Statistics for Physicists

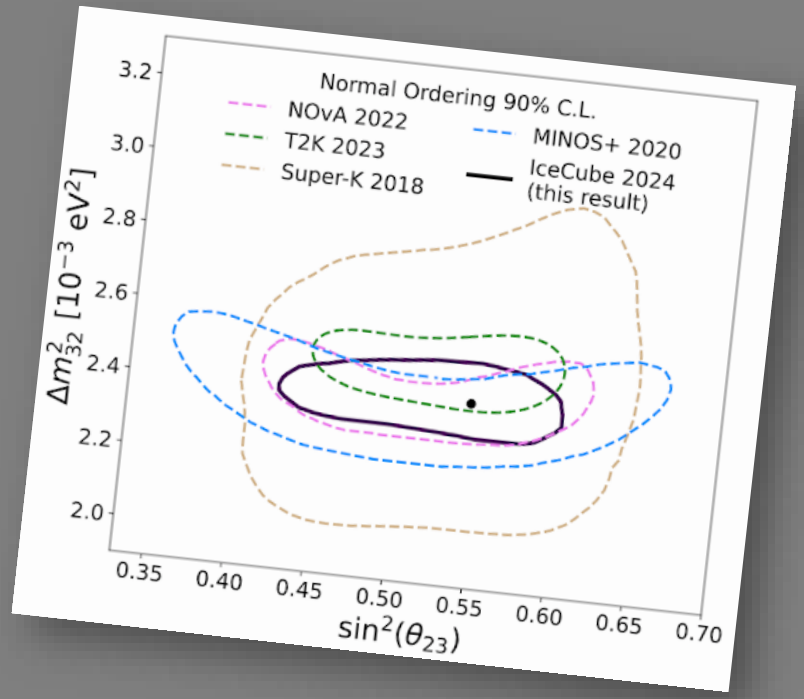
Karlsruhe School of Elementary Particle and Astroparticle Physics (KSETA)

March 2-4, 2026

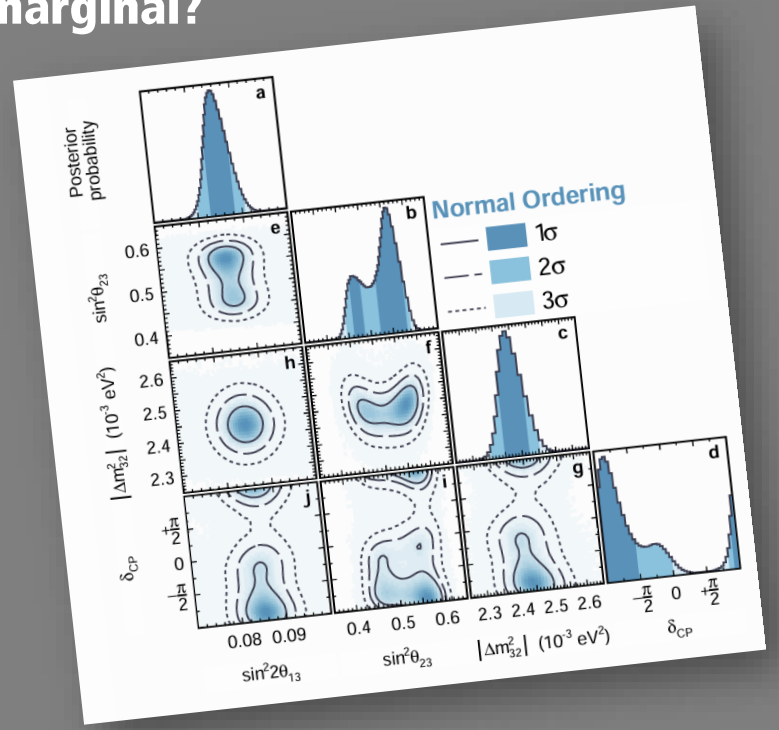
Dr. Philipp Eller (TU Munich)

philipp.eller@tum.de

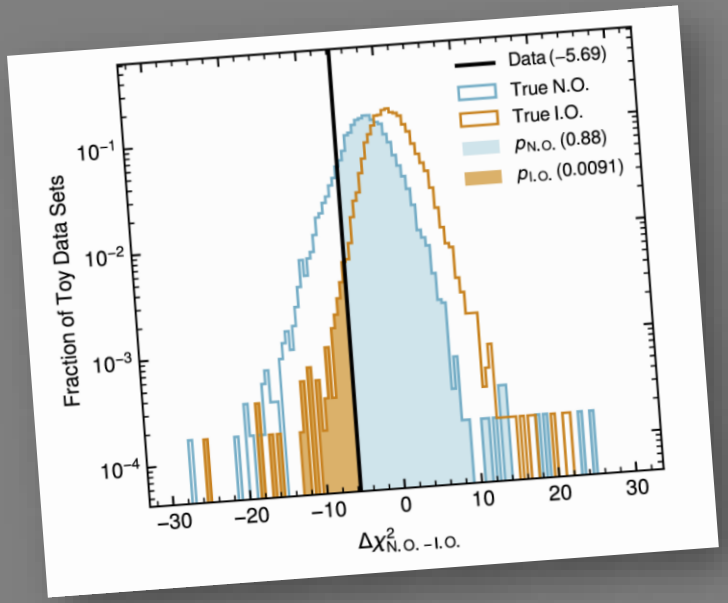
What does „90% C.L.“ mean? Aren‘t these just fancy error bars?



What‘s a posterior? What‘s a marginal?



What‘s a p-value?



What‘s the deal with these 3 and 5 sigma statements?

We report on a measurement of astrophysical tau neutrinos with 9.7 years of IceCube data. Using convolutional neural networks trained on images derived from simulated events, seven candidate ν_τ events were found with visible energies ranging from roughly 20 TeV to 1 PeV and a median expected parent ν_τ energy of about 200 TeV. Considering backgrounds from astrophysical and atmospheric neutrinos, and muons from $\pi^\pm/K^\pm$ decays in atmospheric air showers, we obtain a total estimated background of about 0.5 events, dominated by non-ν_τ astrophysical neutrinos. Thus, we rule out the absence of astrophysical ν_τ at the 5σ level. The measured astrophysical ν_τ flux is consistent with expectations based on previously published IceCube astrophysical neutrino flux measurements and neutrino oscillations.

Overview of Course

Day I	Day II	Day III

• **Probability Basics (90 min)**
  • Probability
  • Distributions
  • Moments
  • Conditional Probability
  • Likelihood
 	
• **Point Estimators (90 min)**
  • Simple Estimators
  • Properties
  • Bias-Variance Tradeoff
  • Maximum Likelihood Estimator
	• **Hypothesis Testing (90 min)**
	  • Single Hypothesis Tests
	  • P-values
	  • Two Hypothesis Tests
	  • Size vs. Power
	  • Neyman_Pearson Lemma
	  • Profile LHR
	
	• **Confidence Level Intervals (90 min)**
	  • Neyman Construction
	  • Interpretation
	  • Wilk‘s and Wald‘s
		• Bayesian Inference **(150 min)**
		  • Closed Form Inference
		  • Role of Priors
		  • Special Priors
		  • Marginals
		  • Model Comparison
		  • Numerical Methods (MCMC)
		
		• **Review of Course (30 min)**
		  • Reiteration of the most important concepts
		  • Interpretation Frequency vs. Bayes

Every now and then (~1x per lecture) we will stop for a hands-on exercise → good to have a pen & paper ready

Probability Basics

What is probability?

- *Rolling a die, what is the probability of getting a six?*
  - Easy: it‘s $\frac{1}{6}$, right?
  - The probability comes from the physical symmetry and construction of the die – a fair die.
  - So it is a property of the die itself to have this internal tendency (or disposition) to produce outcomes
  - In this view, probability is a causal property: the die makes certain outcomes happen with certain tendencies, even if you never roll it

→ This is the „**Propensity**“ interpretation of probability

What is probability?

- *Commuting to work every day: what‘s the probability of being late more than 5 minutes?*
  - This question becomes very difficult in the propensity interpretation, because being late is not a property of the train with a built-in tendency — it depends on a complex mix of circumstances.
  - What we can do, however, is give up on this concept and simply study the relative frequency of things happening.
  - → Over a long period, every day you note down the delay. The fraction of cases where you are more than 5 minutes late will converge to its probability (law of large numbers).
  - A probability exists only because the experiment can be repeated many times.

→ This is the „**Frequeqntist**“ interpretation

What is probability?

- *What is the probability that the wrapped gift contains a game console?*
- This question can neither be addressed with the propensity or the frequentist view. It does not make sense to attribute frequencies to an event that cannot be repeated, and there is no tendency for the package to contain a certain gift. Rather, there is some uncertainty from your incomplete knowledge.
- Also, this should depend on other factors, such as:
  - Do you even like gaming?
  - Is this something you perhaps had expressed a wish for?
  - Is this in the price range you'd expect a gift to be?

→ In the "**Bayesian**" interpretation, probability corresponds to a "degree of believe"

Interpretation

- Depending on the school of thought, probability has different meanings:
- **The „physical" or "objective" probability (Frequentism, Propensity)**
  - Argues in terms of the frequencies of occurrence of particular outcomes
  - In the limit, the frequency will converge to the *true* probability
- **The "evidential" or "subjective" probability (Bayesian)**
  - Assign probabilities not only to outcomes, but also to other unknowns
  - Probability represents a "knowledge" or "degree of believe" about something, not necessarily a frequency of outcomes

Regardless of the interpretation, we can start to mathematically formalizing rules about probability that are true in either interpretation

Probability

The rules of probability are summarized in the **Kolmogorov Axioms**

Axiom 1: Probability of an event x must be positive: $p(x) \geq 0$

- Probabilities of two different events cannot cancel each other out (compare to e.g. amplitudes in QM!)

Probability

The rules of probability are summarized in the **Kolmogorov Axioms**

Axiom 2: Probability of anything happening is unity $p(\Omega) = 1$

- If your world consists of **six** possible events (e.g. in a die), it is guaranteed that one of them happens.
- It's important to enumerate all possibilities

Probability

The rules of probability are summarized in the **Kolmogorov Axioms**

Axiom 3: Probabilities of disjoint events add up.

$$p(x \text{ or } y) = p(x) + p(y)$$

- **Holds only for disjoint events:**
  - **yes:** roll a 3 or a 6
  - **no:** roll an even number or a 4

Parameterized Models

Most commonly, we encounter probabilities of an outcome x that depend on certain parameters θ

$$x \sim p(x|\theta)$$

We say „p of x given θ“

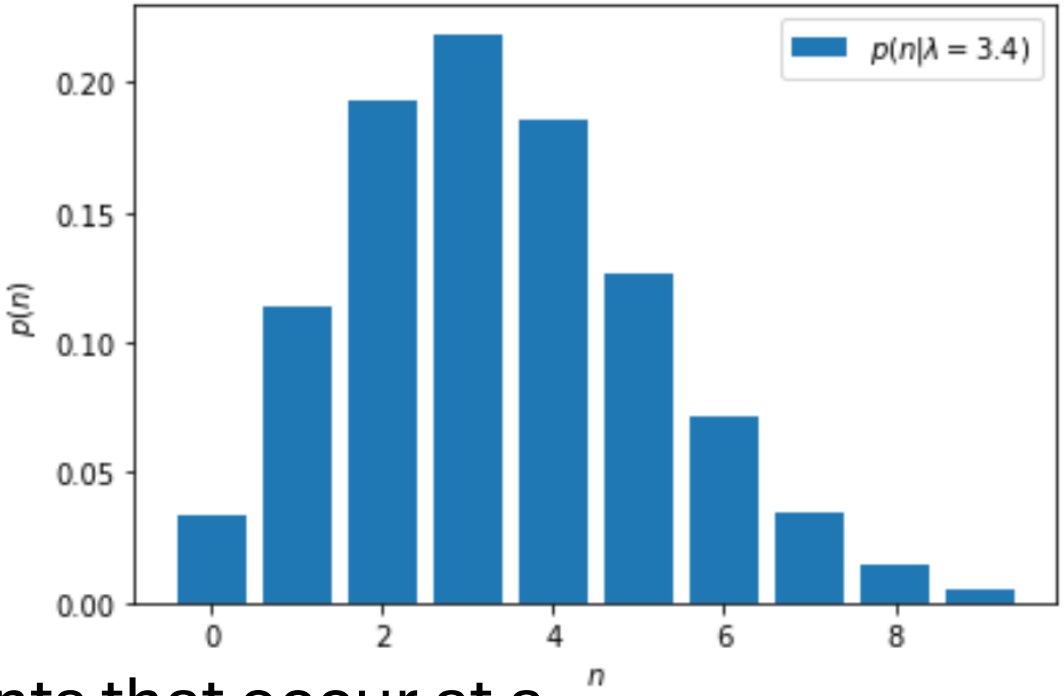
We speak of a „parametric model“ if we can express this in the form of distributions, or at least a data-generating process (forward model)

- Sometimes this is a simple distribution with one, two parameters, …
- … sometimes a very complicated model with many hundrets of parameters!

Examples of a Parametric Model

Poisson Counting Experiment: $\theta = \lambda$

$$p(n|\theta) = \text{Pois}(n|\lambda) = \frac{1}{n!}\lambda^n \exp(-\lambda)$$



Occurs naturally if you are counting events that occur at a known rate within a fixed time window.

Continuous Probabilities

Sometimes, we would rather not specify certain discrete events that are possible, but allow for a continuum

Example: measuring people‘s heights, we might not be interested in the probability of someone being exactly 179.034582… cm

So here, rather intervals are interesting, e.g. between 189 and 190 cm

Since we do not want to define intervals per se, It can be useful to define a probability density over a space

$$P(a < x < b) = \int_a^b p(x)dx$$

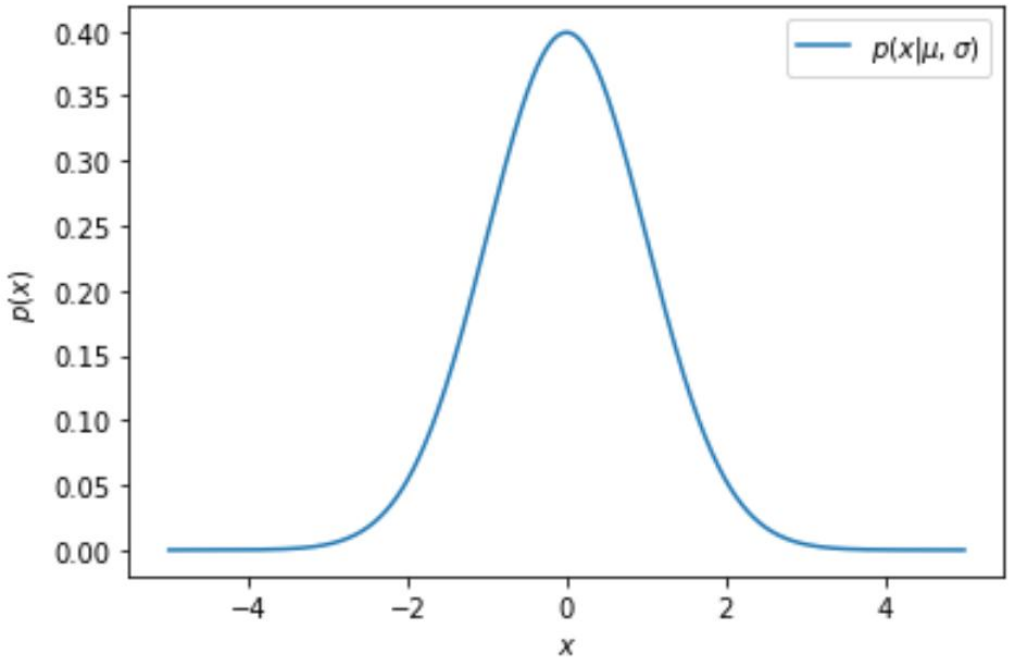
Probability

Probability Density

Examples of a Continuous, Parametric Model

Normal Distribution: $\theta = (\mu, \sigma)$

$$x \sim p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$



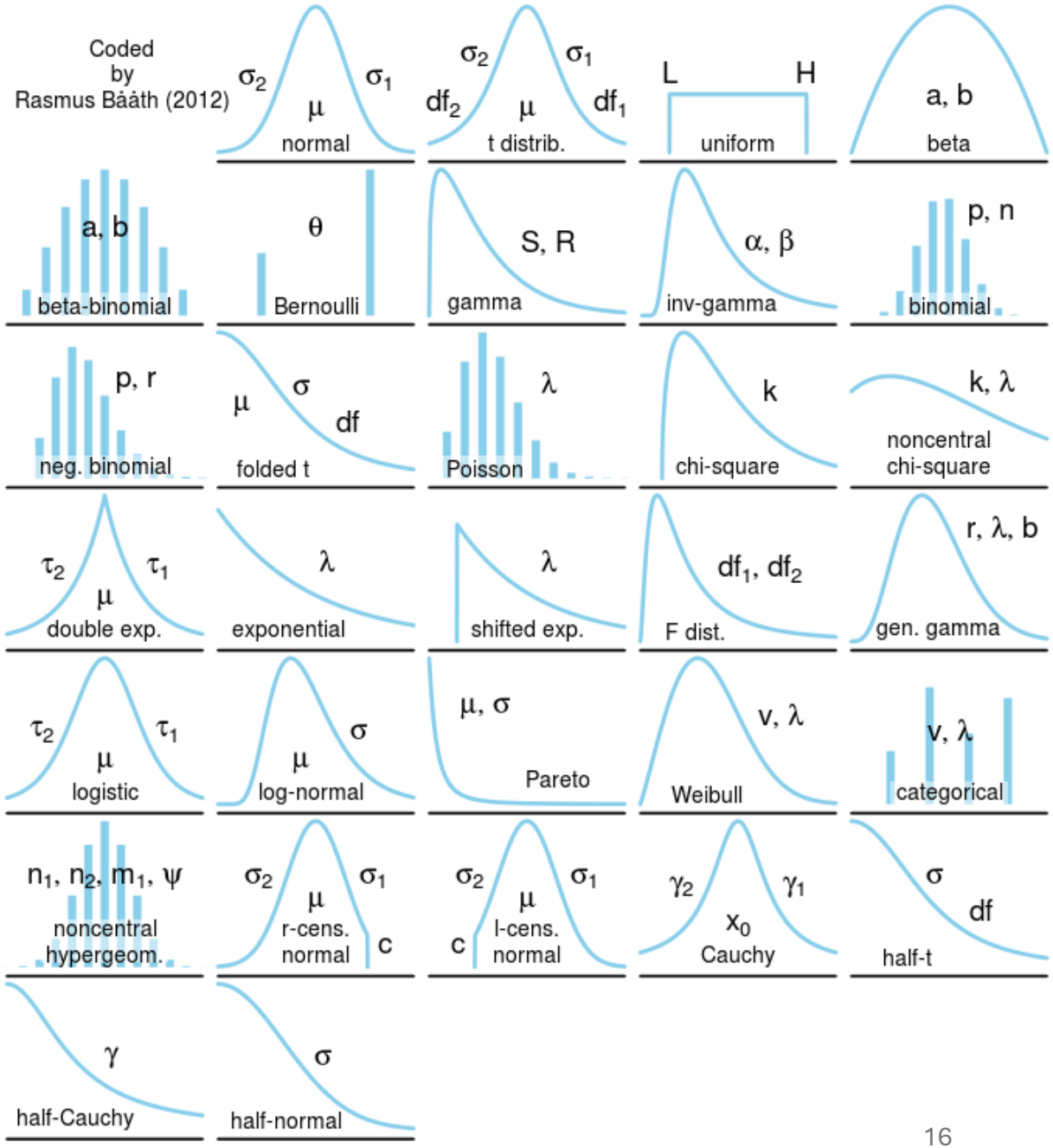
Occurs naturally if the measured quantity is the sum of many individual random processes (e.g. measurements on a detector module)

→ **"Central Limit Theorem"**

Other Distributions

- What distributions do you know / have you heard of?

- See for example: https://en.wikipedia.org/wiki/List_of_probability_distributions



Expected Value and Variance

- **Expected Value** E is the mean of the possible values a random variable can take, weighted by the probability of those outcomes

$$E[x] = \int xp(x)dx$$

→For Poisson: $E[x] = \lambda$

- **Variance** is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value

$$Var(x) = E[(x - E[x])^2]$$

→ For Poisson: $Var(x) = \lambda$

Exercise

Mean and Variance

- Compute the **Mean** and **Variance** of the standard uniform distribution

$$p(x) = U(0,1) = f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{else} \end{cases}$$

- $\mathbb{E}[x] =$

- $Var(x) = \mathbb{E}[(x - \mathbb{E}[x])^2] =$

$$\mathbb{E}[x^2] = \int x^2 p(x) dx = \int_0^1 x^2 dx = \frac{1}{3}x^3\Big|_{x=0}^{1} = \frac{1}{3}$$

$$\mathbb{E}[(x - \frac{1}{2})^2] = \mathbb{E}[x^2 - x + \frac{1}{4}] = \mathbb{E}[x^2] - \mathbb{E}[x] + \frac{1}{4} = \mathbb{E}[x^2] - \frac{1}{4}$$

$$= \frac{1}{3} - \frac{1}{4} = \frac{4}{12} - \frac{3}{12} = \mathbf{\frac{1}{12}}$$

Cumulative Distribution

- The cumulative distribution (cdf) $F(x)$ is defined as $P(X \leq x)$
- So in practice, for a continuous probability distribution $p(x)$ this is $F(x) = \int_{-\infty}^{x} p(t)dt$

- Since probabilities are normalized, $F(x)$ always maps $p(x)$ to the unit interval [0,1]

→ This is very useful, for example, for generating random numbers according to p(x), by transforming random numbers $y \sim U_{[0,1]}$ via the inverse of the cdf $x = F^{-1}(y)$

Likelihood Function

The likelihood is simply $p(x|\theta)$ **viewed as a function of** θ **and fixed** x

The likelihood function:

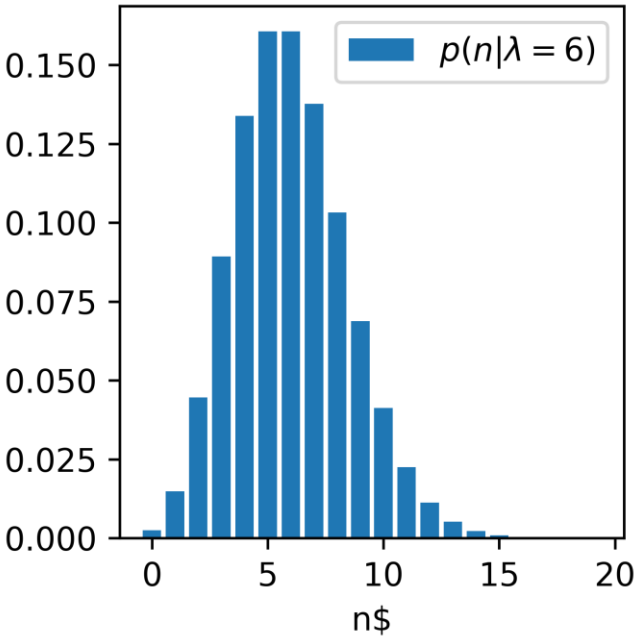
$$L_x(\theta) = p(x|\theta)$$

The more probable the observed data x is under a value θ, the higher the "likelihood value" of θ.

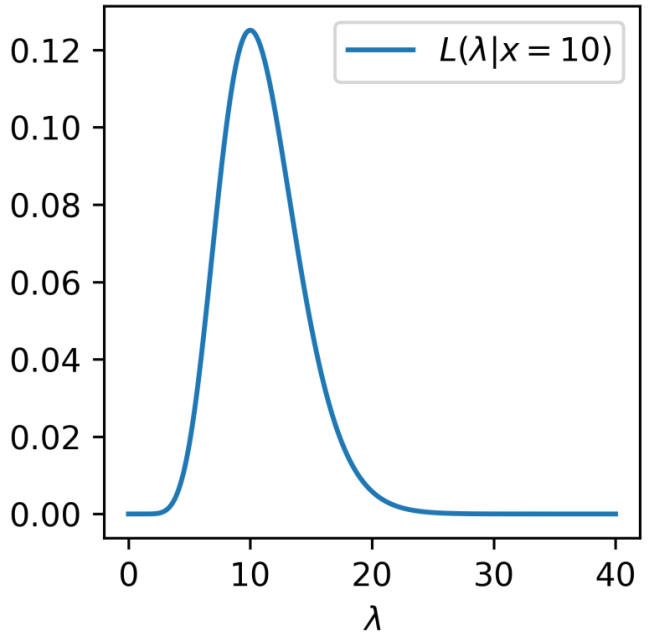
This is **very different from a probability**! We did not specify what θ is, and it cannot be assumed these are random variables

Example: Poisson Likelihood

Consider $n \sim \text{Pois}(n|\lambda)$:



probabilities for observations at a fixed $\lambda = 6$

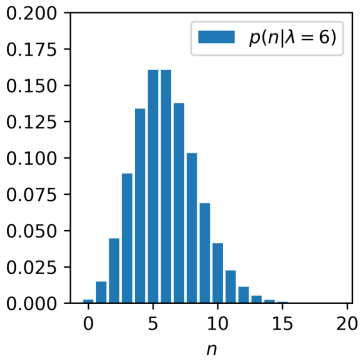


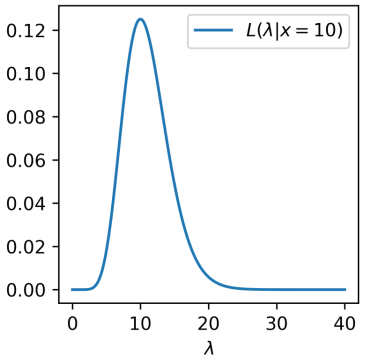
likelihoods for observations for a fixed $n = 10$

Likelihood vs. Probability functions

Important to remember that likelihood and probability functions are different

(naming is a bit unfortunate & doesn't help)





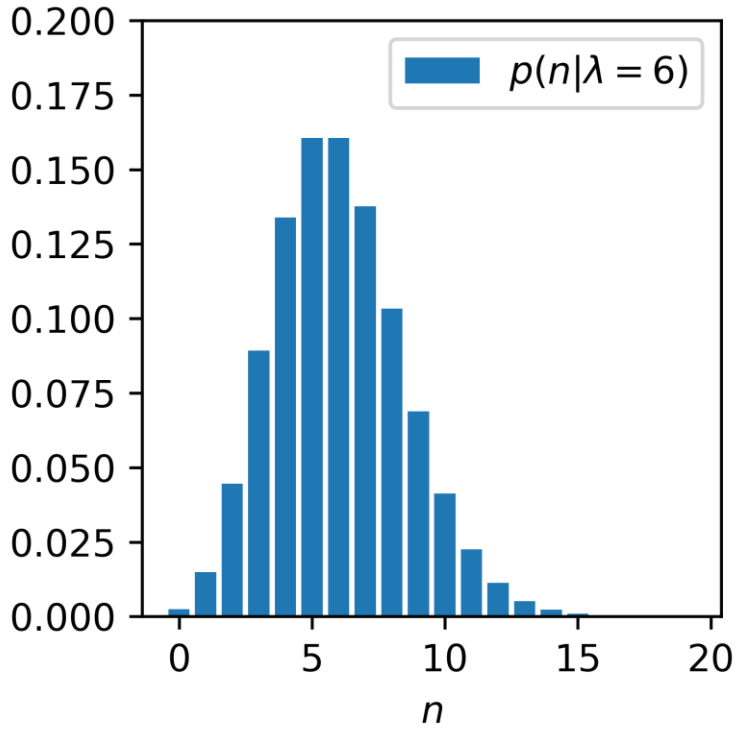
Likelihood vs. Probability functions

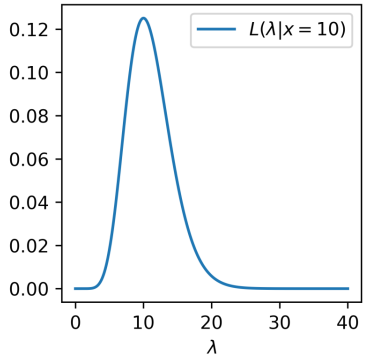
Important to remember that likelihood and probability functions are different

Probability:

fixed parameters θ, variable data x

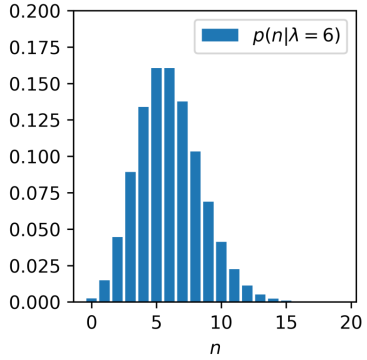
normalized $\int p(x|\theta)\mathrm{d}x = 1$





Likelihood vs. Probability functions

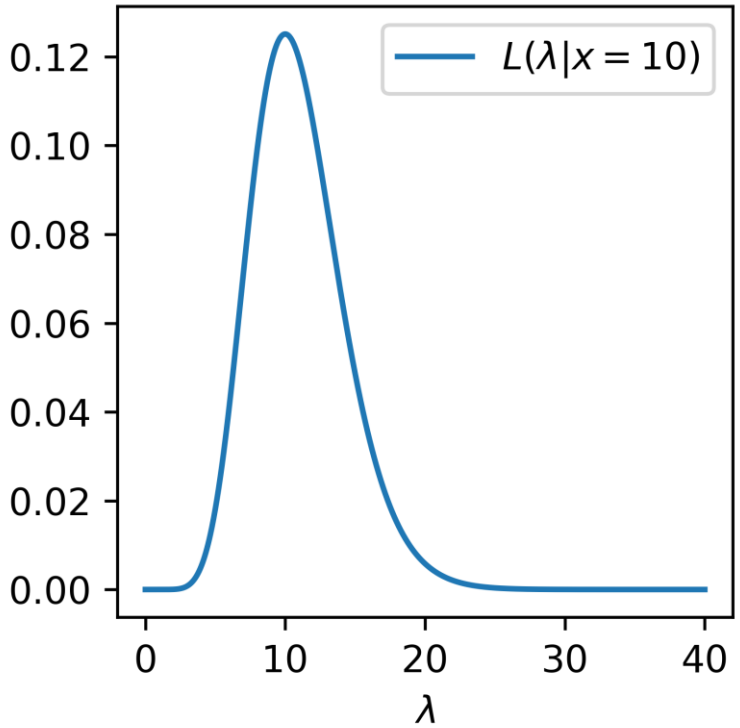
Important to remember that likelihood and probability functions are different



Likelihood:

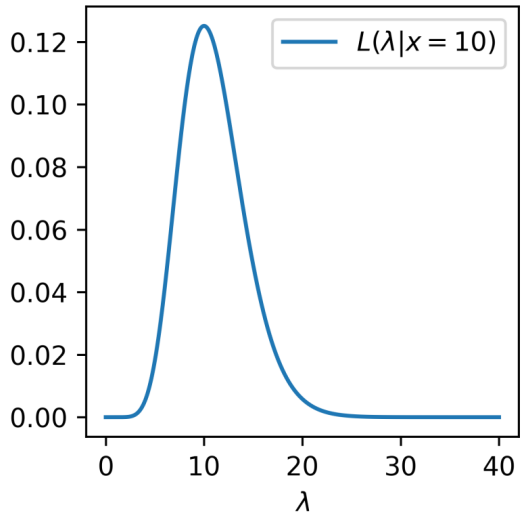
fixed data x, variable parameters θ

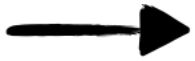
not normalized ~~$\int p(x|\theta)\mathrm{d}\theta = 1$~~

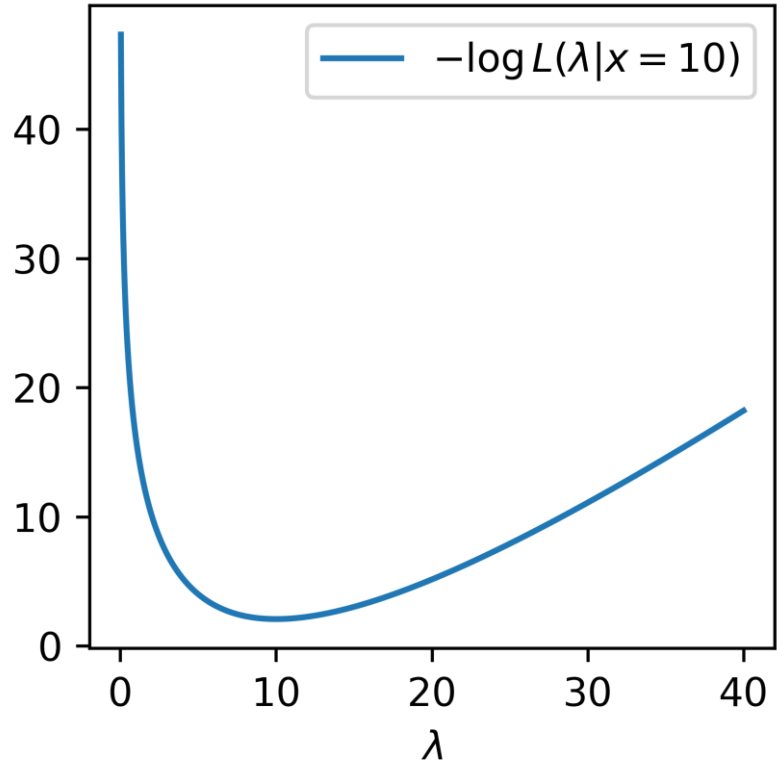


Negative Log Likelihood

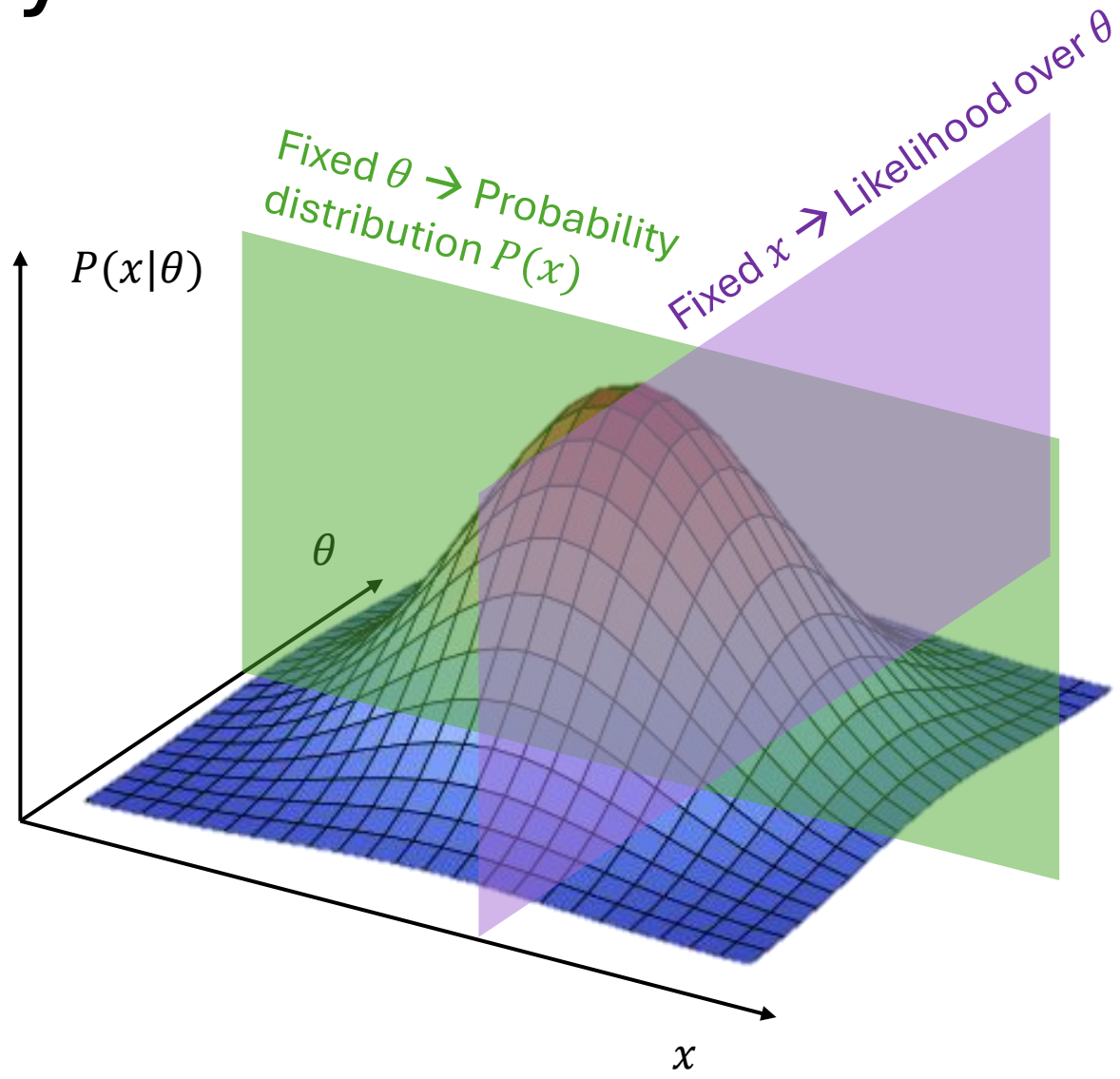
To avoid confusion and general usefulness often we rather use the (negative) log-likelihood LLH (NLL) function $\mathrm{nll}(\theta) = -\log L(\theta)$







Probability vs. Likelihood



Why again is the likelihood not a probability?

- **Reason 1:** Distribution parameters are abstract concepts and not per se random variables

- **Reason 2:** Simple counter-example:

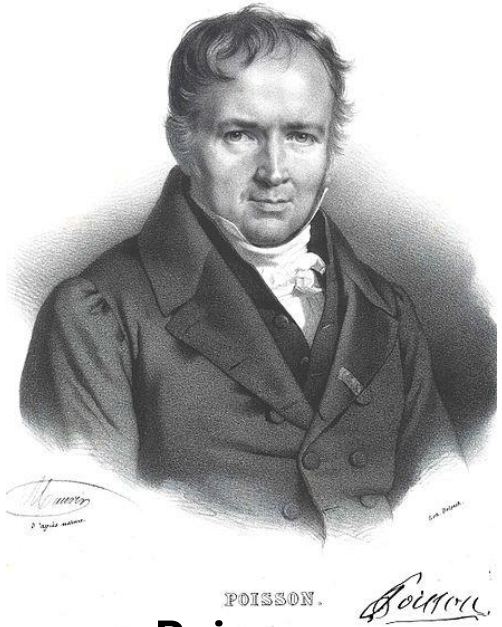
Uniform distribution: $U(a = 0, b > 0) = \frac{1}{b}$

$$\int_0^\infty \frac{1}{b}\,db = \infty \neq 1$$

Beyond simple models

Most realistic models are not simple experiments that happen to have a distribution named after dead people

→ to model a realistic experiment, we need to combine multiple such basic building blocks



Poisson

Gauss

Laplace

Mixture Models

Often, the data you observe may originate from a number of sources

Examples: a "signal" process and a background process with

$$p_{\mathrm{sig}}(x) = p(x|\mathrm{sig}) \qquad p_{\mathrm{bkg}}(x) = p(x|\mathrm{bkg})$$

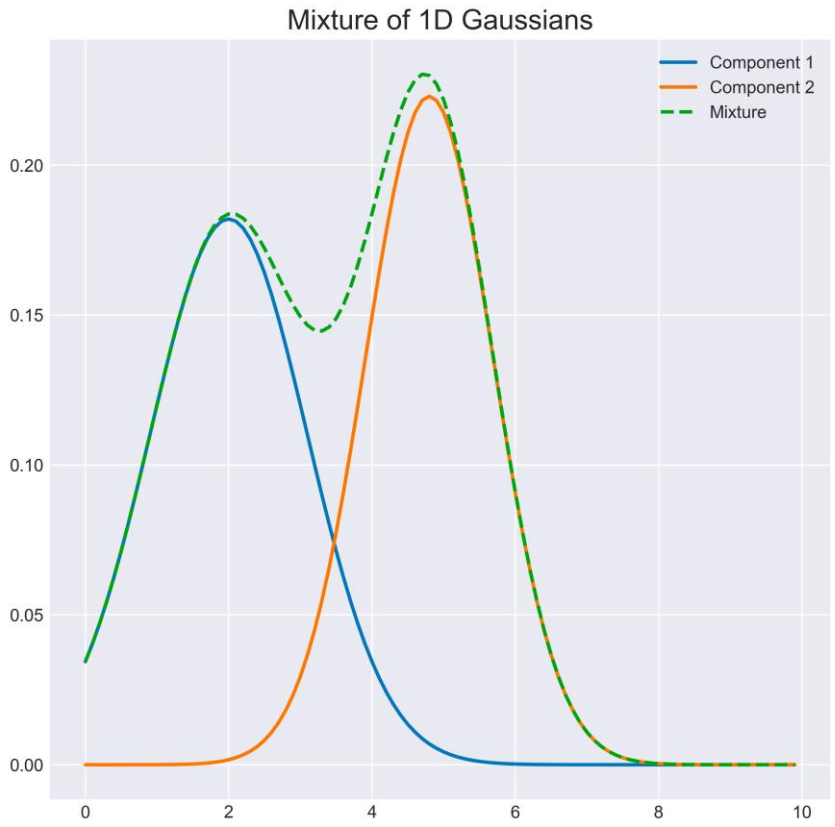
The data density can then be modelled as "mixture"

$$p(x) = p(x|\mathrm{sig})p(\mathrm{sig}) + p(x|\mathrm{bkg})p(\mathrm{bkg})$$

With $p(sig) + p(bkg) = 1$

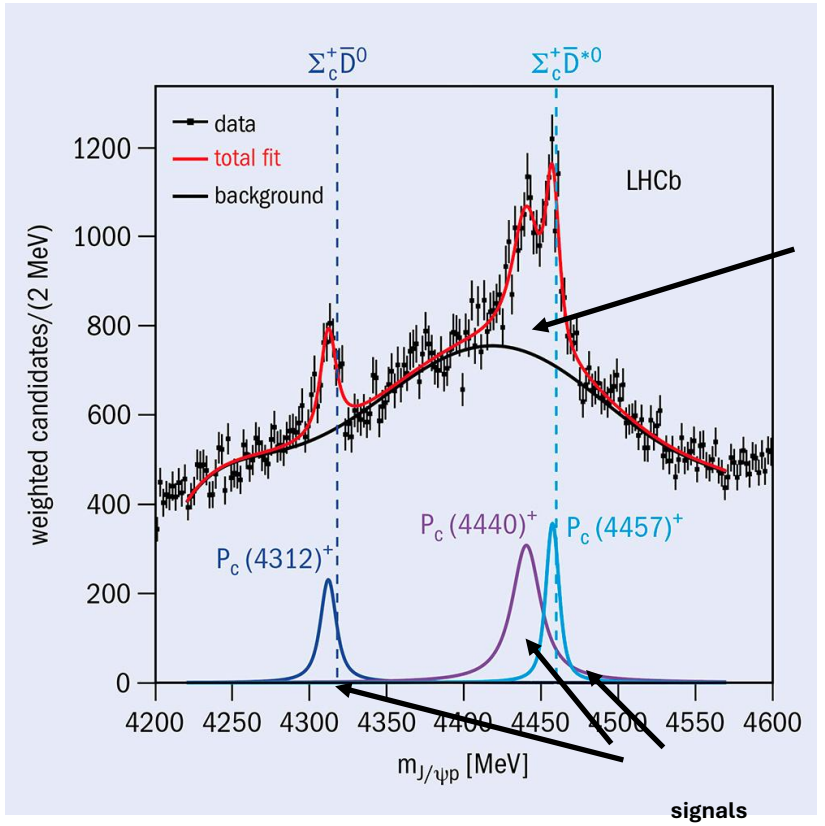
Mixture Models

Example:



Gaussian Mixture Model

[source]



Particle Resonances

[source]

Simultaneous Measurements

Sometimes your experiment consists of multiple *independent* sub-measurements of data x_1 and x_2

The **joint probability** is the product of each measurement's probability

$$p(x_1, x_2) = p(x_1)p(x_2)$$

... equivalently

$$\log p_{\text{global}} = \log p_{\text{E1}} + \log p_{\text{E2}}$$

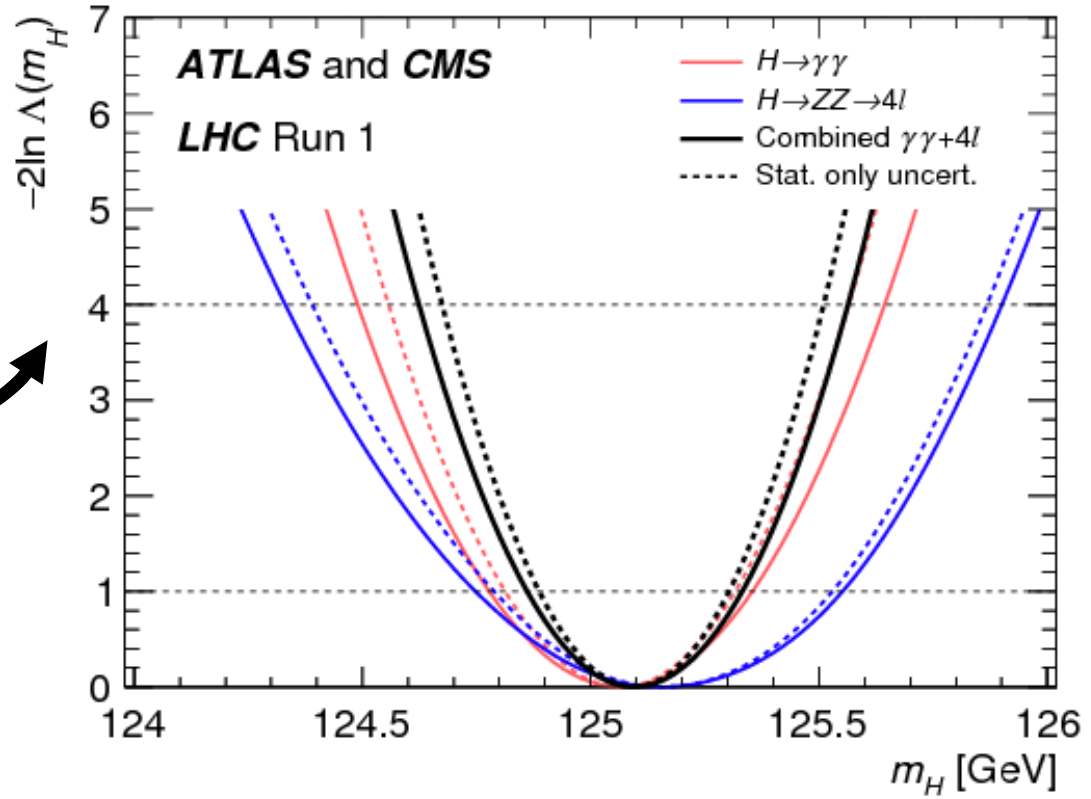
Simultaneous Measurement

Example:

combined Higgs mass measurement of two disjoint datasets

$$\Lambda(m_H) \approx p(\text{data}_{ZZ}|m_H)p(\text{data}_{\gamma\gamma}|m_H)$$

(negative) Log Likelihood

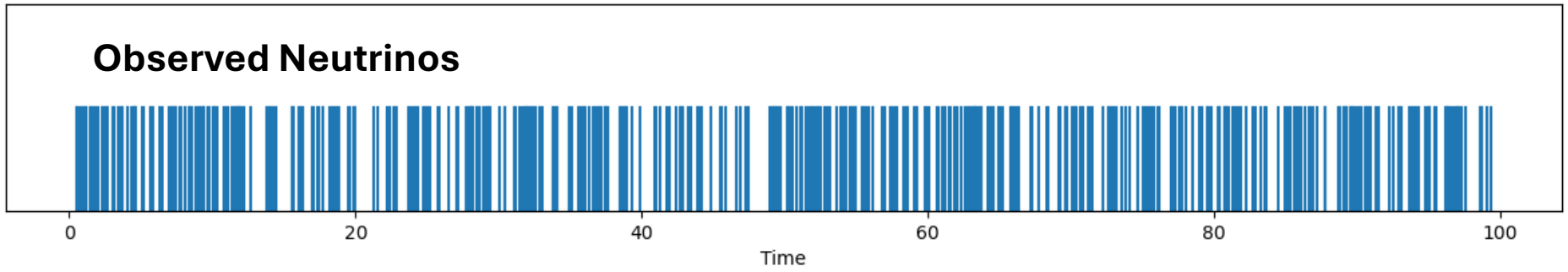


An example experiment:

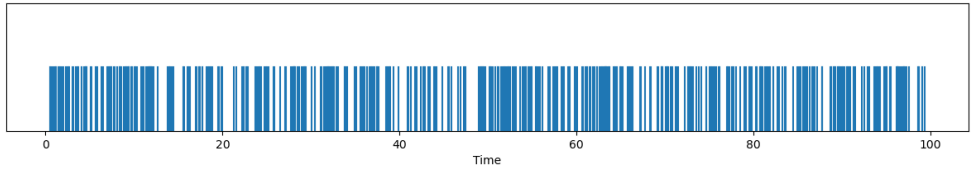
That we‘ll use a few times during the lecture

ν

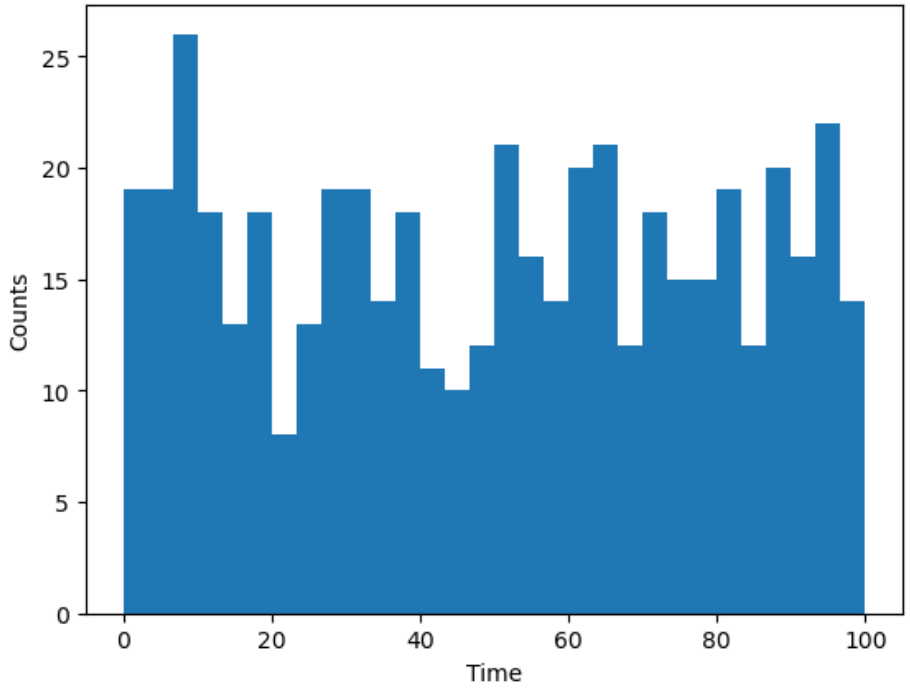
Neutrino Detector



Simple Analysis Procedure

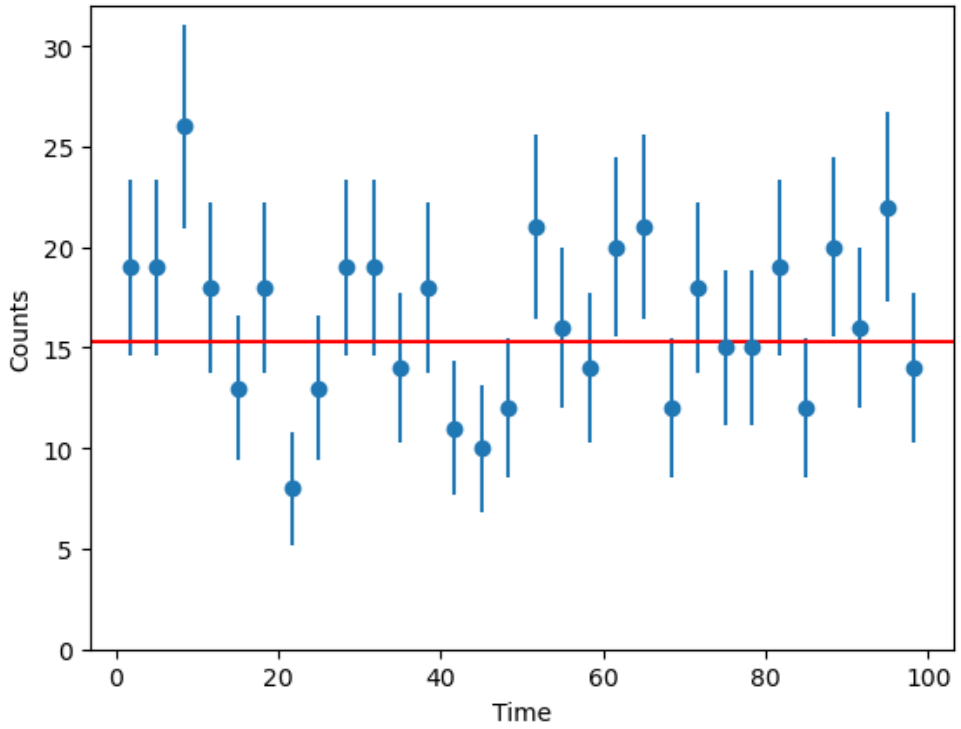


Histogram



Add Errors

Compare with Predicted



Simultaneous Measurements

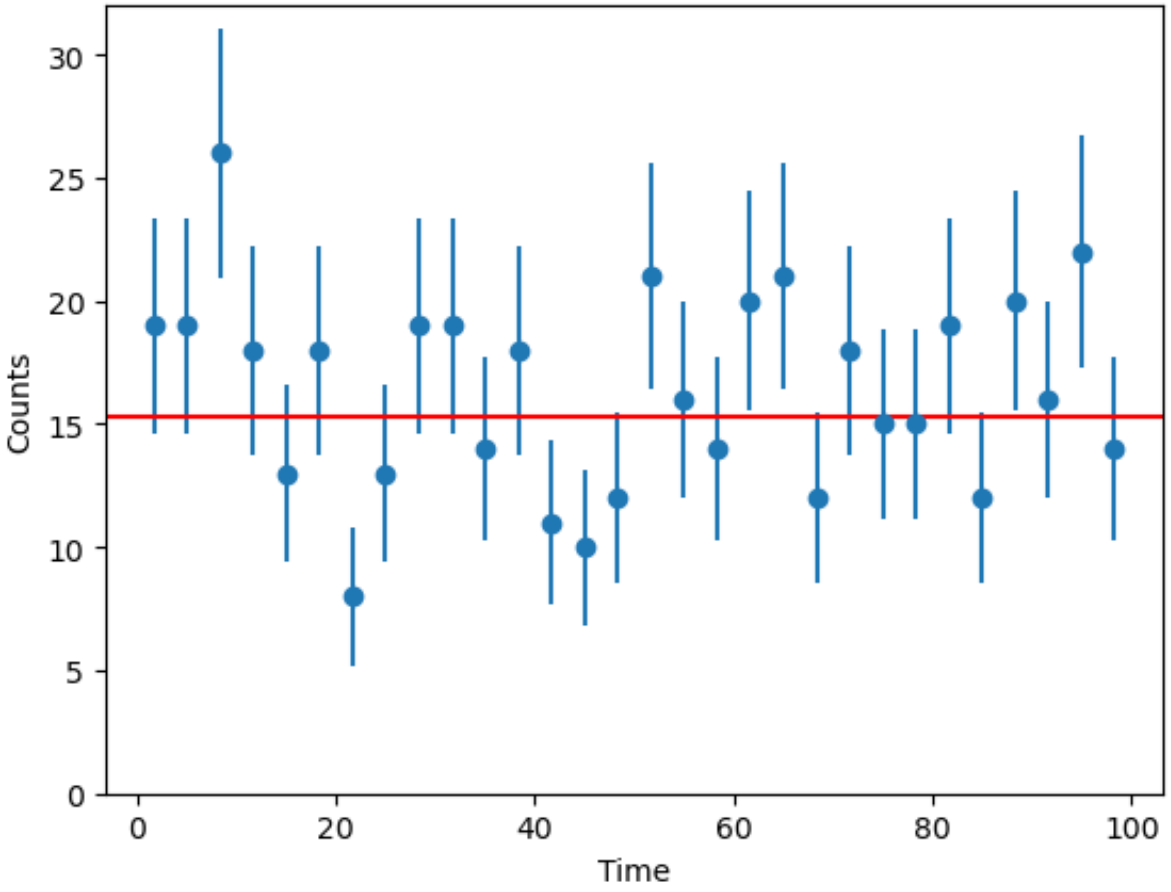
- Our example consists of 30 independent bins, each measuring the same Poisson rate
  - This is also referred to as *i.i.d. (= independent and identically distributed*) in stats. literature

→ So our likelihood is:

$$L(\lambda) = \prod_i p(x_i|\lambda) = \prod_i \frac{1}{x_i!} \lambda^{x_i} \exp(-\lambda)$$

Or more commonly used the LLH is proportional to:

$$LLH(\lambda) = \sum_i (x_i \ln(\lambda) - \lambda)$$



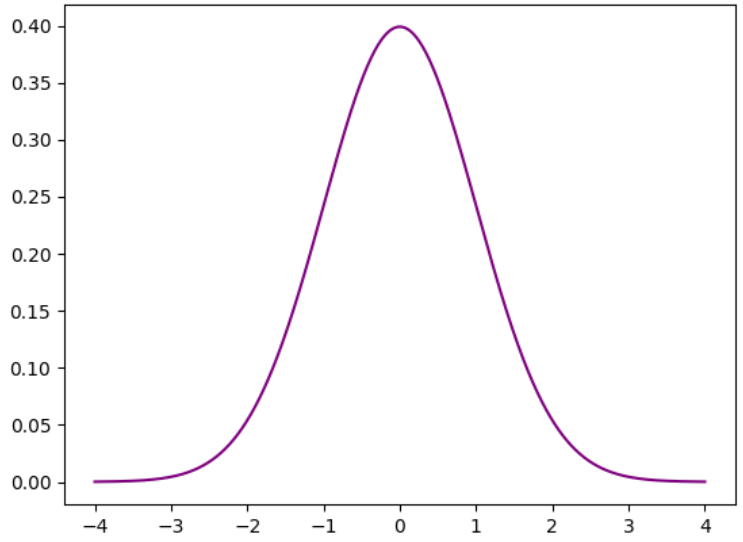
Point Estimators

Model vs. Observation

Abstract Space

i.e. a Model, often containing a number of parameters θ

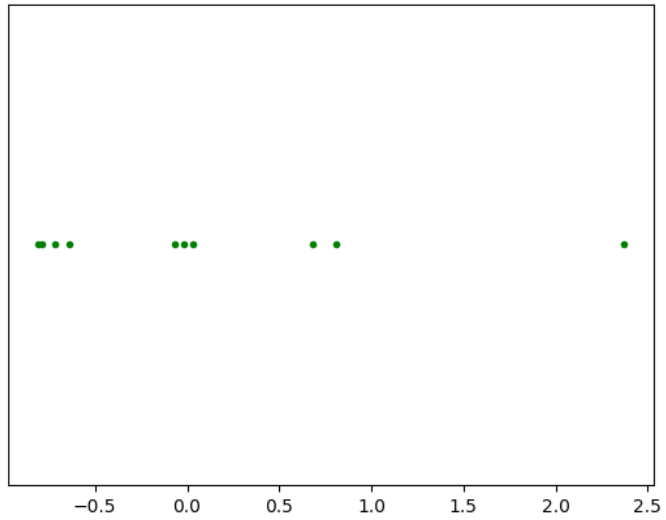
Example: Gaussian with parameters μ and σ



Observable Space

i.e. our "Data" x

Example: Values x = {1.2, -0.7, 0.3, ...}



From Model to Observation

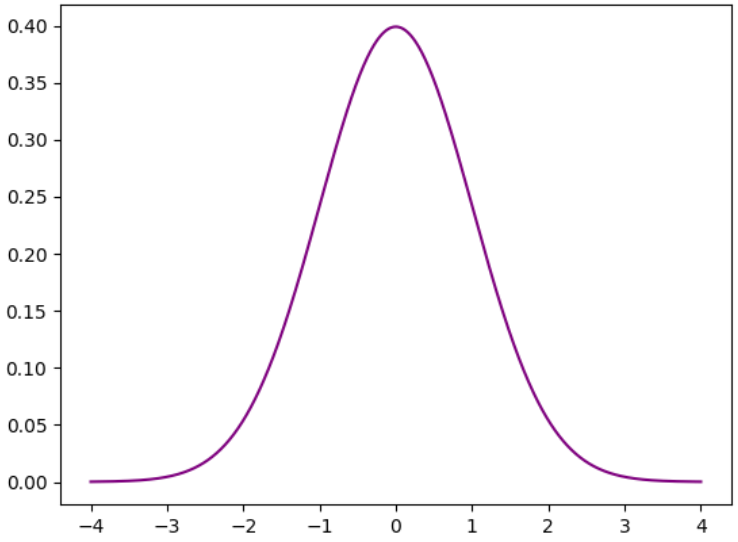
Abstract Space

The **probability** (density) function expresses the probability of observing the data given the model

Observable Space

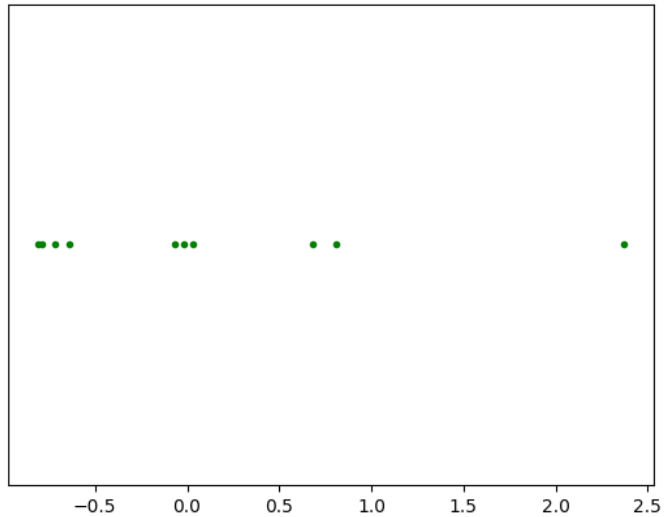
Choose model parameters

$$P(x|\mu = 0, \sigma = 1)$$



Gives Probability P(x)

x is a random variable



Joint probability of observing n data: $\prod_i p(xi)$

And back...?

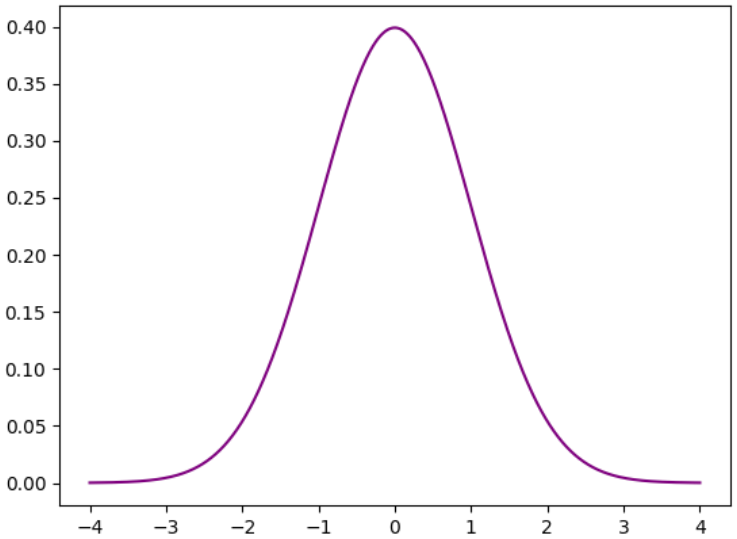
Abstract Space

Observable Space

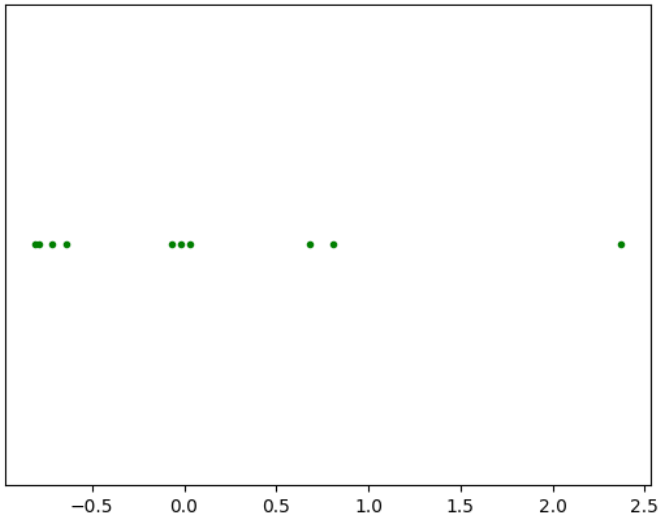
Given a set of observations x

Estimate model parameters

$$P(x = \{1.2, -0.7, 0.3, ...\} | \mu, \sigma)$$



Here P has taken the role of a **likelihood**!
i.e. the probability viewed as a function of its parameters



→ The likelihood allows us to make statements about the model given data

Estimators

Estimators are **functions of the data** (i.e. "statistics") $\hat{\theta}(x)$ that give an **estimate** of a parameter of the underlying model $p(x|\theta_0)$.

Since the data is random, the **estimate** $\hat{\theta}$ is random as well

- no guarantees that any *particular* estimate is close to θ_0
- but we can make statements w.r.t repeated experimentation i.e. long-run properties of estimators

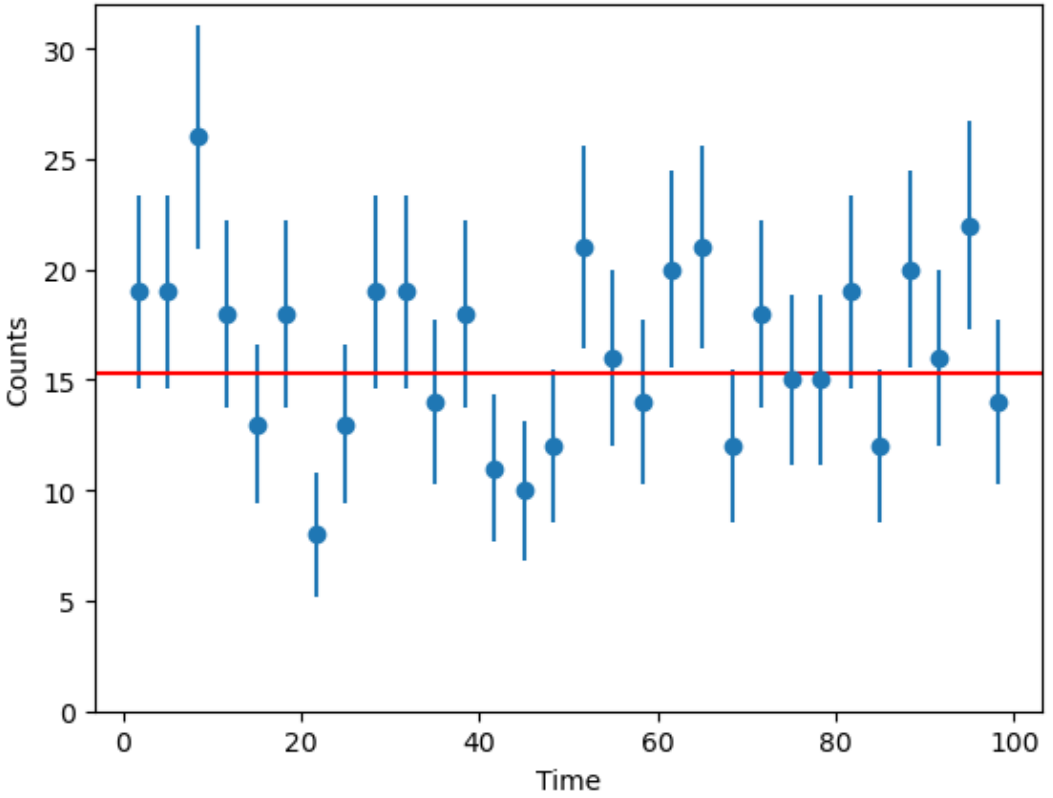
In our example

Task: Get an estimate $\hat{\lambda}$ from our data $\{x_i\}$

Idea: use the weighted mean of all datapoints (weight each by its errorbar $\frac{1}{\sqrt{x}}$)

→This results in the „harmonic mean“

$$\hat{\lambda} = \left(\frac{1}{n}\sum_i \frac{1}{x_i}\right)^{-1}$$

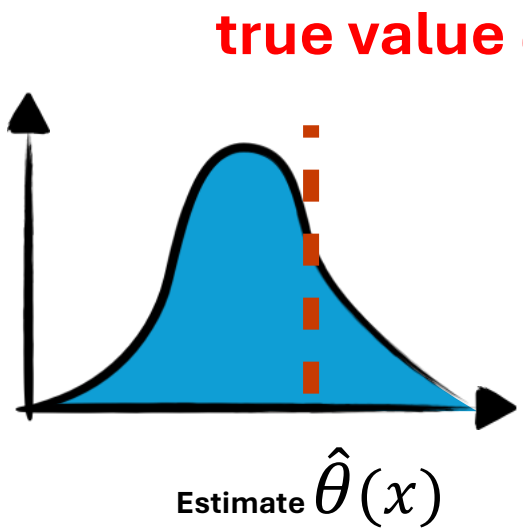


Which results in 4.599 (true rate is 5)

Is this a good or a bad estimate?

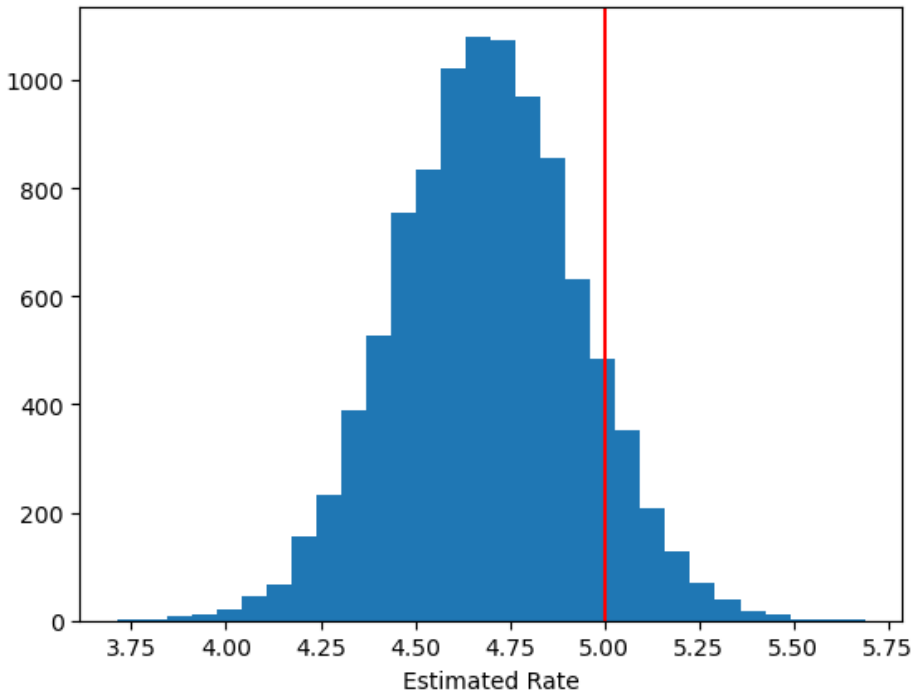
Estimate Distribution

For finite sample sizes we expect estimators to deviate from the true value. Under repeated experiments there's a distribution $p(\hat{\theta}\ (x)|\theta)$



In our example

- We can produce the estimator distribution from sampling possible outcomes with a known rate (here $\lambda = 5$)

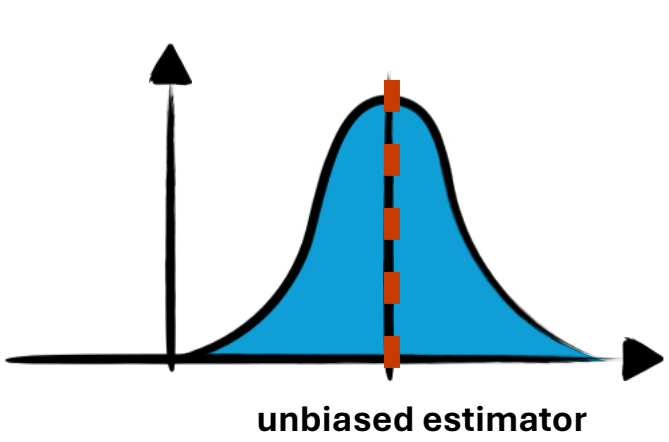


Estimator Bias

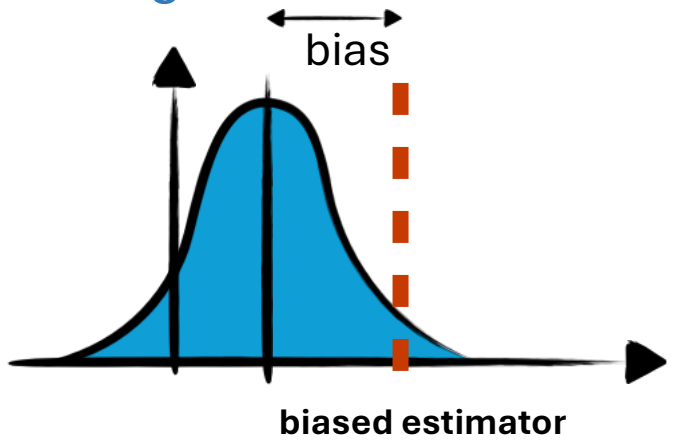
A key metric is the **bias of the estimator:**

- deviation of the expectation value of $\hat{\theta}\ (x)$ from the true value
- generally people prefer unbiased estimators

$$b = \mathbb{E}[\hat{\theta}\ (x)] - \theta_0$$



unbiased estimator

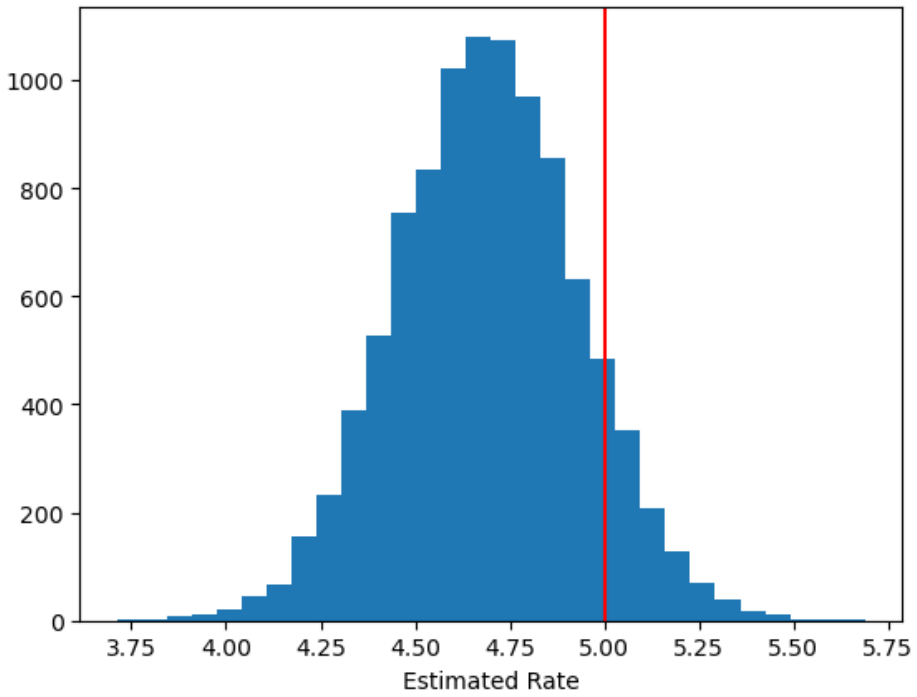


biased estimator

In our example

- We can produce the estimator distribution from sampling possible outcomes with a known rate (here $\lambda = 5$)
- We see the harmonic mean is a **biased** estimator!

(We‘ll fix it later…)



Example: Gaussian Mean

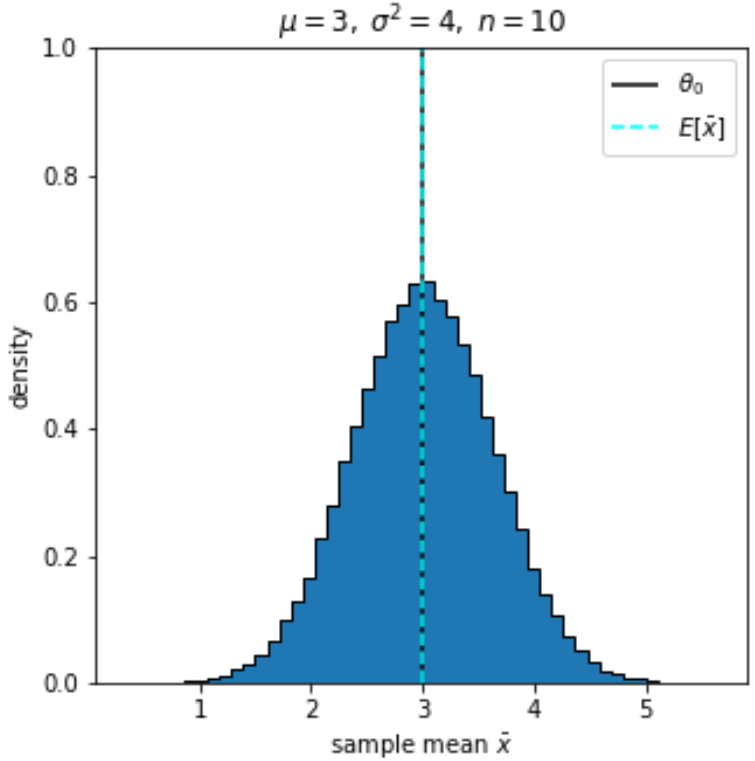
Consider a Gaussian Model with n i.i.d. samples $x_i \sim \mathcal{N}(x_i|\mu, \sigma^2)$

A simple estimator is the "sample mean"

$$f(x) = \overline{x} = \frac{1}{N}\sum_i x_i$$

Distribution of the estimator for

For $n = 10, \mu = 3, \sigma^2 = 4$



Note: $\mu = \mathbb{E}_x \overline{x}$!

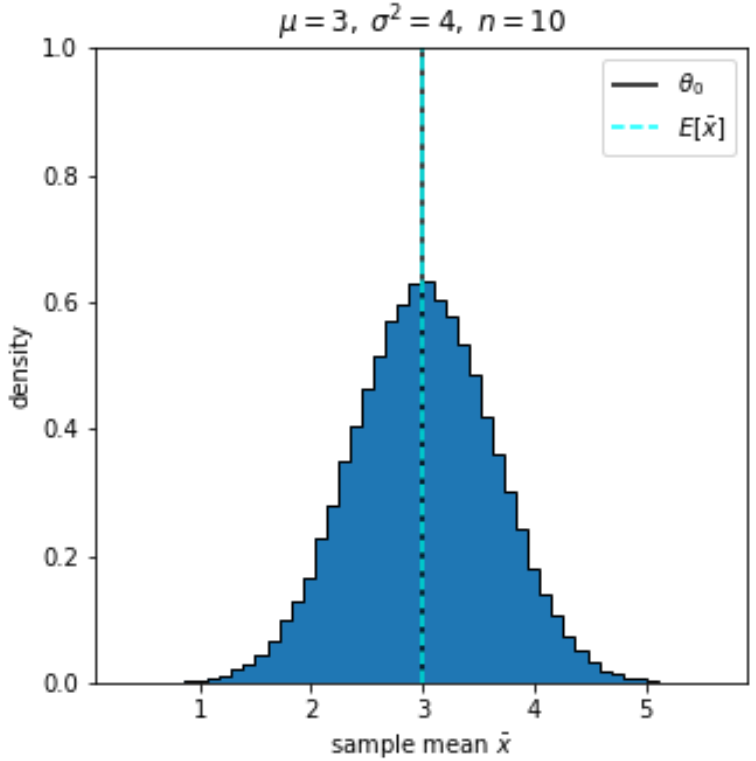
Example: Gaussian Mean

Consider a Gaussian Model with n i.i.d. samples $x_i \sim \mathcal{N}(x_i|\mu, \sigma^2)$

For $n = 10, \mu = 3, \sigma = 1$

$$f(x) = \overline{x} = \sum_i x_i$$

$\overline{x}$ is a and unbiased estimator of μ

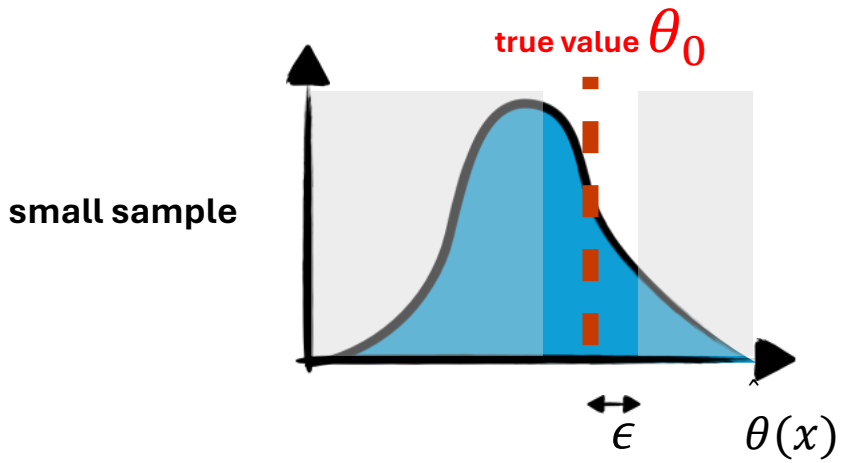


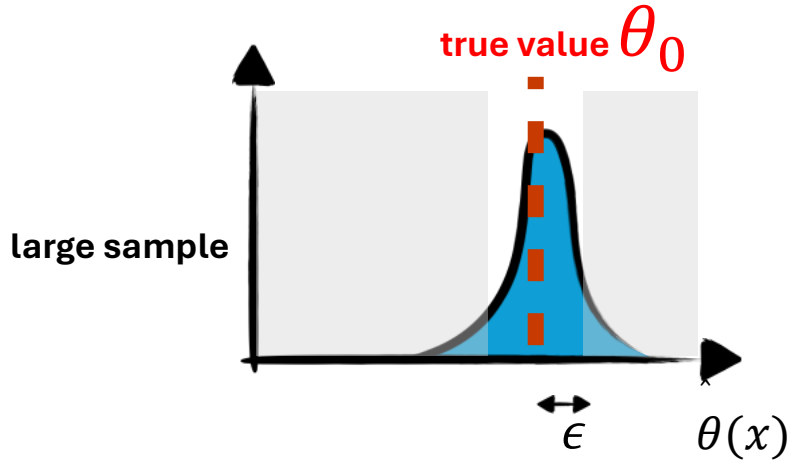
Consistency

A desirable property is that estimators are **"consistent"**

- more data provides you a better estimate on average
- estimation value probability accumulates close to the true value

$$\lim_{n\to\infty} p(|\hat{\theta}\ (x) - \theta_0| > \epsilon) = 0; \forall\epsilon$$





Example: Gaussian Mean

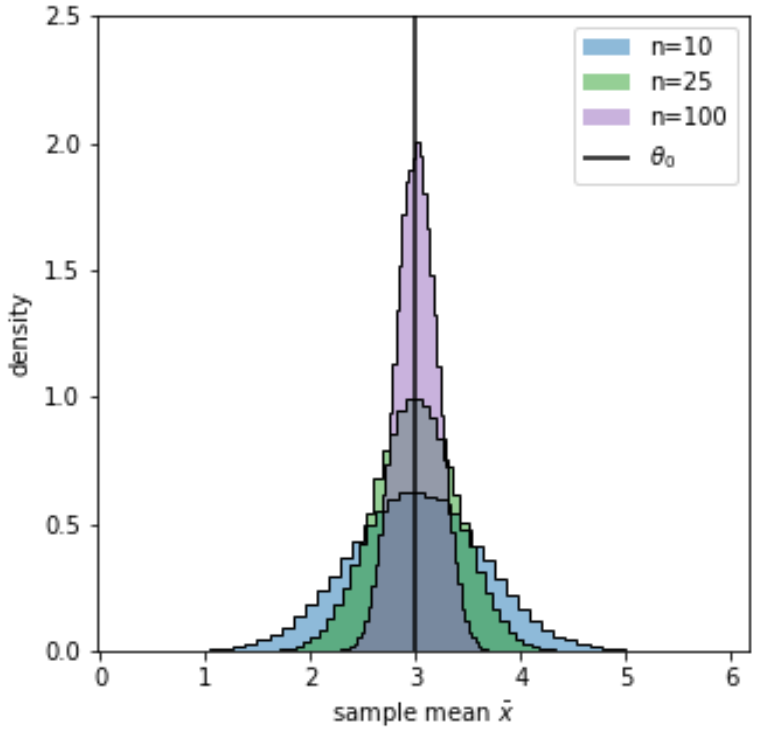
Consistency at play in our Gaussian Example

- sharpening of the distribution around the true value μ

For $n = 10, \mu = 3, \sigma = 1$

$$f(x) = \overline{x} = \sum_i x_i$$

$\overline{x}$ is a consistent estimator of μ



Example: Gaussian Variance

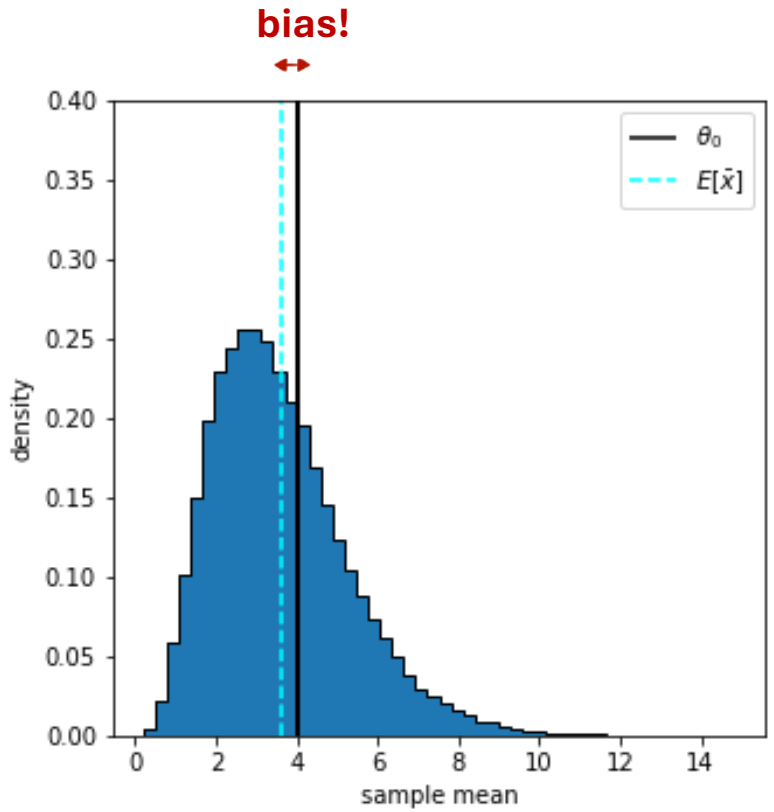
Same Gaussian Setup: but with the **sample variance** as a statistic

For $n = 10, \mu = 3, \sigma = 1$

$$f(x) = s^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2$$

Note: $\mathbb{E}_x[s^2] \neq \sigma^2$!

s^2 is a biased estimator of σ^2



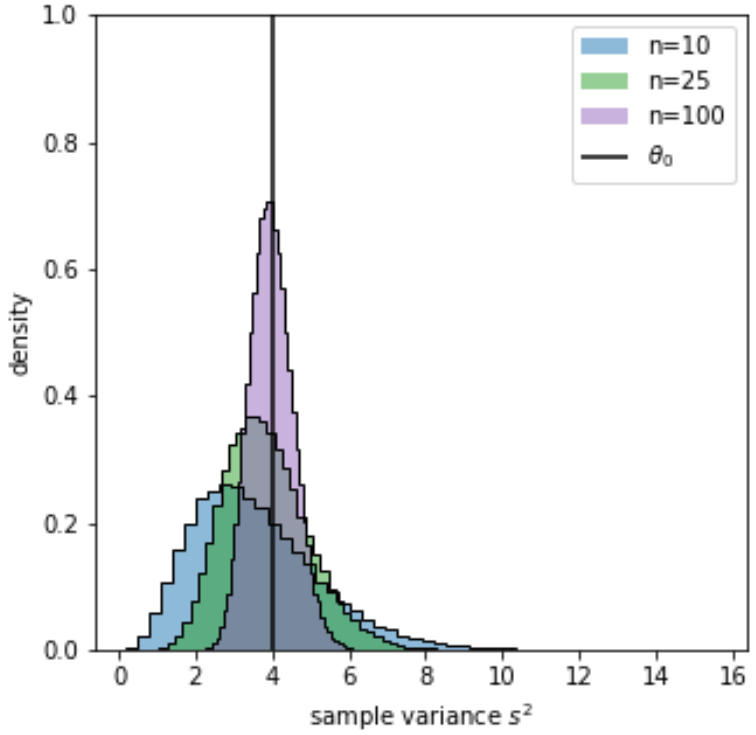
Example: Gaussian Variance

The sample variance is still consistent though:

For $n = 10, \mu = 3, \sigma = 1$

$$f(x) = s^2 = \frac{1}{N}\sum_i (x_i - \overline{x})^2$$

s^2 is a consistent but biased estimator of σ^2

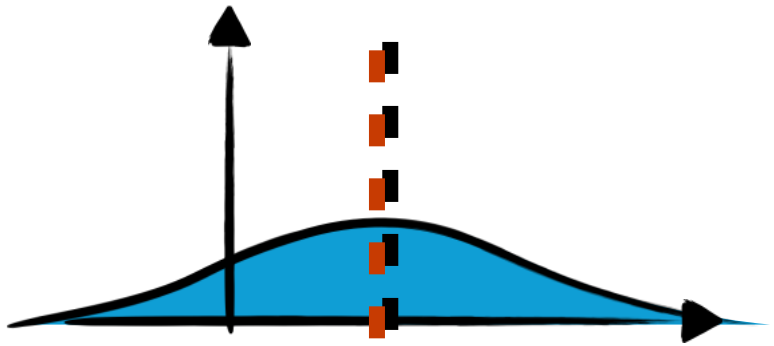


Estimator Variance

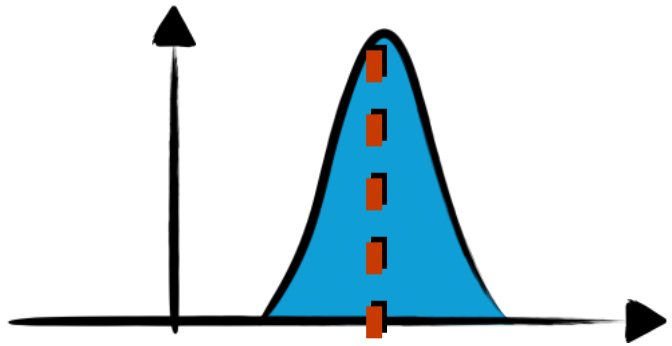
A second metric is the **variance of the estimator:**

- spread of the estimator around its expectation value
- generally lower-variance is preferred over high variance

$$\sigma_\theta = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$



high-variance estimator



low-variance estimator

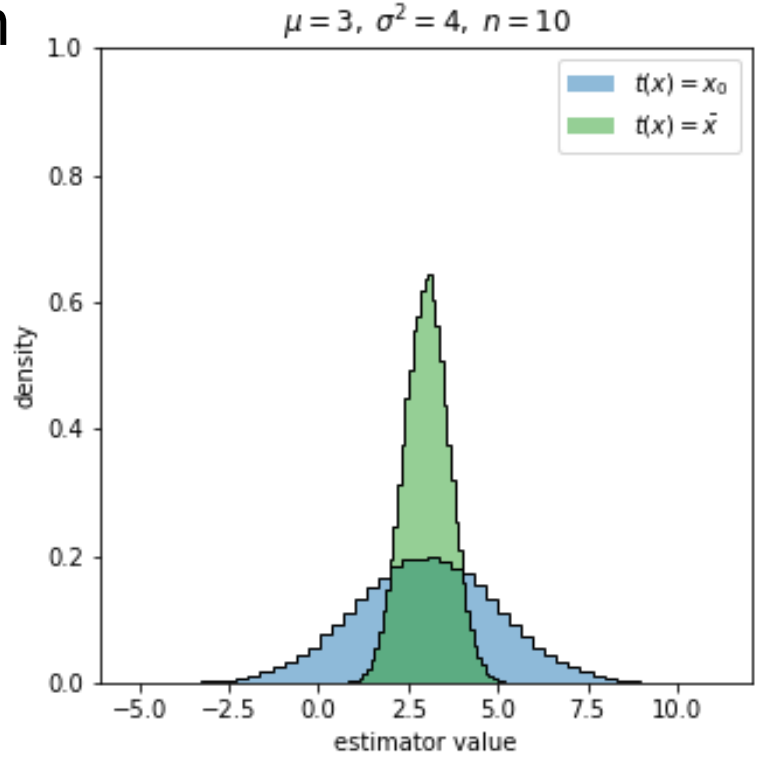
Variance Example: Gaussian Mean

Same Gaussian Model with n i.i.d. samples $x_i \sim \mathcal{N}(x_i|\mu, \sigma^2)$

In a sample $x = (x_1, \ldots, x_n)$ each x_i is an unbiased, consistent estimator of μ.
e.g. $f(x) = x_1$

Why even compute the sample mean $\overline{x}$?

It has much lower variance!



Exercise

Variance of sample mean

Exercise

- Take n samples from a normal distribution, here

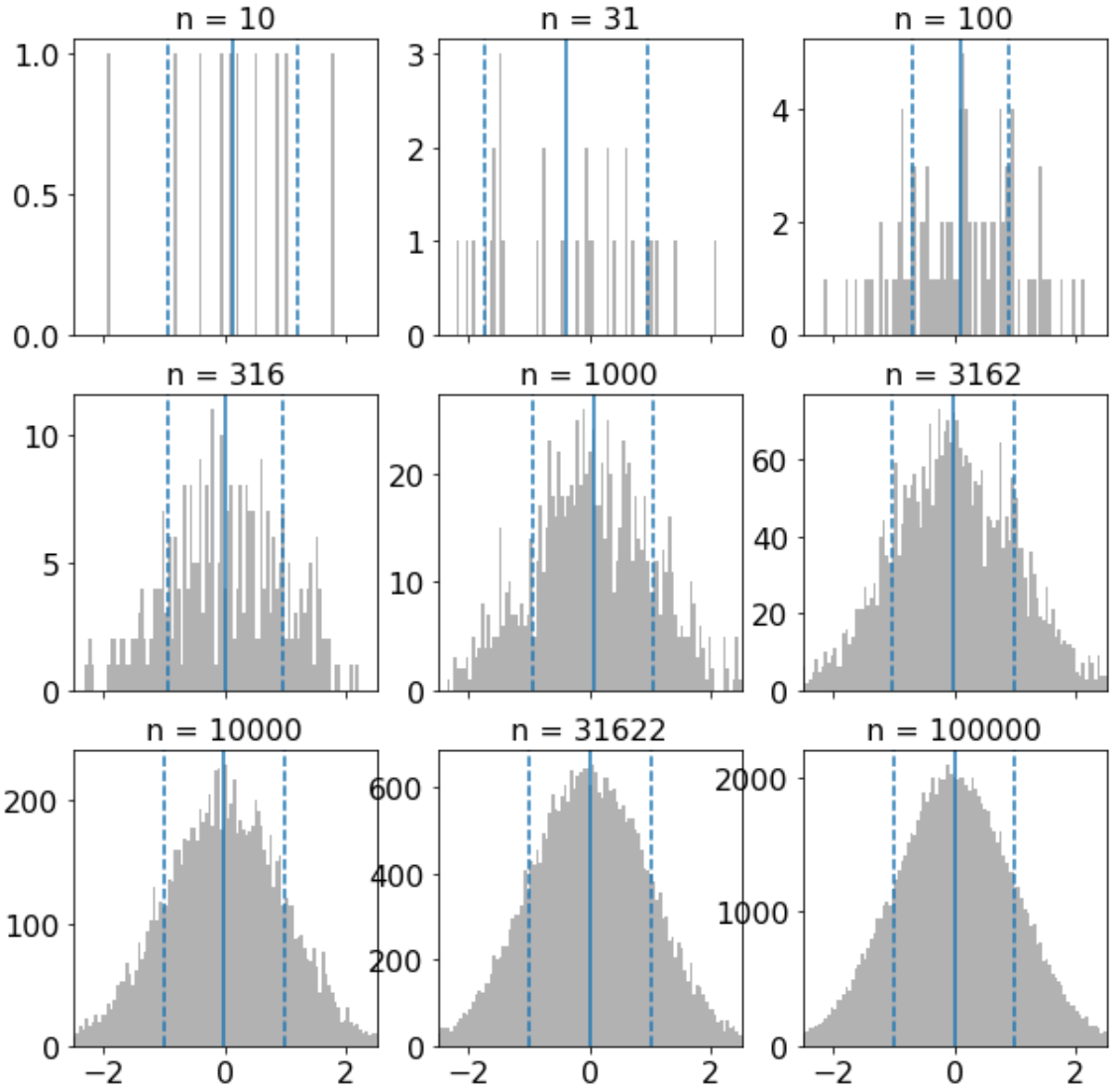
$$\mu = 0$$
$$\sigma = 1$$

(Sample mean +/- sample variance shown in blue)

- What is the variance of the sample mean?

$$\bar{x} = \frac{1}{n}\sum_{i} x_i$$

→ $Var(\bar{x})$ = ?



