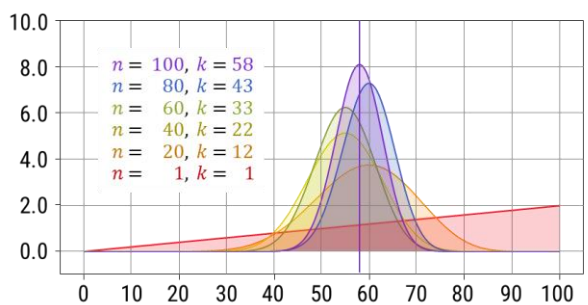
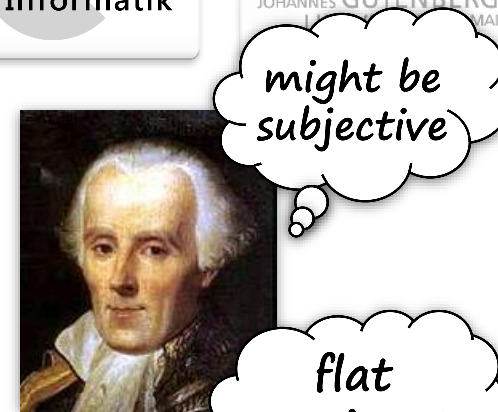
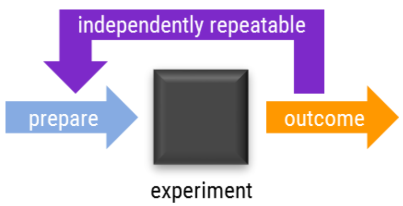
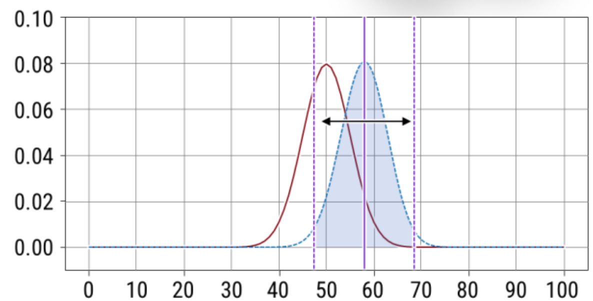
Modelling 2

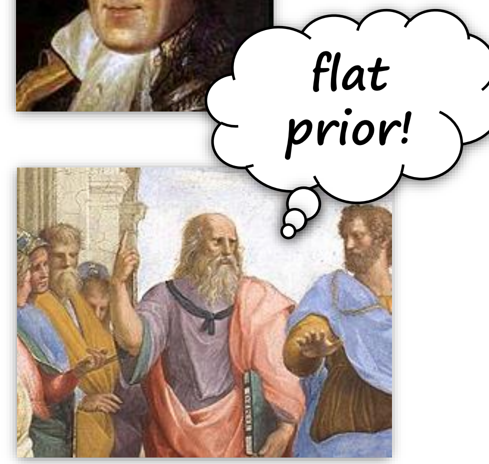
STATISTICAL DATA MODELLING











Chapter 3

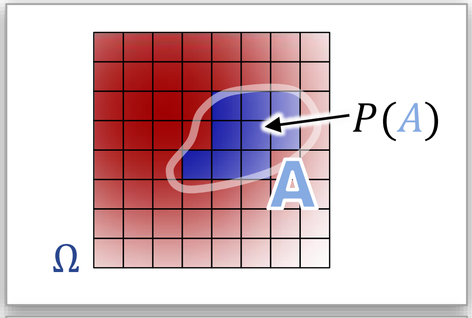
Classical and Bayesian Statistics

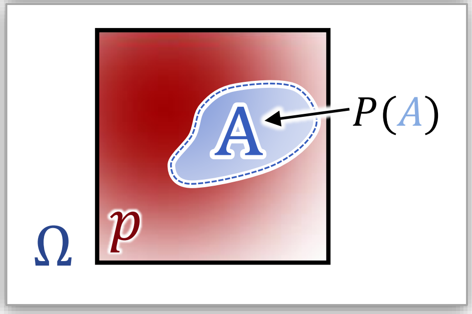
Michael Wand · Institut für Informatik, JGU Mainz · michael.wand@uni-mainz.de

What happened so far...

Probability model

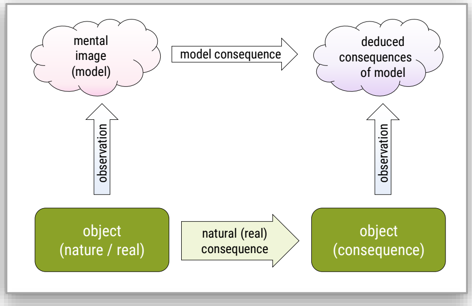
- Additive probability mass / density
- Empirical frequency approaches density with high likelihood

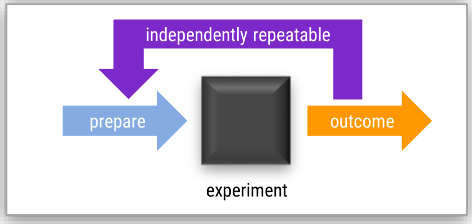




Now: Empirical sciences

- What can we learn from observations?
- How? (Algorithms)





How can we use Probability?

Again, (at least) two schools of though.

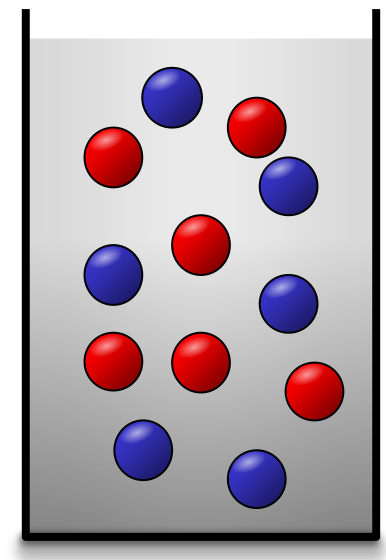
What is Probability?

Question

- *What is probability?*

Example

- A bin with 50 red and 50 blue balls
- Person **A** takes a ball
- Question to Person **B**:
  What is the probability for *red*?



What happened

- Person **A** took a blue ball
- Not visible to person **B**

Philosophical Debate…

An old philosophical debate

- What does "*probability*" actually mean?
- Can we use probabilities for
  - Events with fixed, already determined outcome?
    - But we do not know it for sure
  - Events in the future that will happen only once?

