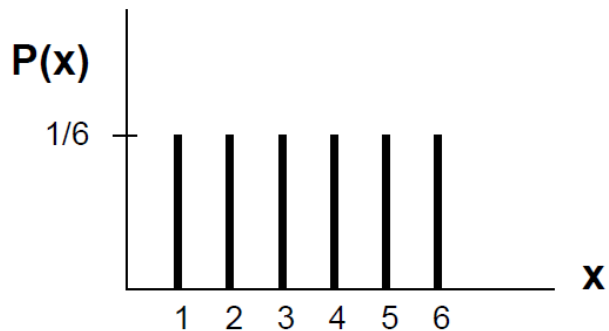
Data Land

- Data (Dice rolls): 1,1,2,1,3,5,3,4,6,…
- Data (Continuous): 0.1, 0.2, 0.23, 0.32, 0.123123, 0.3
- Tables / Histograms **mean and empirical variances**

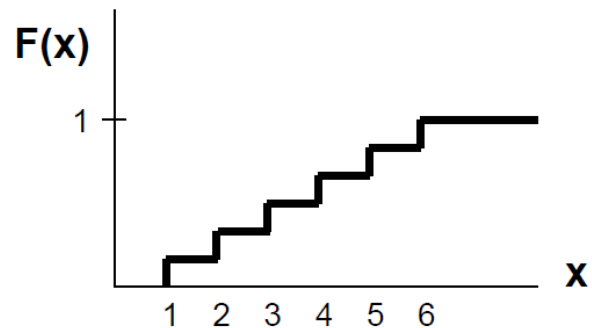
Probability for Discrete Data



Probability mass function (pmf)



Cumulative distribution function (cdf)



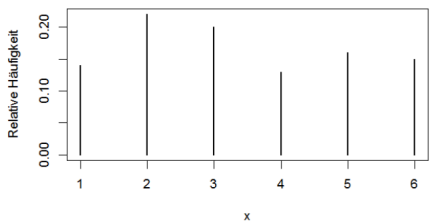
$P(x) = P(X = x)$: probability that the random variable X takes the value x.

$F(x) = P(X \leq x)$: probability that X takes a value $\leq x$.

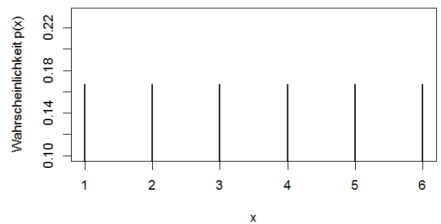
Interpretation of the probability density function (pdf)

Discrete

Hist. of relative frequency

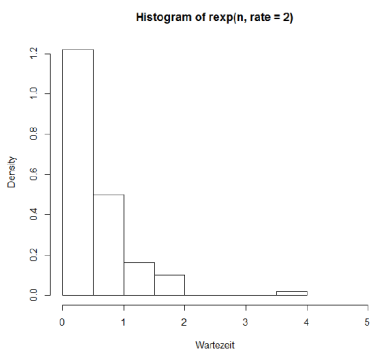


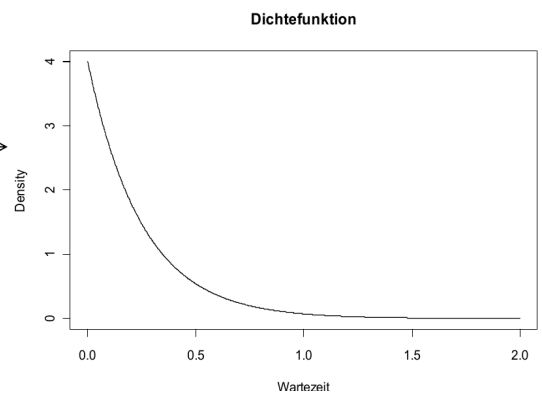
Probability



Continious

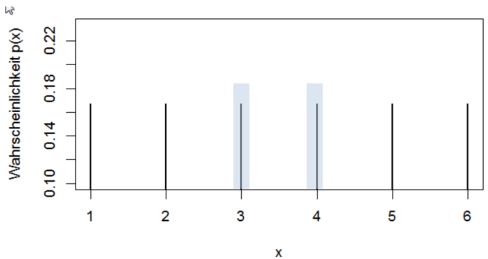
Probability?