Variance of sample mean

$$Var(\bar{x}) = Var\left(\frac{1}{n}\sum_i x_i\right)$$

$$= \frac{1}{n^2} Var\left(\sum_i x_i\right)$$

$$= \frac{1}{n^2}\sum_i Var(x_i) = \frac{1}{n^2} n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

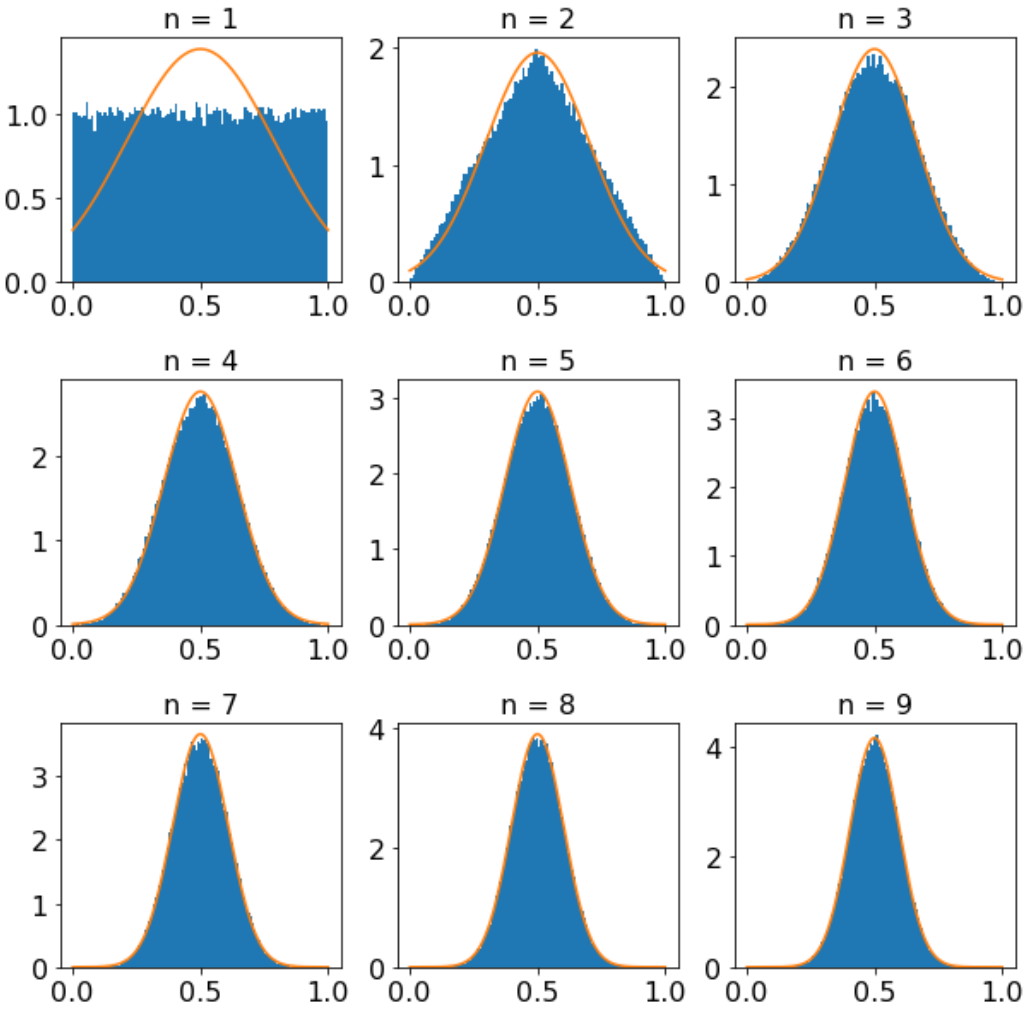
Central Limit Theorem

This exercise connects to the CLT

→ Suppose we make repeated, independent draws x_i of any probability distribution p

- Average (sample mean) $\bar{x} = \frac{1}{n}\sum x_i$
- If p has finite mean and variance, then $\bar{x}$ is asymptotically distributed according to $\mathcal{N}(\mu, \sigma/\sqrt{n})$!

CLT using uniform distribution



Cauchy

What about a Cauchy distribution?

$$p(x|x_0,\gamma) = \frac{1}{\pi}\frac{\gamma}{(x-x_0)^2+\gamma^2}$$

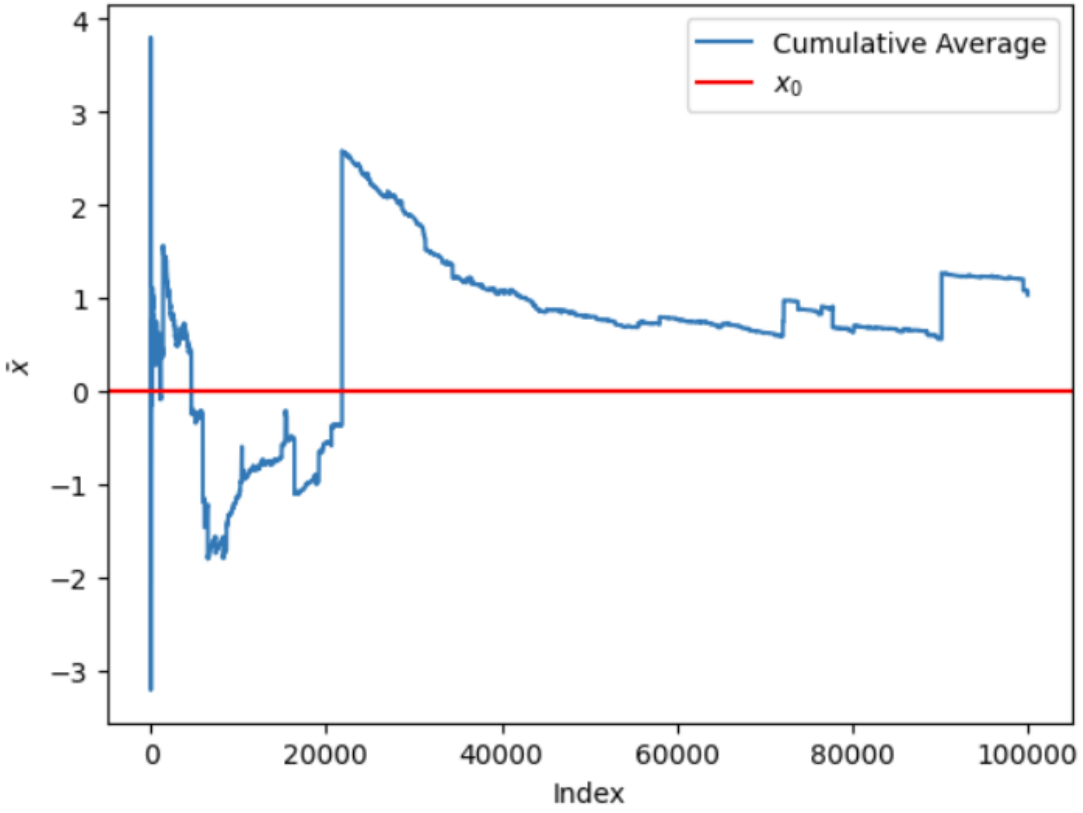
- The distribution does not have finite moments!!
  - $\bar{x}$ is **not** a good estimator for x_0! It is not consistent

```
x = stats.cauchy().rvs(100_000)
```

```
xbar = np.cumsum(x) / np.arange(1, len(x)+1)
```

```
fig, ax = plt.subplots()
ax.plot(xbar, label="Cumulative Average")
ax.set_ylabel(r"$\bar{x}$")
ax.set_xlabel("Index")
ax.axhline(0, c='r', label=r"$x_0$")
ax.legend()
```

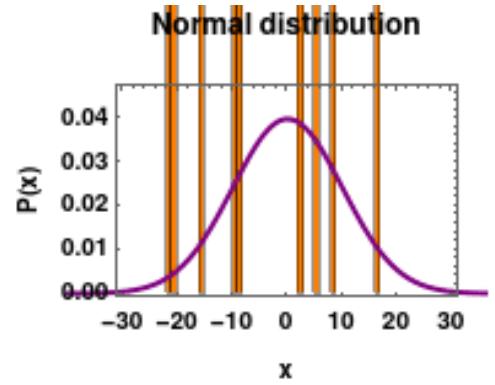
```
<matplotlib.legend.Legend at 0x7f8d226ba5e0>
```

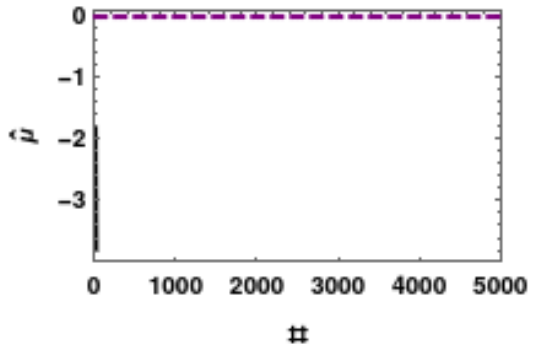


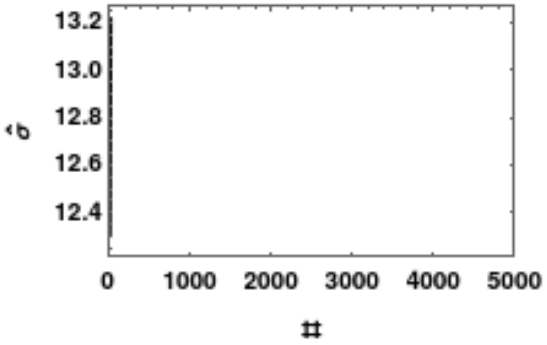
Normal vs. Cauchy

Sample mean | **Sample standard deviation**

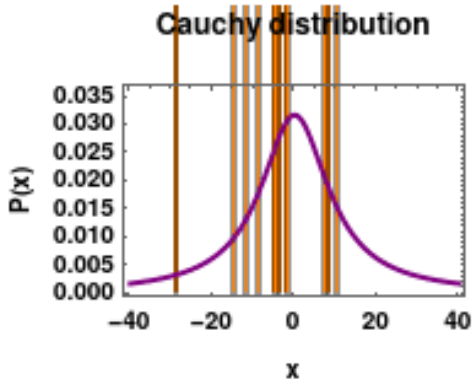
Gaussian

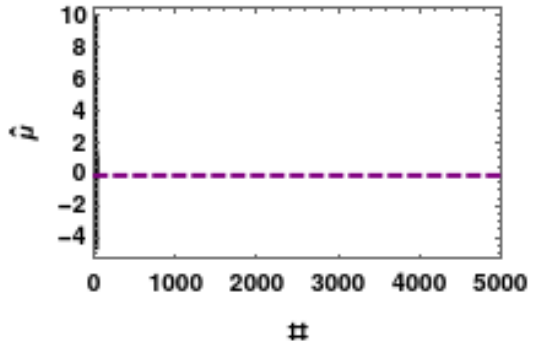


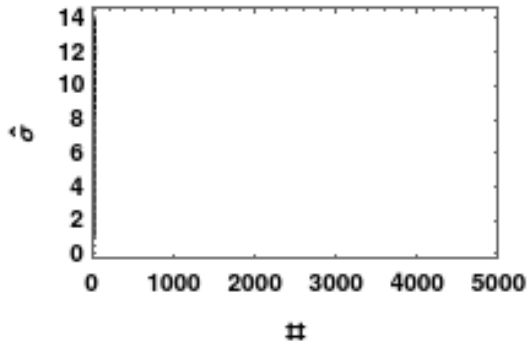




Cauchy







Bias - Variance Decomposition

Both bias and variance contribute to the overall expected deviation, i.e. mean squared error (MSE), from the true value θ_0

$$\begin{aligned}\sigma_{mse} &= \mathbb{E}_x\left[\left(\hat{\theta} - \theta_0\right)^2\right] \\ &= \mathbb{E}_x\left[\left(\hat{\theta} - \bar{\hat{\theta}} + \bar{\hat{\theta}} - \theta_0\right)^2\right] \\ &= var\,\hat{\theta} + bias\hat{\theta}^2\end{aligned}$$

Bias - Variance Decomposition

A common goal is to find an estimator with the lowest mean squared error

$$\sigma_{\text{mse}} = \text{var}\hat{\theta} + \text{bias}_{\hat{\theta}}^2$$

If we restrict our search to unbiased estimators this means

Look for the **minimum-variance** estimator

Minimum Variance

The variance of an unbiased estimator cannot become arbitrarily small.

- It's bounded from below by the amount of "Information" the data can provide about the model parameters
  - lots of information → small variance and vice versa
- Known as the **Cramér-Rao Bound**

$$\mathrm{var}\hat{\theta} \geq \frac{1}{I(\theta)}$$

Fisher Information

The **Information** is measured by the **average square gradient** of the log likelihood function

- average sensitivity to parameter values across all possible x
- also called the **"Fisher Information"** of the model at $I(\theta)$

$$\mathrm{var}\hat{\theta} \geq \frac{1}{\mathbb{E}_x[(\partial_\theta \log p(x|\theta))^2]} = \frac{1}{I(\theta)}$$

Fisher Information

Example: For a Gaussian, $x \sim p(x \mid \mu, \sigma^2)$

$$I(\mu) = \mathbb{E}_x[(\partial\mu \log p)^2] = -\mathbb{E}_x[\partial^2\mu \log p]$$

$$= -\mathbb{E}_x\left[\partial^2\mu\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}\right)\right] = \mathbb{E}_x\left[\partial^2\mu \frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$= \mathbb{E}_x\left[-\partial\mu \frac{(x-\mu)}{\sigma^2}\right] = \mathbb{E}_x\left[\frac{1}{\sigma^2}\right]$$

$$= \frac{1}{\sigma^2}$$

The smaller the variance, the more information you have on the location

Good Estimators

Finding Estimators

We have empirically seen good estimation properties from sample statistics like the sample mean $\overline{x}$ and the sample variance s^2

But we pulled those out of a hat and only for the Gaussian case.

Where do they come from?

Need a robust and generalizable method to produce estimators.

What concept we introduced could be useful?

Maximum Likelihood

An intuitive way to find a good point is to find the parameter $\hat{\theta}\ (x)$ that **maximizes the probability to observe the data we got:**

$$\hat{\theta}_{MLE}(x) = \text{argmax}_{\theta} p(x|\theta)$$

$\hat{\theta}_{MLE}(x)$ **is called the Maximum-Likelihood Estimator of** θ

Example

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- N samples from a Normal distribution:
  - Log-likelihood (omitting constants):

$$\log\left(\mathcal{L}(\mu,\sigma|(x_1,\ldots,x_n))\right) = \sum_i -\log(\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

  - Maximum?

$$\frac{\partial log\mathcal{L}}{\partial\hat{\mu}} = 0$$

$$\sum_{\mathrm{i}} \frac{(x_i-\hat{\mu})}{\sigma^2} = 0$$

$$\sum_{\mathrm{i}} (x_i-\hat{\mu}) = \sum_{\mathrm{i}} x_i - \mathrm{n}\hat{\mu} = 0$$

$$\rightarrow\ \hat{\mu} = \frac{1}{n}\sum x_i$$

MLE for Gaussian Model

Now we see the origin of the mean & variance estimators we used

They are the MLE estimators of the model parameters

$$p(x|\mu,\sigma^2) = \prod_i \mathcal{N}(x_i|\mu,\sigma^2)$$

$$\hat{\mu}_{\text{MLE}} = \overline{x} = \frac{1}{n}\sum_i x_i \qquad \hat{\sigma}^2_{\text{MLE}} = s^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2$$

Numerical Optimization

In general, a closed-form solution for $\hat{\theta}_{\text{MLE}}$ is rarely available, but it's always possible to fall back on **numerical optimization**

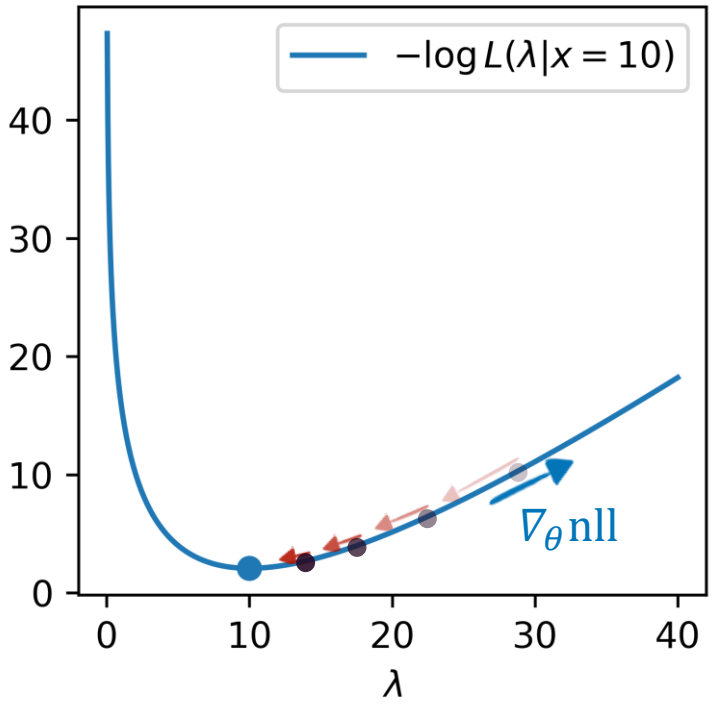
One option via gradient descent

$\hat{\theta} = \theta_{\text{init}}$

while not converged:

$g = \nabla_\theta f(\theta)$

$\hat{\theta} \leftarrow \hat{\theta} - \lambda g$

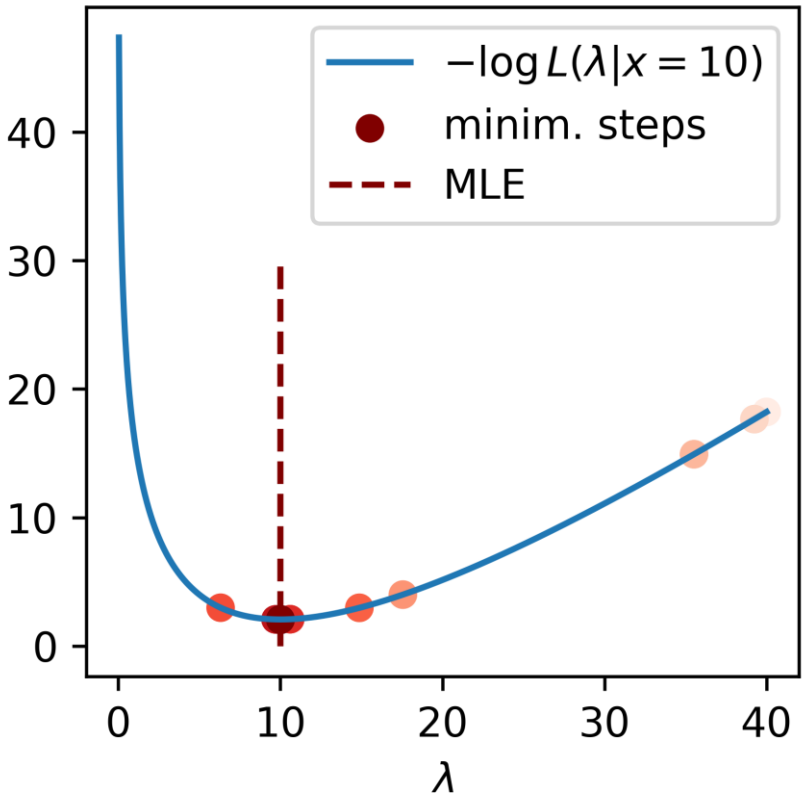


Python Example

```
[69]  ✓ 0.5s

...       fun: array([2.07856165])
          jac: array([2.36034[illegible]3e-05])
      message: 'Optimization terminated successfully'
         nfev: 18
          nit: 9
         njev: 9
       status: 0
      success: True
            x: array([10.00023639])
```

MLE



Properties of MLE

Asymptotic Consistency & Normality

The MLE estimator is not only intuitive but can be shown to have a few nice properties.

- **It's consistent:** probability accumulates near the true value
- The sampling distribution of MLE approaches a normal distribution asymptotically, i.e. for $n \to \infty$

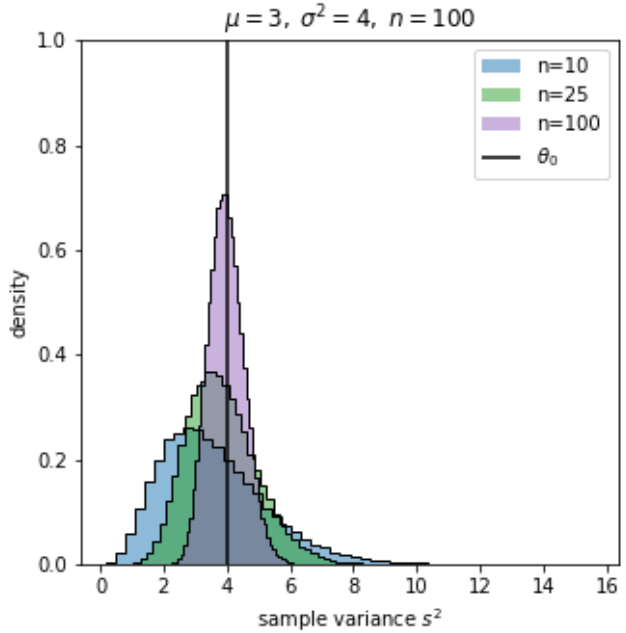
$$\sqrt{n}(\hat{\theta} - \theta_0) \to^{d} \mathcal{N}(0, \text{var}\hat{\theta})$$

= “converges in distribution”

Asymptotic Consistency & Normality

We've seen this already for the Gaussian sample variance

- while the finite-sample $\hat{\theta}$ distributions may not be Gaussian
  it will progressively be "normalized" (again, CLT)



Asymptotically Unbiased

Relatedly: MLE estimators are **asymptotically unbiased**

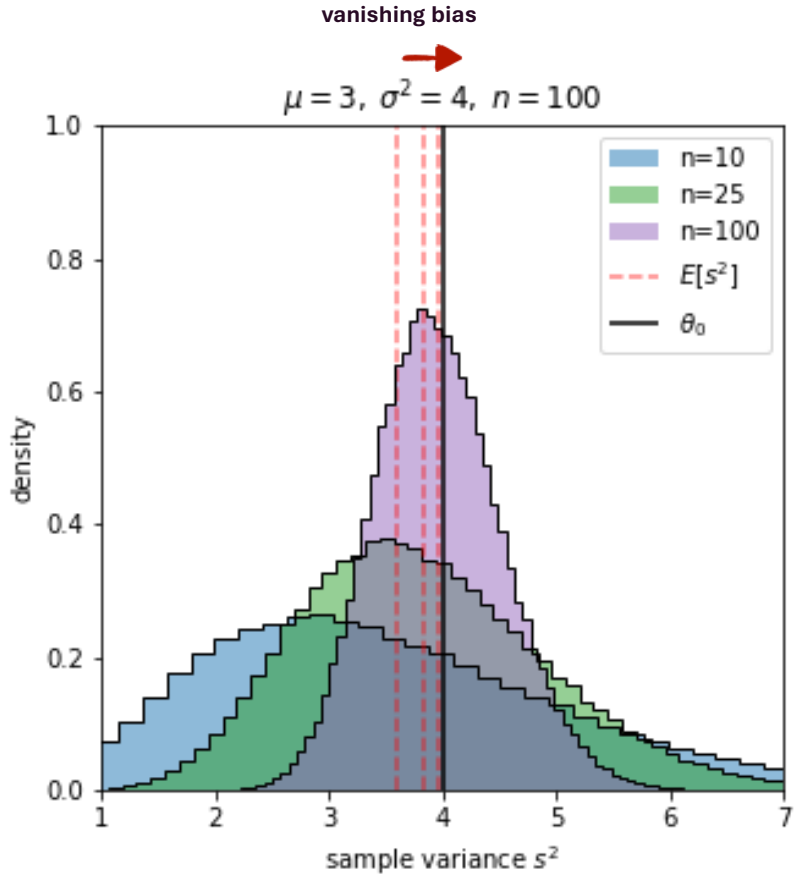
- Note: In general finite-sample MLE are biased

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d \mathcal{N}(0, \mathrm{var}\hat{\theta})$$

vanishing bias

Asymptotically Unbiased

Sample variance is a **biased** MLE estimator, but $\mathbb{E}[s^2]$ moves towards the true value for large samples and the **bias vanishes = asymptotically unbiased**



Asymptotically Efficient

MLE estimators asymptotically **saturate the Cramér-Rao bound**:

- i.e. achieve the minimum possible variance of all unbiased estimators

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d \mathcal{N}(0, \mathrm{var}\hat{\theta})$$

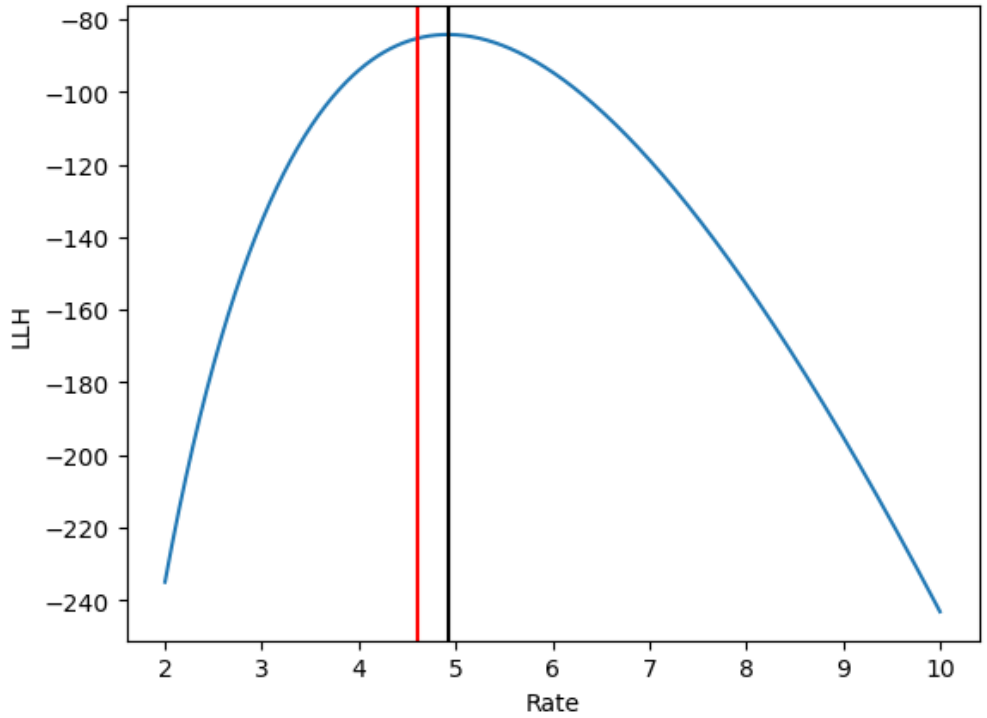
$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d \mathcal{N}(0, I^{-1}(\theta))$$

Inverse Fisher Information

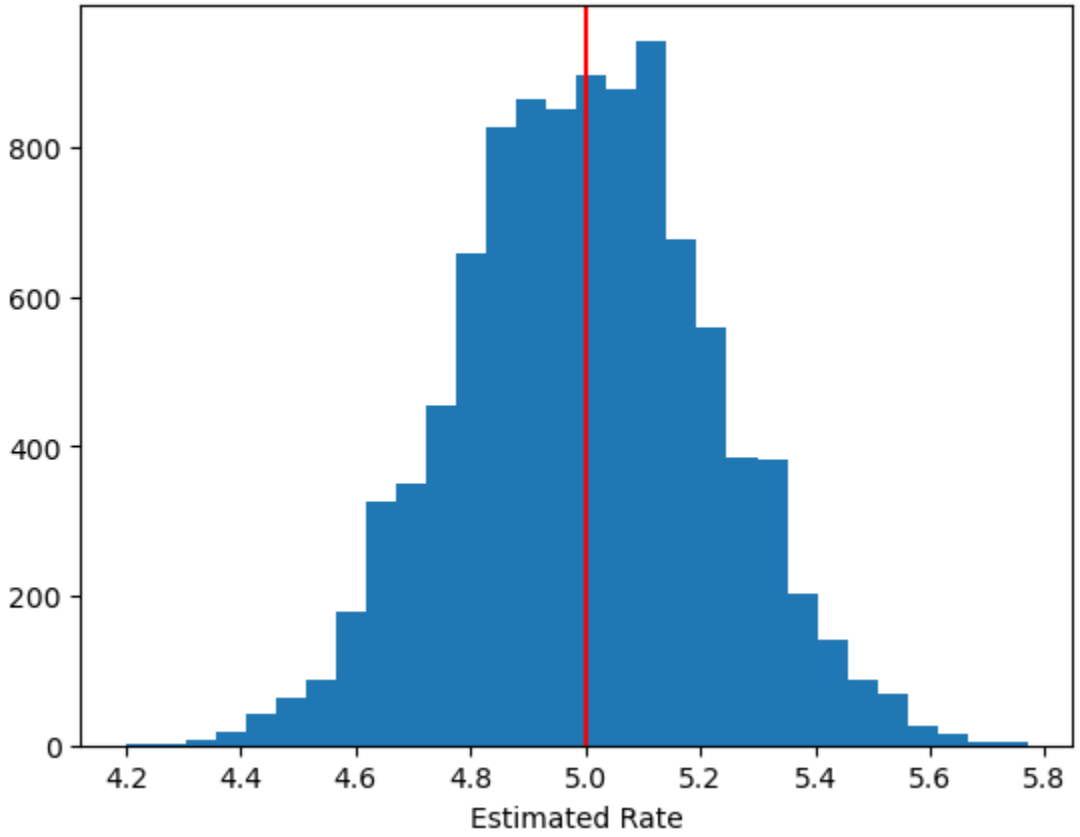
Fixing our Problem: MLE for our Example

- With MLE we get an estimated rate of 4.92

It is much closer to the truth!



And it is not biased!



Recap

- Point Estimation is about finding a single best parameter point to explain the observed data.

- We introduced a few key concepts of estimators in general:
  - Consistency
  - Bias
  - Variance
  - Cramér-Rao bound
  - etc.

- Maximum-Likelihood is the most popular estimation method
  - it has a number of desirable, asymptotic properties (consistency, min. variance,…)

Exercise

Maximum Likelihood Estimation

Find the MLE for the Poisson Likelihood,
and compute the MLE for an observed count of x=7

- Likelihood: $L_x(\lambda) = \frac{\lambda^{\mathrm{x}} \mathrm{e}^{-\lambda}}{\mathrm{x!}}$
- Log-likelihood: $LLH_x(\lambda) = \mathrm{x ln} \lambda - \lambda - \mathrm{ln x!}$
- Find Maximum: $\frac{\partial LLH_x(\lambda)}{\partial \lambda} = \frac{\mathrm{x}}{\lambda} - 1 = 0$
- $\rightarrow \hat{\lambda} = \mathrm{x}$

Statistics for Physicists

DAY 2

Hypotesis Testing

What is Hypothesis Testing?

Hypothesis testing is a framework for assessing whether observed data are consistent with a specified statistical model or assumption – a “hypothesis”

We quantify this with the so-called “**p-value**”

→ A p-value is the probability, under the assumption that the hypothesis being tested is true, of obtaining a result at least as extreme as the one observed in the data.

Example Hypotheses

- A hypothesis that we would like to test could be: „There is no anthropogenic effect on global temperature.”

  Consulting our data, we get a very low p-value, so we reject the hypothesis

- Or, testing a new drug for lowering blood pressure, clinical trial are needed to reject the hypothesis “The drug has no effect.“

  We get a moderate p-value of 0.25 → we cannot reject, because in 25% of such trials, a drug with no effect would yield at least as extreme data

Simple Example:

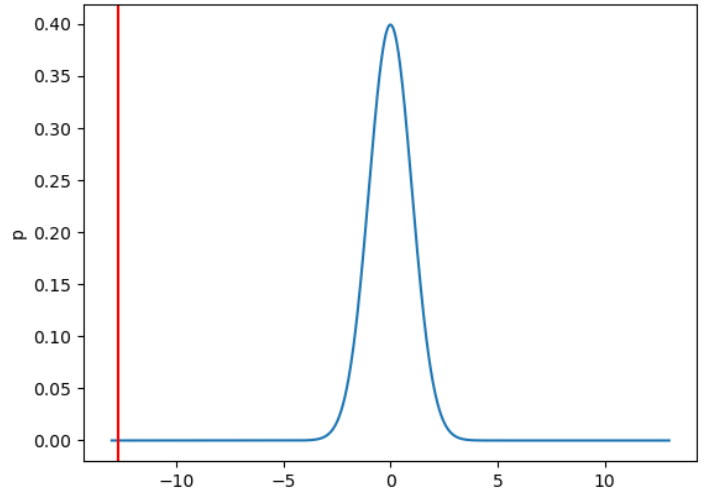
You observe a value of -12.7

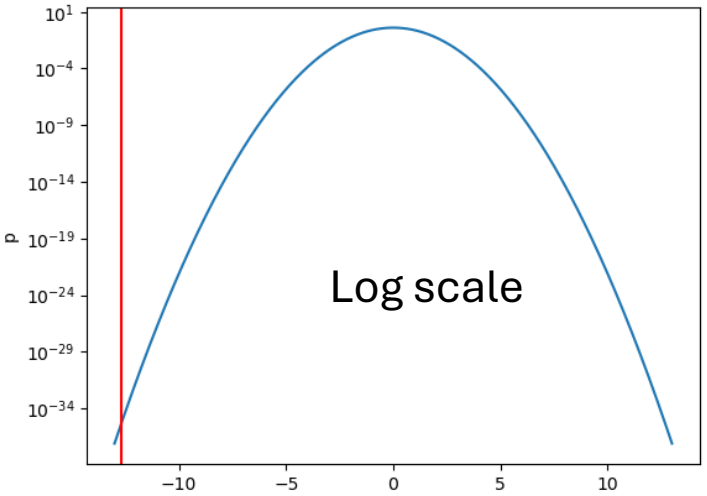
Hypothesis: the data comes from the unit normal distribution

→ Can we reject the hypothesis?
(we call this also the “null hypothesis”)

The probability to get a value at least as extreme, is simply the CDF at -12.7 = **2.96×10^{-37}**

→ at very low p-value, we may want to reject the hypothesis





P-value Properties

- P-values under the „null“ hypothesis are uniformly distributed
  - This is by construction!
  - The statement of „at least as extreme“ uses the cumulative distribution
  - The cumulative transforms the x to $y = F(x)$ with support on (0,1) and the distribution of y is $U(0,1)$

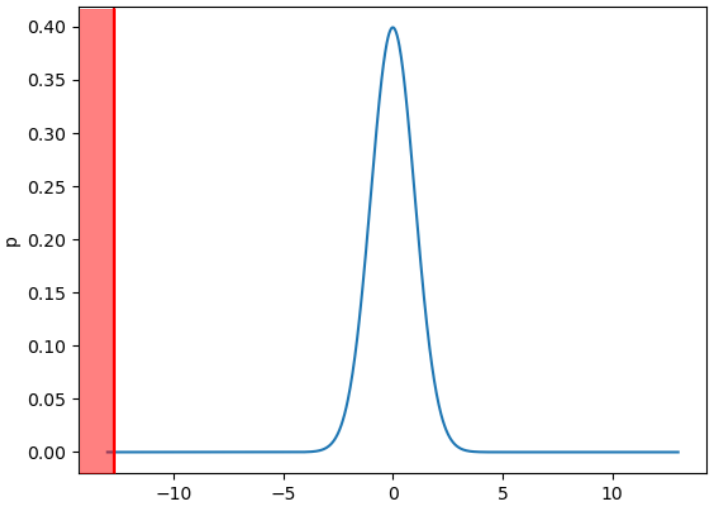
- Sometimes, instead of quoting the p-value itself, it is converted into units of standard deviations
  - The uniform p-value distribution is transformed into a unit normal, and then the number of standard deviations „sigma“ is computed

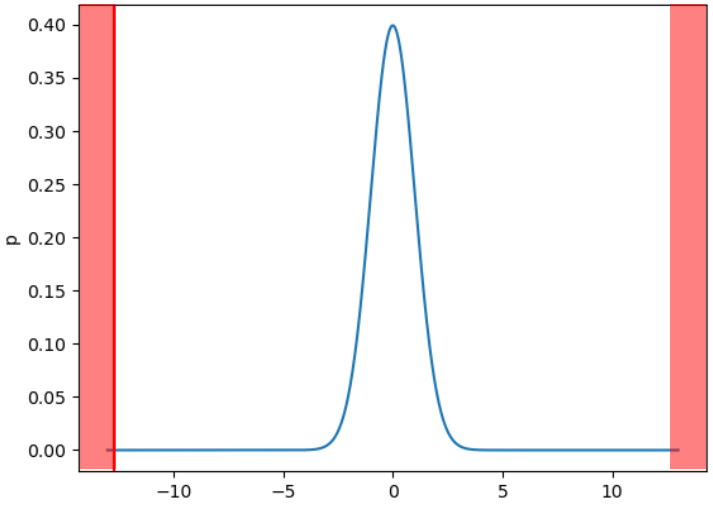
One vs. Two-sided p-values

- In our example, we calculated a one-sided p-value, meaning x has to be at least as negative as -12.7

- Instead, we could also define a two-sided p-value. Meaning in our example, we would encode $|x| > 12.7$

→ p-value = 1.48 x 10^{-37}





P-values in practice

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC ☆

ATLAS Collaboration *

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

ARTICLE INFO

Article history:
Received 31 July 2012
Received in revised form 8 August 2012
Accepted 11 August 2012
Available online 14 August 2012
Editor: W.-D. Schlatter

ABSTRACT

A search for the Standard Model Higgs boson in proton–proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of 126.0 ± 0.4 (stat) ± 0.4 (sys) GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

© 2012 CERN. Published by Elsevier B.V. Open access under CC BY-NC-ND license.

NEUTRINO ASTROPHYSICS

Evidence for neutrino emission from the nearby active galaxy NGC 1068

IceCube Collaboration*†

A supermassive black hole, obscured by cosmic dust, powers the nearby active galaxy NGC 1068. Neutrinos, which rarely interact with matter, could provide information on the galaxy's active core. We searched for neutrino emission from astrophysical objects using data recorded with the IceCube neutrino detector between 2011 and 2020. The positions of 110 known gamma-ray sources were individually searched for neutrino detections above atmospheric and cosmic backgrounds. We found that NGC 1068 has an excess of 79^{+22}_{-20} neutrinos at tera–electron volt energies, with a global significance of 4.2σ, which we interpret as associated with the active galaxy. The flux of high-energy neutrinos that we measured from NGC 1068 is more than an order of magnitude higher than the upper limit on emissions of tera–electron volt gamma rays from this source.

Typical P-values

- **(Particle)Physics**:
  - 3σ (~p ≈ 0.0013): Considered "**evidence**" for an effect
  - 5σ (~p ≈ 3×10^{-7}): Considered a "**discovery**" standard

- **Other fields (Social Sciences, Medicine, Biology,…):**
  - The conventional threshold is **p < 0.05** (5% significance level).
  - Sometimes **p < 0.01** or **p < 0.001** is used for stronger evidence.

- Since scientific results are generally considered 'worthy' of publication when they show significant effects, studies reporting null results are published much less frequently.

Exercise

Hypothesis Tests

Our Null is a Poisson with expected 7 events, but measuring one may get many fewer events

- What is the threshold for an observed count to get a p-value below 5%?

$$P(x \leq 2|\lambda = 7) = 2.96\% \rightarrow \mathrm{x} = 2$$

- What‘s the p-value for an observerd x=0?

$$P(x \leq 0|\lambda = 7) = 0.091\%$$

- Expressed in standard deviations?

$$3.1\sigma$$

Poisson probability distribution		
Random variable, x	P(X = x)	P(X ≤ x)
0	0.0009	0.0009
1	0.0064	0.0073
2	0.0223	0.0296
3	0.0521	0.0818
4	0.0912	0.173
5	0.1277	0.3007
6	0.149	0.4497
7	0.149	0.5987
8	0.1304	0.7291
9	0.1014	0.8305
10	0.071	0.9015

Test Statistic

In order to work with more complex models (i.e. more than a simple observation), we need to define a “test statistic”

A “Test statistic” (TS) is simply put a single numerical value to summarize our data under a hypothesis

→ For example, in our case, we might be tempted to use the likelihood

Our likelihood is:

$$L_x(\lambda) = \prod_i p(xi|\lambda) = \prod_i \frac{1}{x_i!} \lambda^{x_i} \exp(-\lambda)$$

So what‘s the p-value?

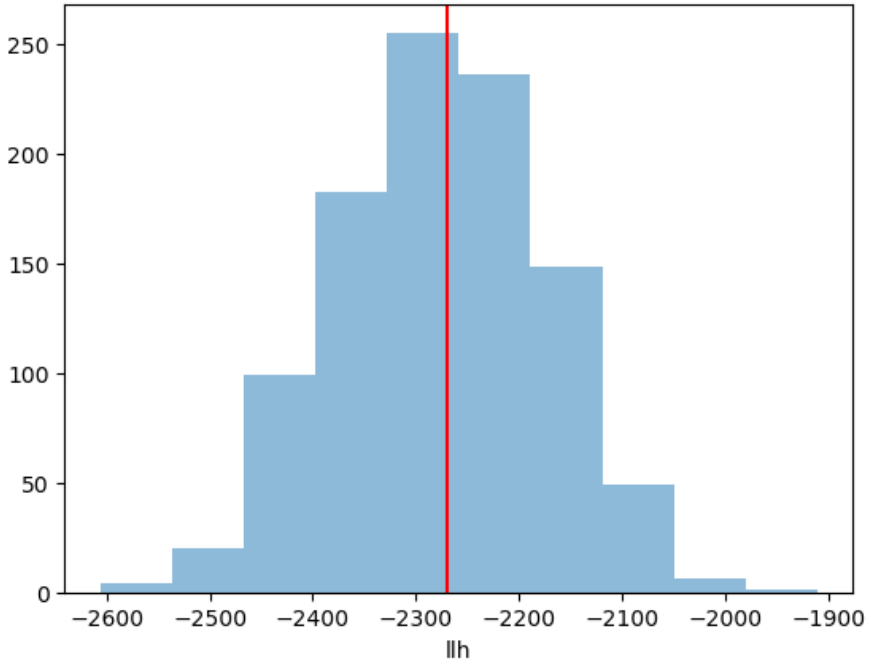
There is typically no analytic distribution that encodes the distribution of the TS (here the likelihood $L_x(\lambda)$) …

But perhaps it’s good enough to approximate $p(TS)$ with many samples of x under our hypothesis :

→ Generate “trials” to see how the TS is distributed, a.k.a. “sampling distribution”

→ estimate p-value (= fraction of trials more extreme than observed value)

In our example: p-value is around 50%, based on 1000 trials



Problems with this Approach

- One should in practice never use the bare likelihood as a test statistic!
- For instance: if we also use the data to estimate parameters of the hypothesis, we run into problems

→Nice write-up about problems using the likelihood: https://www.slac.stanford.edu/econf/C030908/papers/MOCT001.pdf

There are countless alternatives („goodness of fit“ tests): chi2 test, KS test, AD test, …

Introducing: Two hypotheses testing

- We can make the situation much better, if we can specify an alternate hypothesis H_1

- Instead of asking “Is the TS small or large?” we ask: “Is the data relatively more compatible with one hypothesis than another?”