Philosophical Debate…

“Fixed outcome” examples

- Probability for: life on mars
- Probability that the code you wrote is correct

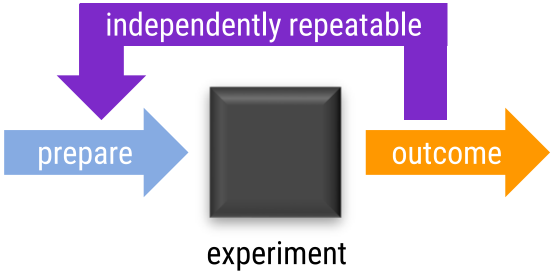
In the future, but not repeatable

- Probability for: rainfall tomorrow
- Probability for: Next season of SciFi-series canceled

Two Camps

Frequentists' (traditional) view

- Well defined experiment
- Probability = relative number of positive outcomes
- Only meaningful as a mean of many experiments



Bayesian view

- Probability expresses a degree of belief
- Mathematical model of uncertainty
- Can be subjective

[https://en.wikipedia.org/wiki/Thomas_Bayes]

Mathematical Point of View

Mathematical definition of probability

- Properties of probability measures
  - Defines rules for computing with probabilities
  - Consistent with both views
- Model building is not math
  - Which original probabilities to set/choose?
  - Question arises when performing empirical science

We will use both

- Bayesian approaches for algorithms
- Frequentist arguments for "objective" error bounds

Operational Perspective

Mathematics

- Same rules, but different models
- Bayesian view is “more liberal”: fewer restrictions

Operational Perspective

What Bayesian statistics permits (in addition)

- Everything can be a random variable
  - Models / model parameter
  - Facts & single outcomes ("does Mars harbor live?")
- Probabilities can be subjective
  - But must be consistent (Kolmogorov Axioms)
  - (Fairly general: Kolmogorov follow from Cox Axioms)

Frequentist: only experimental results "random"

- "Likelihood that a model is correct" not permitted
  (strictly speaking)

Learning from Data

(Maybe in its simplest possible form)

Example: Coin flipping

We found a coin

- Want to determine if/how fair it is

Probabilistic model

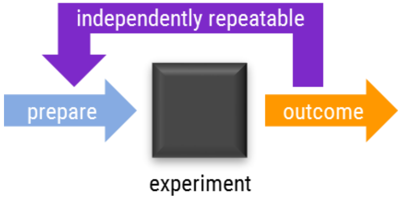
- Throw it once: Bernoulli experiment *(binary outcome)*

$$\Omega = \{0,1\}, \qquad \theta = P(1)$$

- Throw it n times *(independently)*:

Binomial distribution

$$P(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$



- Determine θ from experiment

Note: Quite General

Structurally important example

- Similar
  - Effectivity of medication
  - Likelihood of failure of a mechanical part
- Structure
  - Gaining one bit of information
  - Can repeat independently often

Learn p from Data $\mathcal{D}$

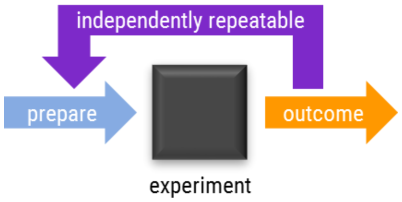
Experiment

- We collect data $\mathcal{D} = (x_1, \dots, x_n) \in \{0,1\}^n$
  - Data is i.i.d. (“*independently identically distributed*”)
- Model

$$k = \sum_{i=1}^{n} x_i\,, \qquad P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Experimental Result

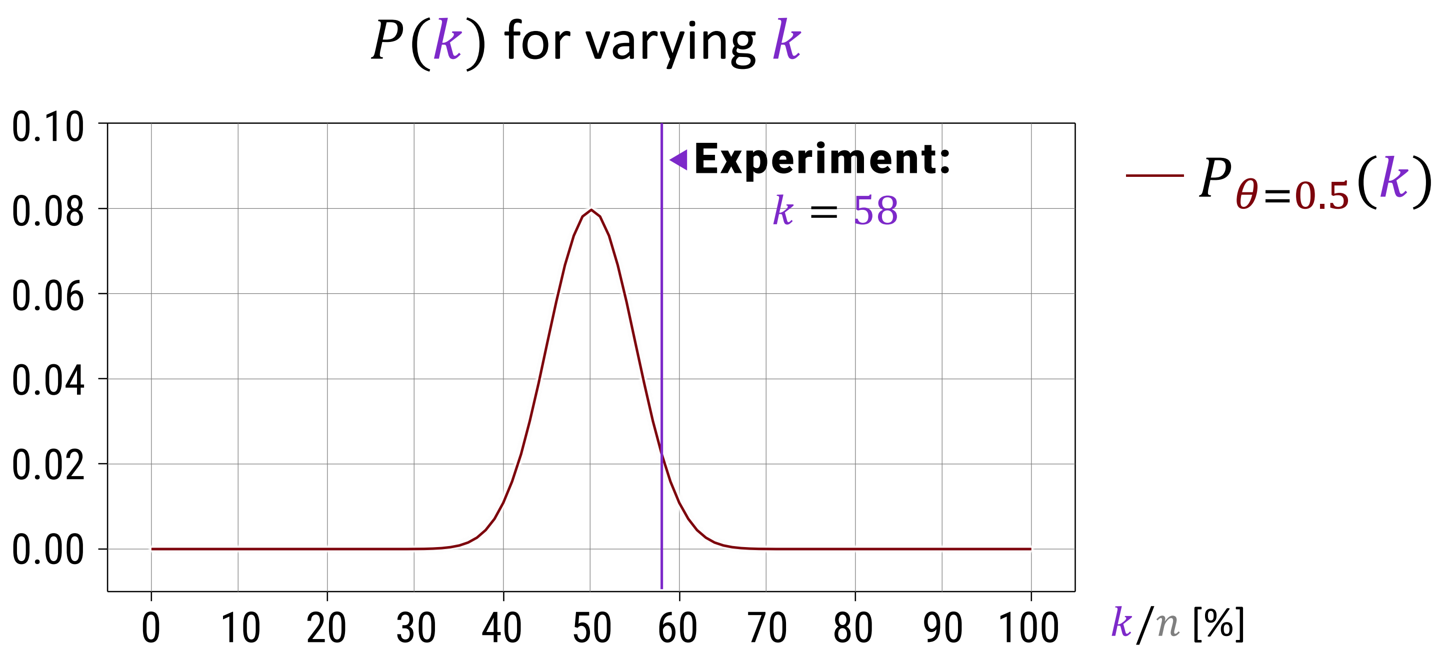
- We observe 58 “1”s for 100 coin tosses



Learning from Data

Part I: **(Classical) Frequentist statistics in action**

Fair Coin Toss: What to expect



Baseline

- $n = 100$
- $\theta = 0.5$ (fair)