Density vs. Probability

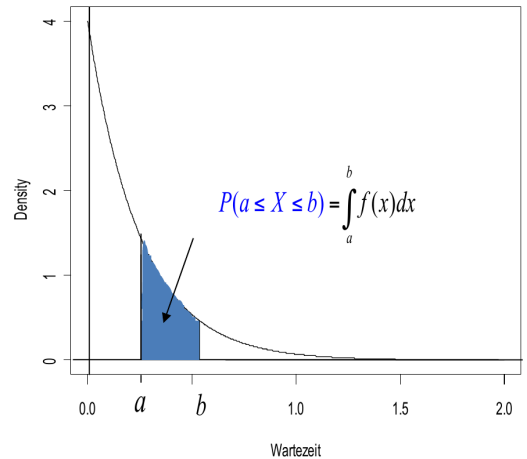
Diskret



$$P(a \le X \le b) = \sum p(X = x_i)$$

Continious

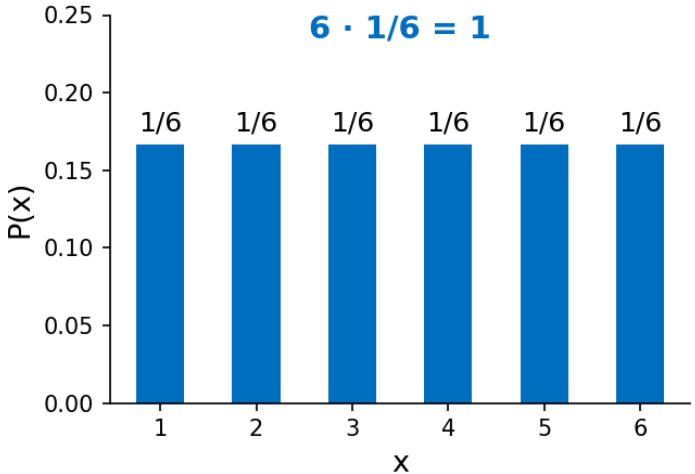
Dichtefunktion

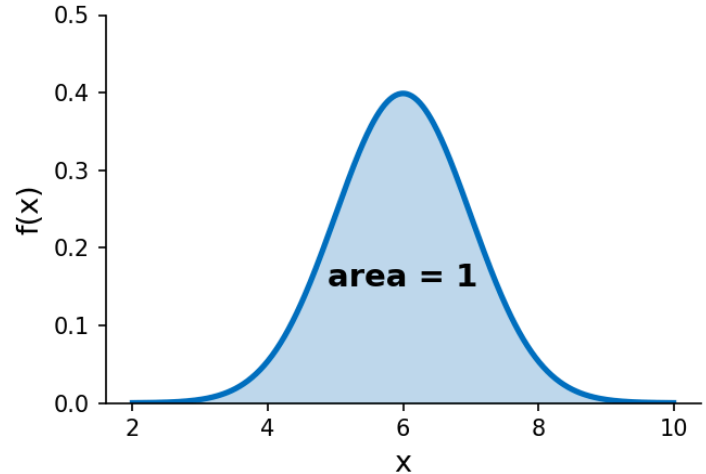


The probability of obtaining a result between a and b is the sum of the probabilities.

The probability of obtaining a result between a and b is the integral of the density function over a to b.

pmf and pdf sum / integrate to one





$$\sum_x P(x) = 1, \qquad P(x) \geq 0$$

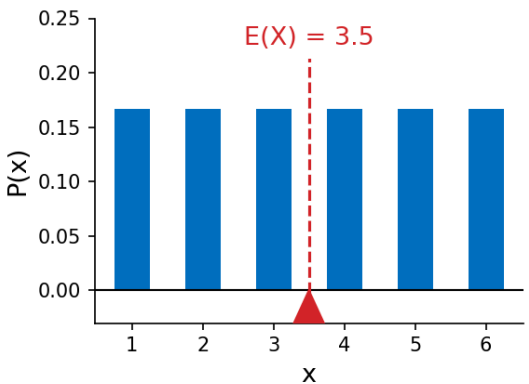
$$\int_{-\infty}^{\infty} f(x)\,dx = 1, \qquad f(x) \geq 0$$

The total probability is 1: some value always occurs.