→ we can formulate a binary decision on which hypothesis to reject/accept

Testing with Sampling Distributions

Simplest Case: consider two hypotheses of one-dimensional data

"the null hypothesis" H_0: **data originates from model** $p_0(x|\theta_0)$

"the alternative hypothesis" H_1: **data originates from model** $p_1(x|\theta_1)$

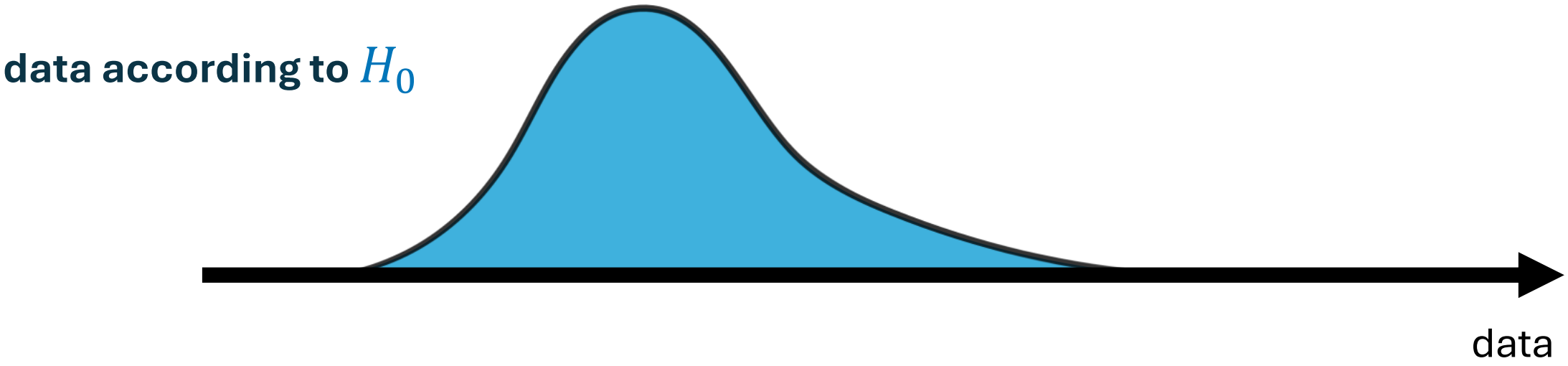
Decision we want to make: **should we reject the "null hypothesis" ?**

Testing with Sampling Distributions

Let's see the data that the hypotheses are predicting

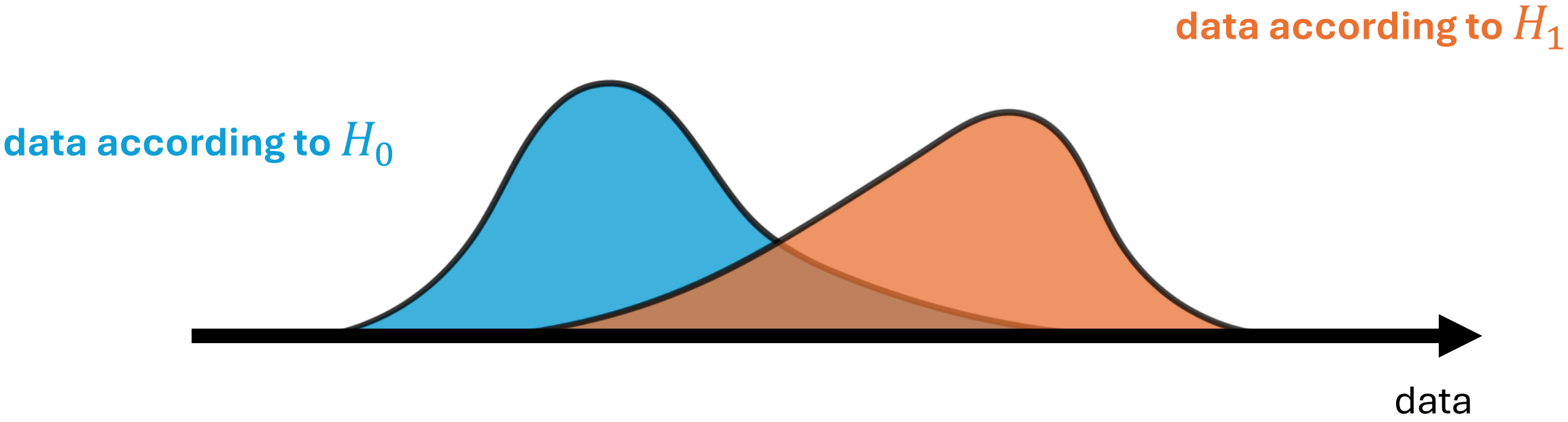
Testing with Sampling Distributions

Let's see the data that the hypotheses are predicting



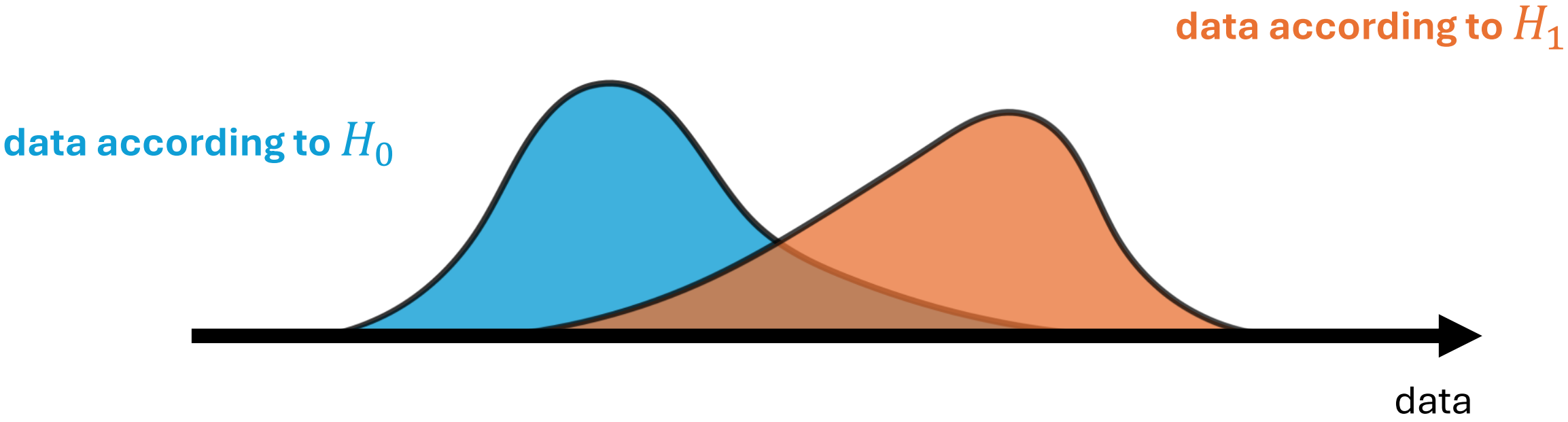
Testing with Sampling Distributions

Let's see the data that the hypotheses are predicting



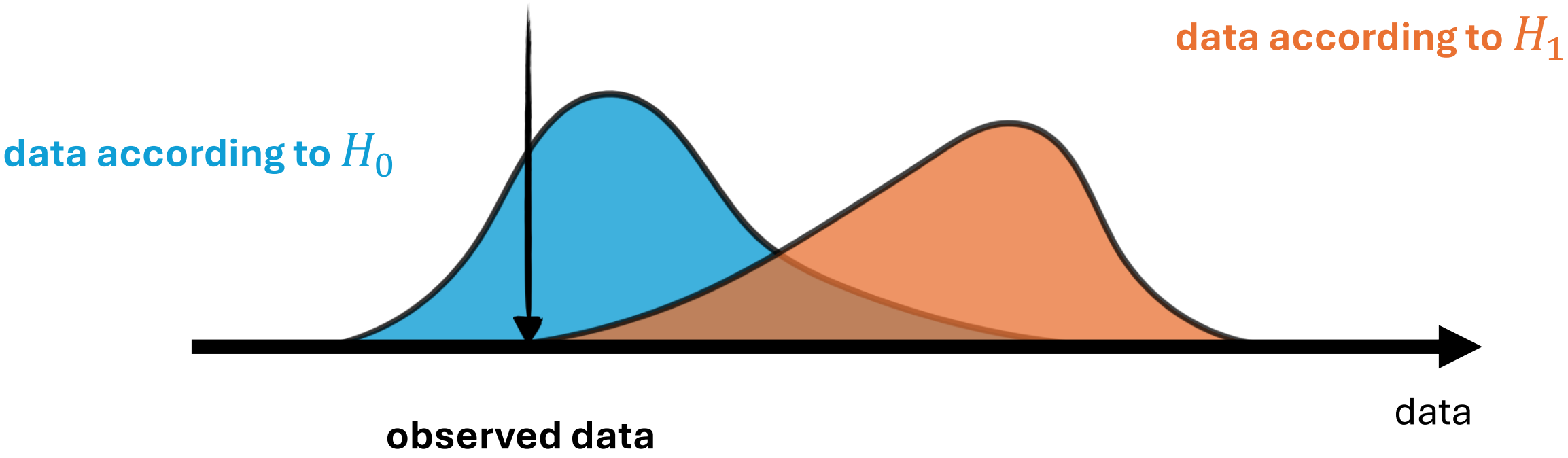
Testing with Sampling Distributions

Let's look at the data **observed** in the real world:



Testing with Sampling Distributions

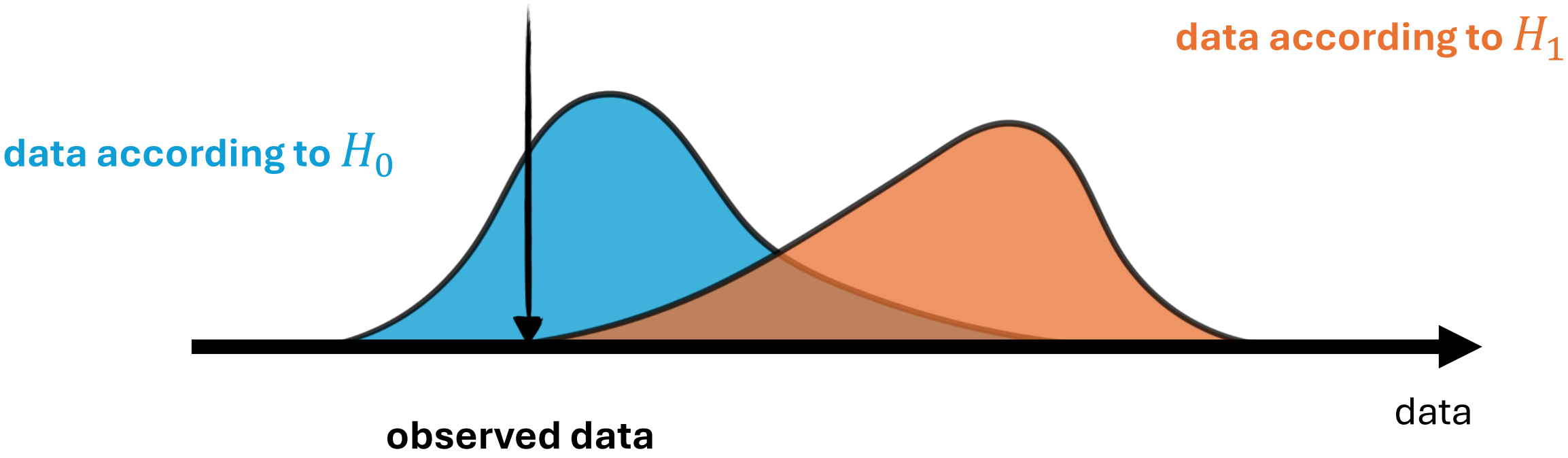
Let's look at the data **observed** in the real world



Testing with Sampling Distributions

Question:

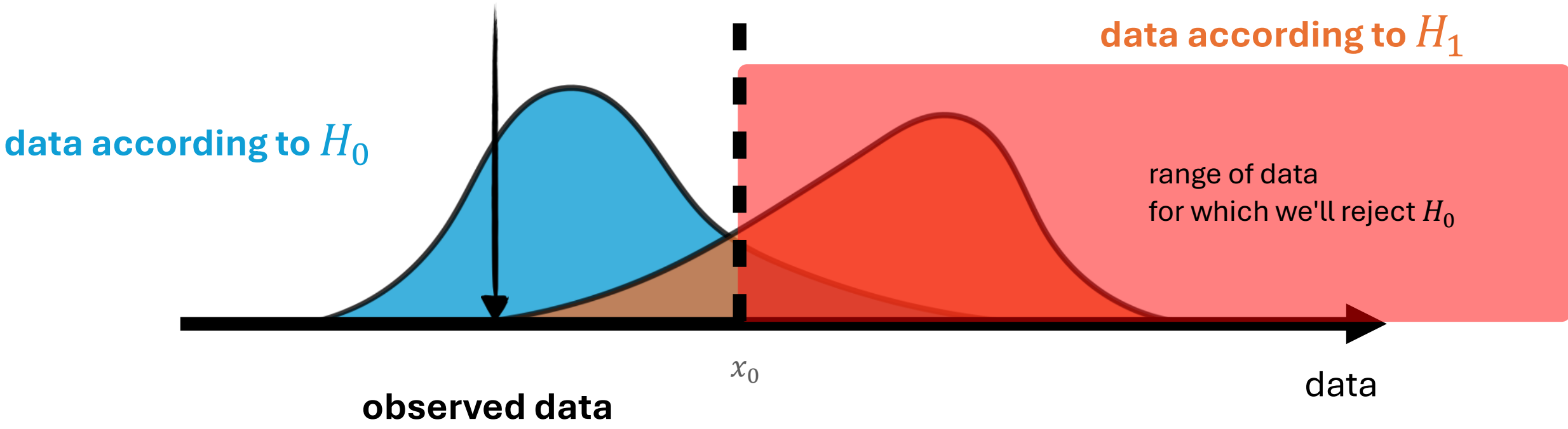
if you had to formulate a rule to reject (or not) H_0 what would it be?



Testing with Sampling Distributions

A reasonable answer:

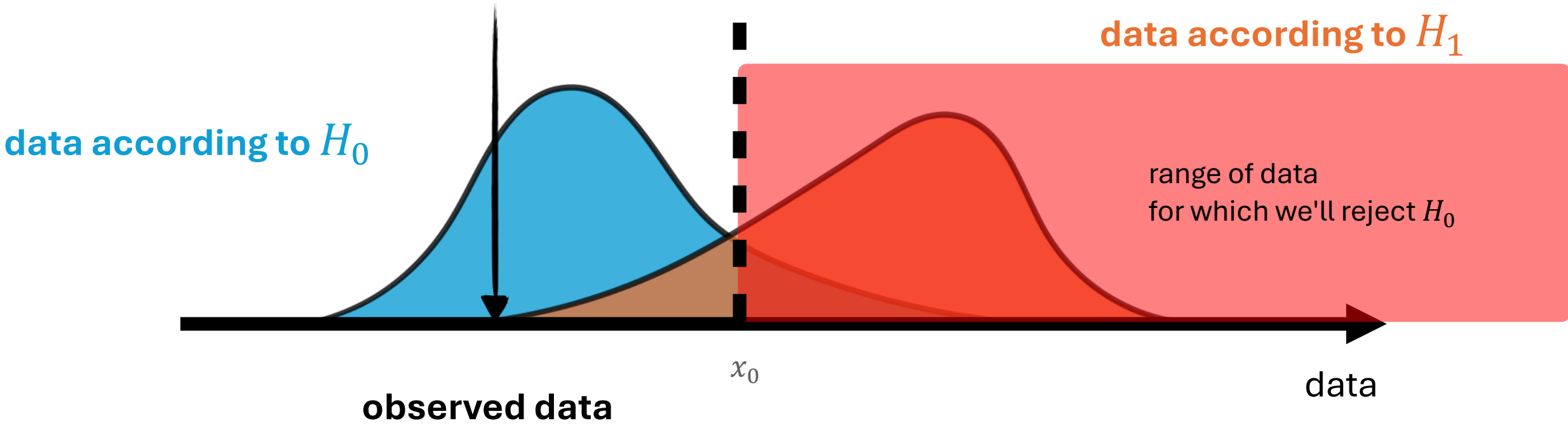
- reject H_0 if data is too far too the right (e.g. data > x_0)



Testing with Sampling Distributions

A reasonable answer:

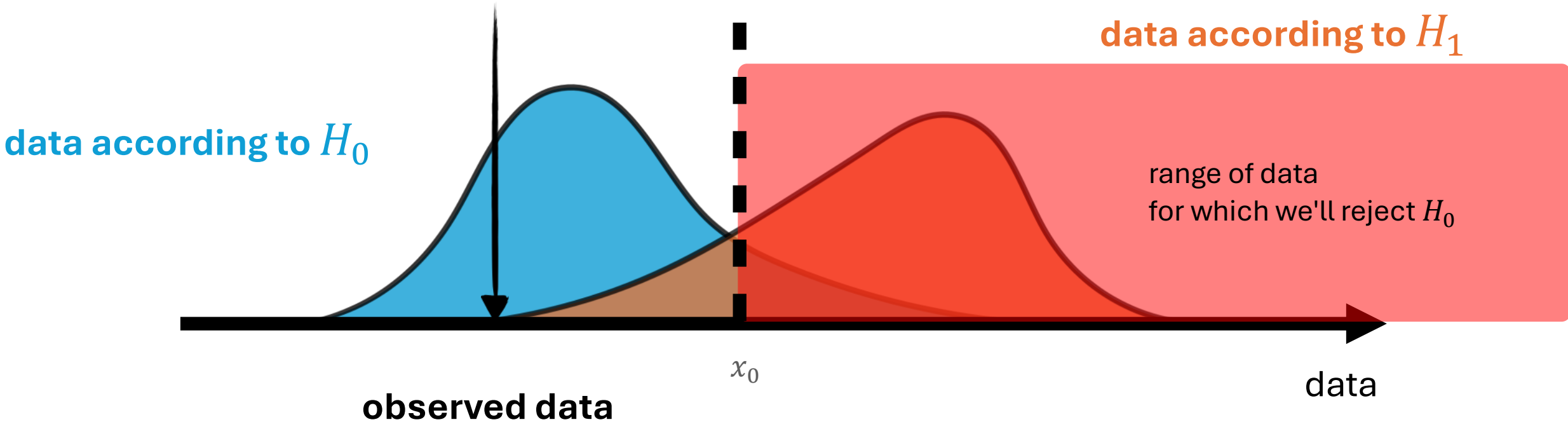
- reject H_0 if data is too far too the right (e.g. data > x_0)
- follow-up question: how do you choose x_0?



Testing with Sampling Distributions

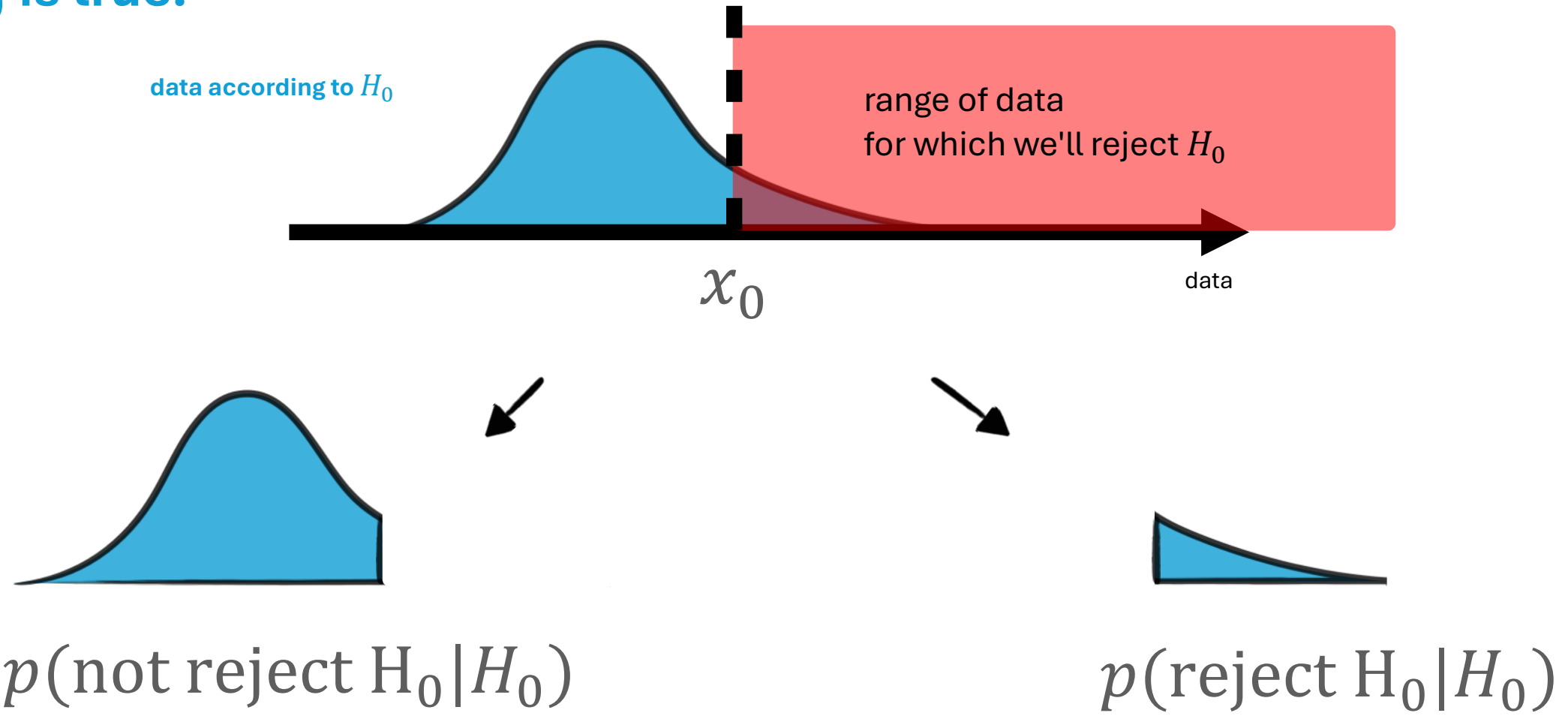
How good is this rule?

- look at the performance for each of possible scenarios
- i.e. probabilities of making the right decision



Testing with Sampling Distributions

If H_0 is true:

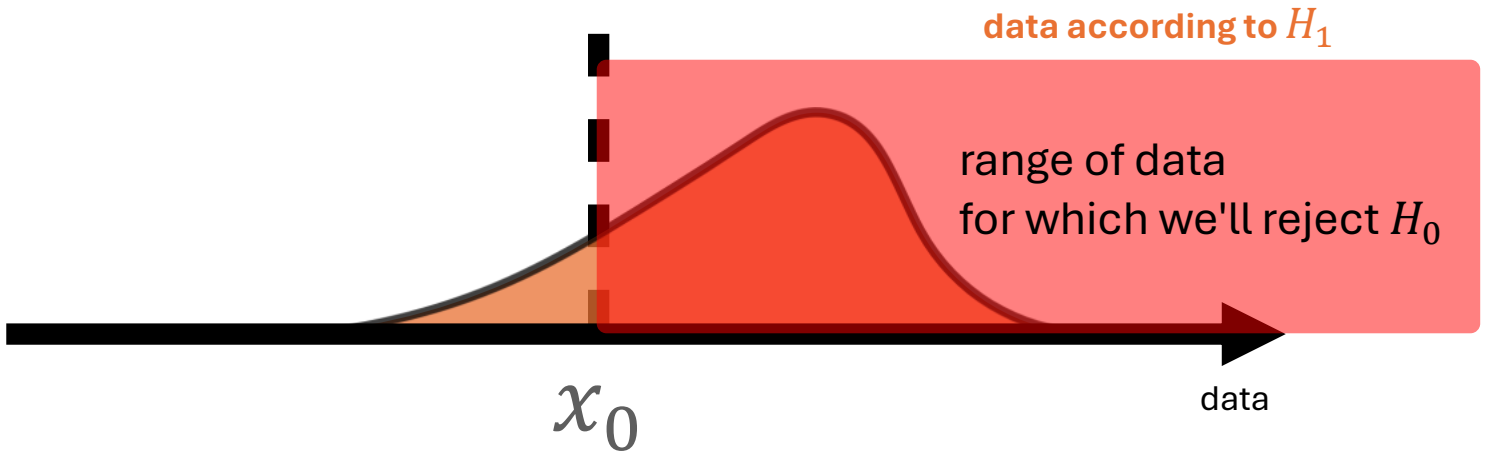


$p(\text{not reject } \mathrm{H}_0 | H_0)$

$p(\text{reject } \mathrm{H}_0 | H_0)$

Testing with Sampling Distributions

If H_1 is true:

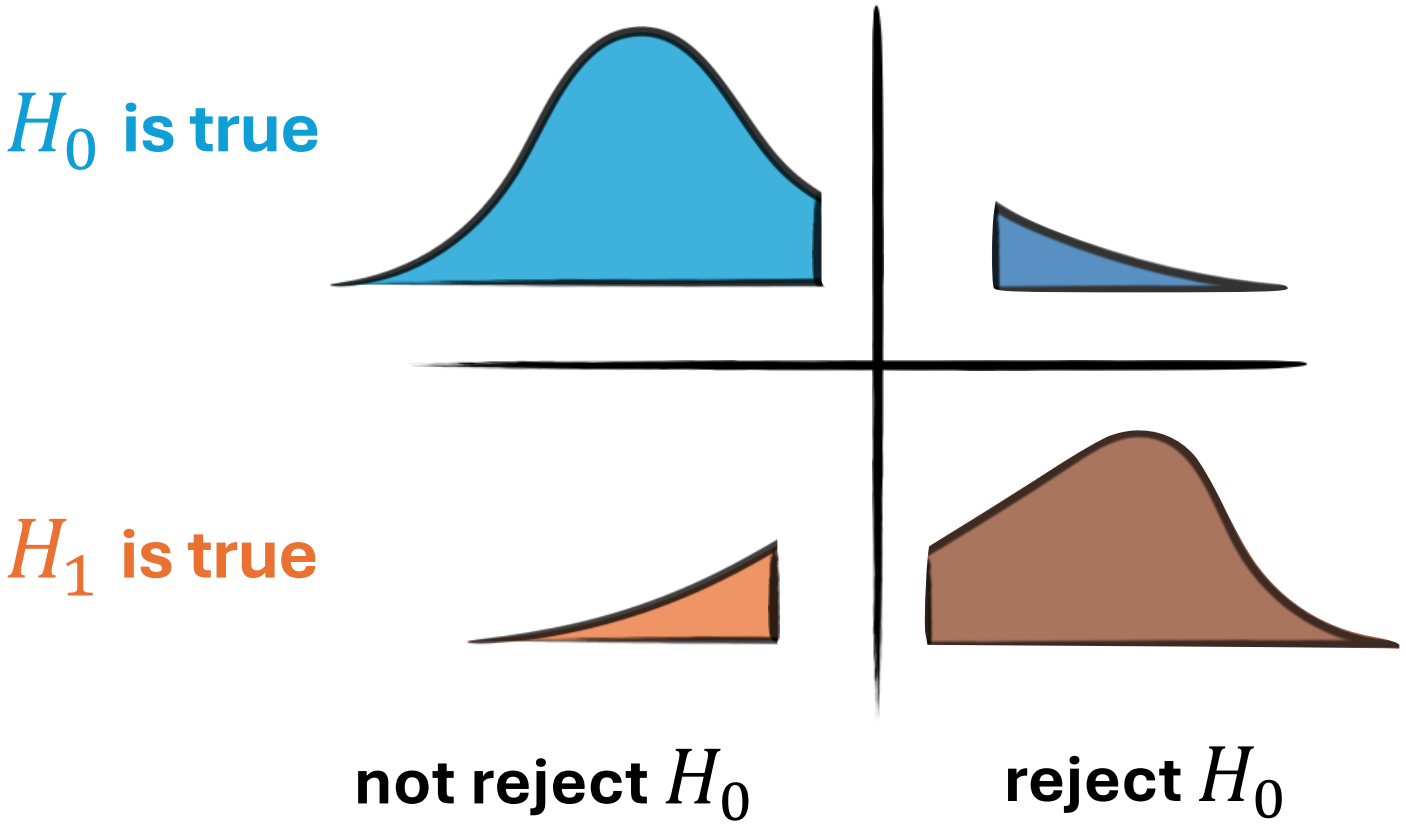


$p(\text{not reject } \mathrm{H}_0 | H_1)$

$p(\text{reject } \mathrm{H}_0 | H_1)$

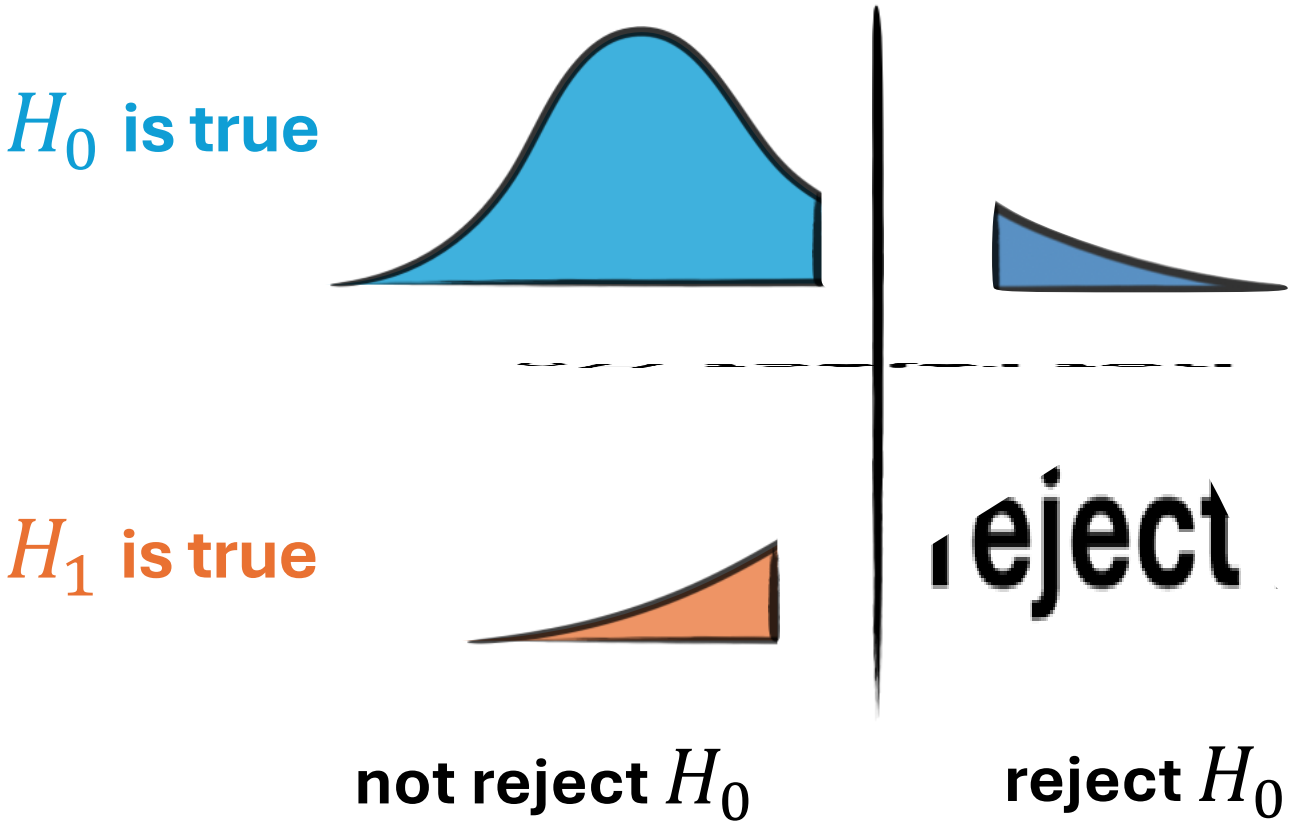
Type I & Type II Error

Two Hypotheses & Two Actions



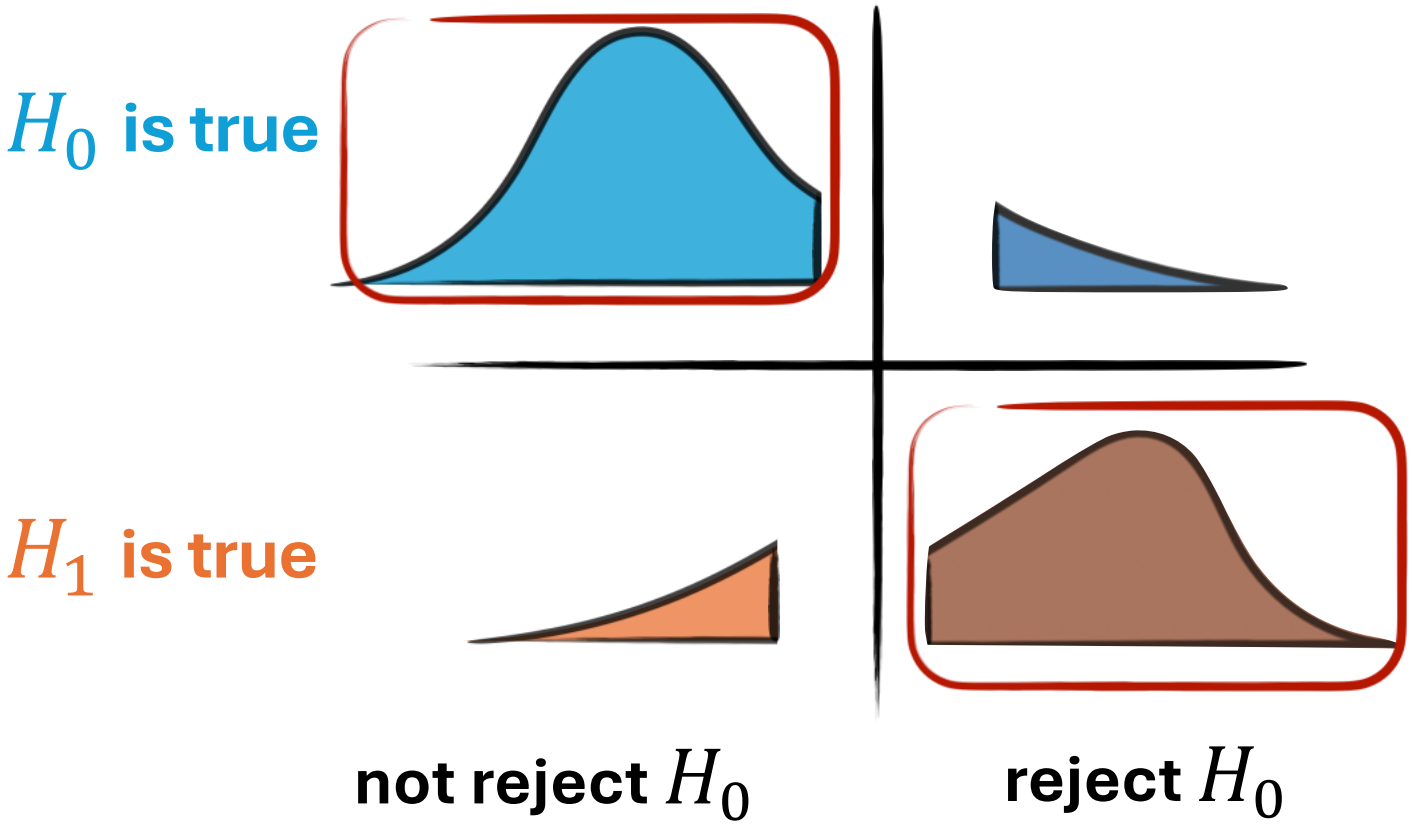
Type I & Type II Error

What are the favorable decisions?



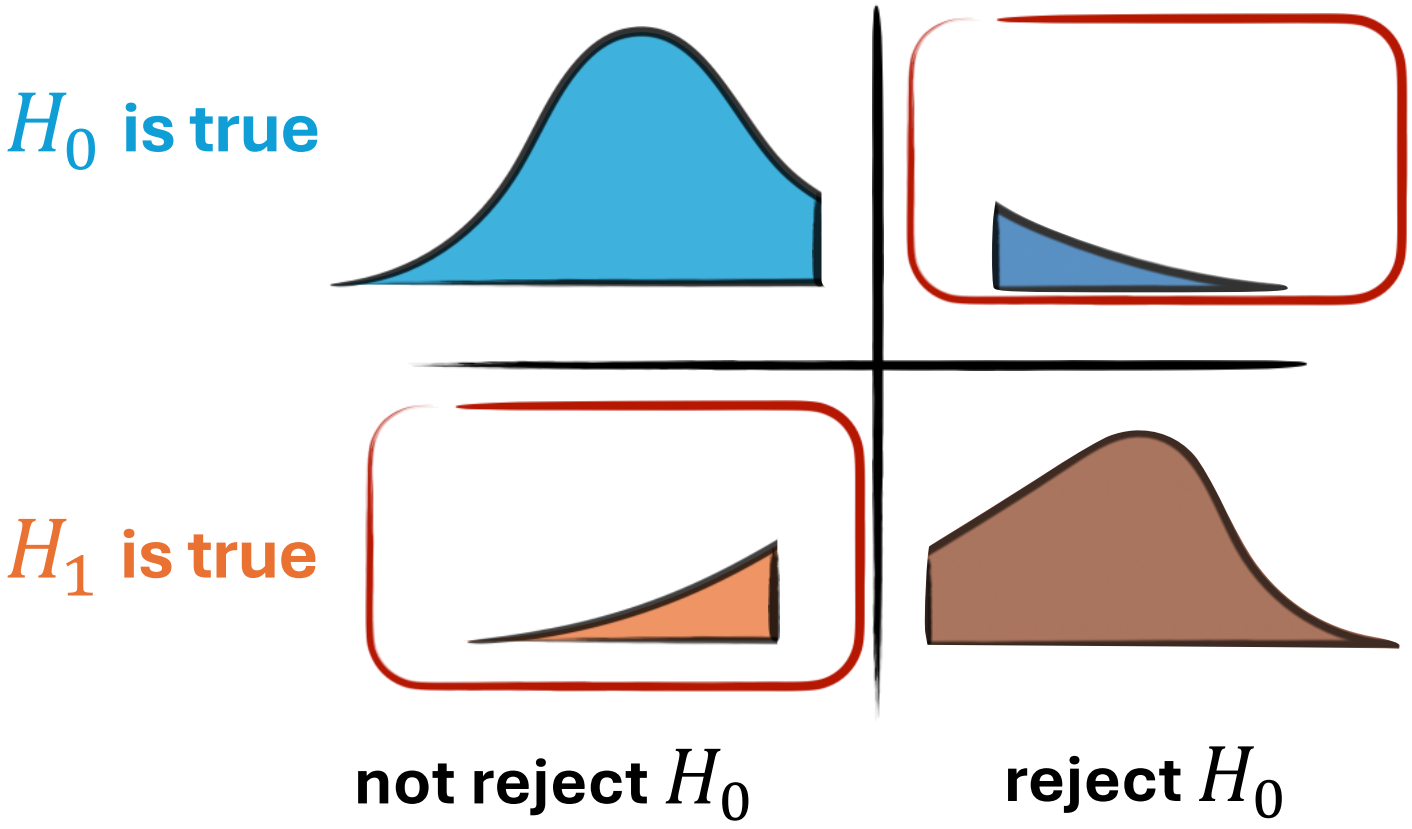
Type I & Type II Error

What are the favorable decisions?



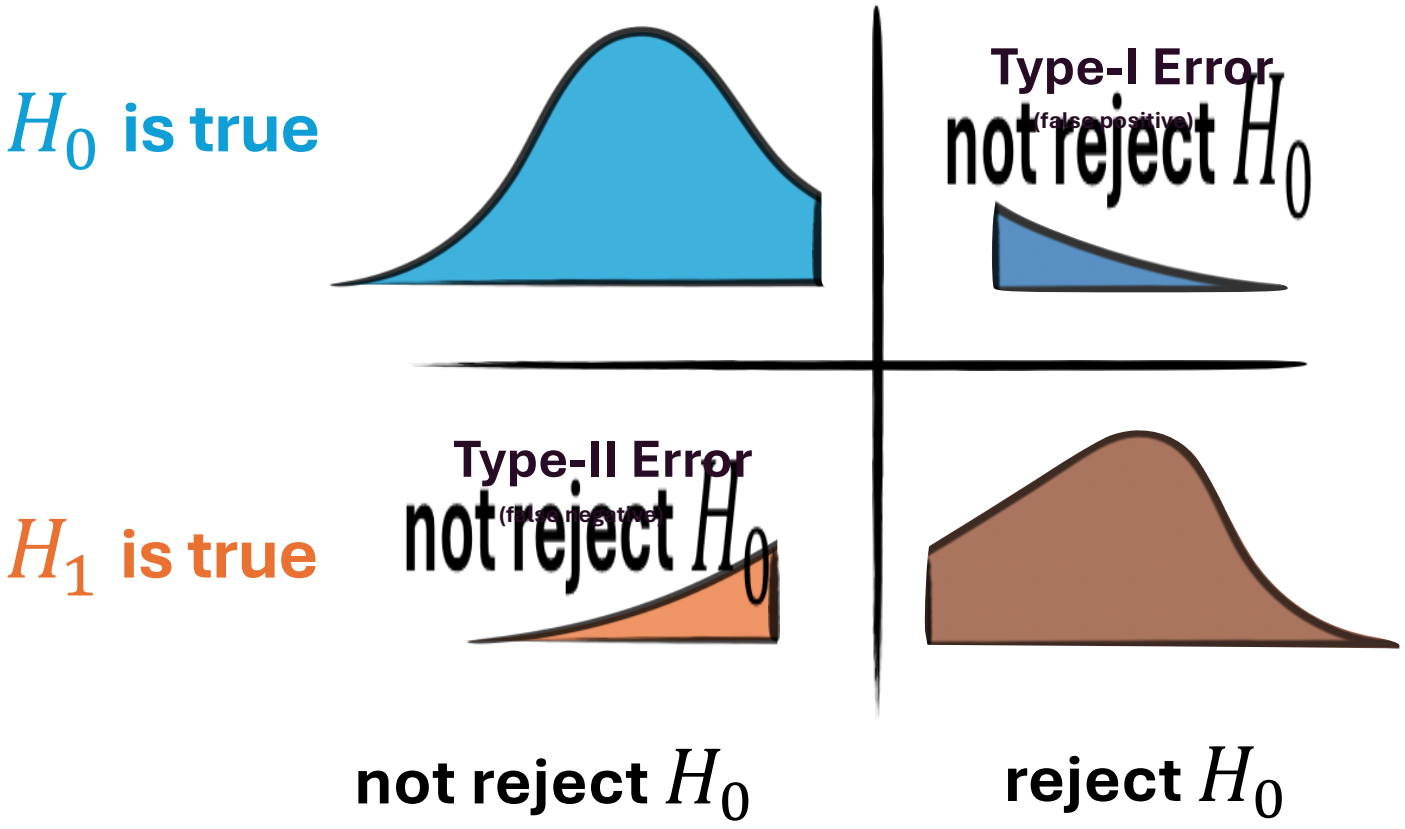
Type I & Type II Error

What are the unfavorable decisions?



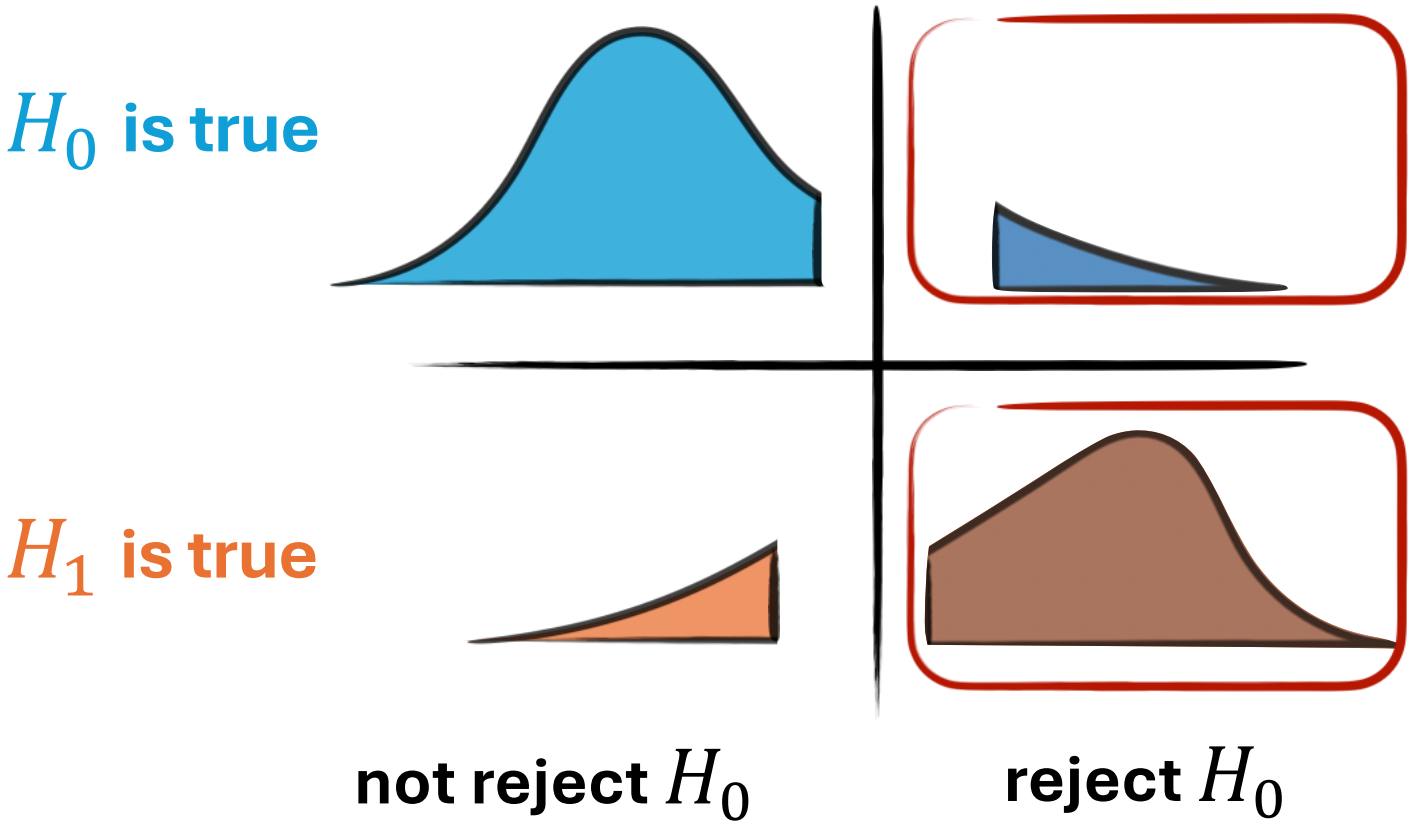
Type I & Type II Error

What are the bad decisions?



Type I & Type II Error

The Performance of the test is fully characterized by two numbers



"the size" of the test
false positive rate

"the power" of the test
true positive rate

Formalizing the Intuition

A statistical test to distinguish two hypotheses H_0 and H_1 with models $p(x|H_0)$ and $p(x|H_1)$ is based on a test statistic $t(x)$

$$t(x)\colon \mathbb{R}^n \to \mathbb{R}$$

The null hypothesis H_0 will be rejected for all x with $t(x) > t_0$

- The set $\omega = \{x\colon t(x) > t_0\}$ is called the **rejection region**

Formalizing the Intuition

The **size of the test** is defined relative to the null hypothesis.

- probability to be in rejection region for H_0

$$\text{size: } p(x \in \omega | H_0) = \int_{t_0}^{\infty} p(t|H_0)$$

The **power of the test** is defined relative to a alternative hypothesis

- probability to be in rejection region given the alternative

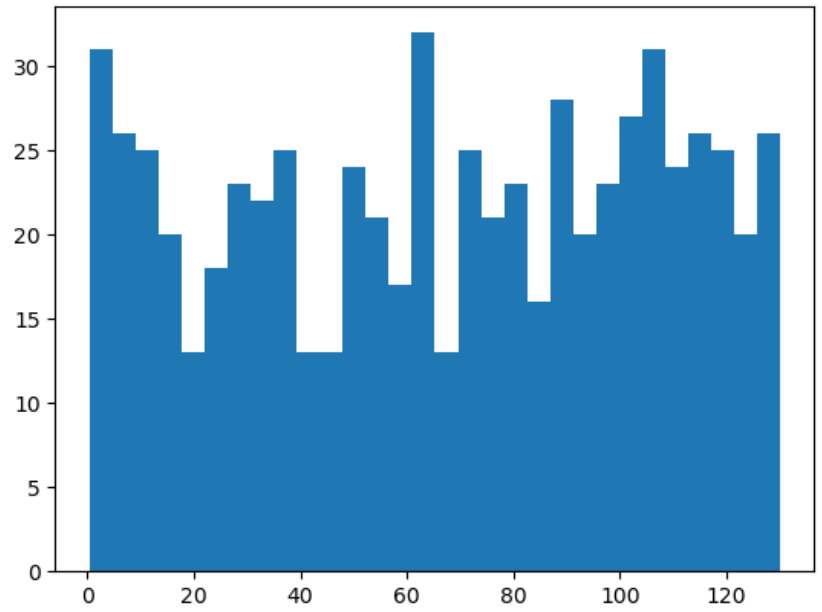
$$\text{power: } p(x \in \omega | H_1) = \int_{t_0}^{\infty} p(t|H_1)$$

Back to our Example

- **Null hypothesis:** rate is constant
- **Alternate hypothesis:** Someone saw a truck arriving and workers entering the reactor at T = 70, and therefore the rate may have changed

- Let‘s test two fixed hypotheses:
  - H0: Rate is constant at 5
  - H1: Rate increased from 5 to 6 at T = 70

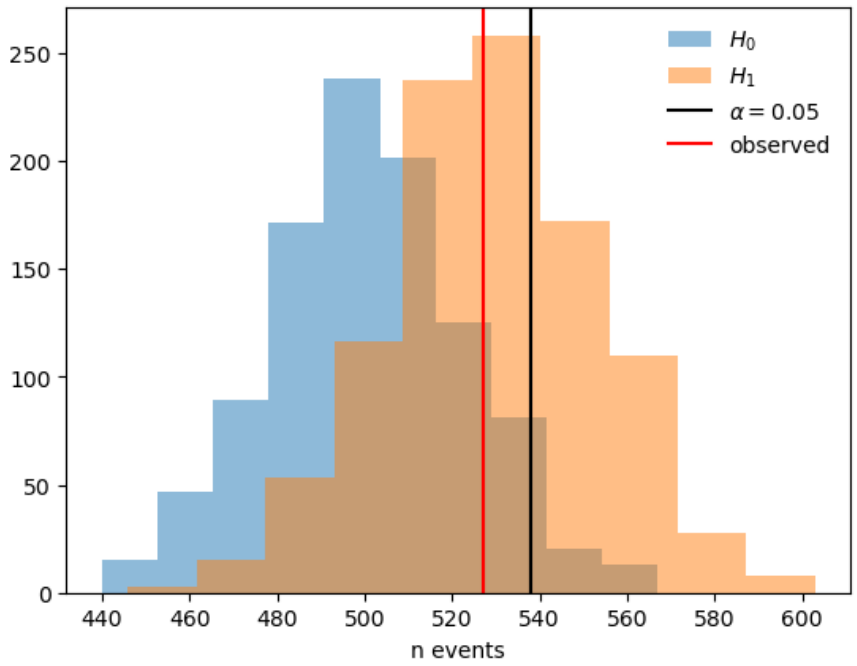
The data you collected that night:



Example Hypothesis test

- We have to choose a test statistic (TS)
  - Let‘s use the total number of observed events (we expect this to be sensitive to de-/increase of the rate)
- And let’s choose the size of the test to be 0.05
- And let’s estimate the p-value via a sampling distribution of 1000 trials

RESULTS:



→ We cannot reject the null Hypothesis at the 5% level (the observed value is not inside the rejection region)

→ The power of the test is rel. large

Comment: p-value vs. size

- p-value = probability of wrongly rejecting the null hypothesis for the **observed value** obtained from the data

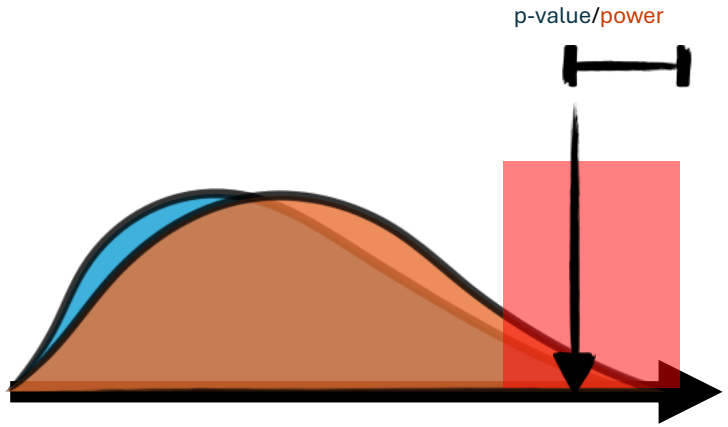
- Size = a value chosen by us for constructing the rejection region. This is independent of the observed data

- The difference is:
  - The size of the test is not a function of the observed data and is constant
  - The p-value is a function of the data and can sometimes be significant (i.e. < than the size), or insignificant

Two examples of low p-values

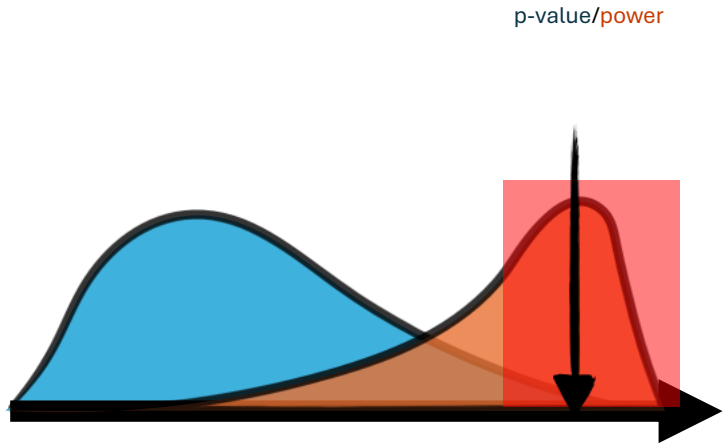
For both experiments the observed data has low p-value (<0.05).