Experiment

- $n = 100$
- $k = 58$

Frequentist Model

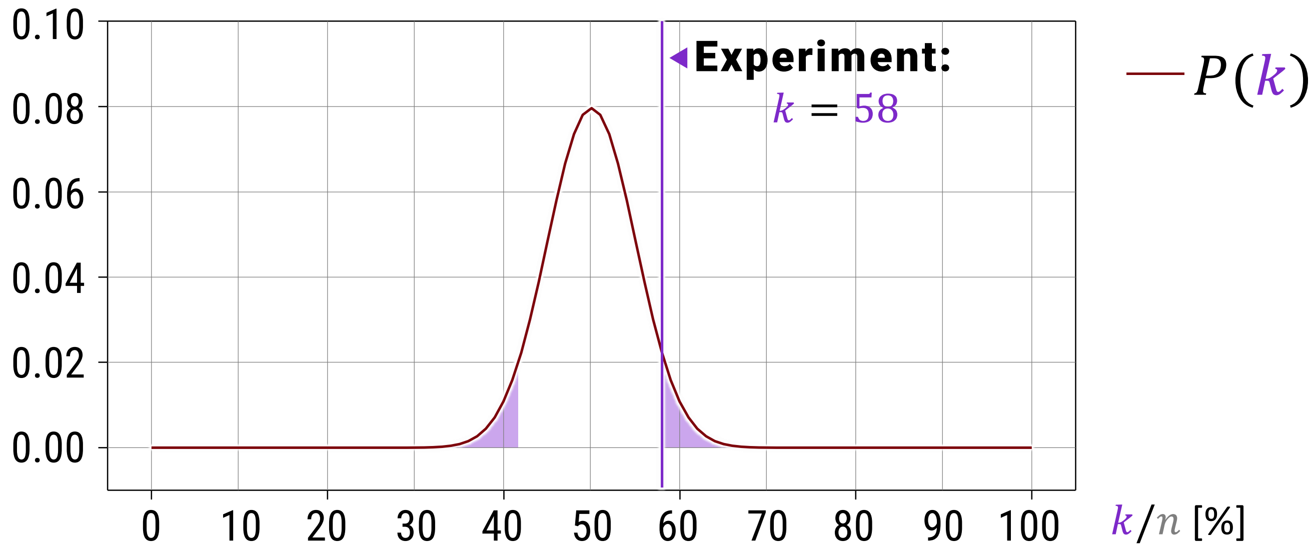
Possible questions

- Is the coin asymmetric (yes/no)?
  - "Two sided test"
- Has the coin been tampered with towards "1"?
  - "One sided test"

Null hypothesis

- The coin is fair ($\theta = 0.5$)
- How likely are different deviations?
  - We look at the two-sided test

Two Sided Test



How often do we observe deviations $\Delta k \geq 8$?

$$P(|k - 50| \geq K) = 2 \cdot \sum_{k=K}^{100} \binom{100}{k} \theta^k (1-\theta)^{n-k}$$

$$\approx 13\%$$

“Conclusion”

Assuming the coin was fair

- Seeing the result we got will happen (on average) to 13% of scientists (“$p = 0.13$”)
  - Likely enough that we usually will not reject fairness
  - Rather insufficient evidence for an unfair coin
- **Traditional cut-offs:** Likelihood of null-hypothesis
  - $p = 0.05$ („significant“)
  - $p = 0.01$ (“highly significant”)
  - $p = 2.7 \times 10^{-7}$ (“discovery” in fundamental physics)

“Conclusion”

Important

- The state of the world is unknown but fixed
  - Never talk about the likelihood of the coin being fair/unfair
  - “Reality” is objective, not probabilistic
- Outcomes of experiments are random
  - Not the “probability of coin is unfair”
  - But: “probability of observing such an outcome”

Of course

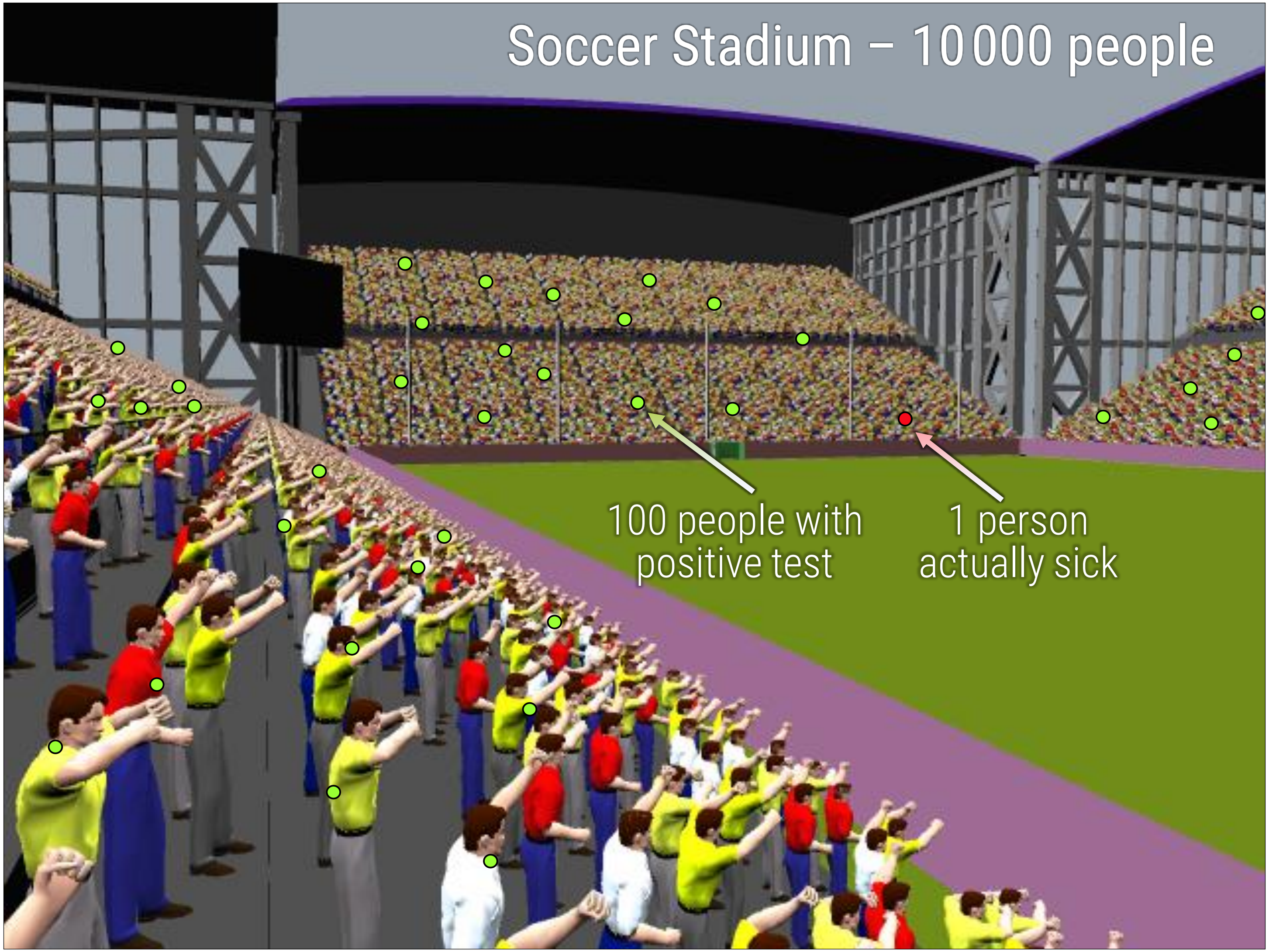
- Want to know the likelihood of the “coin unfair”
- What does $p = 13\%$ (or $p = 1\%$) tell us about it?