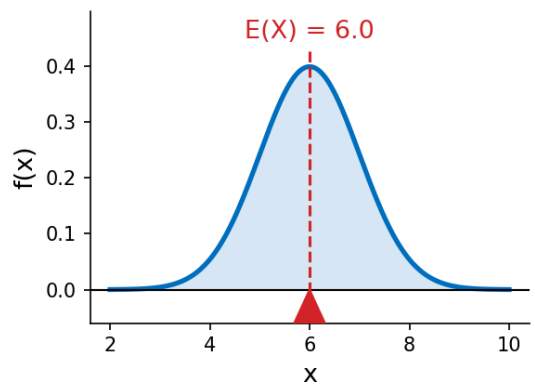
Expectation

What is the expected value when playing dice? What is the expected value of a probability density? Where is the centre of mass?

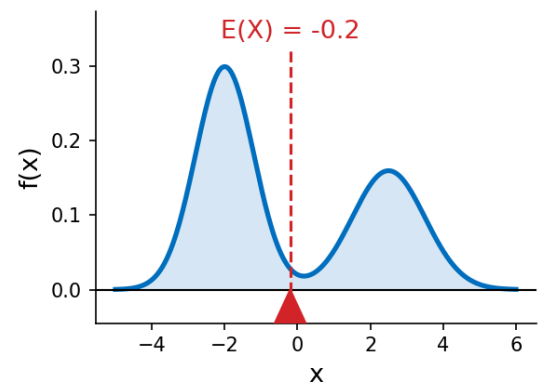
Dice (pmf)



Gaussian (pdf)



Camel (pdf)



$$E(X) = \sum_x x\, P(x) \qquad E(X) = \int x\, f(x)\, dx$$

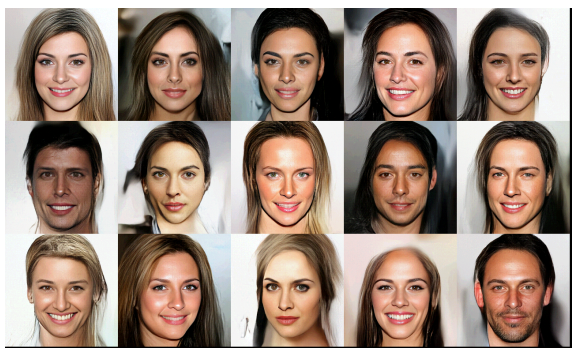
Probability: model

- A coin is thrown $n = 20$ times, the probability for one head is $\pi = 0.2$ how probable is it to have k heads? The parameters are known* $k \sim Binom(n = 20, \pi = 0.2)$
- You drive 100 km, how many liters x of gas do you need? We assume that the data is sampled/drawn from or generated by a normal distribution with $\mu = 6, \sigma^2 = 1$ (the parameters are known)* $x \sim N(\mu = 6, \sigma^2 = 1)$

*We here drop the distinction between X (RV) and x (sample). $X \sim N(\mu = 6, \sigma^2 = 1)$

Probability: model → data (sampling)

- Data varies
  - ipython example of car
    - `np.random.normal(loc=6,scale=1,size=10)`
  - Model with some parameters (knobs are μ and σ) $x \sim$ 
  - x can also be high dimensional data (256x256x3)



15 different x sampled from a deep learning model called glow (more than 100 mio. Parameters, 800 MB)

These are not real people!

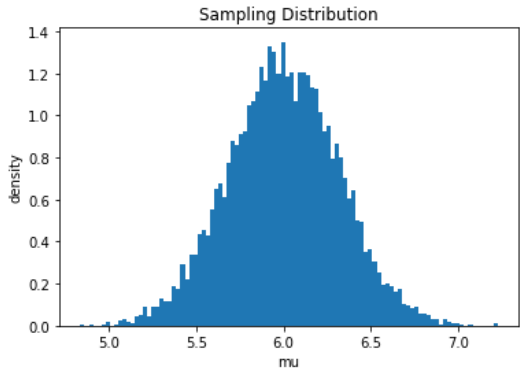
For glow see: arXiv:1807.03039

Inference: data → model (training)

- Parameter Estimation:
  - A coin is thrown n=20 times, you observe k=12 heads, what is your best guess for π?
    - $k \sim Binom(n = 20, \pi = ???)$
  - A car is driven n=10 times 100 km you observe the following 10 values:
    - 6.27045683, 5.94976189, 5.76105195, 5.09243634, 5.42322867, 6.75539123, 6.50091719, 5.02244476, 6.09933231, 6.75138712
  - You assume it is a Normal Distribution
  - $x \sim N(\mu, \sigma^2)$
  - What is your best guess μ and σ ? Any idea?