Should you reject the null hypothesis?



Experiment 1:
(low sensitivity)

observed p-value: 0.03
observed power: 0.04



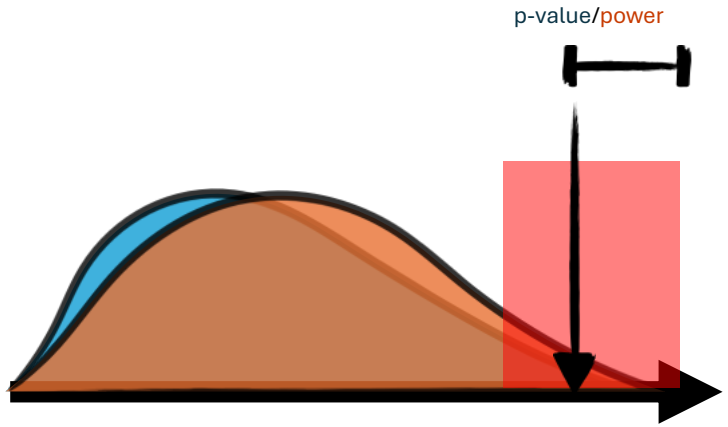
Experiment 2:
(high sensitivity)

observed p-value: 0.02
observed power: 0.5

Two examples of low p-values

Low-power test:

Maybe shouldn't reject H_0 based on p-value alone?



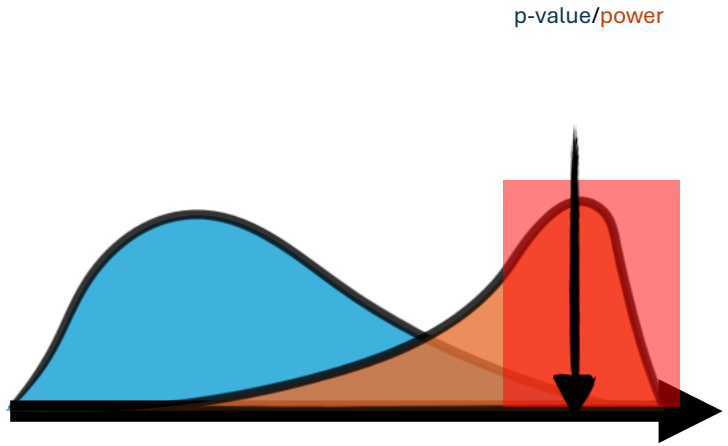
Experiment 1:
(low sensitivity)

observed p-value: 0.03

observed power: 0.04

High-power test.

low p-value is meaningful



Experiment 2:
(high sensitivity)

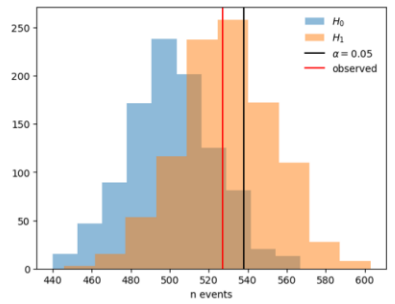
observed p-value: 0.02

observed power: 0.5

The Fundamental Problem in Hypo Tests

Example Hypothesis test

- We have to choose a test statistic (TS)
  - Let‘s use the total number of observed events (we expect this to be sensitive to de-/increase of the rate)
- And let’s choose the size of the test to be 0.05
- And let’s estimate the p-value via a sampling distribution of 1000 trials



→ We cannot reject the null Hypothesis at the 5% level (the observed value is not inside the rejection region)
→ The power of the test is large

→Were our choices optimal?

$$\omega = \{x: t(x) > t_0\}$$

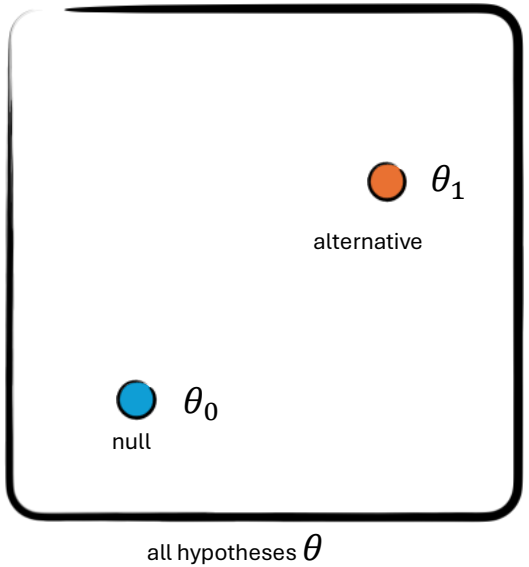
Q1: How to choose the "right test statistic"

Q2: How to choose a good cut-off point

Q3: How to calculate $p(x \in \omega | H)$

Neyman-Pearson Lemma

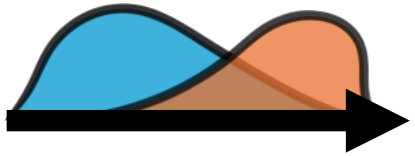
If we want to a null $p(x|\theta_0)$ vs an alternative $p(x|\theta_1)$ we have a very compelling answer



The Neyman-Pearson Lemma:

The Likelihood Ratio: $$t(x) = \frac{p(x|\theta_1)}{p(x|\theta_0)}$$

is the optimal test statistic in the sense that it guarantees maximum power for any given size



Neyman-Pearson Lemma

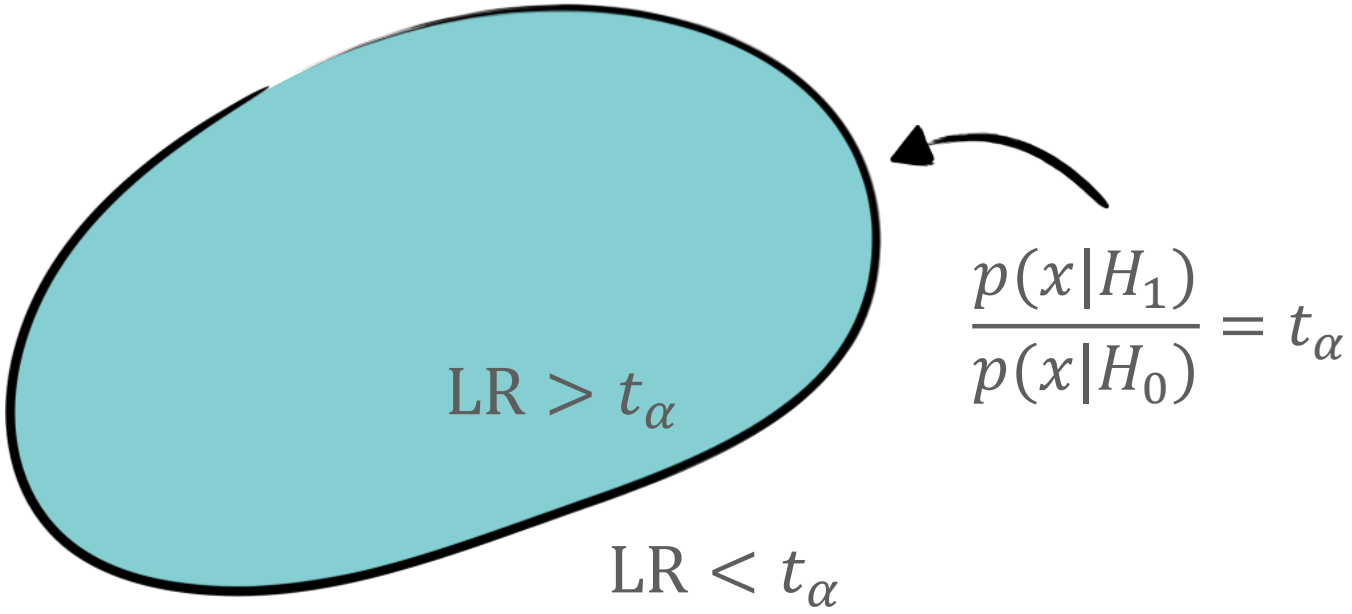
With Likelihood Ratio as test statistic, we reject H_0 when $t(x)$ indicates that data is too H_1-like: **more likely to get x under H_1 than H_0**

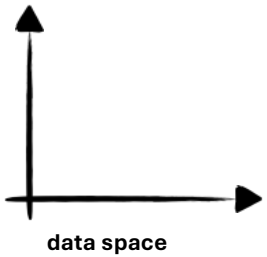
$$t(x) = \frac{p(x|\theta_1)}{p(x|\theta_0)} > t_\alpha$$

or equivalently

$$\lambda(x) = -2\log\frac{p(x|\theta_0)}{p(x|\theta_1)} > \lambda_\alpha$$

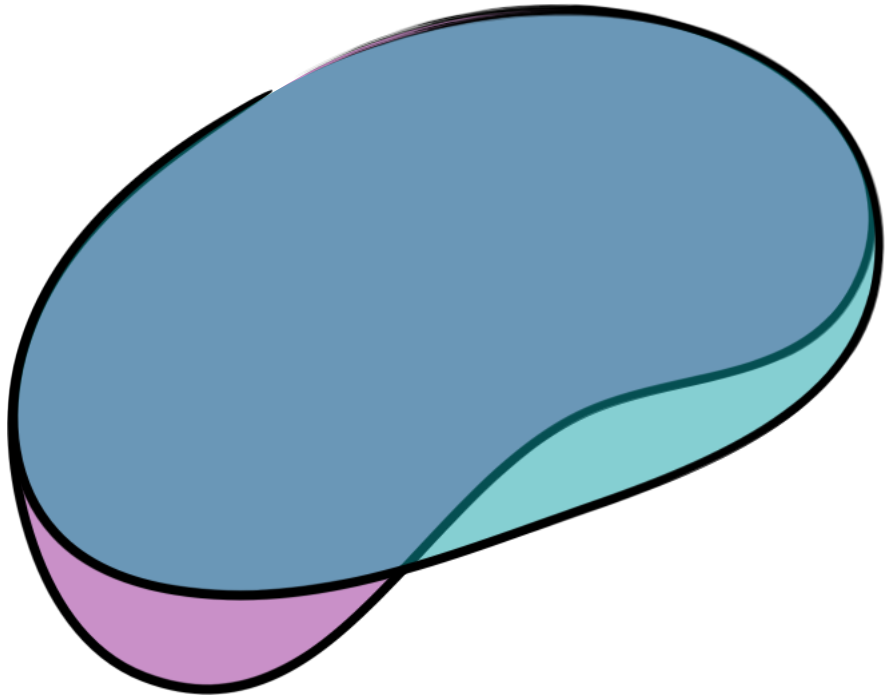
Visual Proof of NP-Lemma

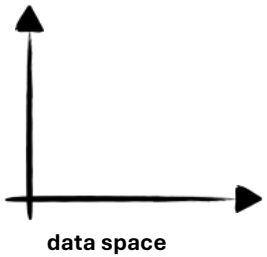




Start with the rejection region as defined by Neyman-Pearson

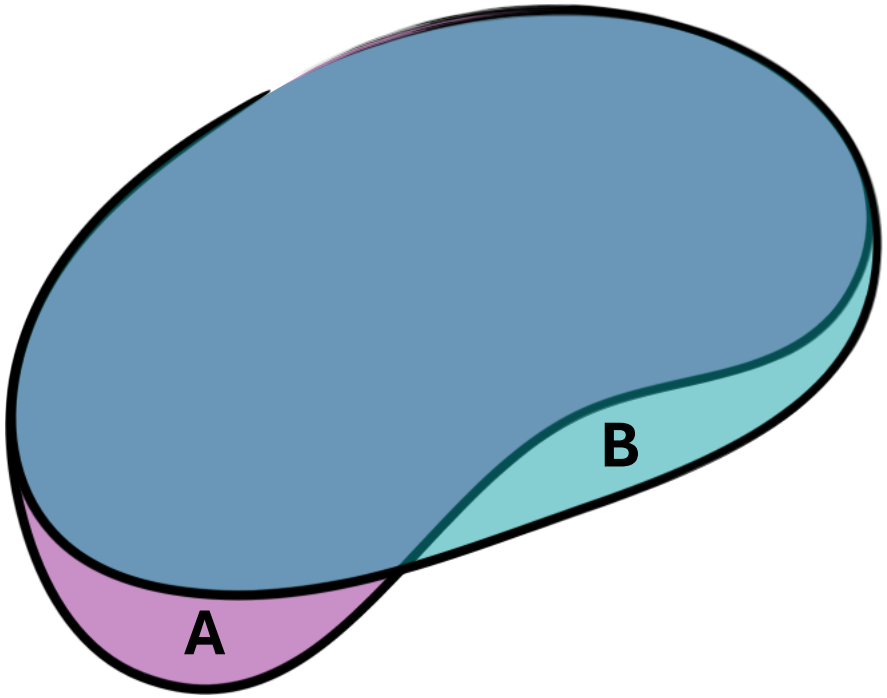
Visual Proof of NP-Lemma

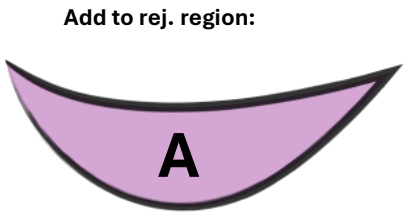


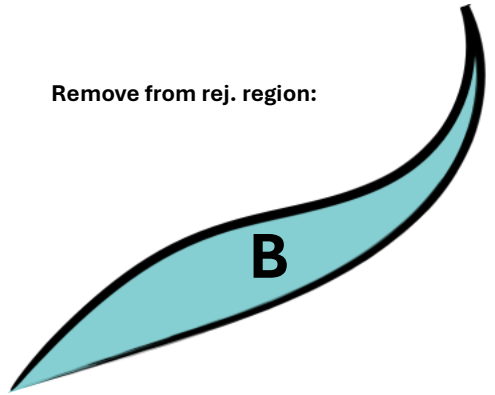


Consider a potential alternative region and see how it's power compares

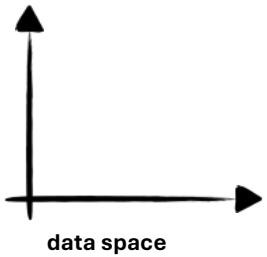
Visual Proof of NP-Lemma



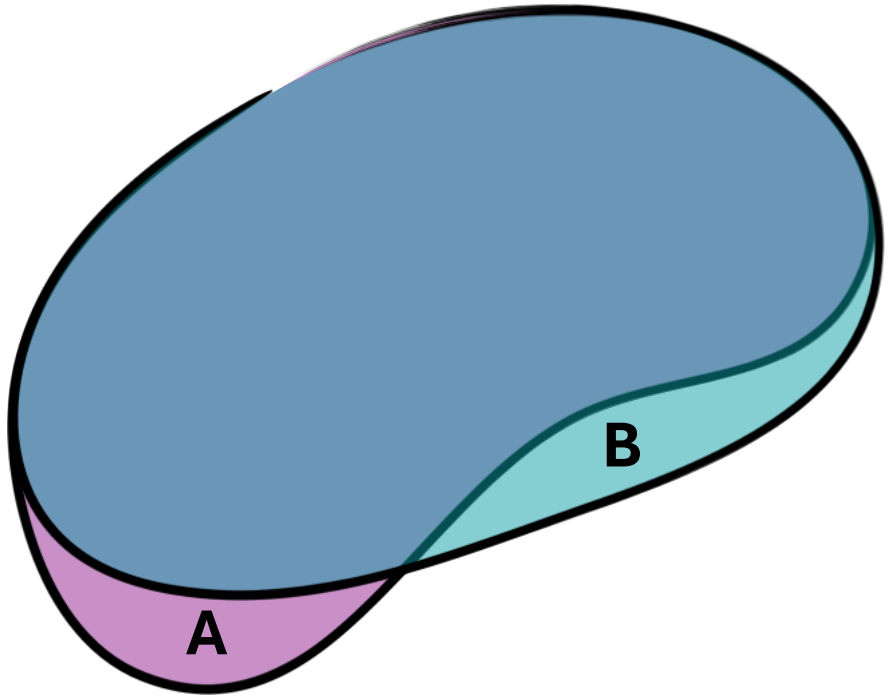




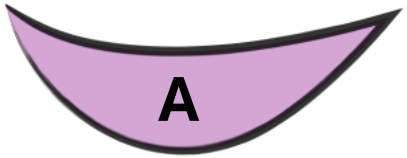
Focus on the areas that are different between the two



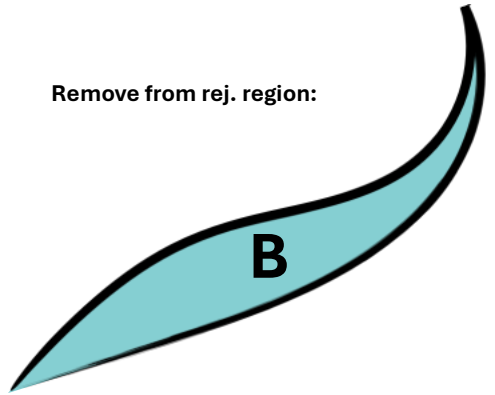
Visual Proof of NP-Lemma



Add to rej. region:

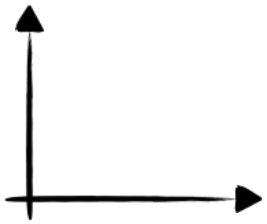


Remove from rej. region:

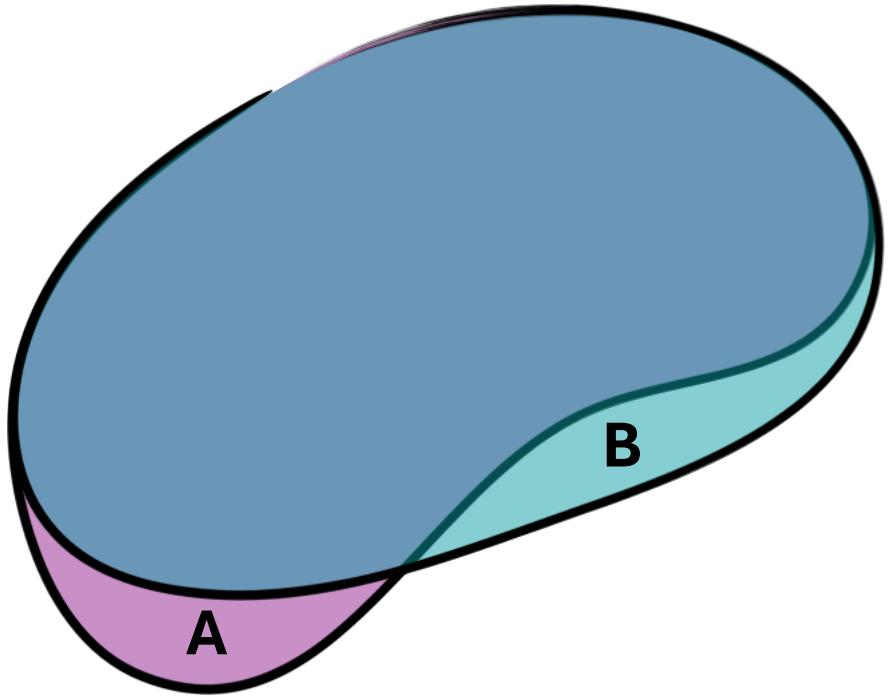


① $p(A|H_0) = p(B|H_0)$

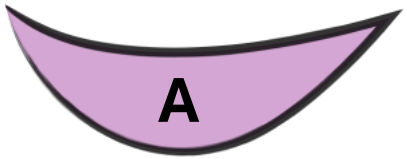
test size should stay constant



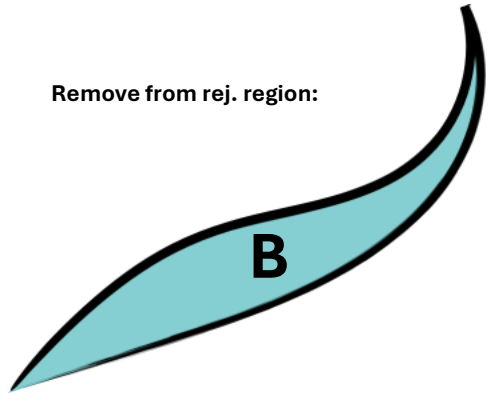
Visual Proof of NP-Lemma



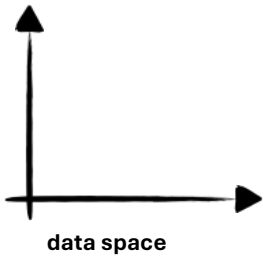
Add to rej. region:



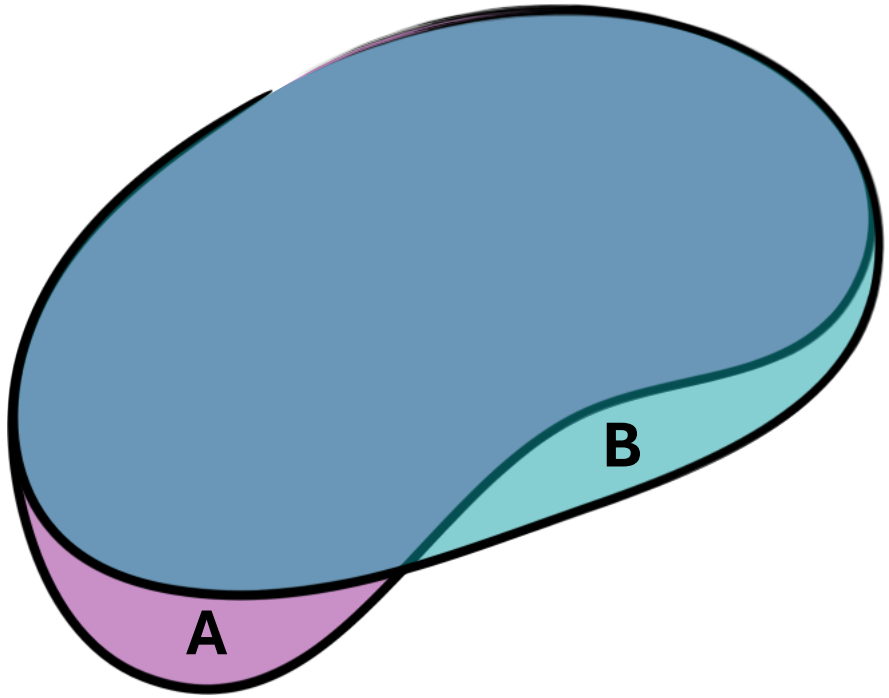
Remove from rej. region:



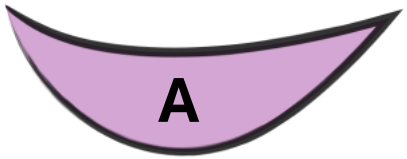
1. $p(A|H_0) = p(B|H_0)$ — test size should stay constant
2. $p(B|H_1) > t_\alpha p(B|H_0)$ — removed region is <u>over</u> threshold k_α



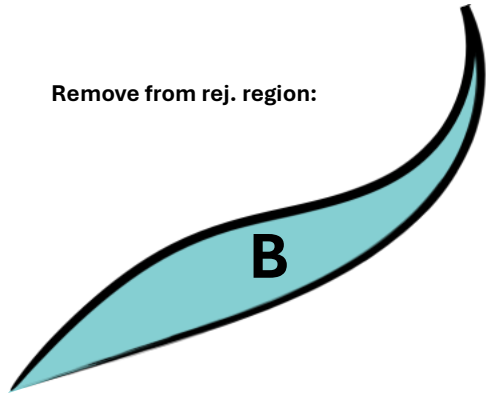
Visual Proof of NP-Lemma



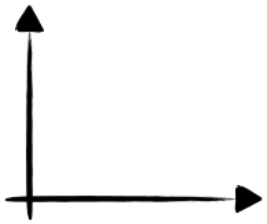
Add to rej. region:



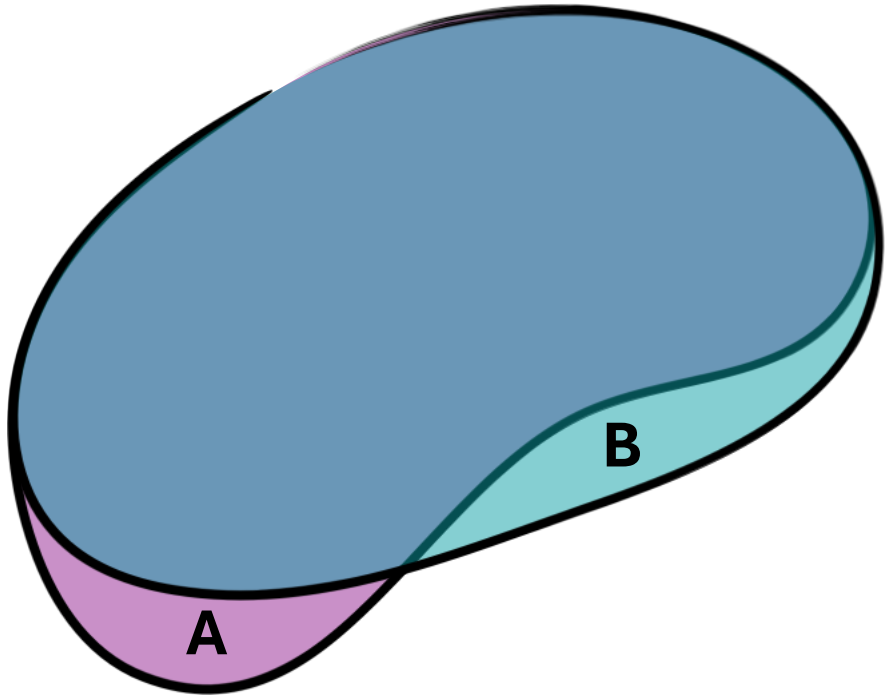
Remove from rej. region:



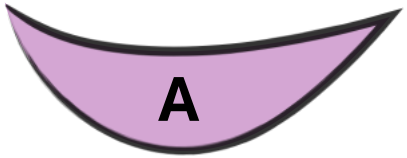
1. $p(A|H_0) = p(B|H_0)$ — test size should stay constant
2. $p(B|H_1) > t_\alpha p(B|H_0)$ — removed region is <u>over</u> threshold k_α
3. $p(A|H_1) < t_\alpha p(A|H_0)$ — added region is <u>under</u> threshold k_α



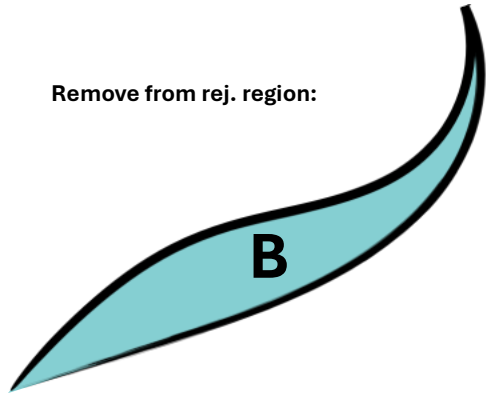
Visual Proof of NP-Lemma



Add to rej. region:



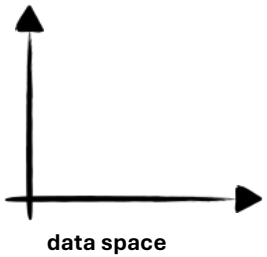
Remove from rej. region:



① $p(A|H_0) = p(B|H_0)$ — test size should stay constant

② $p(B|H_1) > t_\alpha p(B|H_0)$ — removed region is <u>over</u> threshold k_α

③ $p(A|H_1) < t_\alpha p(A|H_0)$ — added region is <u>under</u> threshold k_α

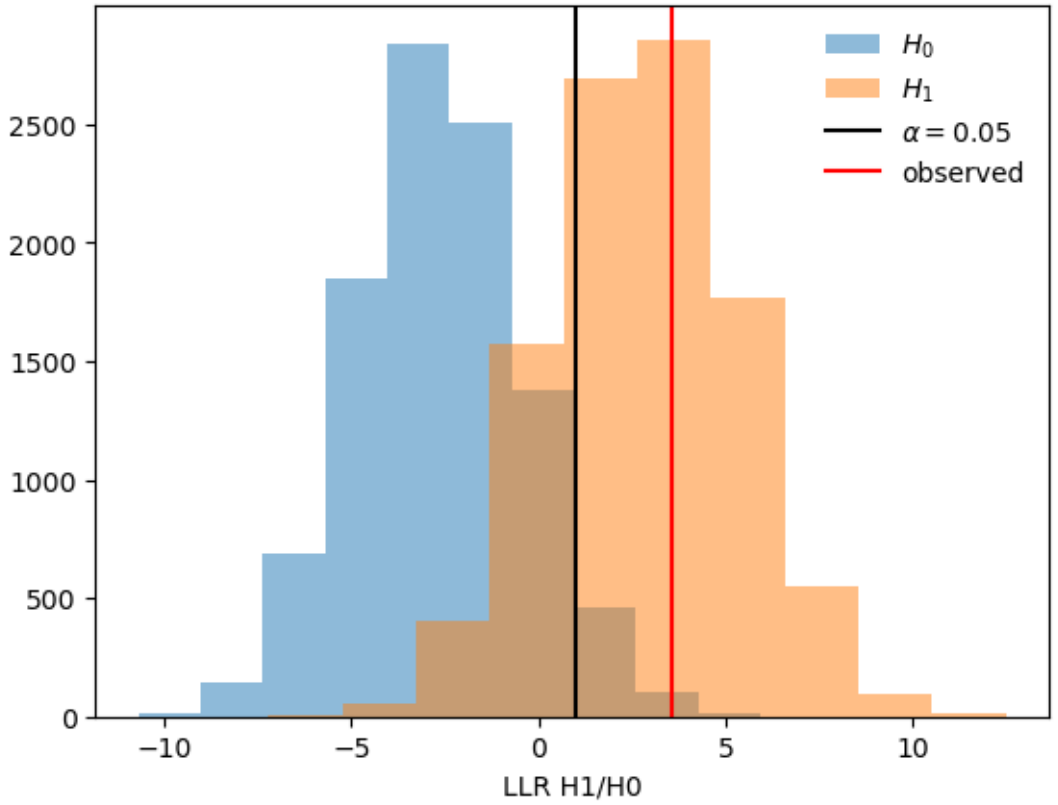


$$\overset{③}{p(A|H_1)} < \overset{①}{t_\alpha p(A|H_0) = t_\alpha p(B|H_0)} \overset{②}{< p(B|H_1)}$$

new region has less power than NP-region!

Using Likelihood Ratio for our example

- Using Neyman-Pearson, i.e. the LHR as TS, we can reject the null at > 5%!
- P-value deep inside the rejection region



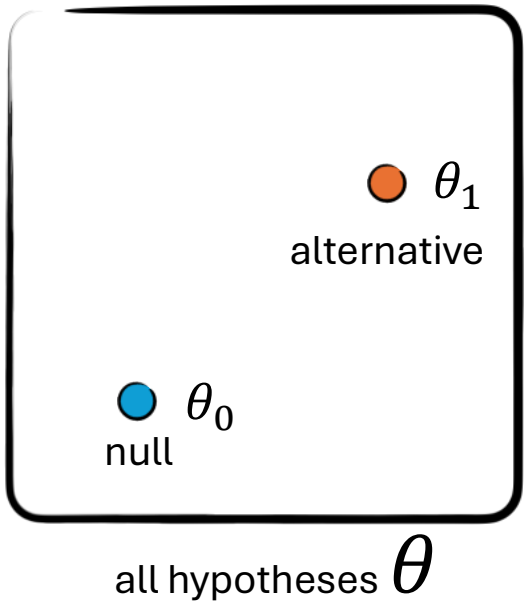
Complex Hypotheses

Complex Hypotheses

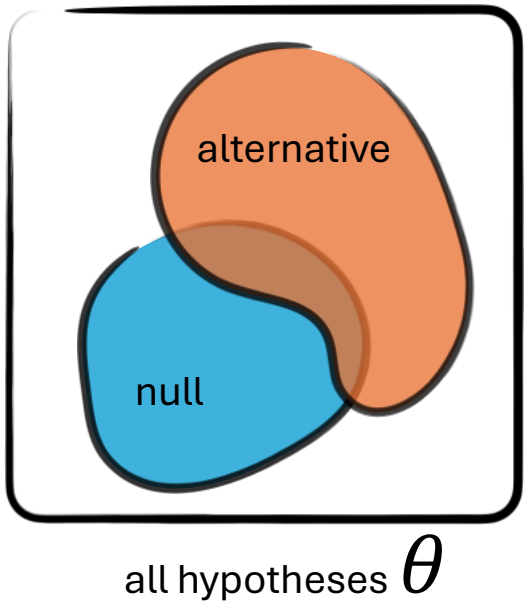
Neyman-Pearson Lemma is nice, but very often our Hypotheses are **not simple in the sense that we've defined.**

- each hypothesis may consist of multiple parameter values
- referred to as "complex hypotheses"

Simple:



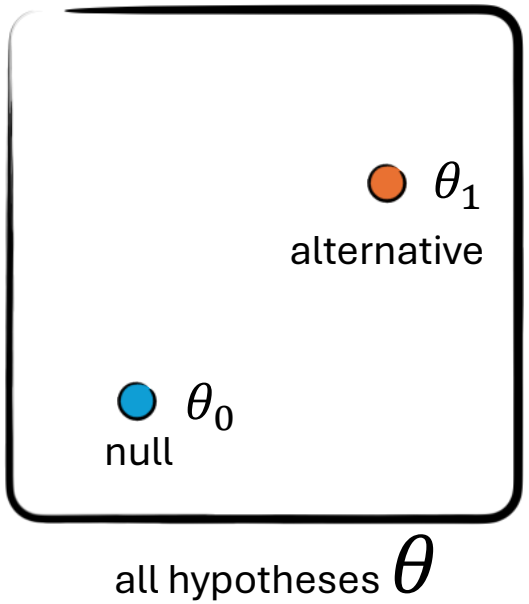
Complex:



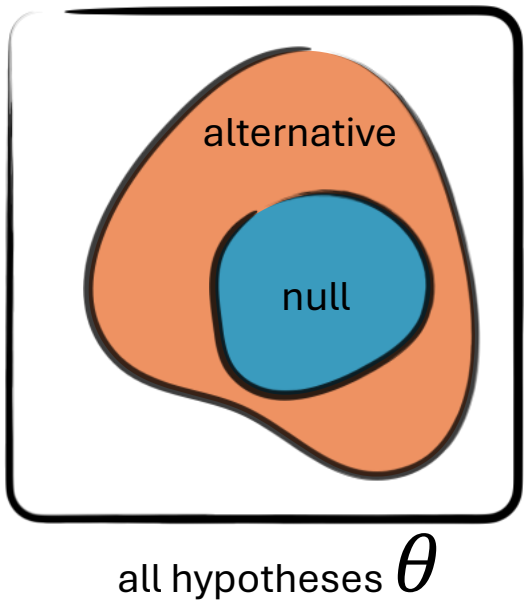
Nested Hypotheses

Hypotheses are "nested" if the null hypothesis set is a subset of the alternative hypothesis set. This is quite common in physics.

Simple:



Complex & Nested:



Common Complex Hypotheses

As we know, realistic models have many parameters including

nuisance parameters just added to model the data

- want to formulate hypothesis in terms of parameters of interest instead of the nuisance parameters → nested models

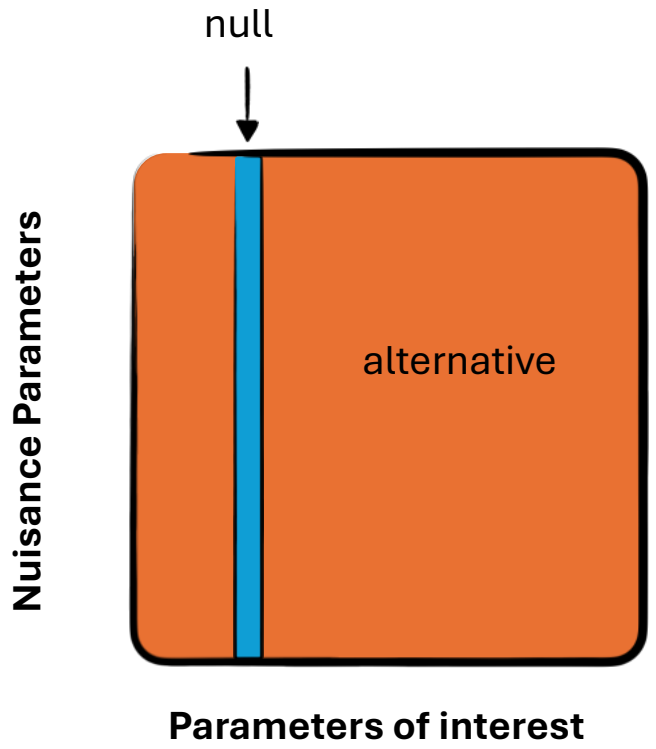
nuisance parameters

$$\theta = (\mu, \nu)$$

core physics
"parameters of interest"

Common Complex Hypotheses

- **Null $H_0: \theta = (\mu = \mu_0, \nu)$**
  all models w/ a given value for POIs are part of this hypothesis
- **Alternative H_1: any θ**



Making our Example more complex

- Nominal rate of neutrinos that we measure from the reactor "r"
  - We had fixed this to $r = 5\ events/time$ (and $r = 6\ events/time$ after the truck arrived)
  - We could also decide to make this a free measurement parameter!

- A second parameter could encode the fraction of nuclear fuel left after the truck arrived "f„
  - If $f = 1.0$ → no change
  - If $f > 1.0$ → fuel was added
  - If $f < 1.0$ → fuel was removed

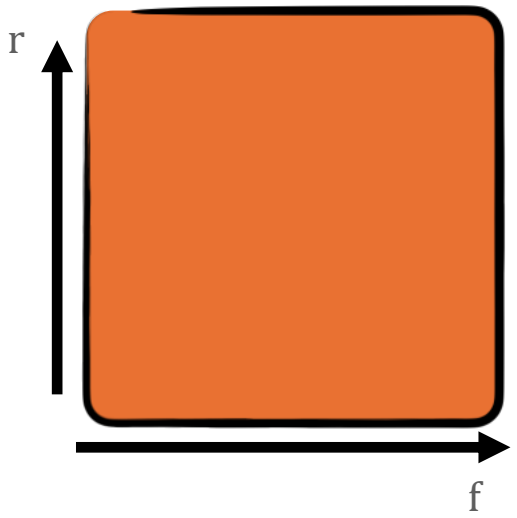
Nested Model

Take our experiment with a measurement of neutrinos before an after the truck arrived

$$p(n|r,f) = \mathrm{Pois}(n_{\text{before truck}}|r)\mathrm{Pois}(n_{\text{after truck}}|f \cdot r)$$

Parameter of Interest: fuel fraction f

Nuisance Parameter: reactor rate r



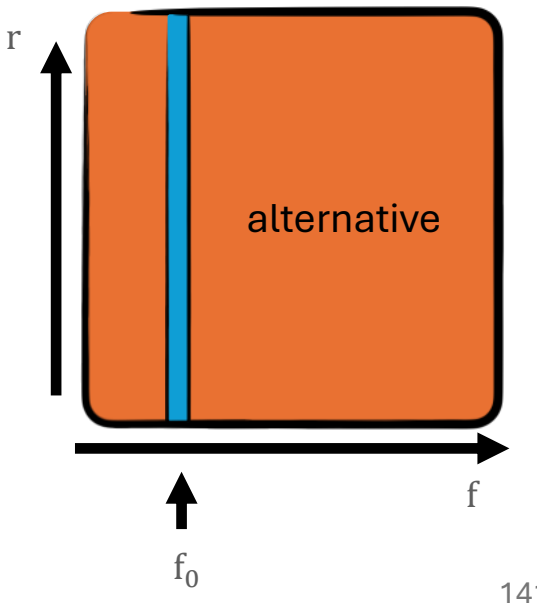
Nested Model

Take the counting experiment with a main measurement of signal and background a separate control measurement of the background

$$p(n|r,f) = \mathrm{Pois}(n_{\text{before truck}}|r)\mathrm{Pois}(n_{\text{after truck}}|f \cdot r)$$

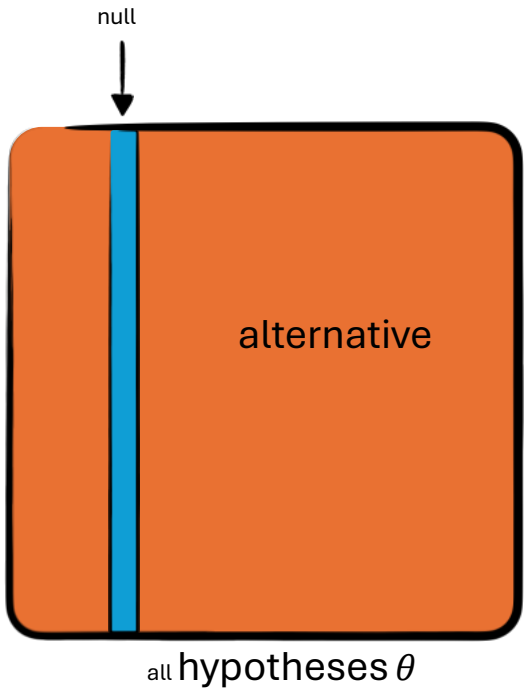
Null hypothesis: "No change in rate $f = f_0 = 1$"
(i.e. for any reactor rate r)

Alternative: "any other scenario"



So how do we do tests for complex Hypos?

For tests of complex hypotheses we can follow a simple strategy using our **point estimation toolbox**



So how do we to tests for complex Hypos?