Example

Slightly more involved example

- Person feels unwell
  - Doctor runs several tests for rare (and “bad”) disease
- Test outcome “positive”. Statistically,
  - **Sick person:** Test always gives the correct answer
  - **Healthy person:** False positive with $p = 1\%$
- But, we also know
  - Disease is rare, only 1 in 10.000 patients has it
    - …of patients seeing a doctor…
    - …with these symptoms…
    - …not looking at any testing.

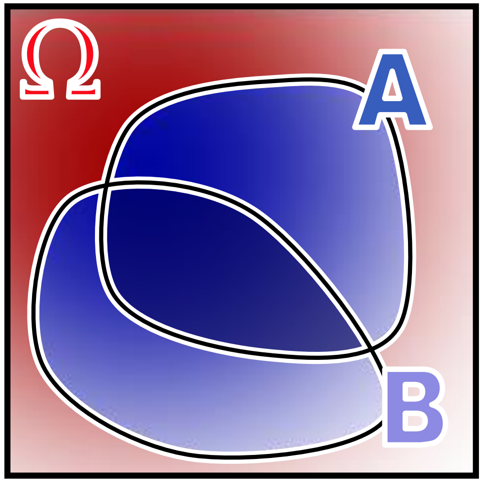
Intuition


Soccer Stadium – 10 000 people
100 people with positive test
1 person actually sick

How to Combine Likelihoods?

Bayes' rule

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}$$



Derivation

- $\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B)$
  $\Pr(A \cap B) = \Pr(B \mid A) \cdot \Pr(A)$

$$\Rightarrow \Pr(A \mid B) \cdot \Pr(B) = \Pr(B \mid A) \cdot \Pr(A)$$

Joint Probabilistic Model

Test Characteristics

	test neg	test pos	
not sick	0.99	0.01	$\leftarrow P(test \mid \overline{sick})$
sick	0.0	1.0	$\leftarrow P(test \mid sick)$

Disease Characteristics

not sick	0.9999	$\leftarrow P(sick)$
sick	0.0001	$\leftarrow P(\overline{sick})$

Joint Model

$$P(test, sick) = P(test \mid sick) \cdot P(sick)$$

Joint Probabilistic Model

Applying Bayes' rule

$$P(\text{sick}|\text{testPos}) = \frac{P(\text{testPos}|\text{sick}) \cdot P(\text{sick})}{P(\text{testPos})}$$

$$= \frac{P(\text{testPos}|\text{sick}) \cdot P(\text{sick})}{P(\text{testPos}|\text{sick})P(\text{sick}) + P\left(\text{testPos}\middle|\overline{\text{sick}}\right)P\left(\overline{\text{sick}}\right)}$$

$$= \frac{1.0 \times 0.0001}{1.0 \times 0.0001 + 0.01 \times 0.9999}$$

$$= \frac{0.0001}{0.0001 + 0.009999} \approx 0{,}009902$$

$$\approx 0{,}01$$ ← *most likely healthy*

New Conclusion

What did we do?

- Better model
  - Larger, more realistic probability space
  - Full model $p(test, disease)$
- Conclude that disease is unlikely even $p = 0.01$ test
  - Avoid "prosecutor's fallacy"

Still Frequentist

- This is still a frequentist model
- We just modeled correctly how experiments "repeat"

When does this turn Bayesian?

Other cases

- **Test results:** (all at $p \leq 0.05$)
  - Customers prefer green gummy bears over red
  - There is a new elementary particle
  - There is live on mars
  - There is live on mars, and it loves watching our sitcoms
- **We cannot assign prior probabilities here**
  - p(“live on mars”) is not frequentist

When does this turn Bayesian?

“Sagan principle”

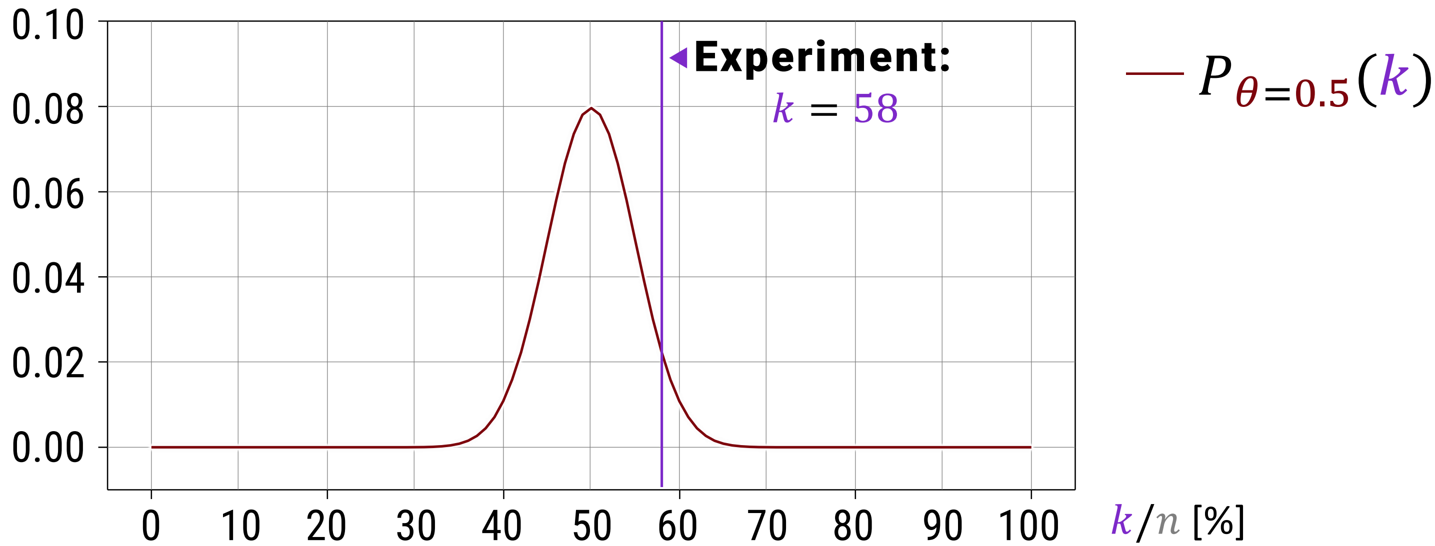
- “Extraordinary claims need extraordinary evidence”
  - Plausibility goes into judgement
  - P(“live on mars”) is “very low”
  - P(“live on mars watches Alf”) is even lower
- This is Bayesian now
  - Subjective
  - Probability for “facts”
    - They are true or false, strictly speaking
    - We only model our “believe”