Goodness of estimation (Sampling Distribution)

- Is taking the mean of the data good estimate for the expectation?



- The expectation of estimate should be the real value (**unbiasedness**)
- Sample size $n \to \infty$ estimate should be true value (**consistent**)
  - a bit complicated to formulate that correctly (we don't care here)
- The spread of the sampling distribution should be small (efficiency)

Demo

https://github.com/ioskn/mldl_htwg/blob/master/intro/intro.ipynb

The power of sampling

- Often simple questions in stats cause problems.
- If possible sample
- Example
  - X ~ unif(0,1)
  - How is X^2 distributed?
  - What is the mean?

Excercise (with change of variable)

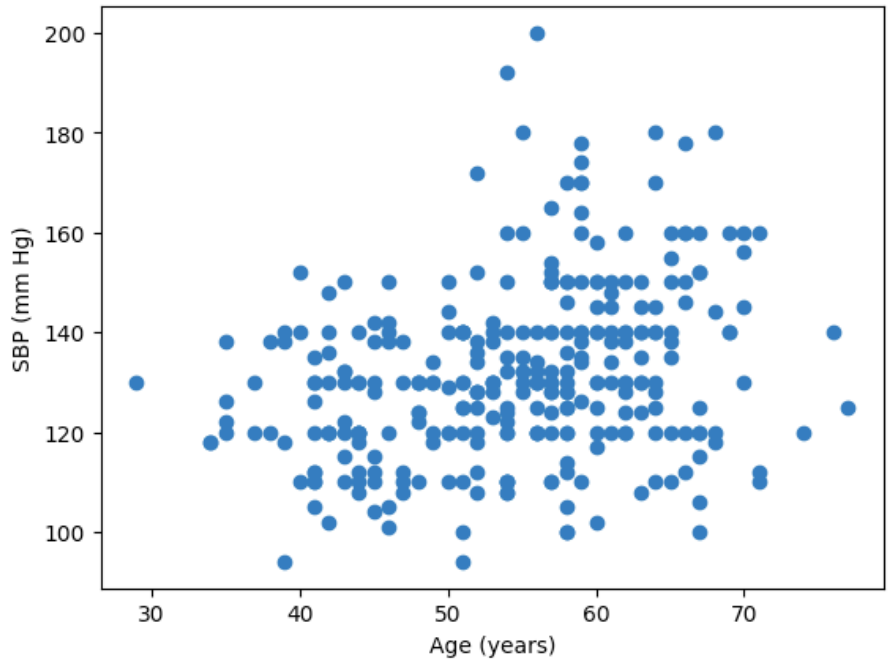
Demo (Power of Sampling)

https://github.com/ioskn/mldl_htwg/blob/master/intro/intro.ipynb

Correlation

We start in Data Land

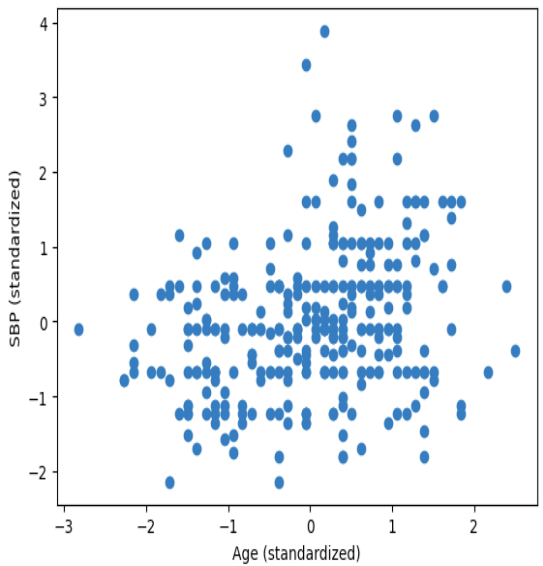
Is there a correlation between age and blood pressure?