For tests of complex hypotheses we can follow a simple strategy using our **point estimation toolbox**

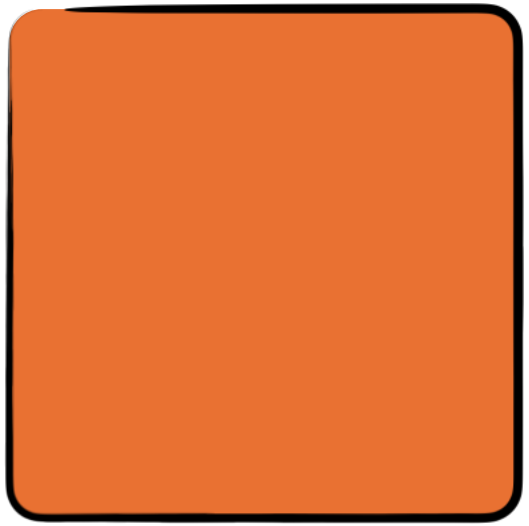
Idea: pick the"model" within the hypothesis set that **that is most compatible with the observed data** (i.e. the MLE within the set) as a "proxy" to measure compatibility of the set at large

$$\theta_{\text{proxy}}^{H_0} = \text{argsup}_{\theta \in H_0} p(x|\theta)$$

(i.e. $p(x|\theta_{\text{proxy}}^{H_0}) = \sup_{\theta \in H_0} p(x|\theta)$)

An Intuitive test for Complex Hypotheses

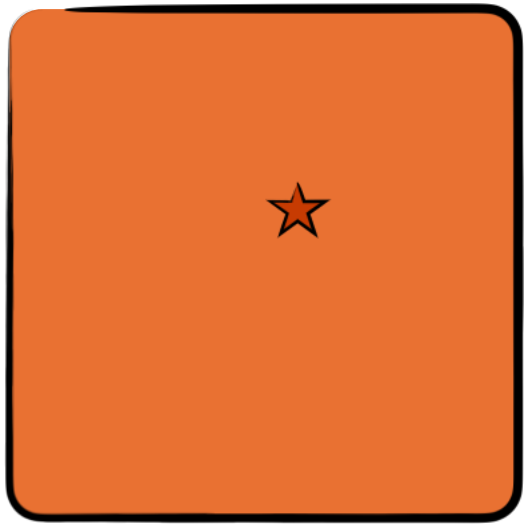
Take the alternative ("unrestricted") hypothesis H_1:



An Intuitive test for Complex Hypotheses

Take the alternative ("unrestricted") hypothesis H_1:

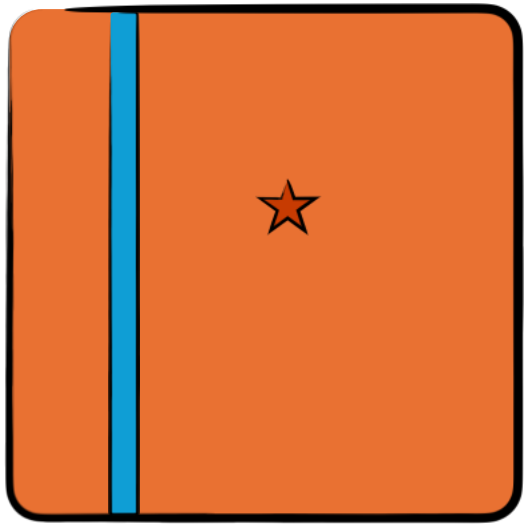
- the proxy model is the MLE model wrt to H_1



$$\theta^{H_1}_{\text{proxy}} = \underset{\theta \in H_1}{\text{argsup}} p(x|\theta)$$

An Intuitive test for Complex Hypotheses

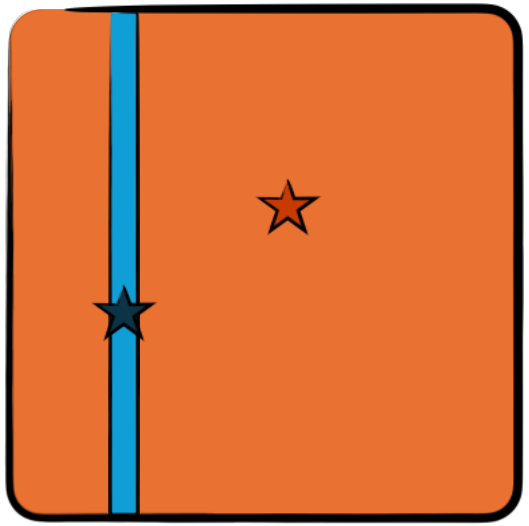
Take the null ("restricted") hypothesis H_0:



An Intuitive test for Complex Hypotheses

Take the null ("restricted") hypothesis H_0:

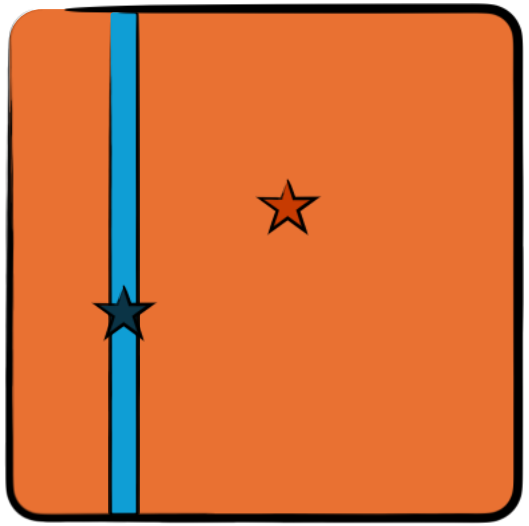
- the proxy model is the MLE among all possible values in H_0



$$\theta_{\text{proxy}}^{H_0} = \underset{\theta \in H_0}{\text{argsup}} p(x|\theta)$$

An Intuitive test for Complex Hypotheses

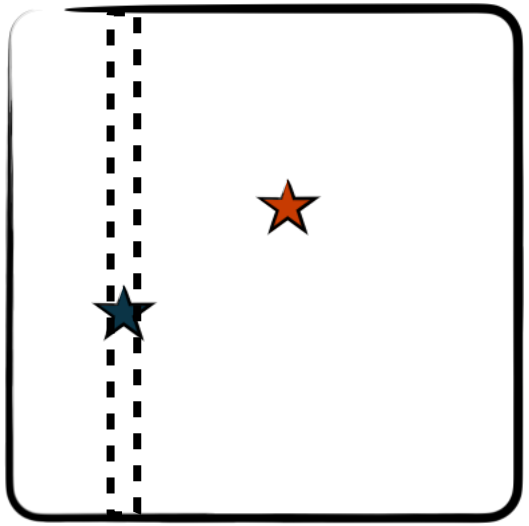
With the two "proxy models" we can **go back to the simple hypothesis test case**, where we know a good test statistic:



An Intuitive test for Complex Hypotheses

With the two "proxy models" we can go back to the simple hypothesis test case, where we know a good test statistic through N-P Lemmas:

The Likelihood Ratio



$$L = -2 \log \frac{p(x|\theta^{H_0}_{\text{proxy}})}{p(x|\theta^{H_1}_{\text{proxy}})}$$

Likelihood-Ratio Tests

Our intuitive procedure is known as Likelihood-Ratio Test (LRT) for composite hypotheses: a generalization of Neyman & Pearson's test

$$t(x) = -2\log\frac{p(x|\theta_0)}{p(x|\theta_1)}$$

simple Hypothesis case (Neyman-Pearson)

$$t(x) = -2\log\frac{\sup_{\theta\in H_0} p(x|\theta)}{\sup_{\theta\in H_1} p(x|\theta)}$$

composite Hypothesis case

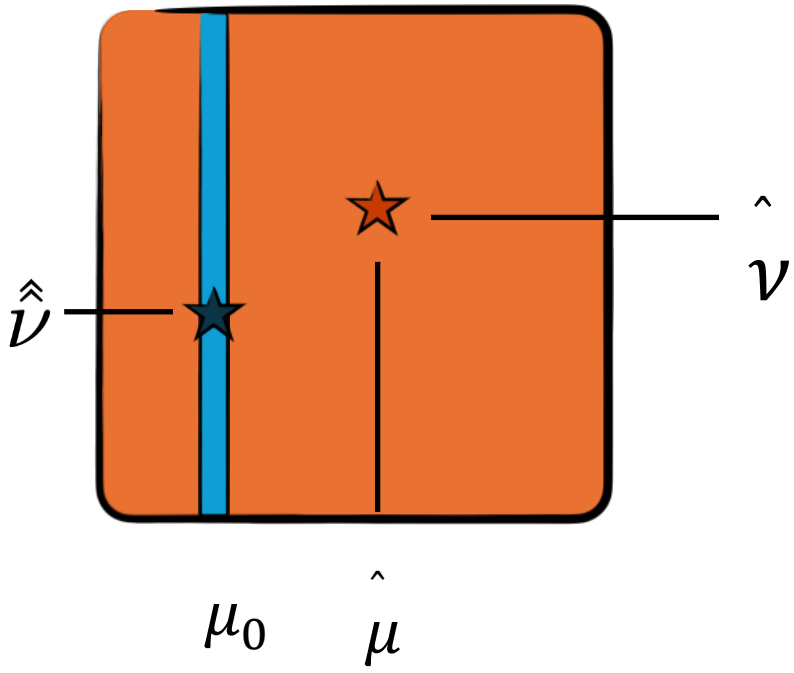
L'hood-Ratio Tests or nested Hypotheses

For the "nested hypotheses" case this is also called the **profile-likelihood ratio statistic**

optimal ν-values at fixed value of μ

$$\lambda_{\mu_0}(x) = -2\log\frac{L(\mu_0, \hat{\hat{\nu}})}{L(\hat{\mu}, \hat{\nu})}$$

globally optimal values



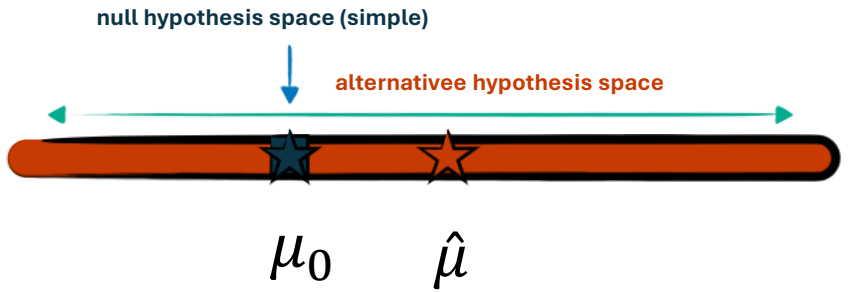
Model w/o nuisance parameters

The asymptotically best test statistic is just the likelihood function normalized to the MLE

To test a **specific null hypothesis H_0: μ_0 - a simple hypothesis**

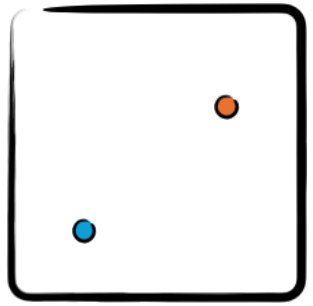
- compare its Likelihood value to the MLE value
  reject if it's too large

$$\lambda_{\mu_0}(x) = -2\log\frac{L(\mu_0)}{L(\hat{\mu})}$$



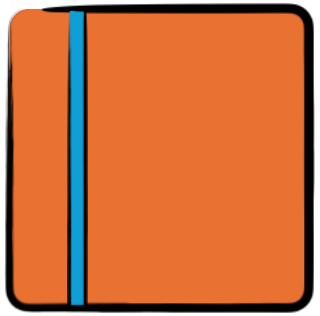
Optimality of LRT

The LRT is asymptotically optimal in a number of ways, e.g. it has **best average power for tests** for the nested hypotheses :



$$t(x) = -2\log\frac{p(x|\theta_0)}{p(x|\theta_1)}$$

Likehood Ratio is always optimal



$$\lambda_{\mu_0}(x) = -2\log\frac{L(\mu_0, \hat{\hat{\nu}})}{L(\hat{\mu}, \hat{\nu})}$$

LRT is asymptotically optimal*

Sampling Distributions of LRT

Sampling Distribution of LRT

One of the biggest advatages of the LRT: it's not only intuitive & optimal but also its sampling distribution are asymptotically known for nested hypotheses!

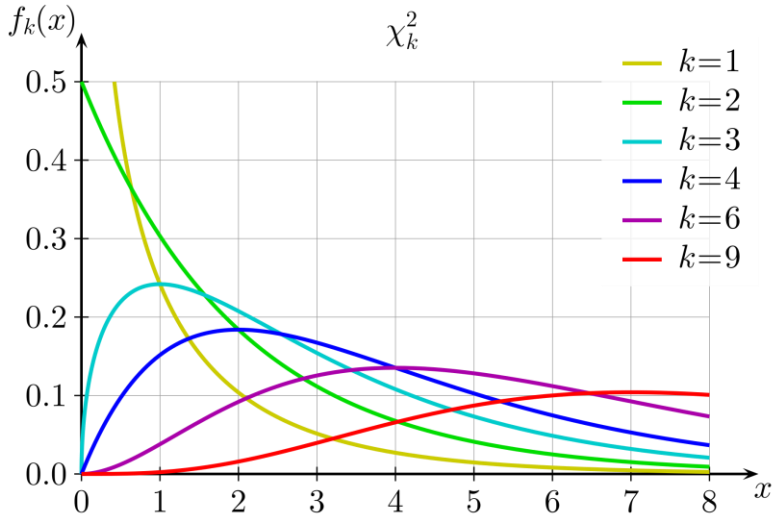
The LRT statistic for r parameters of interest is follows asymptotically

the non-central χ^2 distribution with r degrees of freedom

$$p(t_\mu|\mu') = \chi^2_{\mathrm{nc}}(n, \Lambda^2_{\mu'})$$

Wilk‘s theorem

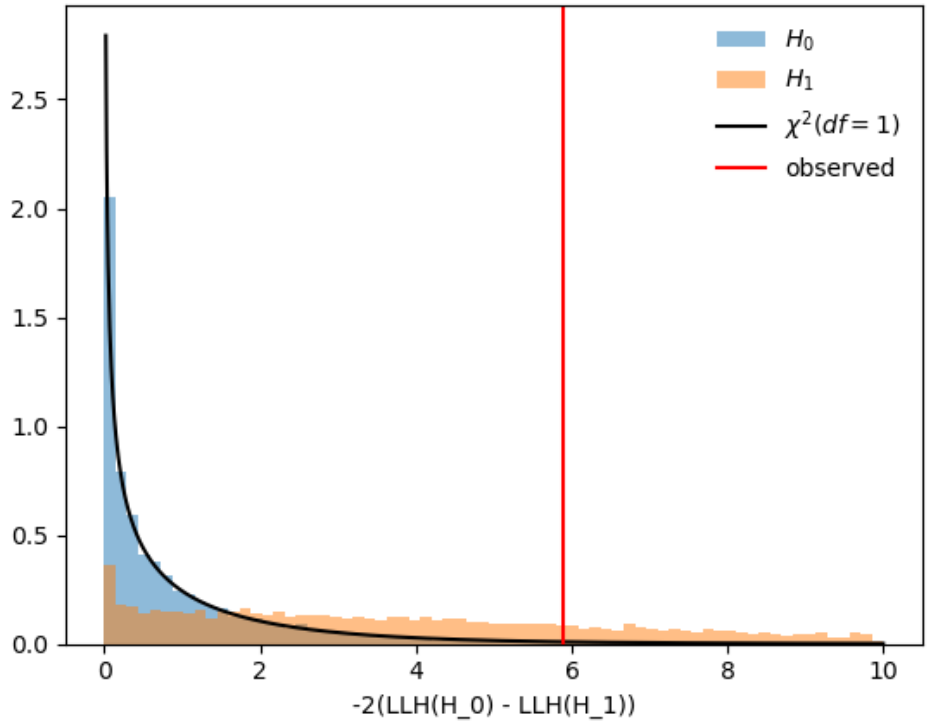
- The LRT test statistic ($-2\Delta LLH$) under the **null hypothesis** is distributed as the (central) chi2 distribution!

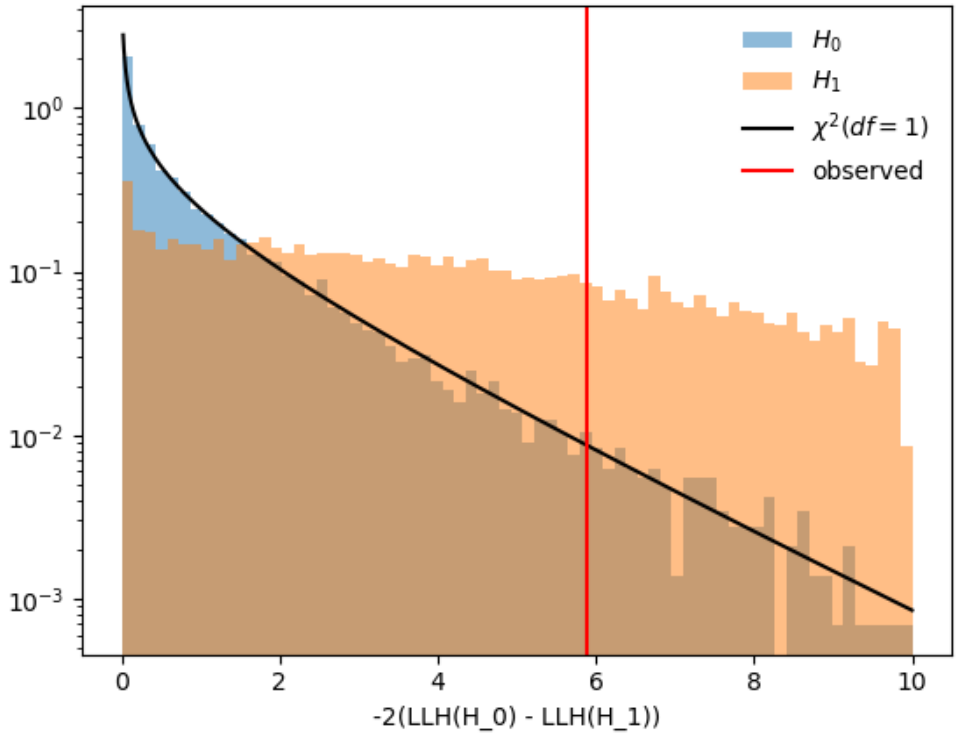


- The degrees of freedom (df) is the difference in free parameters between H0 and H1 (In our case df=1)

Sampling distributions in our Example

- P-value from trials: 1.59%
- P-value from chi2: 1.52 %





Summary

- Hypothesis testing is one of the main workhorses in frequentist analysis
- idea is to check compatibility of data with (a set) of models
- done via looking at the data in a way that differentiates models through **test statistics** and comparing observed data (p-values) to **predicted distribution** of data of the hypotheses
  - In general: need to use MC to find sampling distribution shape
- **Likelihood-ratio-based tests** often are the most powerful way to perform such a test
- **Neyman-Pearson**: LRT is uniformly most powerful for simple hypos
- Asymptotically: we can derive exact sampling distributions
- Wilk's Theorem: distribution for null is χ^2

Confidence Level Intervals

Recap: Point Estimators

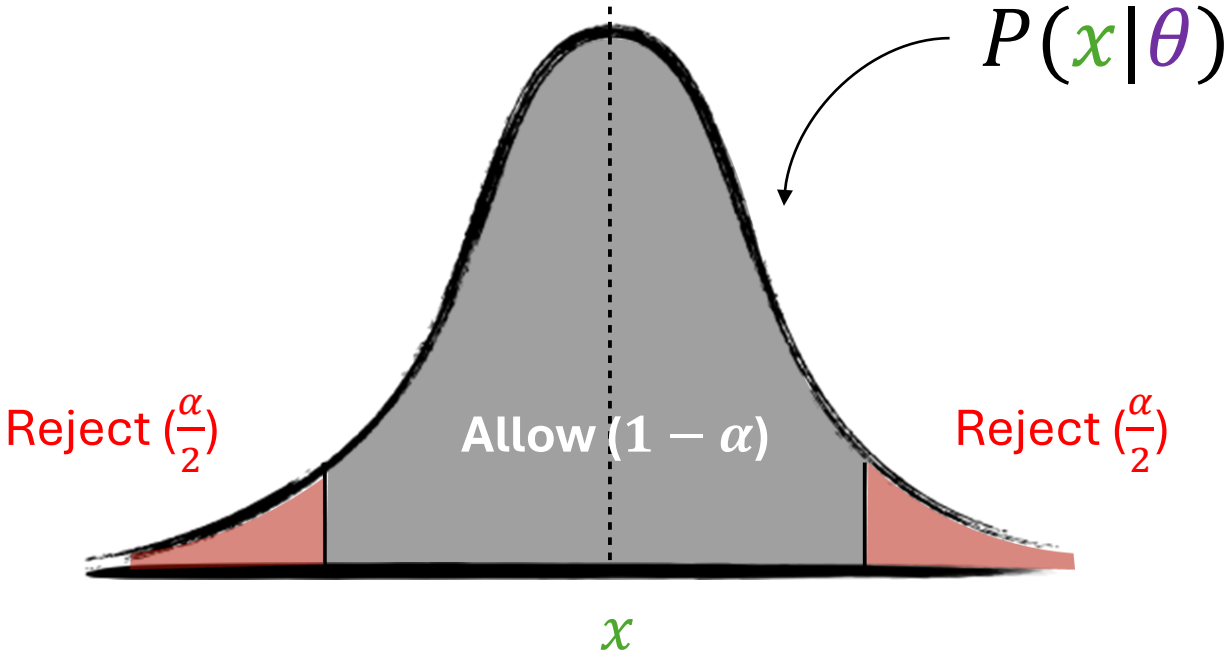
- Remember:
  - The principle of maximum likelihood gives us desirable estimators
    → Asymptotically efficient and unbiased
- For example in the case of a Gaussian:
  - **Sample mean** as the MLE of the location $\boldsymbol{\mu}$ of the distribution
  - **Sample standard deviation** as the MLE of the scale $\boldsymbol{\sigma}$ of the distribution

- But often, we want to make some statements about "uncertainty"
  - However, in the frequentist picture, there exists no concept of a probability distribution $p(\theta)$!
  - Instead, there is just one true value θ_{true}

→ How can we then build meaningful intervals?

Start with intervals in observed x

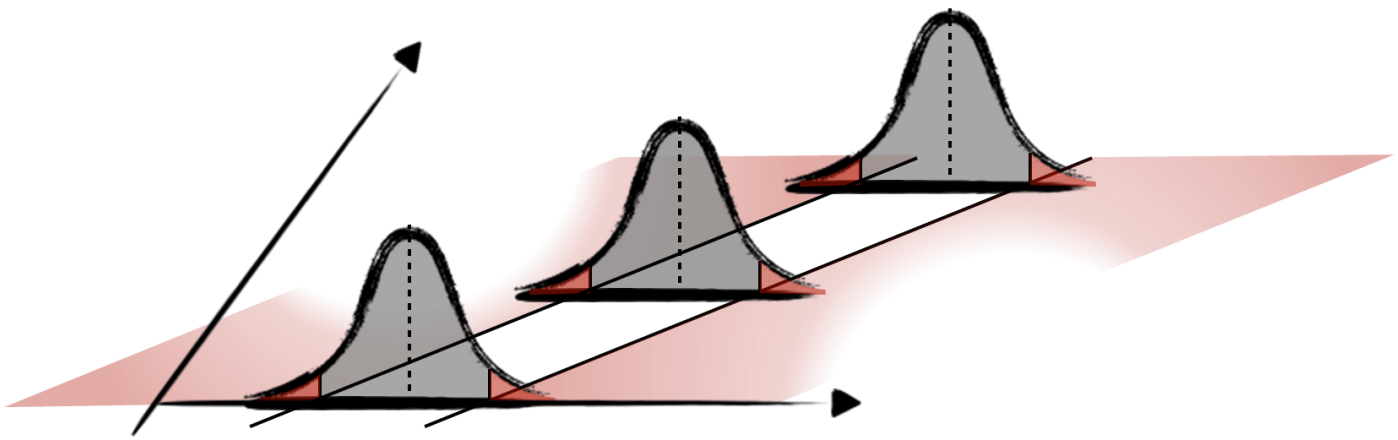
- We can build intervals of x for fixed parameters θ
- E.g. two-sided interval with $\alpha = 0.1$

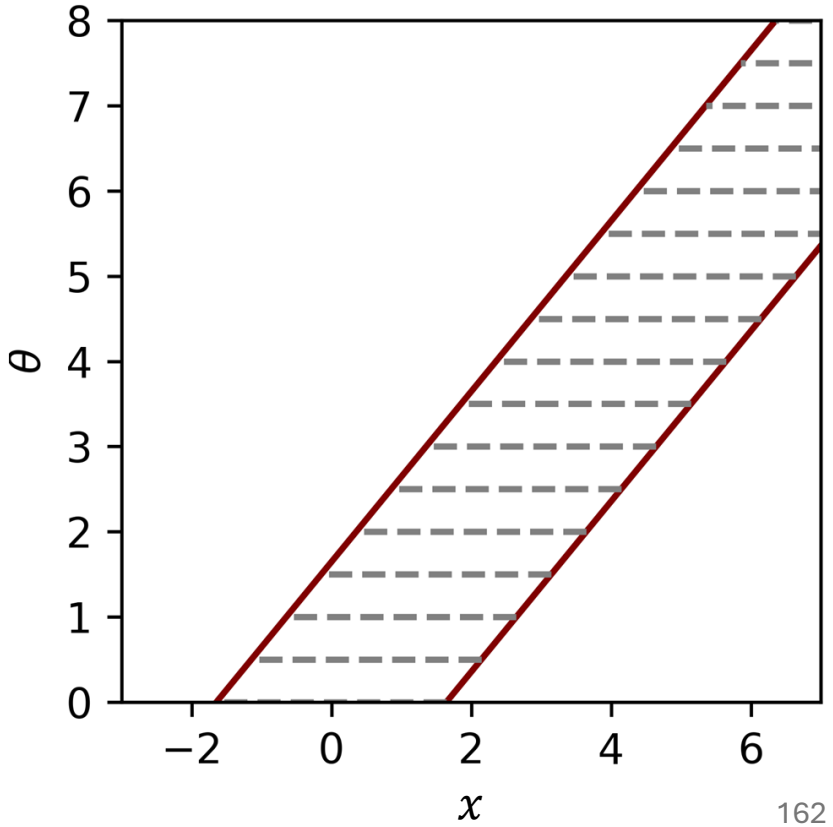


Band plot walk-through: Step 1

- We can construct such intervals in x for any choice of θ!

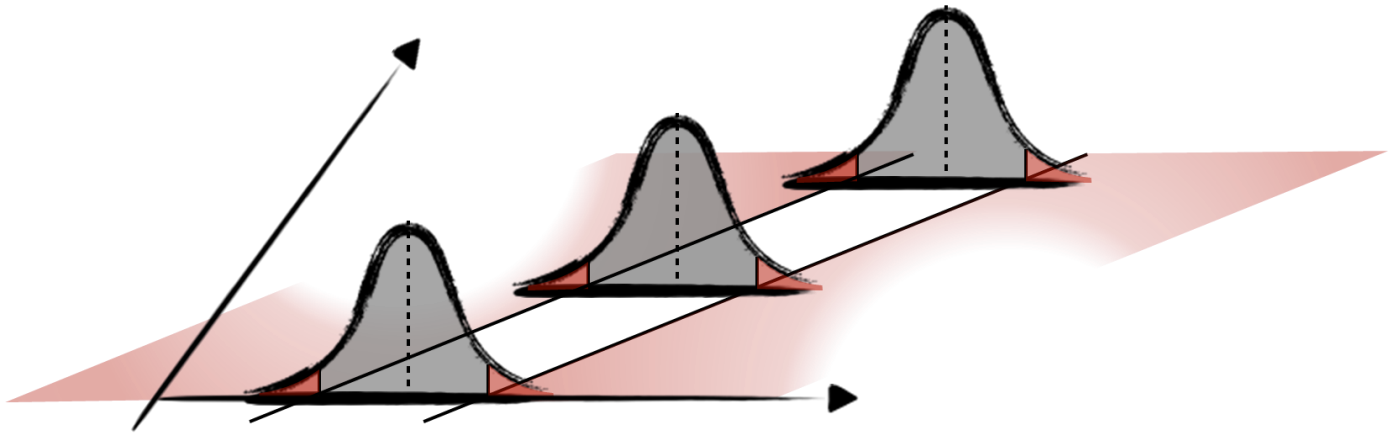
→ This is called the “Neyman” band plot

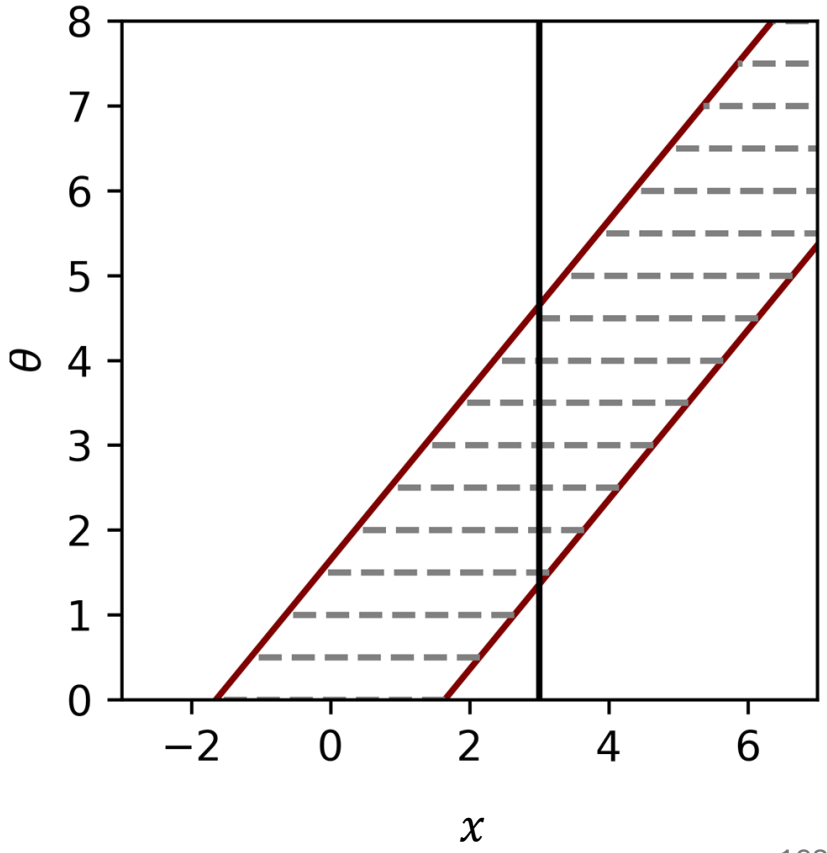




Band plot walk-through: Step 2

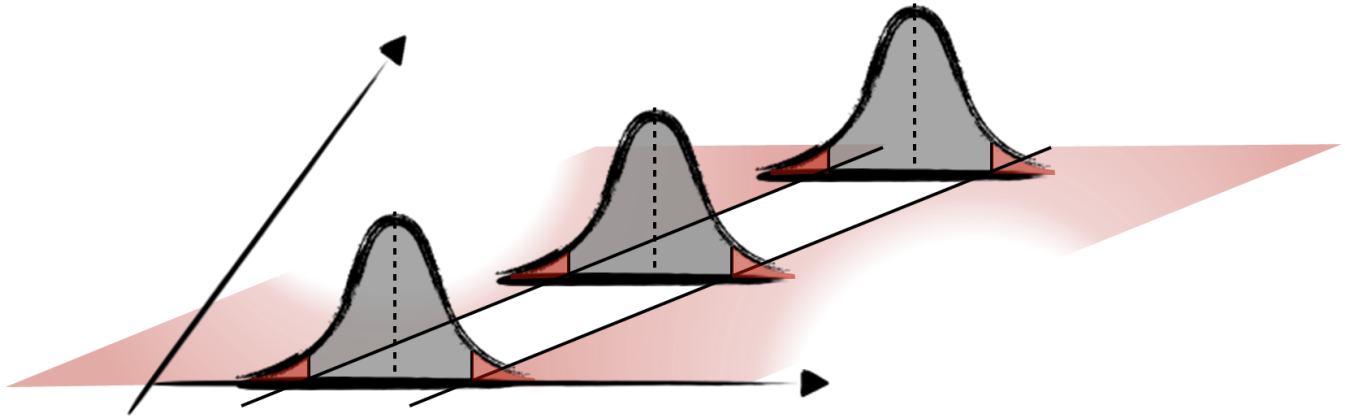
- Now fix the observed value at your measured x

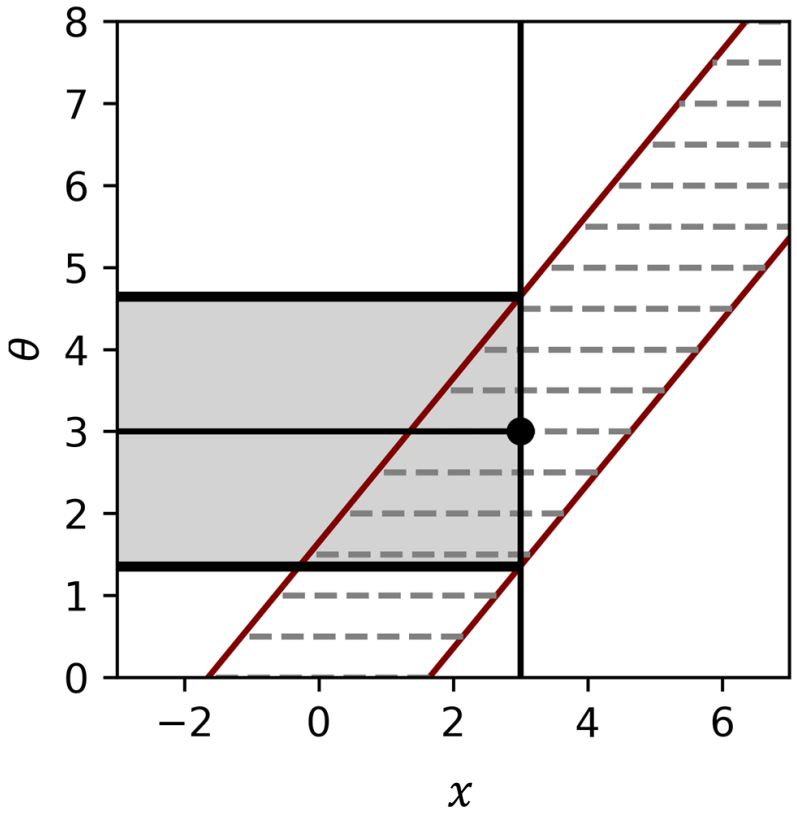




Band plot walk-through: Step 3

- Read of corresponding interval in θ





Exercise

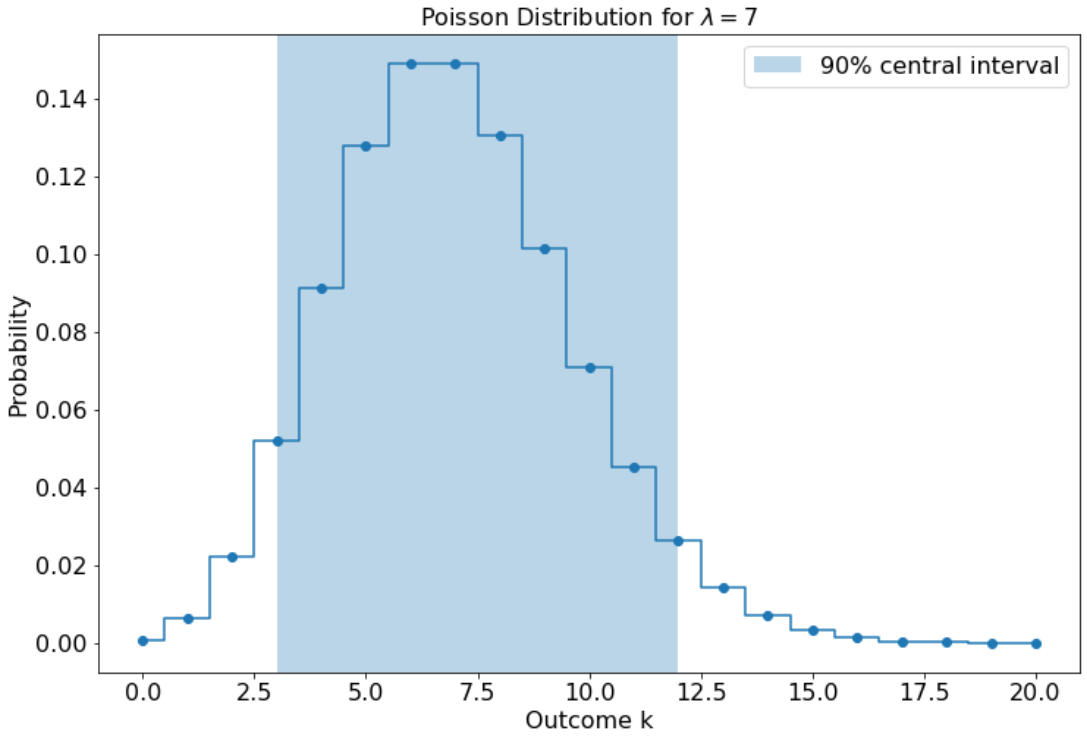
Neyman Construction

Construct the 90% Neyman Band for a Poisson distribution

Get the C.L interval for an observed x=7

Interval for fixed parameter

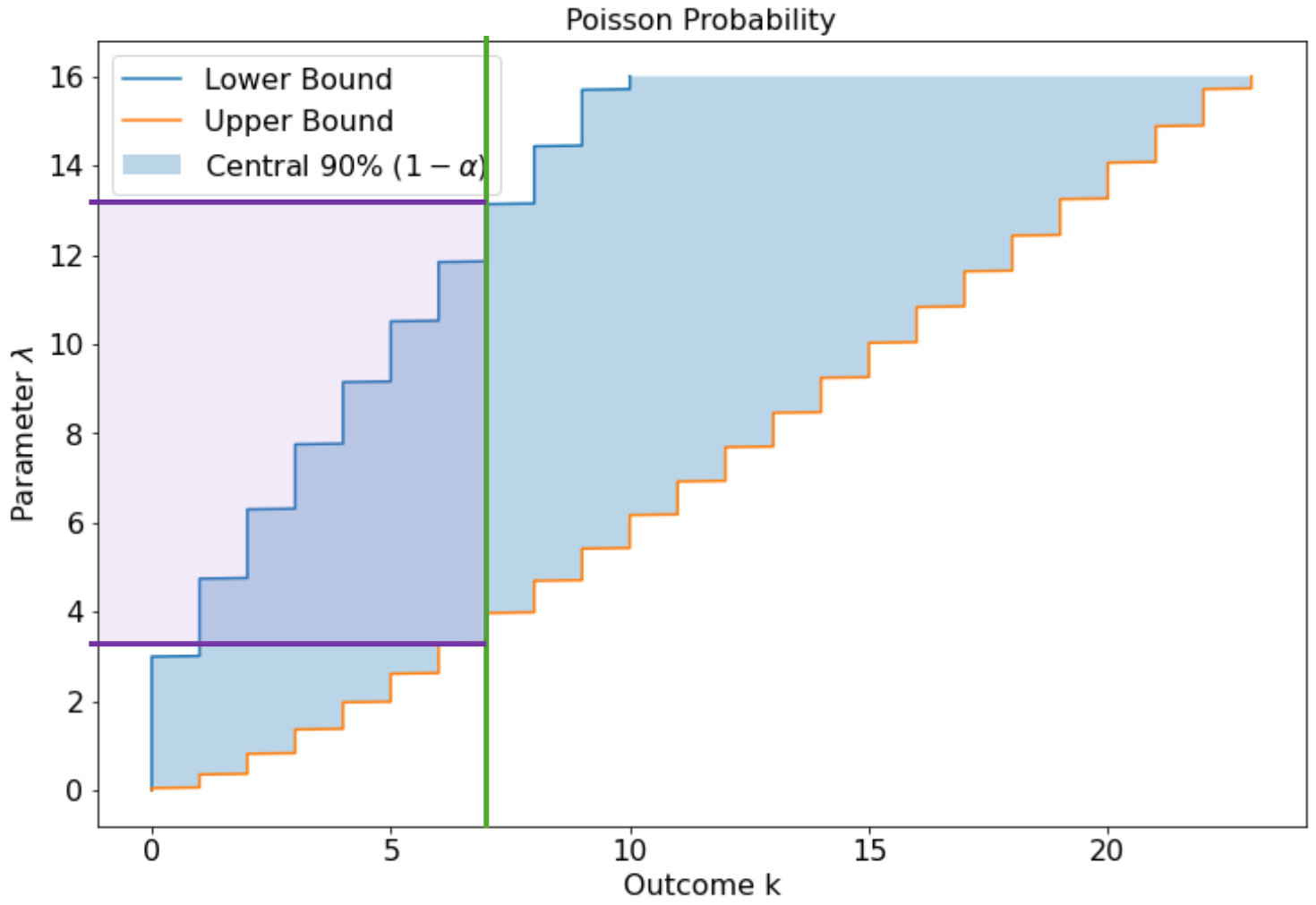
- Let's suppose we perform a counting experiment and we measure an outcome of $k = 7$
- We know for a fixed λ the range of possible outcomes k
  - → i.e., the probability distribution
  - And we know how to construct intervals that contain $1 - \alpha$ of the probability mass



Neyman Construction

- Poisson band plot:
  - Construct the interval in k for all possible λ
    Here we chose $\alpha = 0.1$
  - Fix the random variable to our measured value $k = 7$
  - Read off the interval

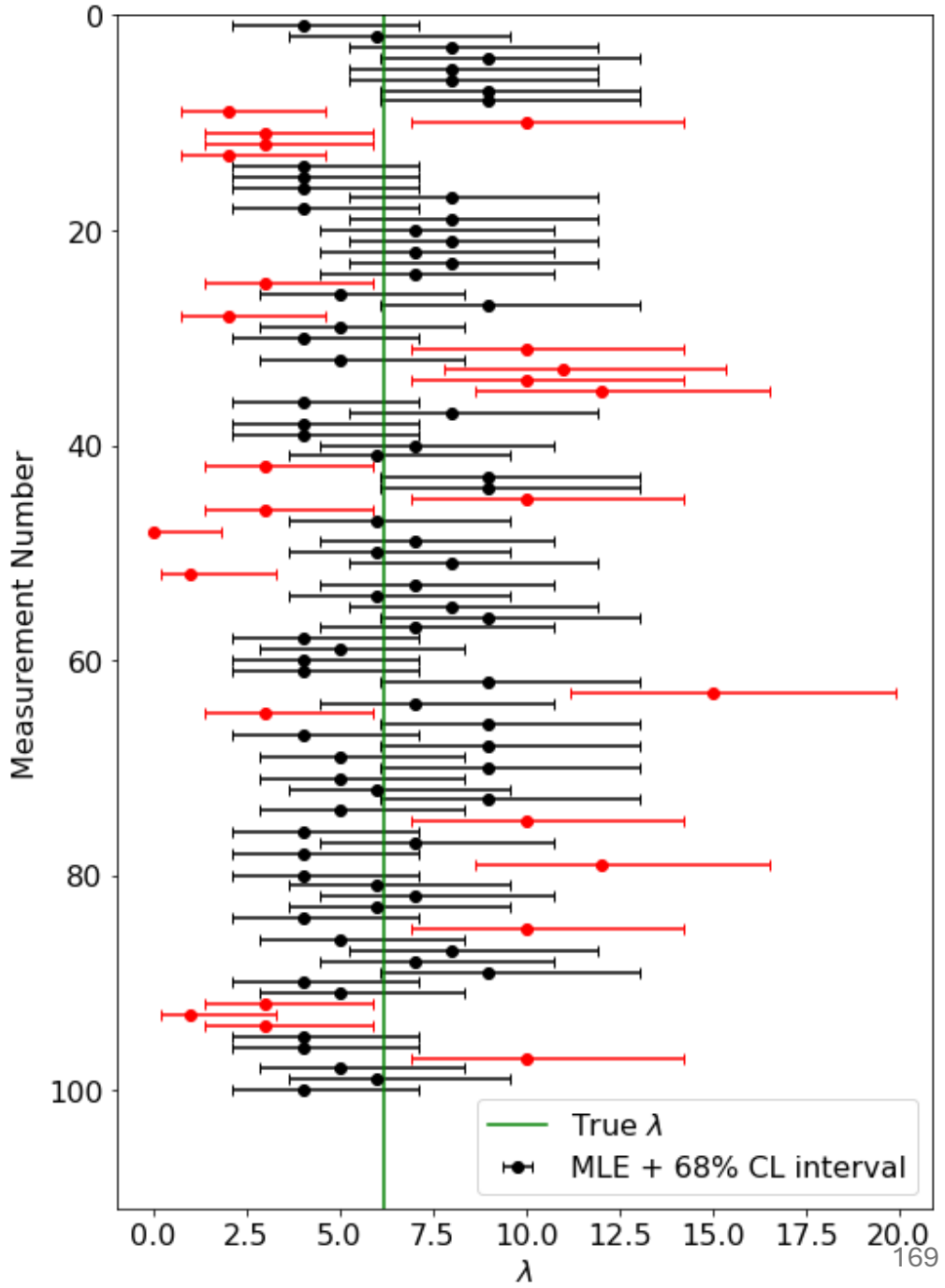
$\lambda \in [3.3, 13.1]$ @ 90% C.L.



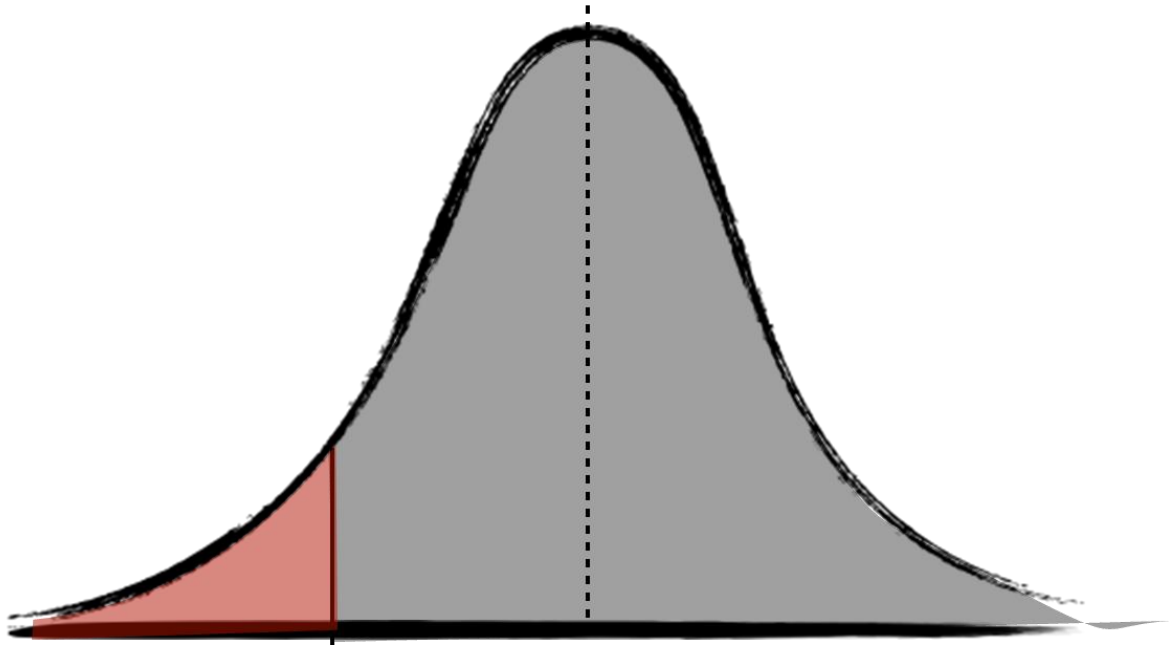
Interpretation:

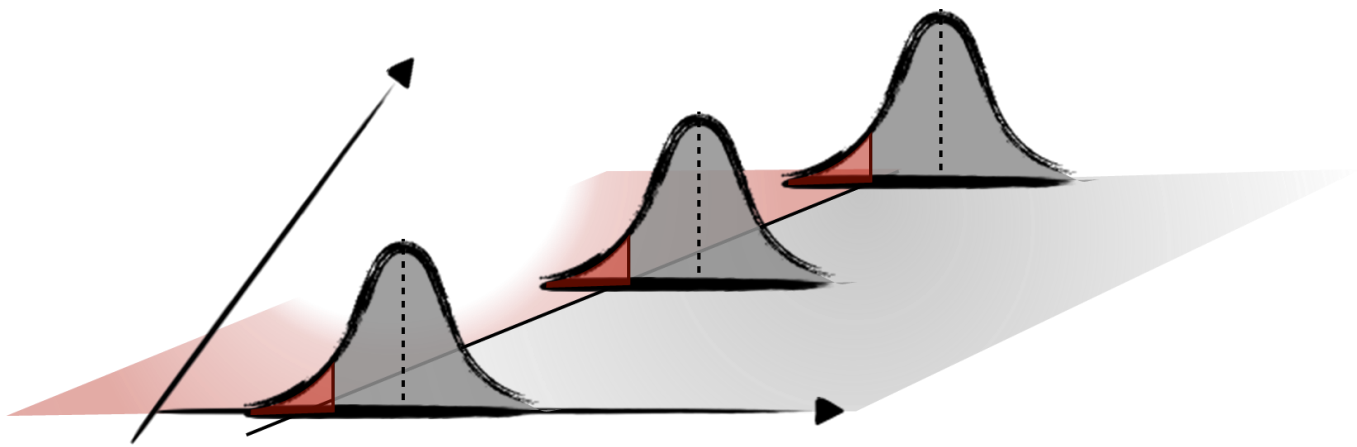
In repeated experiments, the CL interval contains the true value at least $1 - \alpha$ of times

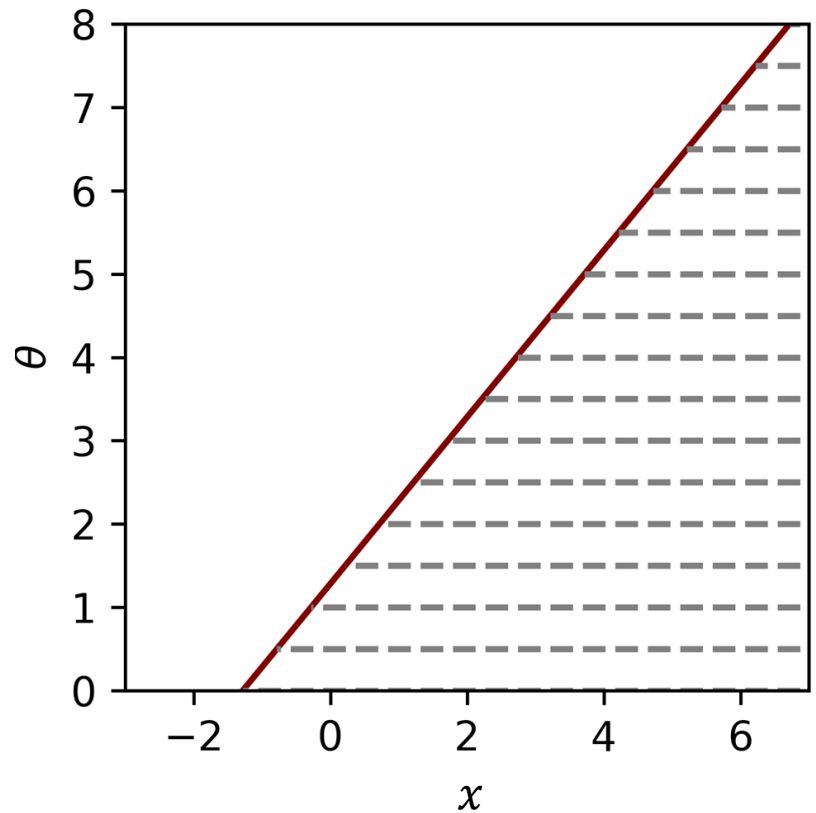
- Here: 100 Poisson experiments with $\lambda = 6.2$
- Slight *overcoverage*
  - Actually only 25% measurements fall outside
  - Typical problem for discrete distributions