Back to…

Learning from Data

Part I: **(Classical) Frequentist statistics in action**

Coin Toss Experiment



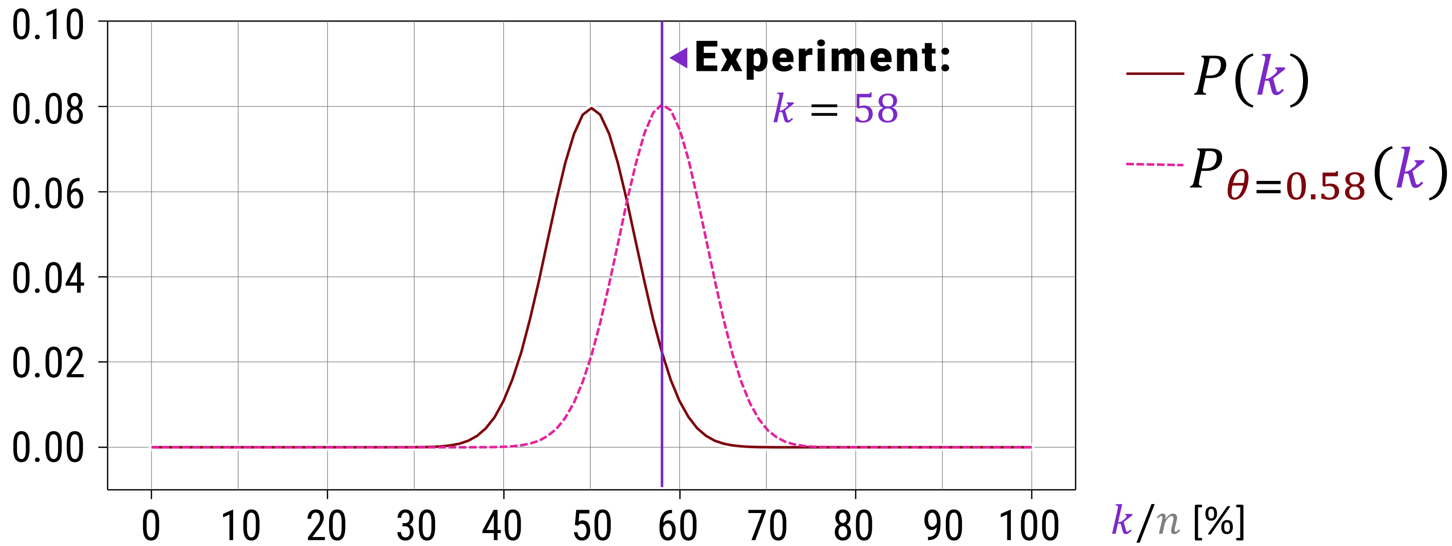
Baseline

- $n = 100$
- $\theta = 0.5$ (fair)
- $P(k)$ for varying k

Experiment

- $n = 100$
- $k = 58$

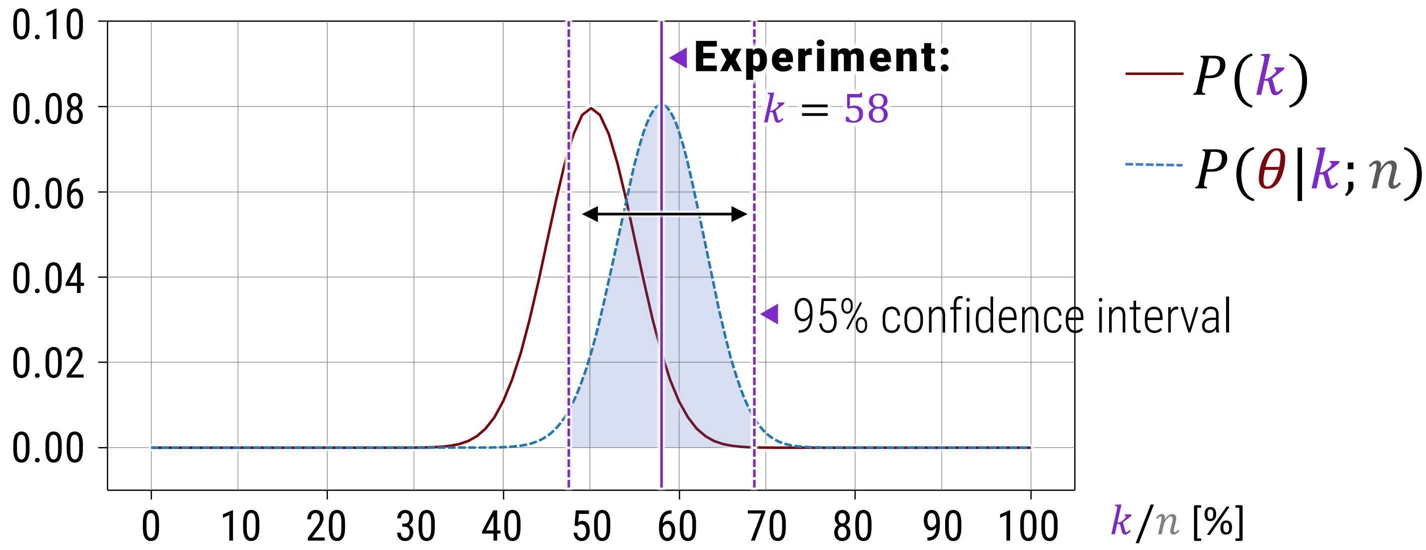
Coin Toss Experiment



Maximum Likelihood Estimator

- Estimate model parameter
- MLE: highest likelihood for observation: $\theta = 0.58$
- 95% “confidence interval” $k \in [48,68]$

Coin Toss Experiment



Maximum Likelihood Estimator

- 95% “confidence interval” $k \in [48,68]$
- Assuming $\theta = 0.58$ is the true model,
  95% of experiments will see outcomes $k \in [48,68]$
  - Not likelihood or spread of true value

Learning from Data

Part II: **Bayesian statistics in action**

Bayesian Variant

We now redo everything

- Bayesian framework
- Parameter "θ" is a random variable
  - Reminder: θ is the probability of "1"