Data from: https://archive.ics.uci.edu/dataset/45/heart+disease

Z-Transformation

- We need to remove the units (”Apple with Oranges”)
- We scale the data



- From the scaled data we calculate

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z(x_i) z(y_i)$$

- See:

https://github.com/ioskn/mldl_htwg/blob/master/intro/cor.ipynb

Pearson Korrelation

- Ausgeschrieben

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z(x_i)\, z(y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})/sd_x \cdot (y_i - \bar{y})/sd_y = \frac{\operatorname{cov}(X,Y}{sd_x\, sd_y}$$

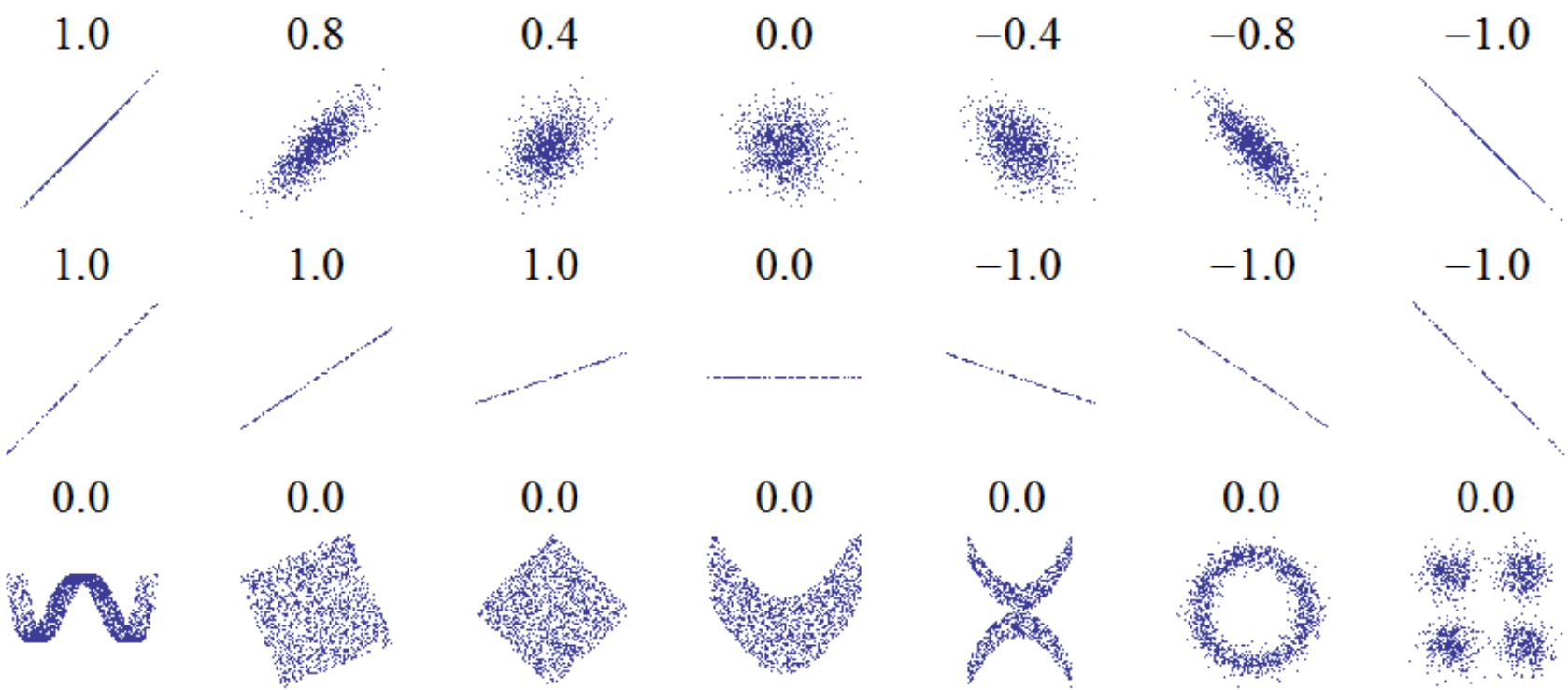
$$\operatorname{cov}(X,Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