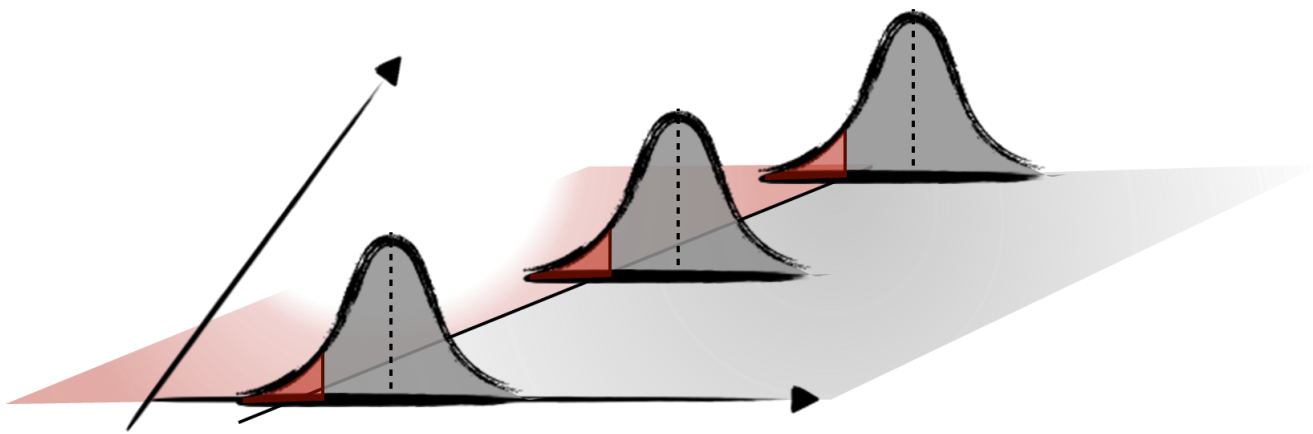


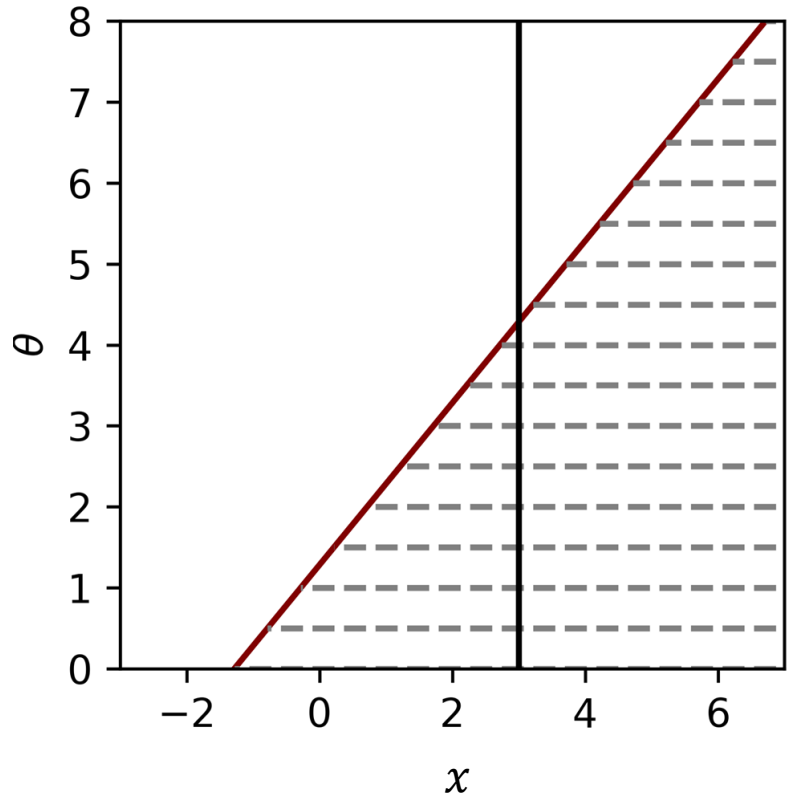
One-sided Intervals

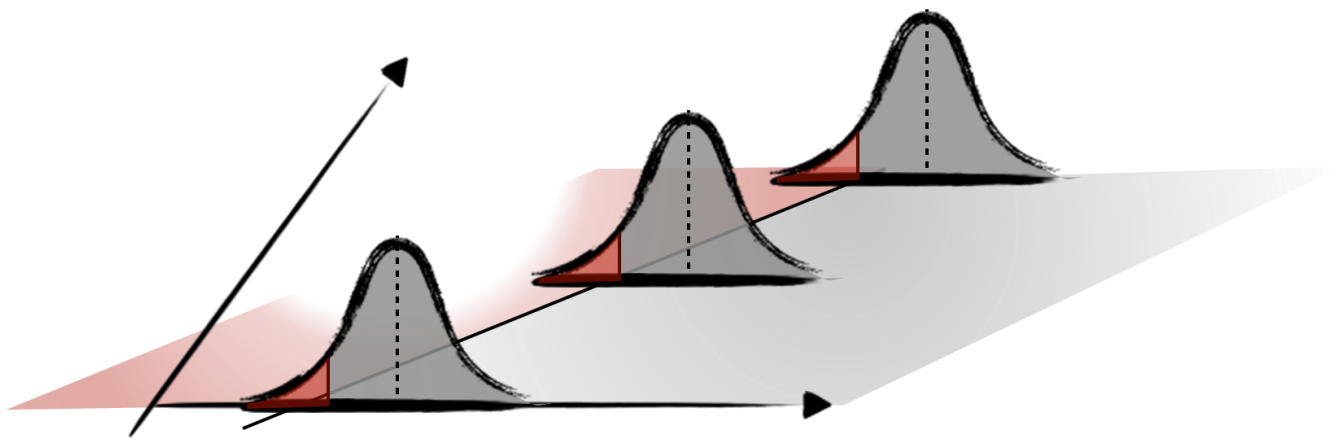


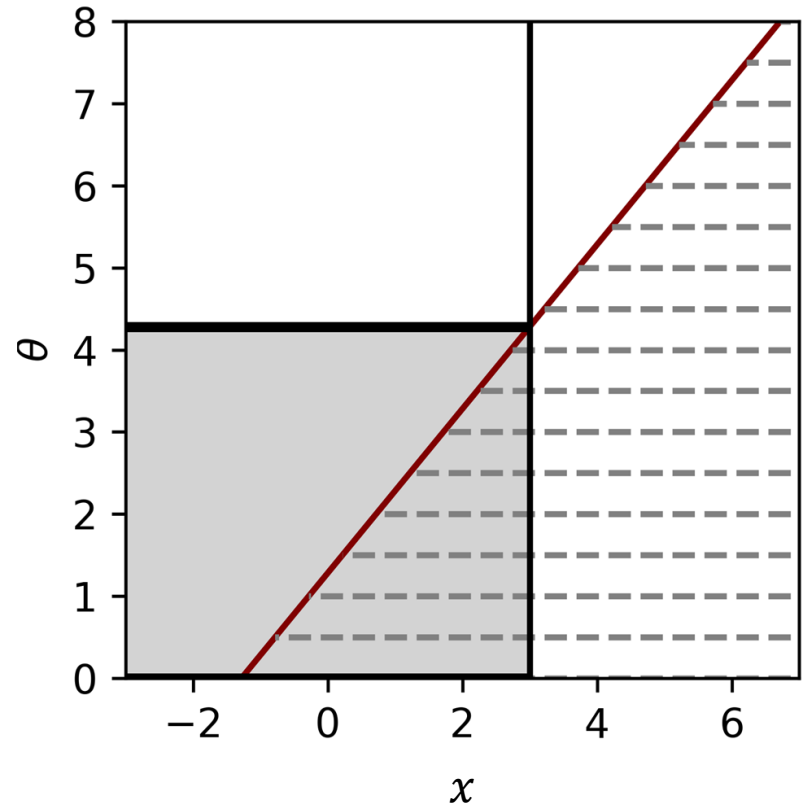



8
7
6
5
4
3
2
1
0
θ
−2
0
2
4
6
x




8
7
6
5
4
3
2
1
0
θ
−2
0
2
4
6
x



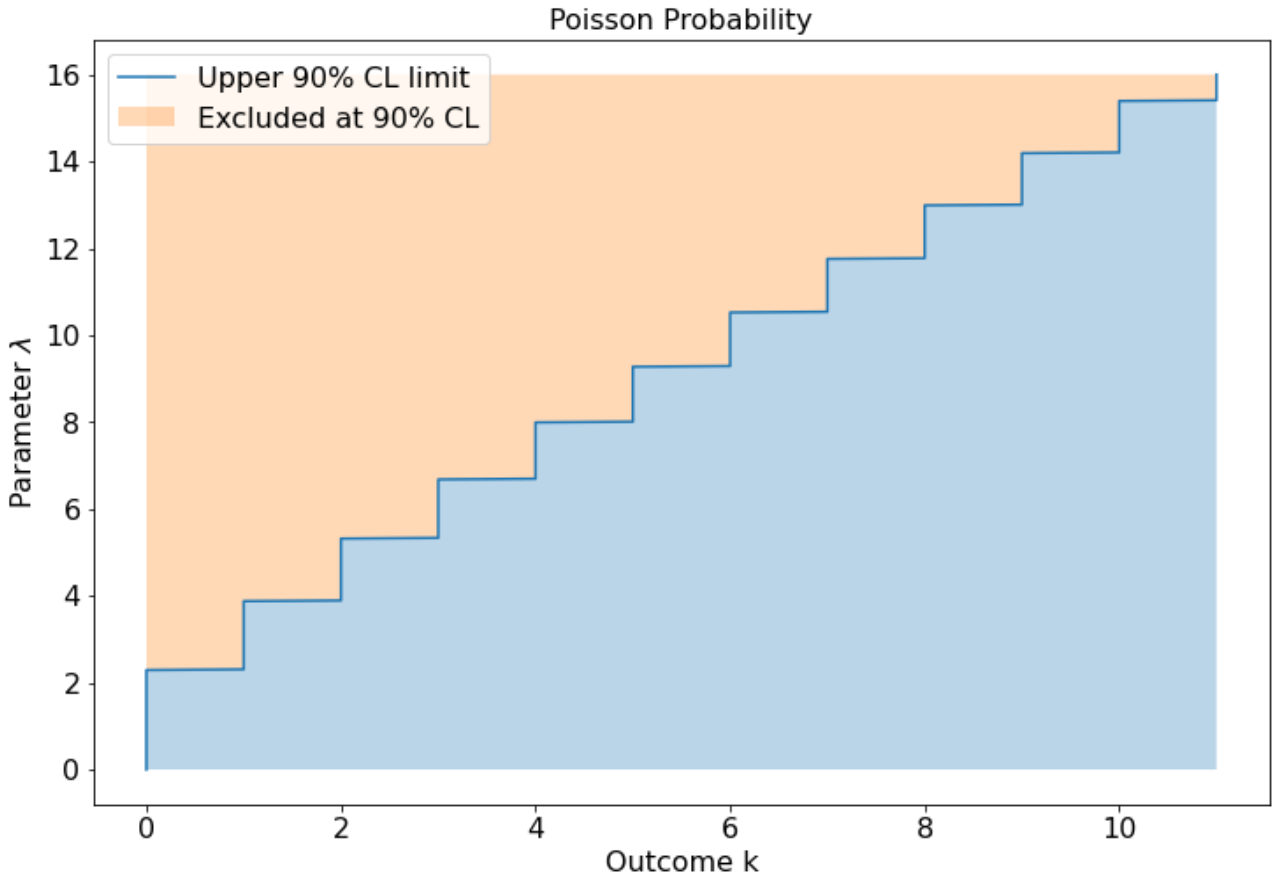

8
7
6
5
4
3
2
1
0
θ
−2
0
2
4
6
x

Example: Poisson

- Upper limit: maximum value of λ for which the true value is smaller at least $1 - \alpha$ of times

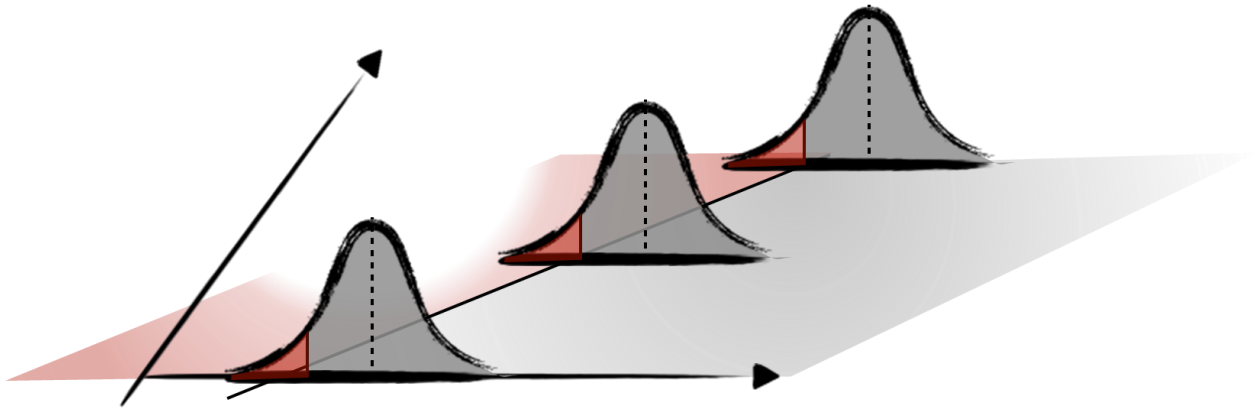
- For $k = 7$: $\lambda < 11.7$ @ 90% C.L.

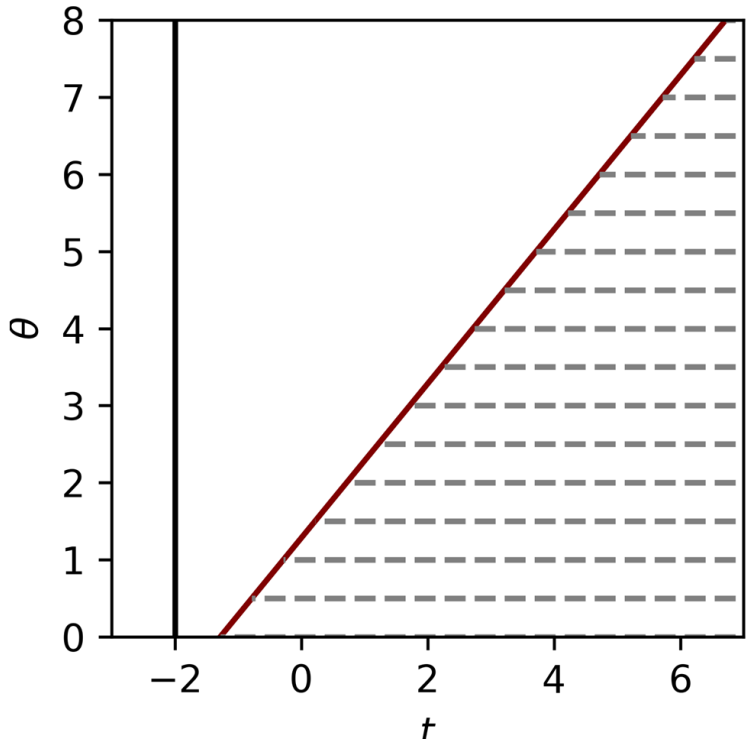
- Lower limit: analogous



Empty intervals..?

- What happens if we observed x = -2?





C.L. intervals **do not** give you the probability that a certain parameter value is true!

Interpretation

Meaning of Frequentist Intervals

Interpretation of frequentist intervals is different than Bayesian intervals

What does $\theta < \theta_{\text{up}}$ or $\theta \in [\theta_-, \theta_+]$ mean here?

Bayesian: represents **belief** of what the value of the parameter is.

(i.e. "given the data & my priors I believe θ to be within $[\theta_-, \theta_+]$)

Frequentist: a **summary of the obs. data** "in the language of the model"

"my hypothesis tests deem $\theta \in [\theta_-, \theta_+]$ compatible w/ observed data"

Randomness of Intervals

Intervals are random - same argument as for Bayesian Intervals:

- the computed intervals $I_\theta^\alpha(x)$ are **"random objects"** because they are derived from random data x
- they may or may not include the **true value** of the data source

Coverage

For Bayesian analysis coverage is usually not focus of attention.

- more about modeling your beliefs
- less focus on relation of your belief to a possible "true value"

But in in Frequentist analysis coverage is **taken very seriously**

- it is **the** main defining property of an interval estimation method
- partly due to focus is on repeated experimentation where you may have get many intervals & long-run frequencies matter

Coverage for Freqentist Intervals

Often we do tests with test size of $\alpha = 0.05$.

→ i.e. probability of Type-I error (falsely rejecting H_0) is 5%

Implies: if we use $\alpha = 0.05$ in our construction

→ the probability of the intervals to cover the true value is 95%
referred to as "95% confidence level intervals" (95% C.L.)

Confidence in what?

95% Confidence Level Interval: **what exactly are we confident about?**

- Not a statement of confidence what the true value of θ is
- Rather confidence in whether intervals we construct are covering

Compare to "Credible Intervals" (C.I.) (Bayesian)

- Actually is a statement about your belief regarding the true value of θ

Correct Statements for Intervals

Bayesian Intervals (credible intervals):

"Given the data I believe there is a 95% probability that $\theta \in [\theta_-, \theta_+]$"

- Coverage is usually not analyzed

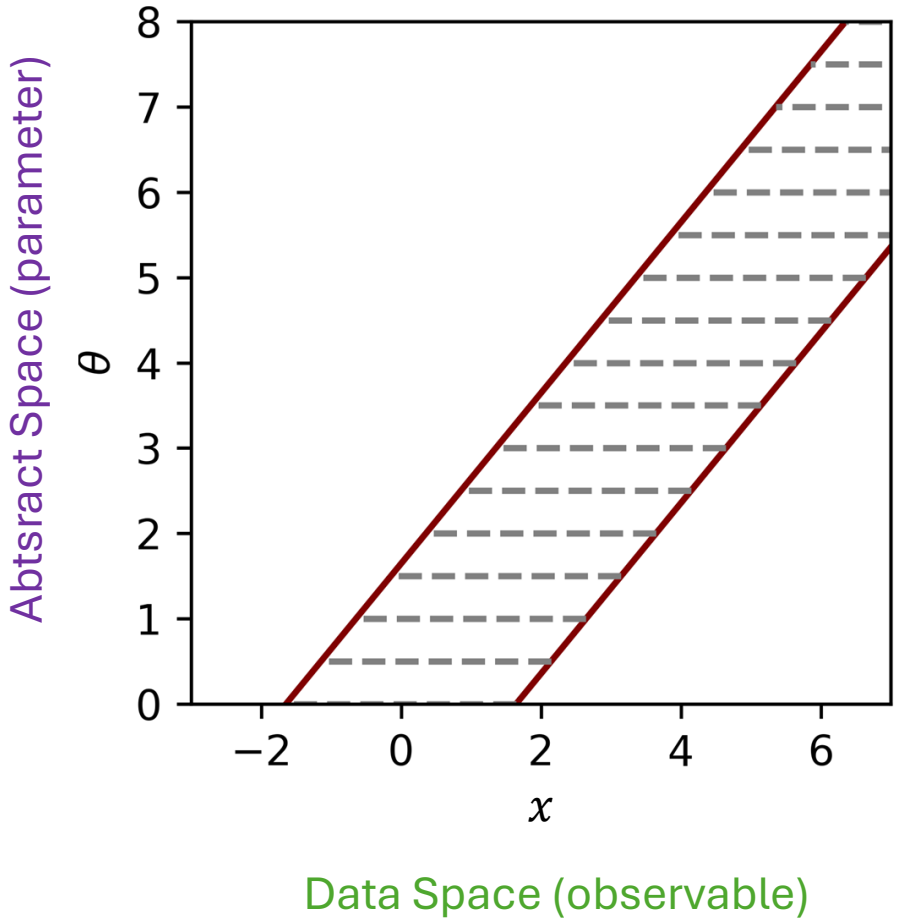
Frequentist Intervals (confidence intervals):

"Given the data all $\theta \in [\theta_-, \theta_+]$ are not rejected by a test of size 0.05"

- Coverage is 95% by definition: Intervals under repeated experiments will include true value of θ 95% of the time

Recap

- Neyman Band Construction
  - Allows us to draw intervals in θ containing the true value at least $(1-\alpha)$ times in repeated trials (= frequency)
  - Construction:
    - For fixed values of θ construct intervals in x
    - Fix x at observed value
    - Read off intervals in θ

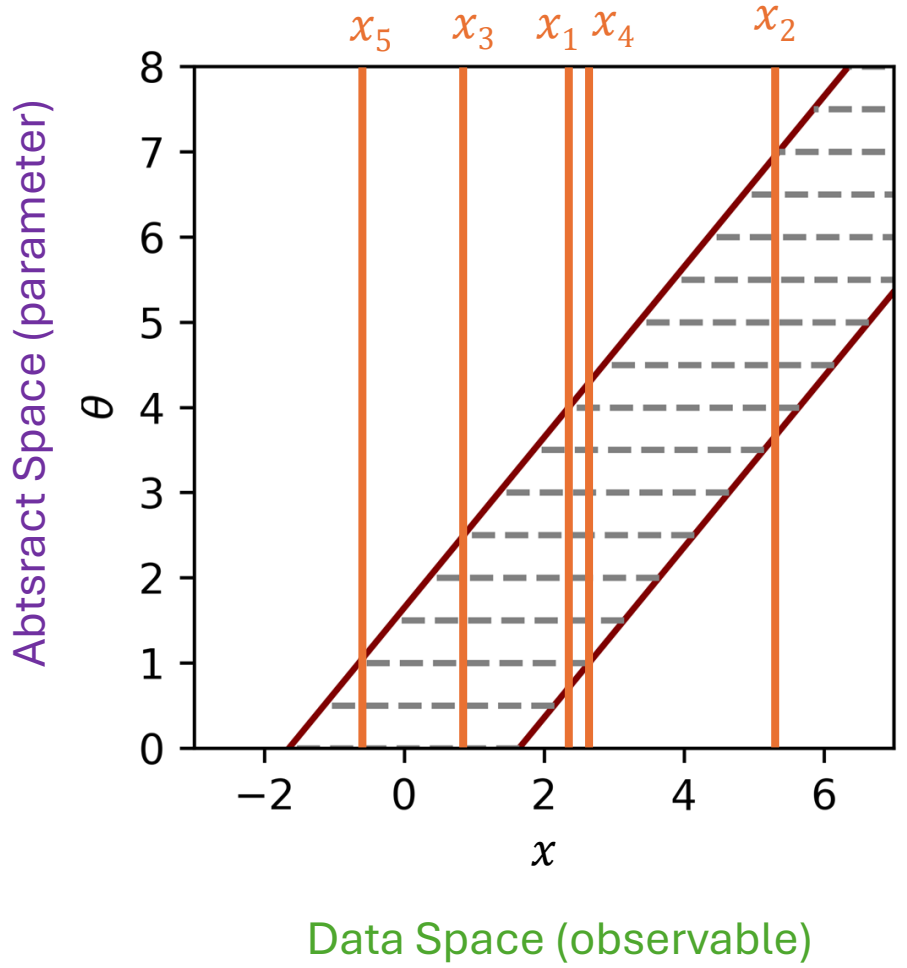


Beyond Simple Models

What if I have more data…?

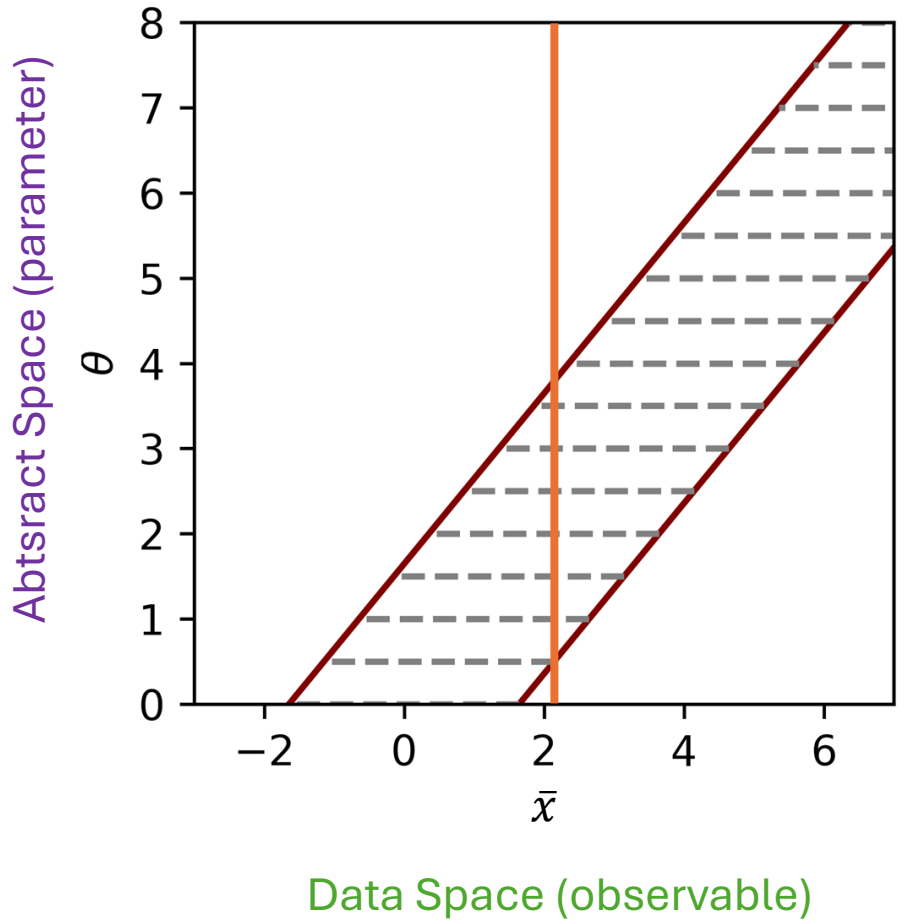
- Let's say I measure n examples from a Gaussian distribution $\{x_1, x_2, \ldots, xn\}$
- I get n different intervals $I_1, I_2, \ldots, In$

- But what you really want is combined answer
  - We'll encounter a "natural" way how such multiple data can be combined in the Bayesian case next year ("update of knowledge")



Alternative 1

- We could combine our data into a single value:
  - For the Gaussian case, maybe a good choice would be the sample mean $\bar{x} = \frac{1}{n}\sum_i x_i$
- This is still an observable in the data space ☺
- Follow the exact same procedure to construct corresponding Neyman Bands ☺
- We need to define a good quantity by hand ☹



Alternative 2

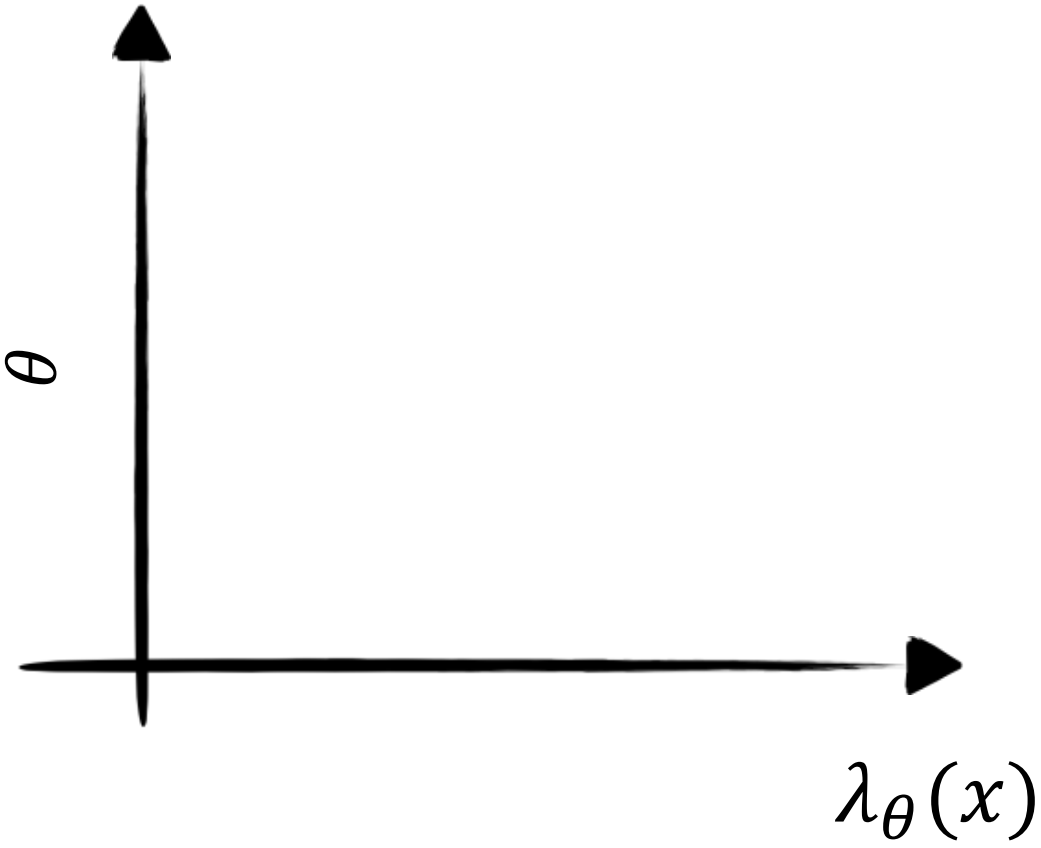
- In the discussion of point estimators, we found that the maximum **likelihood** estimator (MLE) is a desirable way how to summarize our observations
- In the discussion of hypothesis tests, we found that **likelihood**-ratio tests (LRT) have several desirable properties (e.g., Neyman-Pearson Lemma says LRT is most powerful TS)
- → Can we use the **likelihood** also for intervals..?

LRT as TS for Neyman Band

- LRT test statistic:

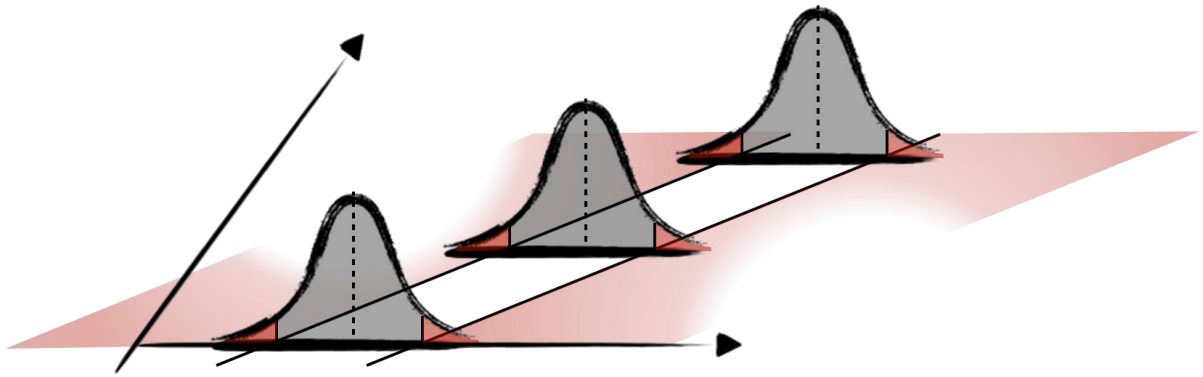
$$t_\theta = \lambda_\theta(x) = -2 \log \frac{p(x|\theta)}{p(x|\hat{\theta})}$$

- Instead of drawing intervals in x for every θ we construct intervals in t_θ for every θ

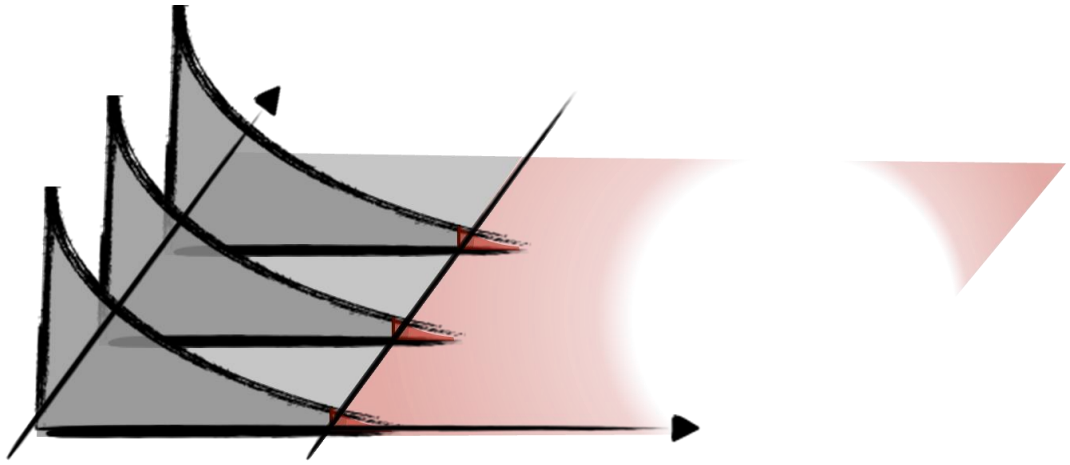


What do the Rejection Regions look like?

- When taking the LRT ratio test, the null hypothesis distribution is always the same, regardless of θ:
  → the χ^2 distribution (Wilk's Theorem)



Intervals in x

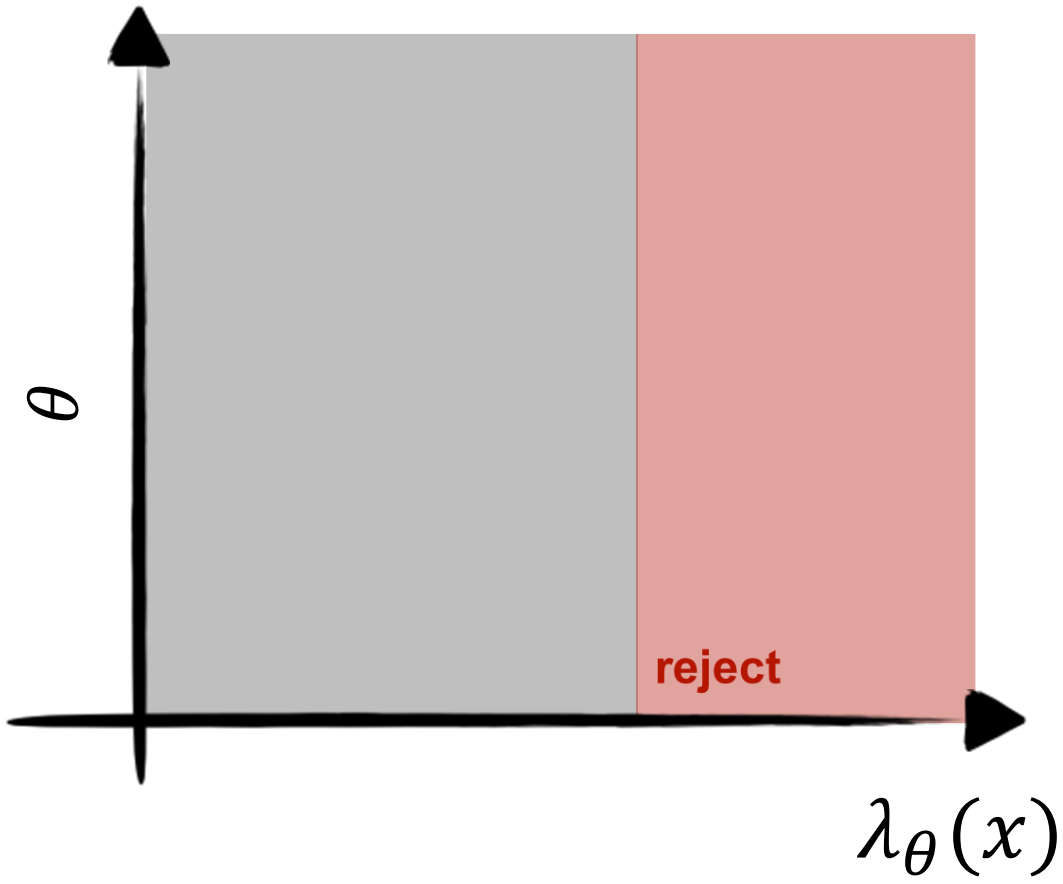


Intervals in $\lambda_\theta(x)$

What do the Rejection Regions look like?

- If null distribution is the same, the rejection region is the same for all parameter values.

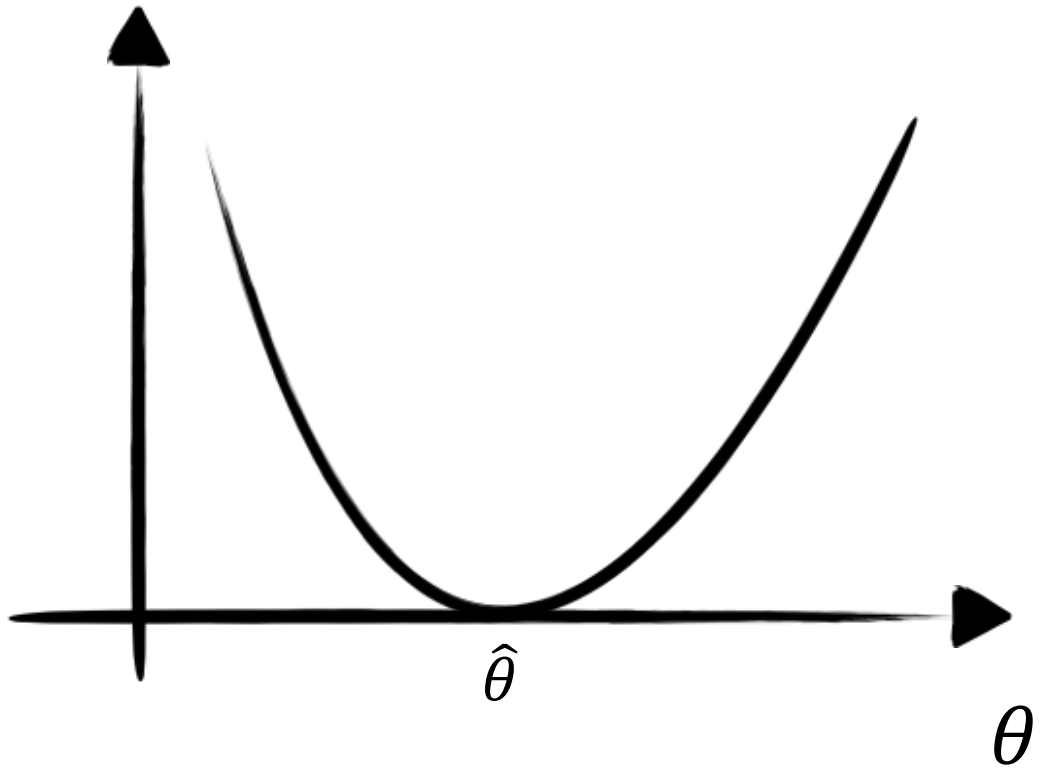
- And we know from hypothesis testing already that the distribution follows a $\chi 2$



How does the observed look like?

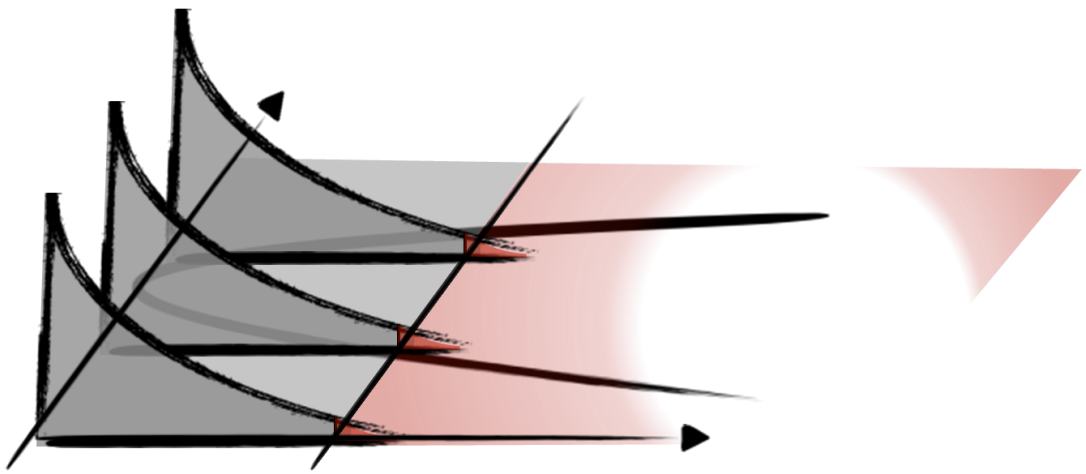
- Now we switched from x to $\lambda_\theta(x)$
  - This itself is a function of θ!
- From Wald/Wilk we know its asymptotic form:

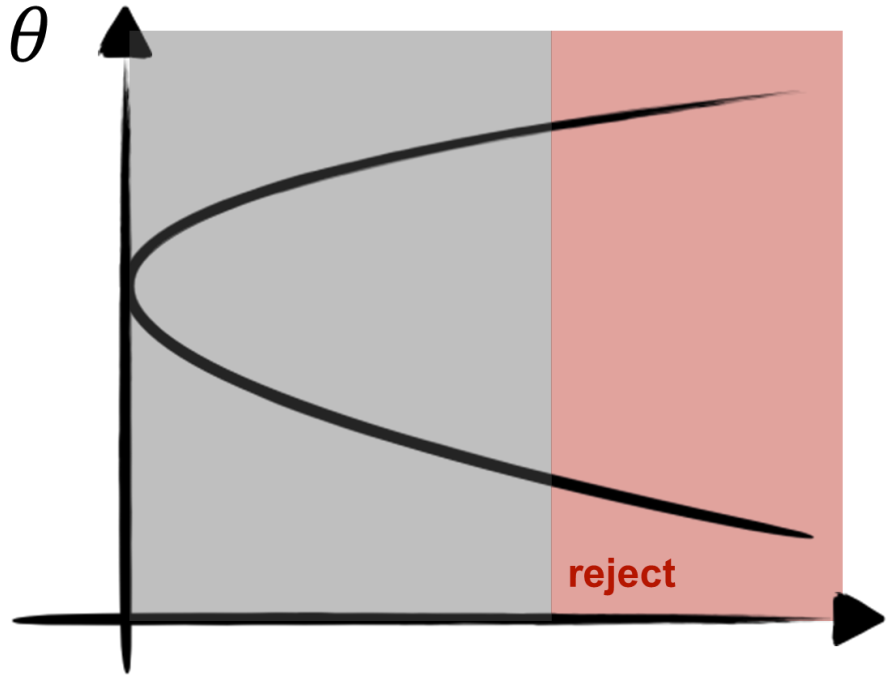
$$t_\theta = \lambda_\theta(x) = \frac{(\theta - \hat{\theta})^2}{\sigma_{\hat{\theta}}}$$



which is a parabola around the MLE

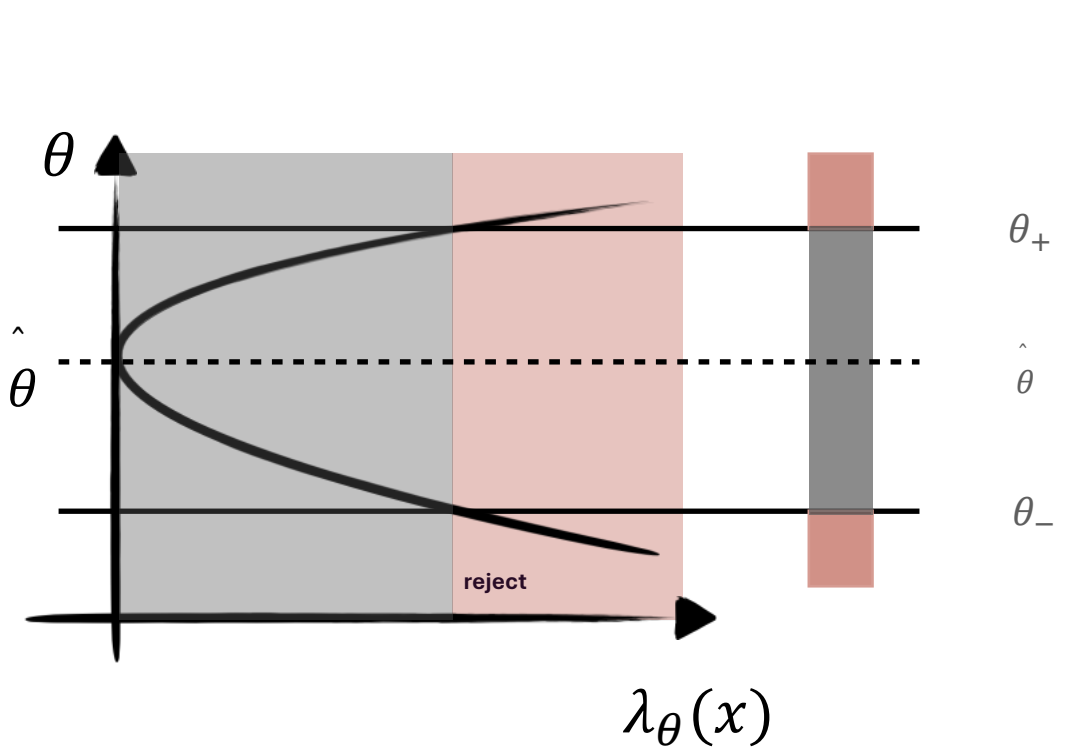
Putting it together

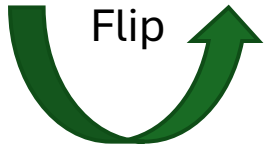


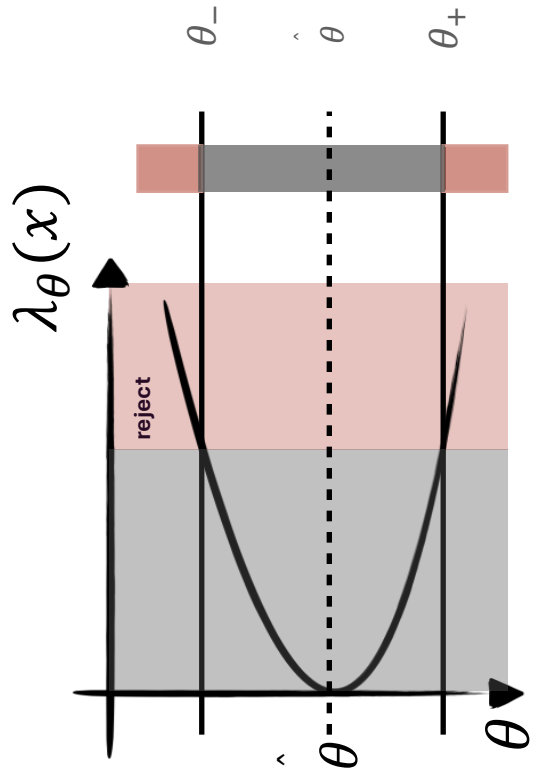


Constructing the Interval

The rejected parameter values are the parameters for which the (profile) likelihood ratio is sufficiently worse than the MLE likelihood







Standard 1σ intervals

Common Interval (with 68% coverage) achieved for the following rejection region

$$\lambda_\theta(x) = -2(\mathrm{LL}(\theta_{\mathrm{lim}}) - \mathrm{LL}(\hat{\theta})) = 1$$

$$\mathrm{NLL}(\theta_{\mathrm{lim}}) - \mathrm{NLL}(\hat{\theta}) = \frac{1}{2}$$

→ Based on a chi2 with d.o.f.=1 (asymptotic)

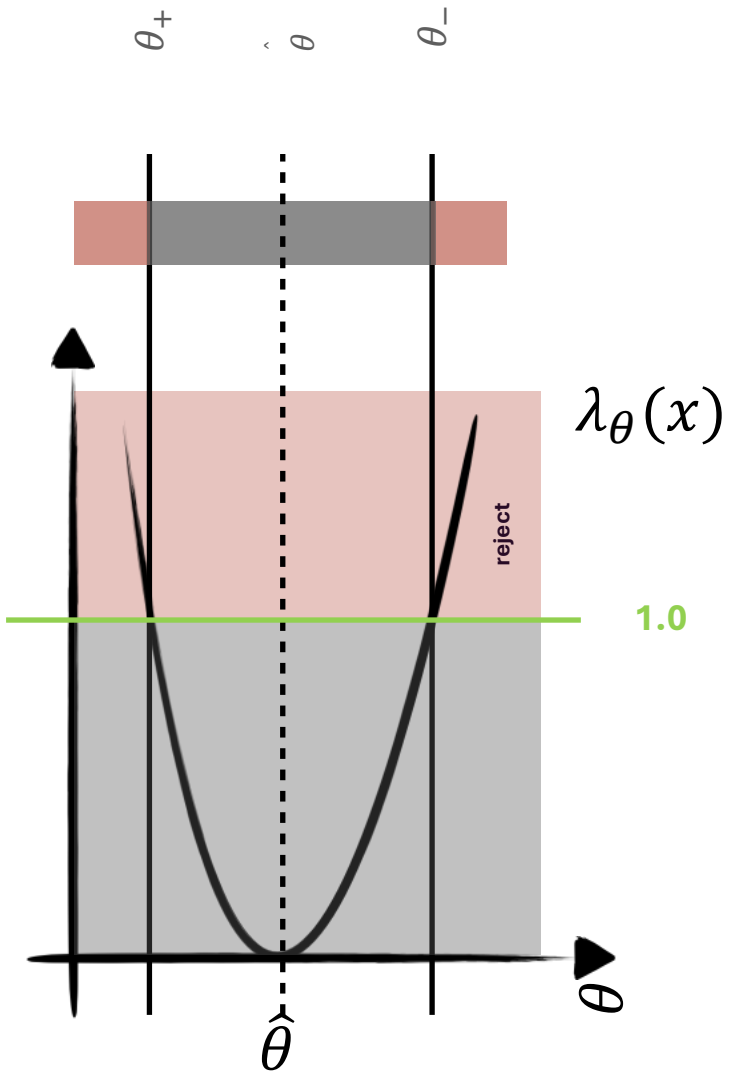


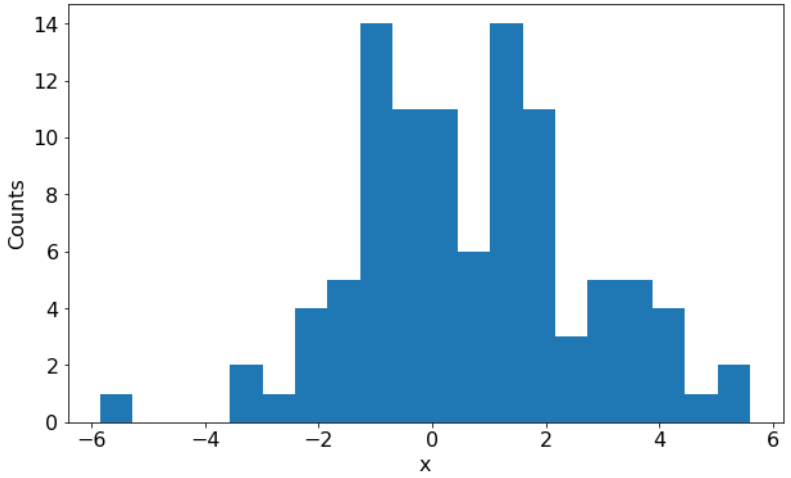
Table 9.5 The values of the quantile Q_γ for different values of the confidence level $1-\gamma$ for $n = 1, 2, 3, 4, 5$ fitted parameters.