Bayesian model

$$P(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

- No fundamental change
- Just consider θ as random variable now

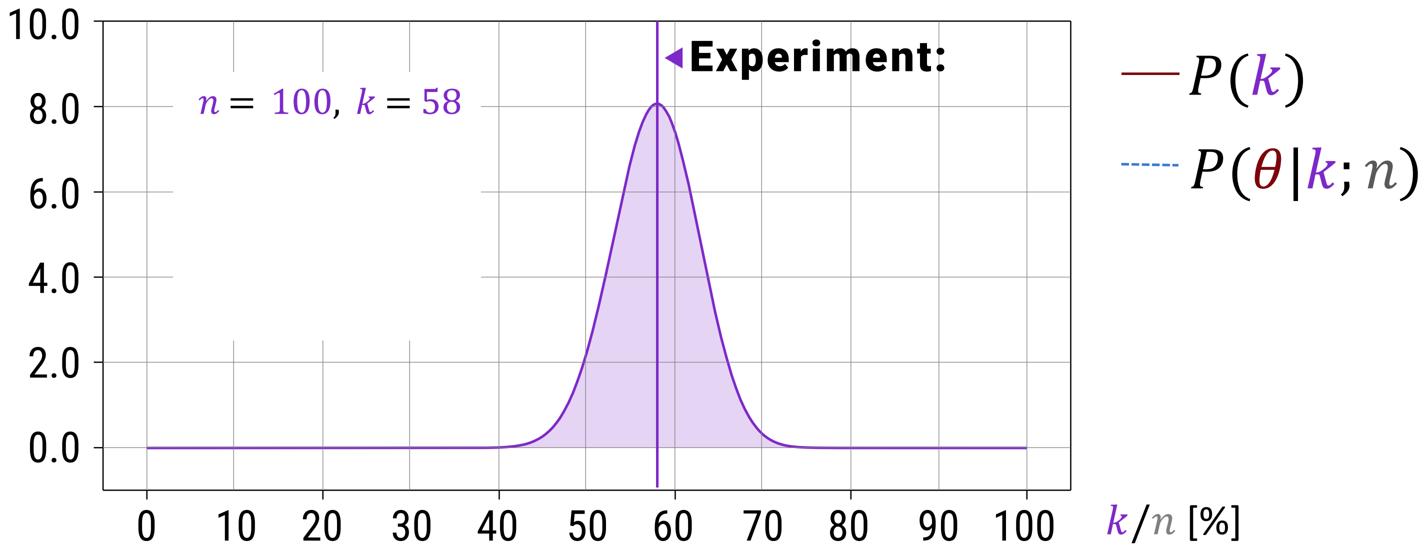
Bayesian Variant

Inference Model

- Use Bayes rule

$$\underbrace{P(\theta|k)}_{\text{"posterior"}} = \frac{\overbrace{P(k|\theta)}^{\text{"likelihood"}} \cdot \overbrace{P(\theta)}^{\text{"prior"}}}{\underbrace{P(k)}_{\substack{\text{"evidence"} \\ \text{("marginal likelihood")}}}}$$

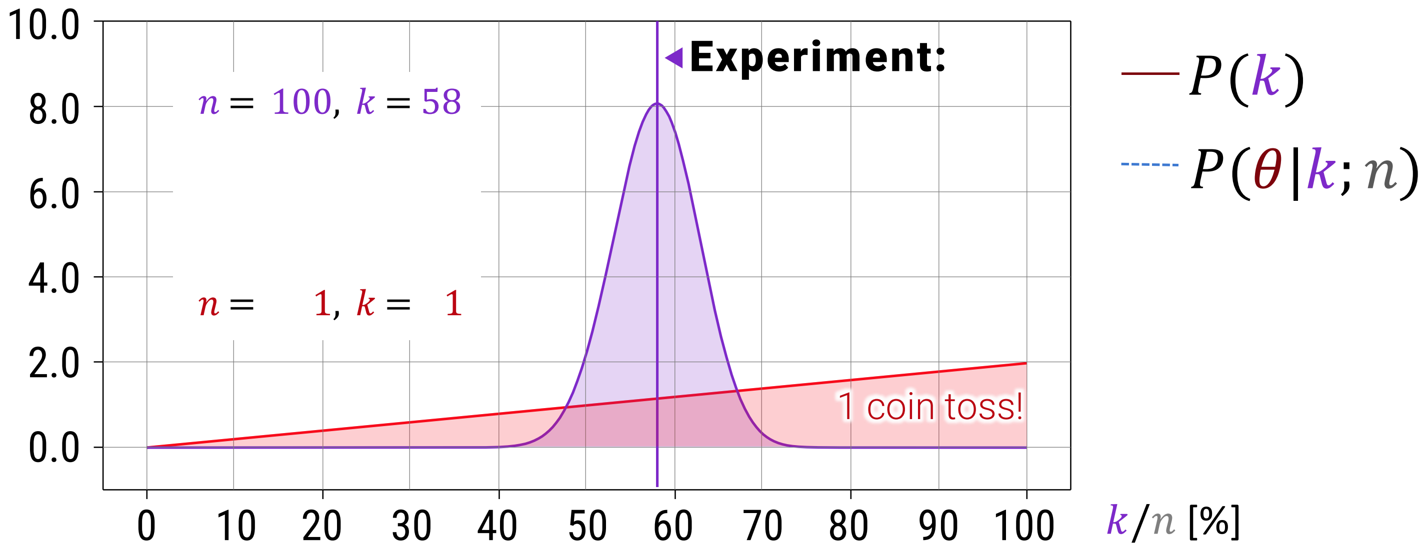
This is how it looks like



Bayesian approach

- Yields probability density over parameters
  - Allows to use uncertainty
- Principle: Keep uncertainty as long as possible!