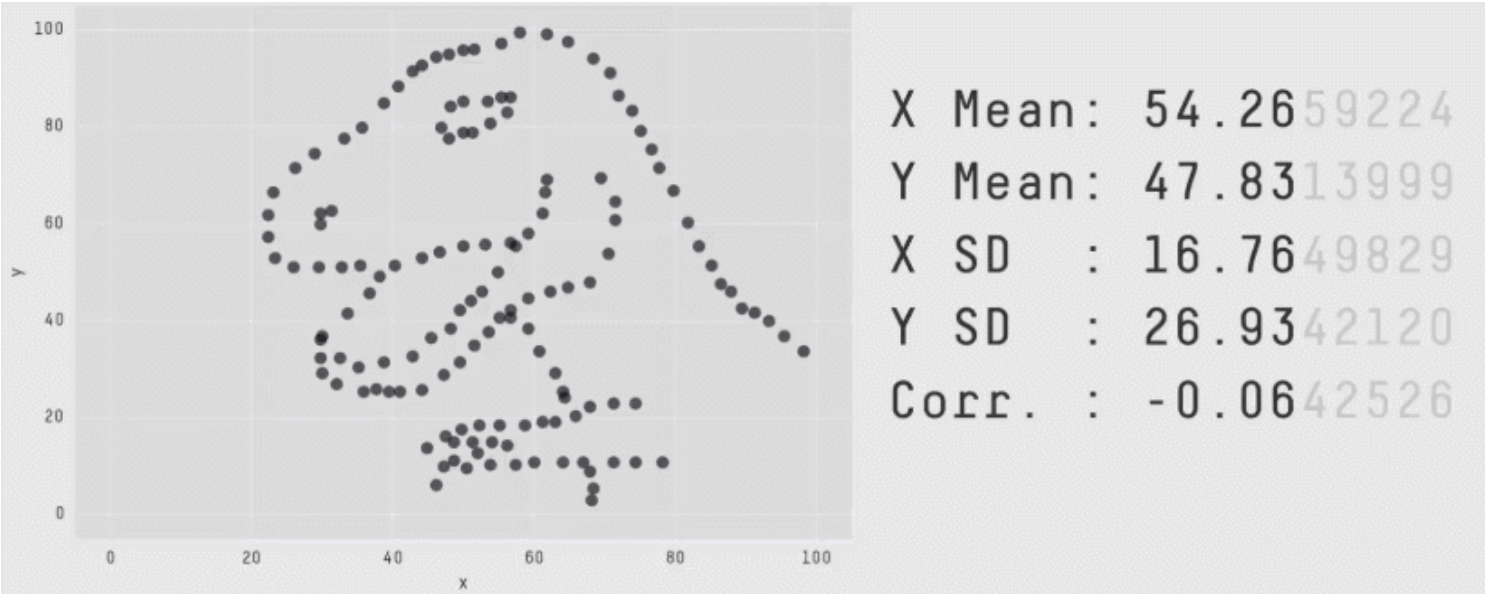
- Kovarianz ist Erweiterung der Varianz auf zwei Variablen

$$\operatorname{cov}(X,X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (x_i - \bar{x}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Scatterplots for various empirical correlation coefficients



Correlation only capture linear dependencies



MW 0, STD 1, Corr 0
https://www.r-bloggers.com/the-datasaurus-dozen/

Limits of correlation analysis

- Correlations are not the only form of statistical dependency. It often happens that two variables are uncorrelated but still are dependent on each other.
  - Use scatterplots!
- Estimates of correlation coefficients are quite unstable, especially for small sample sizes. In addition, the correlation coefficient strongly reacts to outliers. There are statistical tests for checking whether the estimated coefficient is significantly above 0.
- In multidimensional problems, the data cannot be displayed in a single scatterplot. In order to still get some idea about the correlational structure of our data, we display the scatterplot of each pair of variables in the form of a scatterplot matrix (not perfect) or do PCA

Correlation: In Probability Land

- Definition of correlation (not to be mixed up with correlation coefficient)

$$C_{xy} = E(x \cdot y) = \sum_x \sum_y x \cdot y \; p(x,y) \qquad \text{discrete}$$
$$C_{xy} = E(x \cdot y) = \iint x \cdot y \; p(x,y) \, dx \, dy \qquad \text{continuous}$$

- Estimation (E→MW) of the correlation on N data points

$$\widehat{C_{xy}} = MW(x \cdot y) = \frac{1}{N} \sum_1^N x_i \cdot y_i$$

- If x and y are statistically dependent, they typically are positive or negative at the same time (they are covariant). Consequently, their product is large and positive which results in a positive correlation.