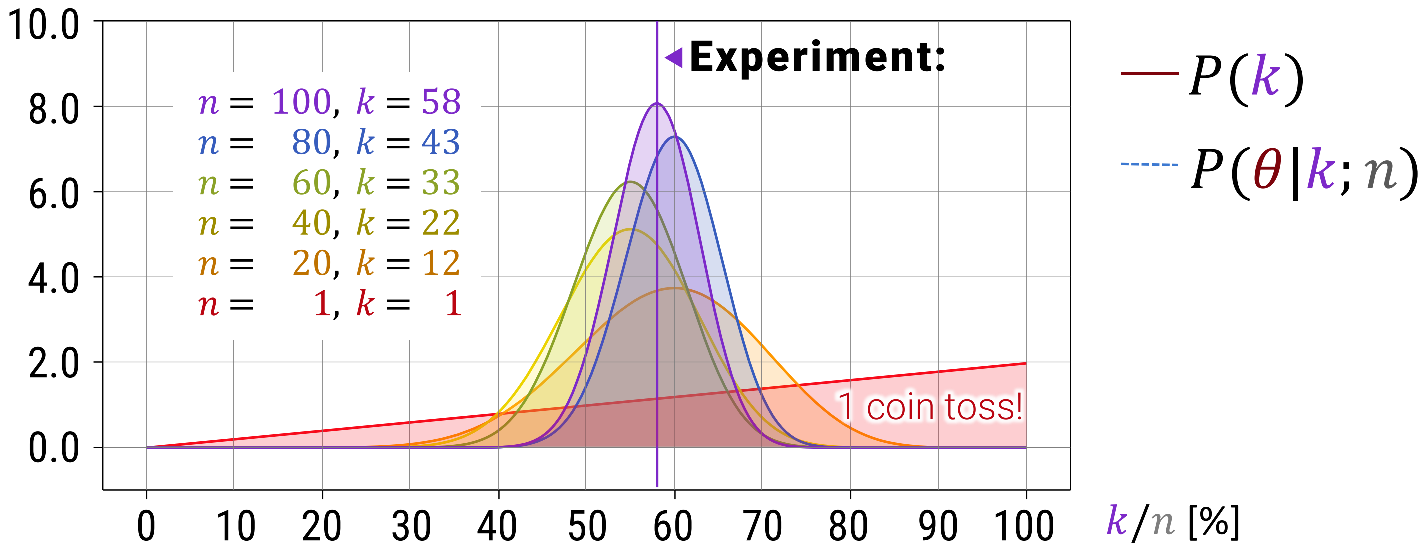
This is how it looks like



Bayesian approach

- Yields probability density over parameters
  - Allows to use uncertainty
- Principle: Keep uncertainty as long as possible!

This is how it looks like



Bayesian approach

- Yields probability density over parameters
  - Allows to use uncertainty
- Principle: Keep uncertainty as long as possible!

But we want a value!

Inference

- Maximum à posteriori

$$\theta_{\text{est}} = \arg\max_{\theta} P(\theta|k)$$

same as
maximum likelihood
for
flat prior $p(\theta)$

- "True Bayesian": Marginalization

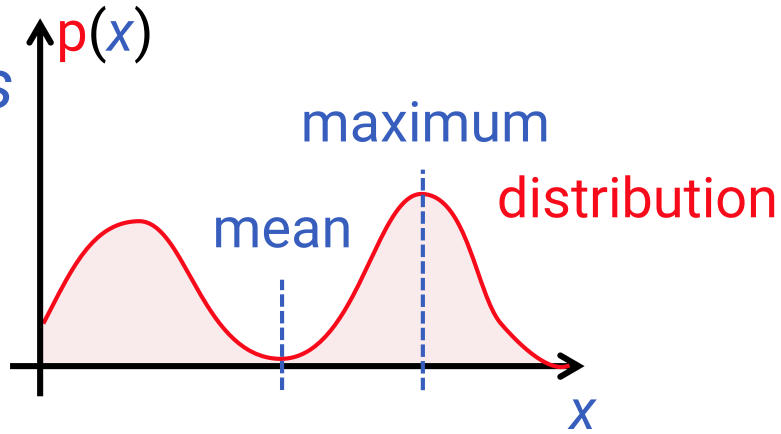
$$\theta_{\text{est}} = \int_{\theta=0}^{\theta=1} \theta \cdot P(\theta|k)\, d\theta = \mathbb{E}_{\theta}[P(\theta|k)]$$

This is
"Bayesian style"

Two Types of Inference

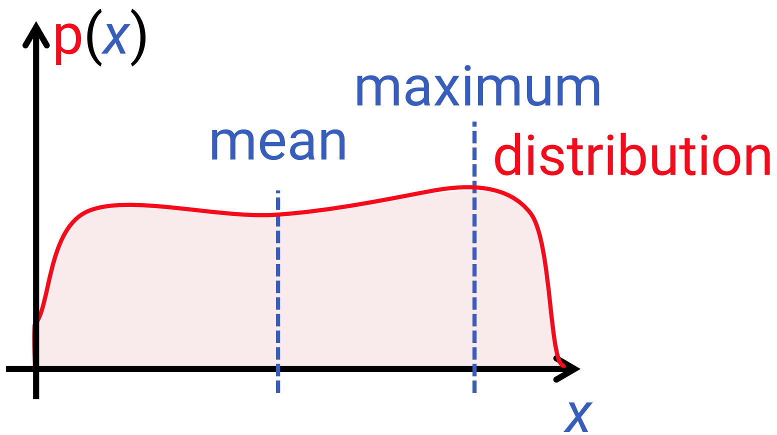
"Estimation"

- Output *most likely parameters*
  - Maximum density
    - "Maximum likelihood"
    - "Maximum a posteriori"
  - Mean of the distribution



"Bayesian inference"

- Output probability density
  - *Distribution for parameters*
  - More information



- Marginalize to reduce dimension

Bayesian Variant

In our example

- Use Bayes rule

$$P(\theta|k) \sim \frac{P(k|\theta)\overbrace{P(\theta)}^{\substack{\text{(uninformative)}\\ \textbf{flat prior}}}}{\underbrace{P(k)}_{\substack{\textbf{constant}\\ \text{(after experiment)}}}}$$

$$= \binom{n}{k} \theta^{k} (1-\theta)^{n-k}$$

- Point of maximum density = expectation = 0.58
  - Simple binomial distribution
  - No priors used

MLE? MAP? BI?

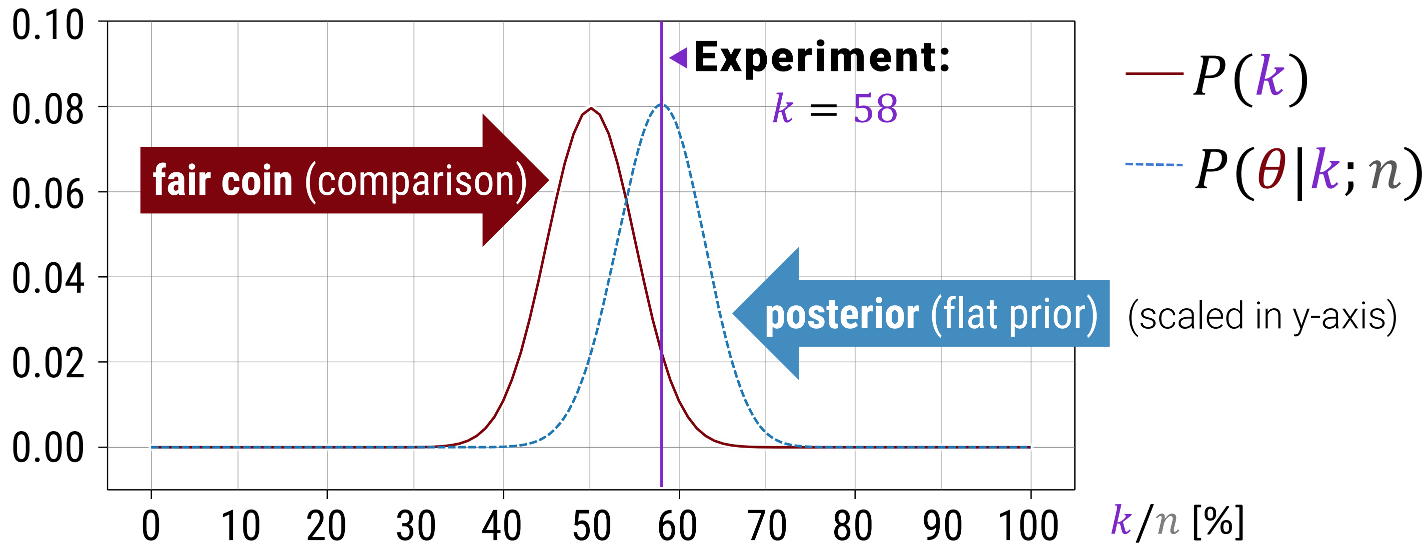
Maximum likelihood vs. a posteriori

- Prior needed if problem is ill-posed
  - Not enough information from data
  - And vice-versa: MLE ok for highly constrained models

Marginalization vs. Maximum A Posteriori

- No difference for simple distributions (Gauss, Binom)
  - Pronounced differences possible in complex models
- “Full Bayesian” inference usually reduces overfitting
  - Integrating over models favors simple models
  - Unfortunately, it is often very (too) costly

Fair Coin Toss: What to expect



Baseline

- $n = 100$
- $\theta = 0.5$ (fair)

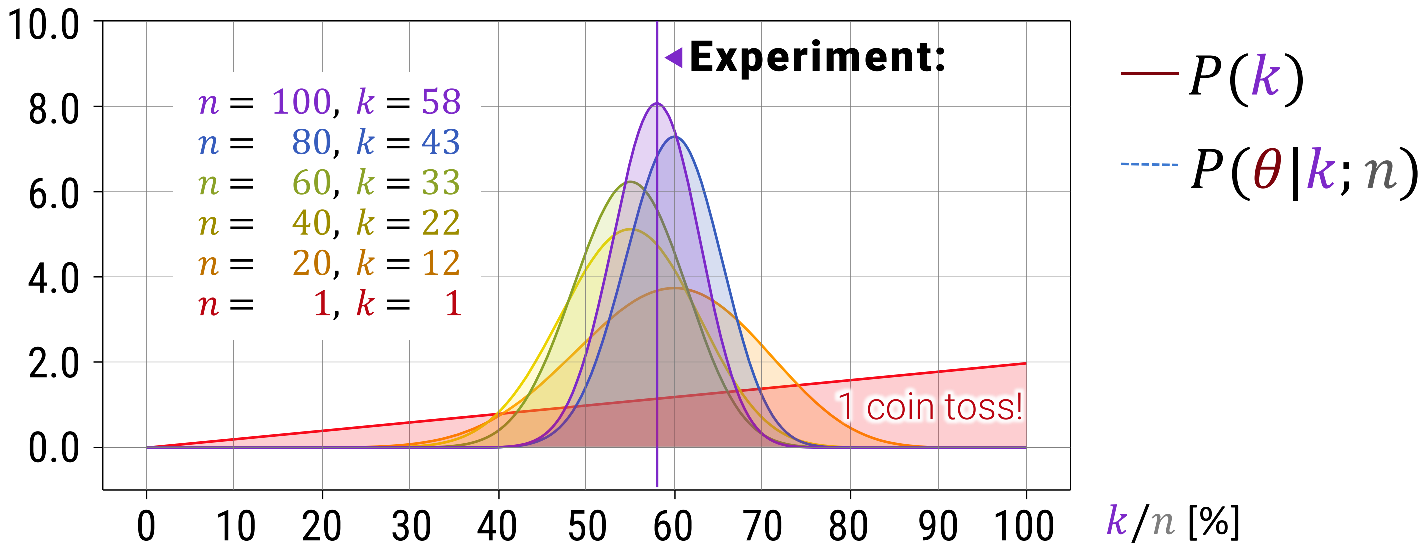
Experiment

- $n = 100$
- $k = 58$

Conclusion

- $\theta = 0.58$ most likely (MLE/MAP/Mean same in this case)

Uncertainty!



Principle

- Keep uncertainty as long as possible!

Summary

Bayesian & Frequentist Statistics

Bayesian features

- Any knowledge can be probabilistic
  - Also: models & model parameters ("$p(\theta)$")
  - No need for repeatable experiment
- Knowledge can be subjective
  - Hand-crafted "priors", not learned from data

Disadvantages

- Model parameters as random variables "$p(\theta)$" implies the use of priors
  - Explicit or implicit – no way around knowledge modeling
- Frequentist: use "only" knowledge from data

What is it good for?

Bayesian vs. classical (frequentist)

- No “subjective” priors: Often same results
  - But Bayesian approach lets us keep uncertainty along
  - “Feels easier to use”
- Bayesian: general prior knowledge
  - Different results if we had assumed coin “likely fair” or “likely biased towards 1” or the similar

What is it good for?

My personal / subjective impression

- Bayesian vs. frequentist techniques all plausible
- Differences arise for subjective priors
  - Unavoidable when modeling distributions over parameters
  - “Uninformative priors” are not always (never?) possible

When frequentist?

- Prove objective effect
  - E.g.: Show that result in a scientific paper is “significant”
  - E.g.: Measure accuracy of a (ML-) model
- Subjective probabilities harm credibility

What is it good for?

When Bayesian?

- Modeling knowledge
  - Of a subjective agent
  - Learn knowledge from data (over time)
  - Quantify and encode uncertainty
- Ill-posed problems
  - When data cannot provide all the information
  - Regularization needed!
  - Regularly the case in ML-applications
    - Try explaining “cat images” without prior assumptions
- “AI” and “machine learning”
  - Any complex result impossible without priors

#goBayesian

How do we do it?

Bayesian Principles

Model building

- Specify a complete model $p(x_1, \dots, x_d)$ $(\Omega = \mathbb{R}^d)$
  - Always needed – not specifically Bayesian
  - We can – in principle – compute any event probability
- Use Bayes' rule to fuse probabilistic knowledge
  - Combine observations and prior knowledge
- Use statistical priors to encode "helpful" information
  - If there is not enough data, you need priors
    - In ML, we always need priors!
  - Btw: This is also true for frequentism
    - Priors are build implicitly into the parametrization
    - But do not distort "confidence" values

Bayesian Principles

Inferring knowledge

- "Learning" models
- Inferring "predictions" from fixed models

If too costly, use MAP with an appropriate prior

What to do

- Marginalize over all irrelevant variables
  - This might include model parameters
  - Reduces potential for overfitting
- Result is the function or value that remains
  - Function: free variables of interest remain
  - Value: expectation of the model over "everything"

More to come

We will practice this in the next video.