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Cowan, p. 135

Putting it to work

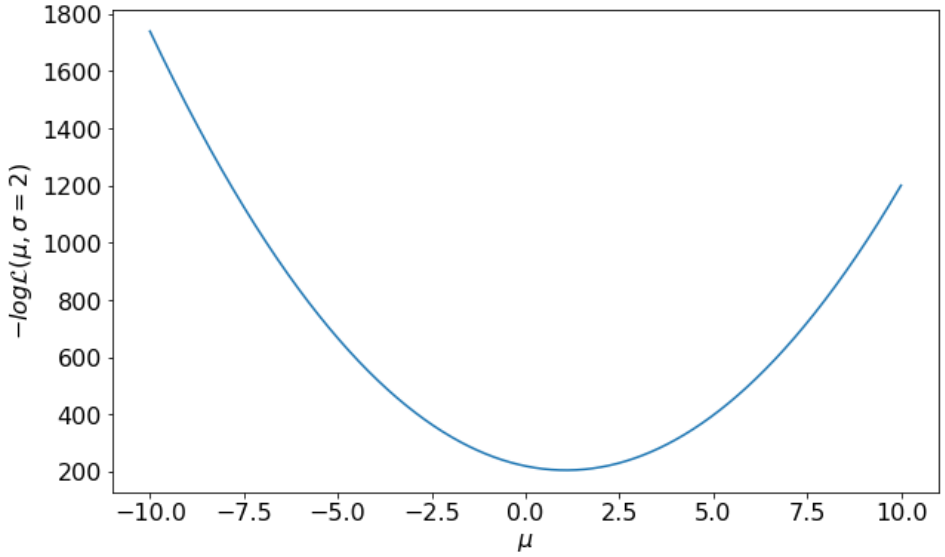
- Let's suppose we draw 100 samples from a normal distribution with μ=1 and σ=2



- Likelihood is obtained from the joint probability

$$\mathcal{L} = \prod_i p(xi|\mu, \sigma)$$

- Let's keep fixed $\sigma = 2$ and look at μ

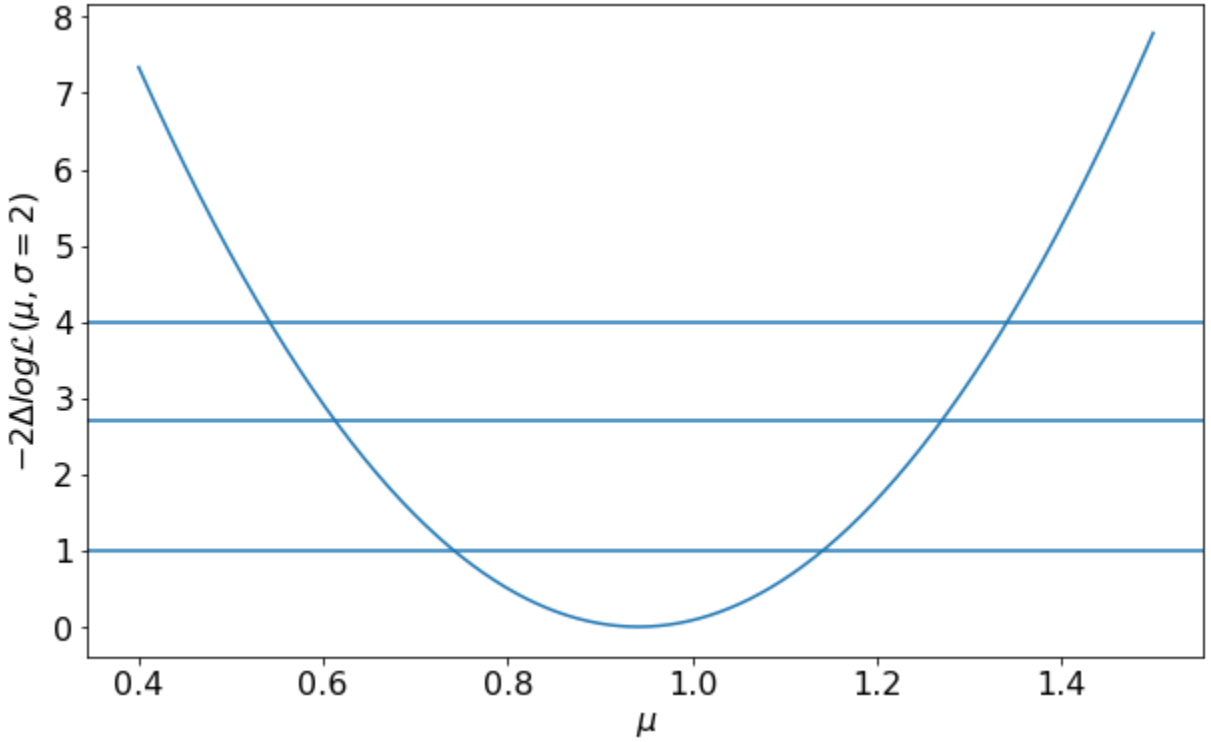


Intervals

- Construct the CL intervals
  - Here I chose 1 sigma (68%), 90% and 2 sigma (95%)
- The 1 sigma interval is around:
  $\mu = 0.95 \pm 0.2 \ @ \ 68\% \ C.L.$
- Also, remember CLT: from n measurements with $\sigma = 2$ we expect to get an error on μ of $\frac{\sigma}{\sqrt{n}}$ → Checks out!!

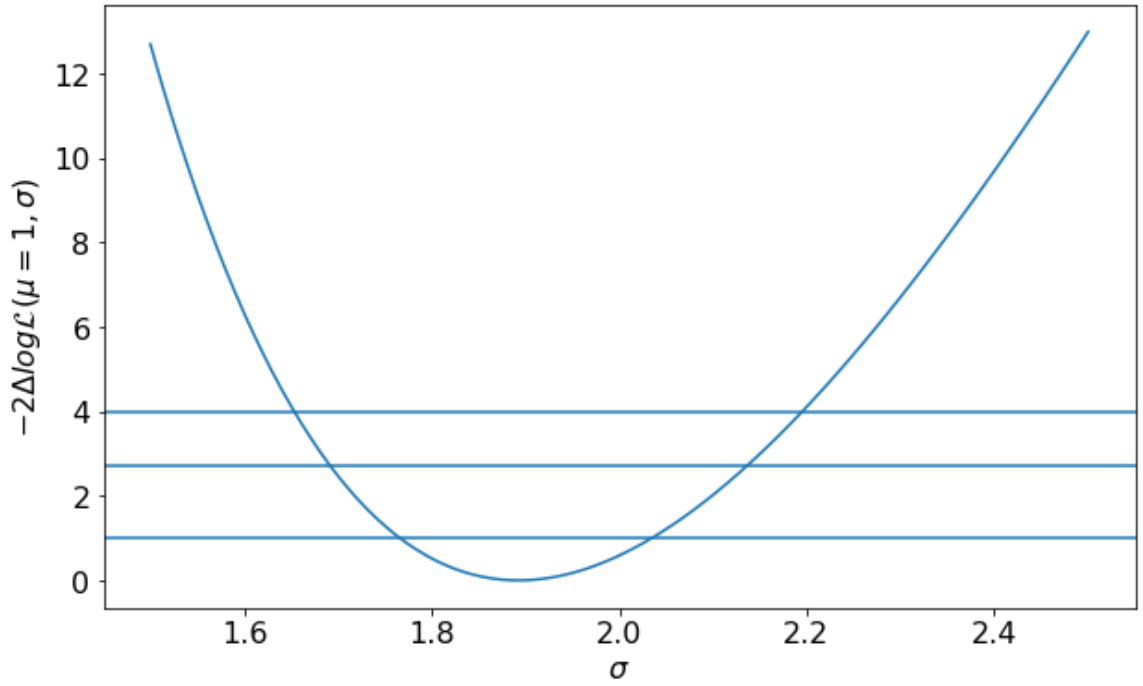
Table 9.5 The values of the quantile Q_γ for different values of the confidence level $1-\gamma$ for $n = 1, 2, 3, 4, 5$ fitted parameters.

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1



What about sigma?

- Follow same procedure

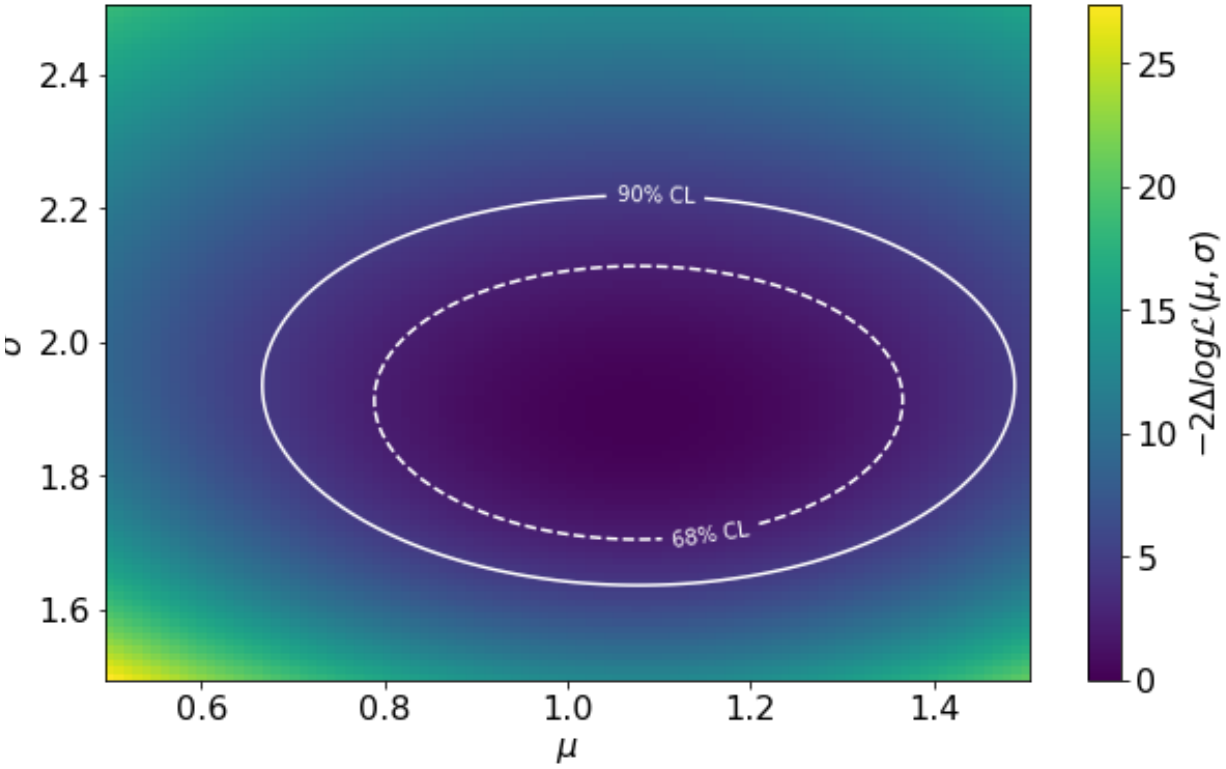


Both parameters!

- We can play the same game, but vary both parameters!!
- Need to use now χ^2 distribution with d.o.f. = 2 for critical values!

Table 9.5 The values of the quantile Q_γ for different values of the confidence level $1-\gamma$ for $n = 1, 2, 3, 4, 5$ fitted parameters.

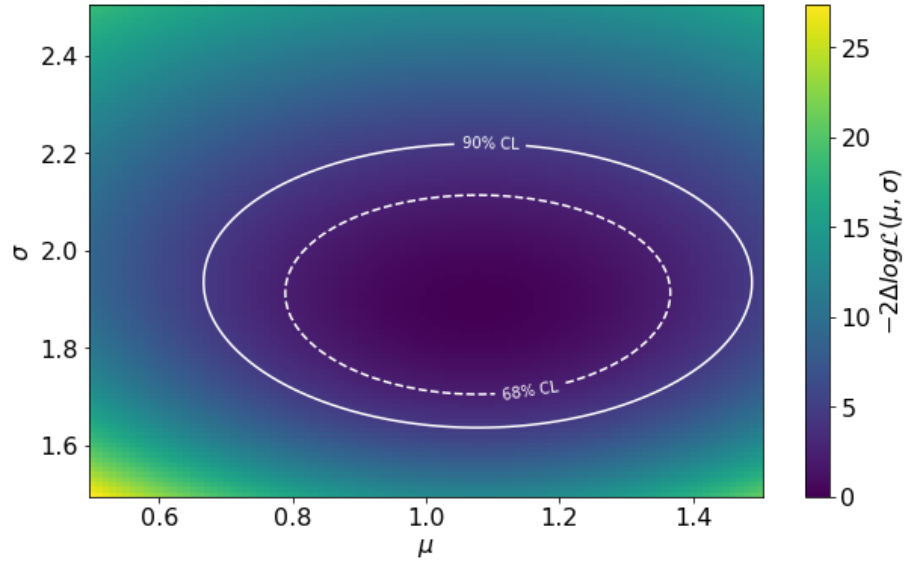
$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

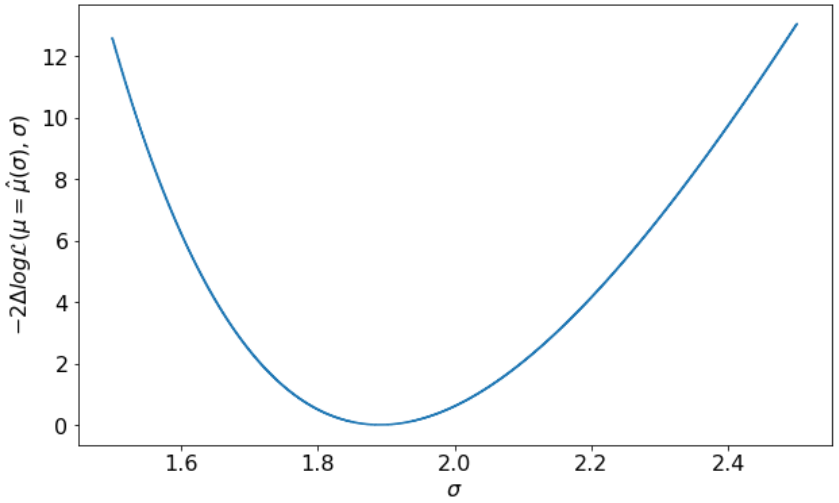


Profile LLH

- Let's say we do not know the true parameter μ nor σ
  - So we can follow the procedure from above
  - But maybe we're only interested in μ (or σ)

- We can follow a procedure called the *profile* likelihood
  - It sets all other parameters to their MLE as a function of the parameter of interest





Summary Intervals

- Neyman Construction is about repeated Hypothesis Tests scanning the parameter space of the model
  - Confidence Interval: all points that are not rejected by a test (e.g. p=0.05) are inside the interval
  - well defined "coverage" properties:
    - If data comes from θ_0 the interval will include θ_0 95% of the time
    - does not mean: 95% belief that θ_0 in any given interval
- Using the LRT as a test statistic
  - Provides a general framework for constructing intervals
  - Asymptotics are known (based on Wald / Wilk's)
  - Intervals = Contours of the log-likelihood function
- Care needed at boundaries of parameter space
  - A principled test (i.e. likelihood-ratio test) solves a number of issues from more ad-hoc test strategies: flip-flopping, empty intervals

Statistics for Physicists

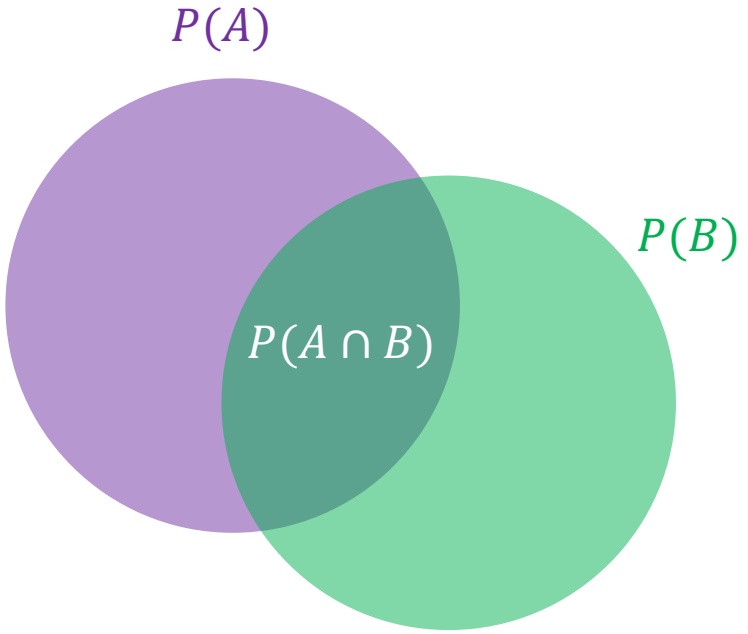
DAY 3

Bayesian Inference

Conditional Probability

- Let‘s define the conditional probability of A occuring, assuming B did occure

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$



$\rightarrow p(A|B)p(B) = p(A \cap B)$

Bayes' Theorem

since $p(A \cap B) \equiv p(B \cap A)$:

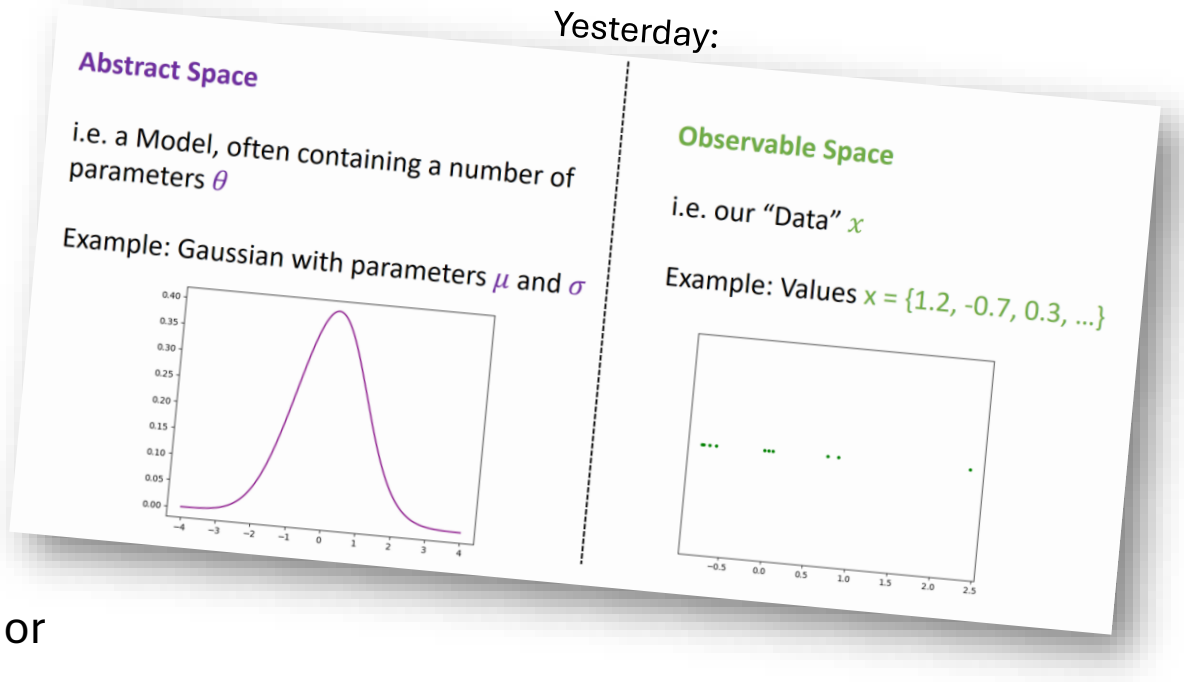
$$p(A|B)p(B) = p(A \cap B) = p(B|A)p(A)$$

→ we get Bayes' theorem:

$$\boxed{p(A|B) = \frac{p(B|A)p(A)}{p(B)}}$$

In the context of Models and Data

- A = our abstract model M with parameters θ
- B = our data x



Likelihood

Posterior

Prior

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}$$

Evidence (Marginal Likelihood)

Bayesian Probability

- Let‘s make the stark assumption, that we can upgrade θ to be a random variable!
- This means there exists a distribution $p(\theta)$!
  - In the Frequentist world, there is no such assumption
  - This is the price we have to pay for Bayesian Inference

- Conceptually very different, and once we accept this,
  - It means there also exists a joint distribution $p(x, \theta)$
  - And that we encode our degree of believe about the value of θ in the form of the distribution $p(\theta)$
  - Before we look at the data x we call it the prior distributuion $p(\theta)$
  - After inference (i.e. consulting the data x), we call it the posterior $p(\theta)$

Law of total Probability

- What is the "Evidence": $p(x|M)$?
  - It is the probability of the data x under the Model M
  - We'll discuss later what its interpretation is and how we can make use of it (→ Model comparison)
- For now, we can use the "law of total probability"

$$P(B) = \int P(A, B)dA$$

to express it as $p(x) = \int p(x, \theta)d\theta = \int p(x|\theta)p(\theta)d\theta$

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)} = \frac{p(x|\theta, M)p(\theta|M)}{\int p(x|\theta, M)p(\theta|M)d\theta}$$

Example: Fair coin?

We can employ Bayes' theorem for a parameter inference problem

Let's study the example of tossing a coin

→ Probability of one outcome:

- Heads: p (e.g. 0.5)
- Tails: $q = 1 - p$ (e.g. also 0.5)

- Multiple coin tosses:

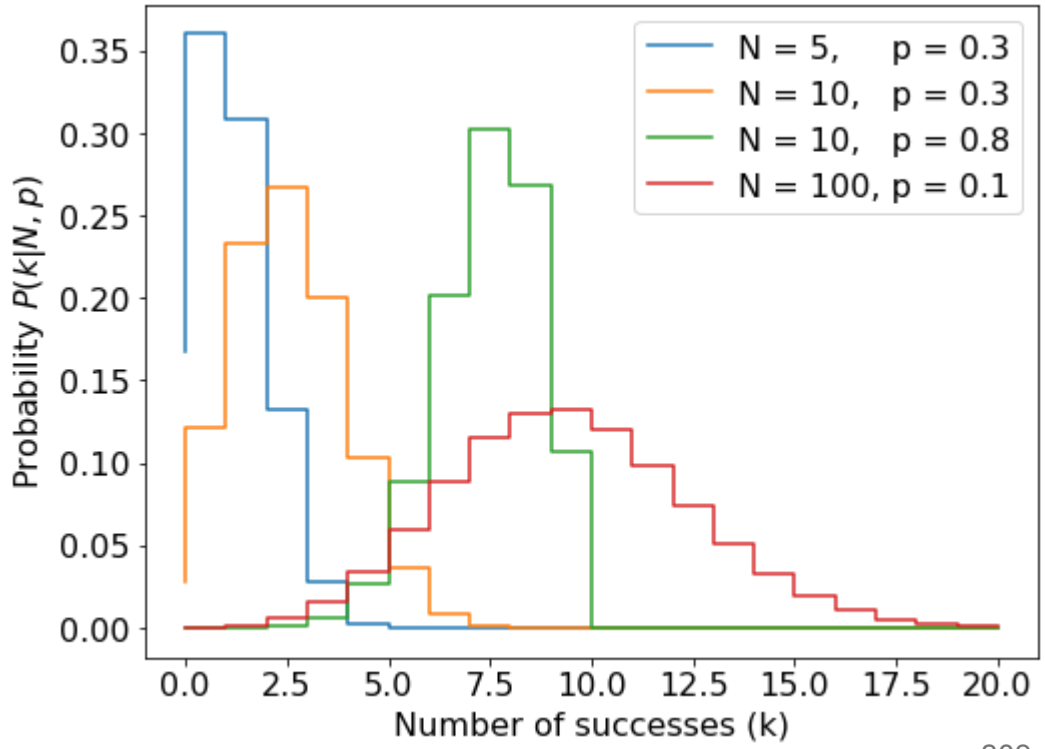
→ Binomial Distribution (next slide)

No country for old men
(https://www.youtube.com/watch?v=OLCL6OYbSTw)

Binomial Distribution

- $P(k|N,p) = \frac{N!}{k!(N-k)!}p^k q^{N-k} = \binom{N}{k}p^k(1-p)^{N-k}$
  - p: probability of a success
  - N: number of independent trials
  - k: number of successes

Let's try to infer the parameter p given our observed number of successes k in N coin tosses!



Putting Bayes to work

$$p(p|k,N) = \frac{p(k|p,N)p(p)}{p(k)}$$

- Now $p(k|p,N)$ we know, it is the binomial probability distribution $\binom{N}{k}p^k(1-p)^{N-k}$
- $p(p)$ we have to choose! Let's be very conservative and assume a uniform distribution between 0 and 1 (→ $p(p) = 1$ for p in [0,1])
- The marginal p(k) can be calculated from the integral over p

$$p(p|k,N) = \frac{p^k(1-p)^{N-k}}{\int_0^1 p^k(1-p)^{N-k}\,dp}$$

Integral is a standard beta function:

$$\beta(k+1, N-k+1) = \int_0^1 p^k(1-p)^{N-k}\,dp = \frac{k!\,(N-k)!}{(N+1)!}$$

Giving the posterior:

$$p(p|k,N) = \frac{(N+1)!}{k!\,(N-k)!}p^k(1-p)^{N-k}$$

Posterior

$$p(p|k,N) = \frac{(N+1)!}{k!\,(N-k)!} p^k (1-p)^{N-k}$$

- The posterior looks very much like the binomial distribution we started out from
  - **Except, it is now a function of p instead of k! It's a beta distribution**
  - And has an additional factor $(N+1)$
  - This is now a properly normalized pdf

- Let's analyze the posterior:
  - The mode of the posterior lies at $p^* = \frac{k}{N}$
    → This is the same as we would get from maximum likelihood (which is expected as we started from a flat prior)

First Moments of Posterior

- Expectation Value:

$$E[p] = \int_0^1 pp(p|N,k)dp = \int_0^1 \frac{(N+1)!}{k!\,(N-k)!} p^{k+1}(1-p)^{N-k}dp = \cdots$$

(using properties of the beta function) $\ldots\ E[p] = \frac{k+1}{N+2}$

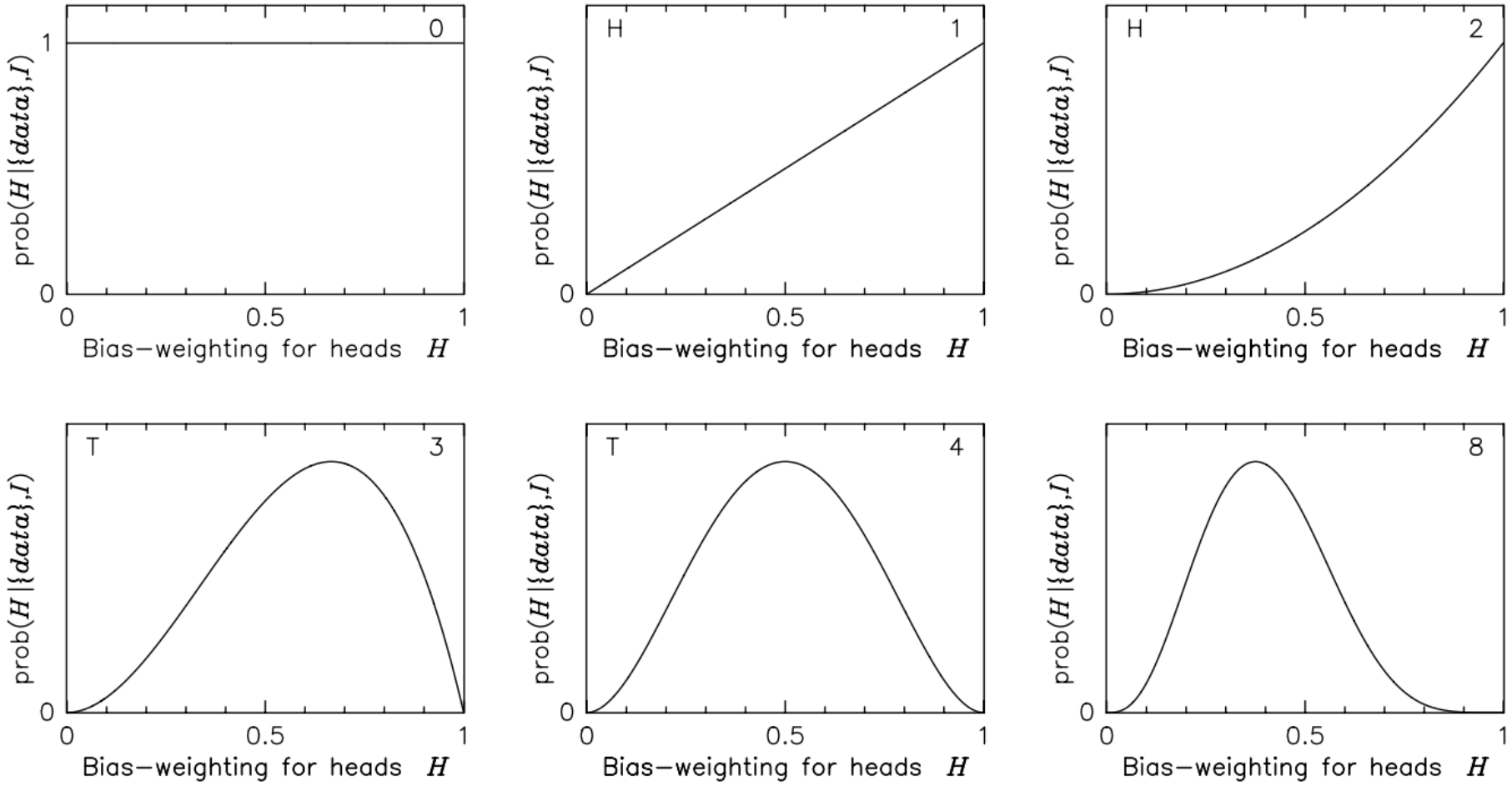
Variance:

$$V[p] = \frac{(k+1)(N-k+1)}{(N+2)^2(N+3)}$$

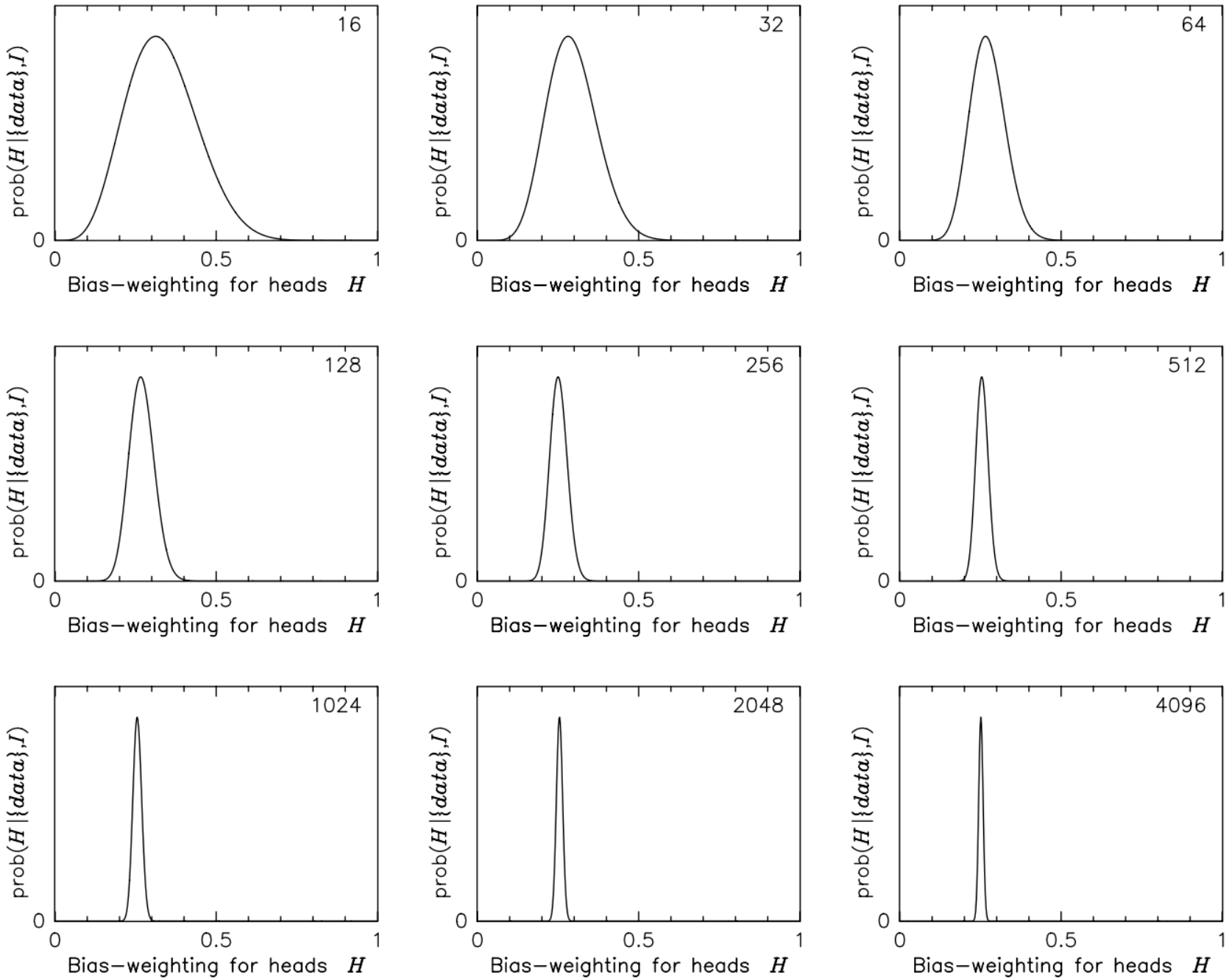
→These are also valid for $N = k = 0$, in which case we get the $E = \frac{1}{2}$ and $V = \frac{1}{12}$

✓ (these are the mean and variance of a standard uniform = our prior)

Visualization of Posterior



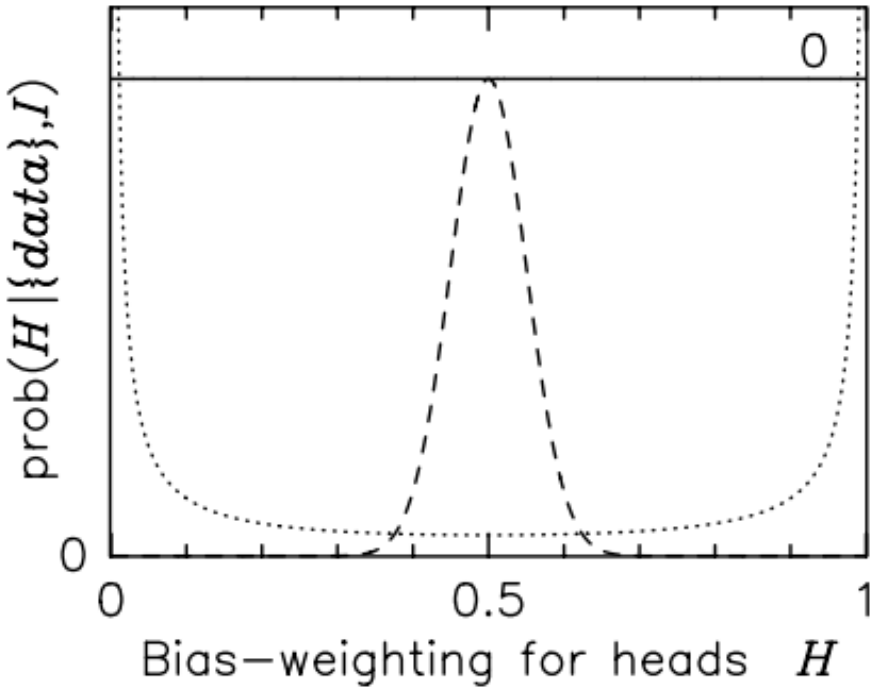
(from D. S. Sivia – Data Analysis, A Bayesian Tutorial)


16
32
64
128
256
512
1024
2048
4096
prob(H|{data},I)
0
0.5
1
Bias−weighting for heads H

Priors

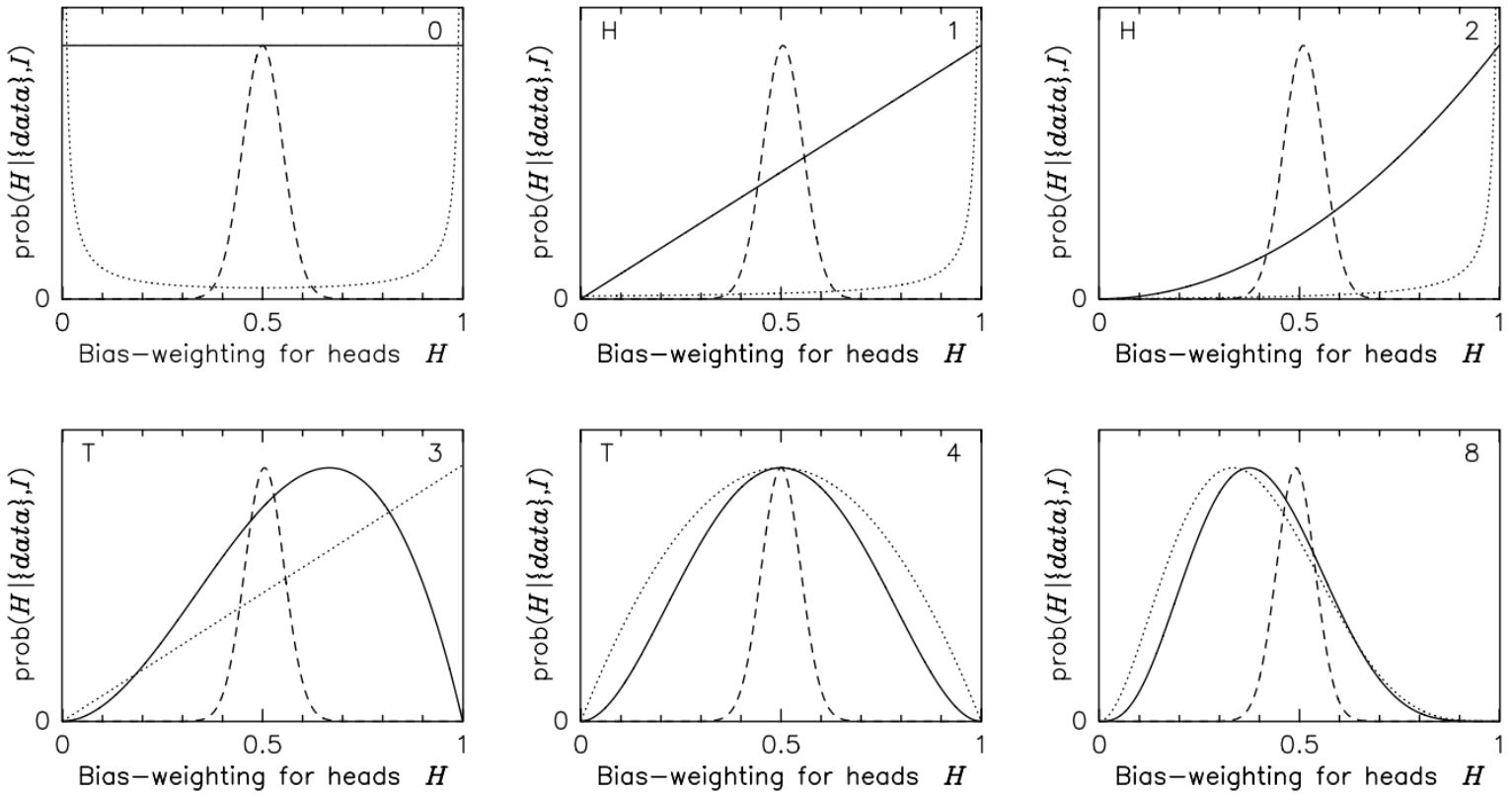
Importance of priors

- We had to choose the prior $p(p)$ ourselves and tried to be “unbiased” by taking a uniform distribution
- Another fair assumption would be something centered around 0.5 with some tails towards 0 and 1, because we could start with the assumption that coins are usually quite fair
- Or we could choose another extreme that is heavily biased towards 0 and 1



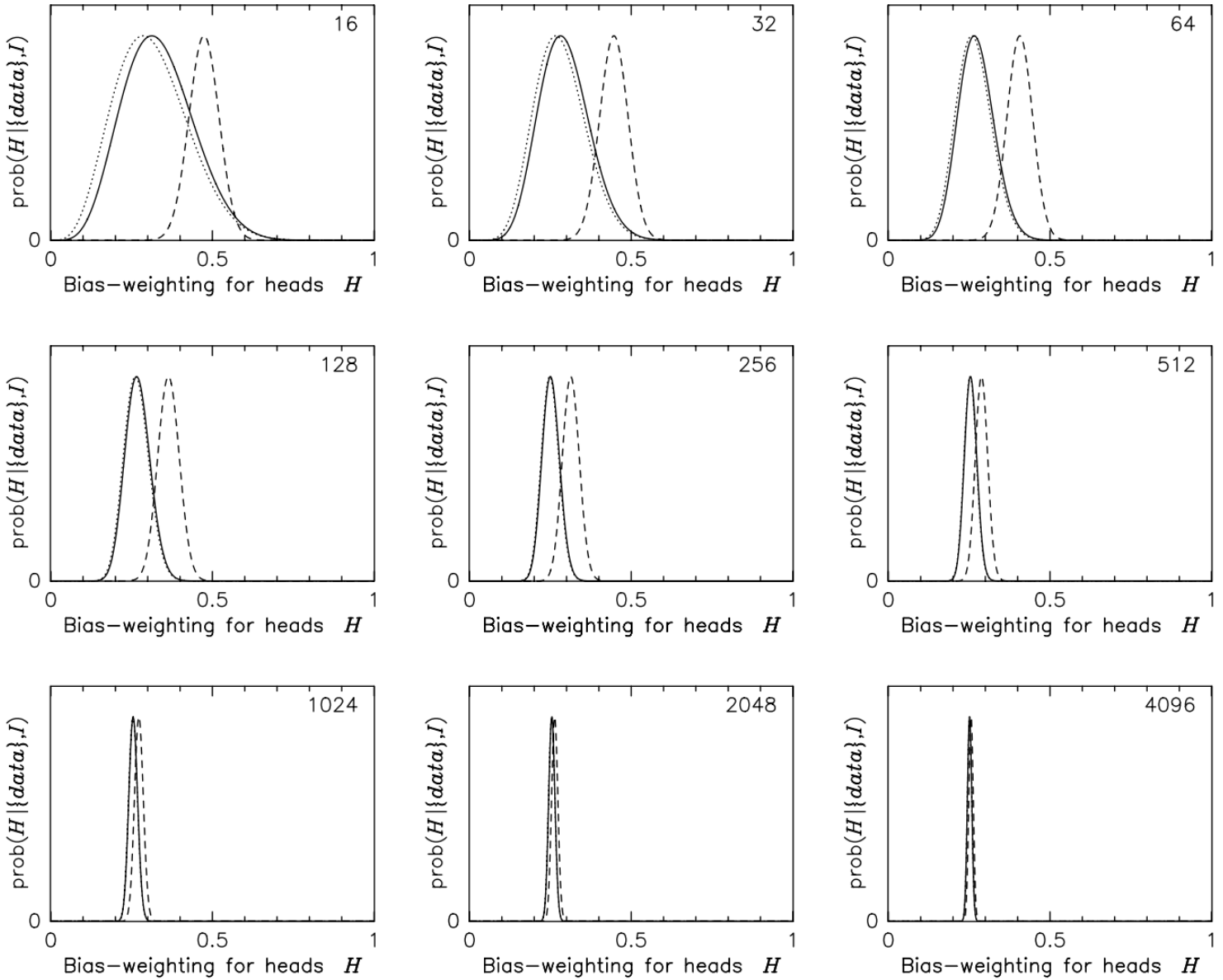
Effect of priors

- If we only analyze a few coin tosses, the effect of the prior is large



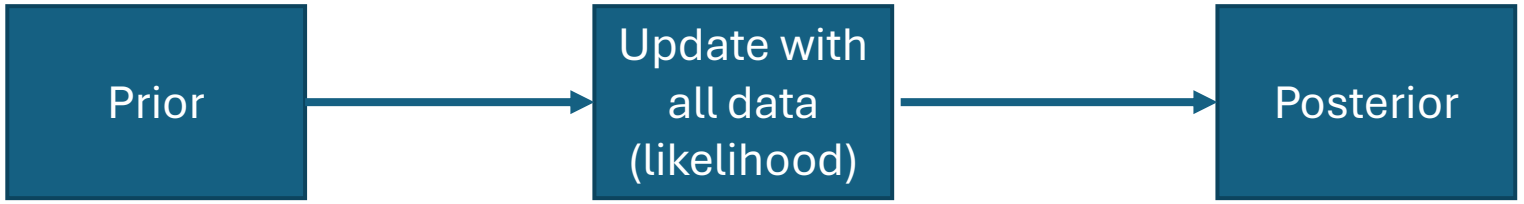
But when analyzing a larger dataset, the effect of the prior diminishes

→ the data is more informative than the prior, and prior choices don't matter as much anymore



Update of knowledge

- What if instead of analyzing all data at once, we successively performed one coin toss after the other, and use the posterior from one as the prior of the next?



Coin toss update of knowledge

Instead of Analyzing N tosses with successes k, we split it up into N_1 tosses with successes k_1 and N_2 with k_2
(where $N = N_1 + N_2$ and $k = k_1 + k_2$)

We know the result of analyzing the first batch already:

$$p(p|k_1, N_1) = \frac{(N_1 + 1)!}{k_1!\,(N_1 - k_1)!} p^{k_1}(1-p)^{N_1 - k_1}$$

Plugging everything in

$$p(p|k_2, N_2) = \frac{p(k_2|p, N_2)p(p)}{p(k_2)} = \frac{p(k_2|p, N_2)p(p)}{\int p(k_2|p, N_2)p(p)\, dp}$$

$$= \frac{p^{k_2}(1-p)^{N_2-k_2}p^{k_1}(1-p)^{N_1-k_1}}{\int p^{k_2}(1-p)^{N_2-k_2}p^{k_1}(1-p)^{N_1-k_1}dp} = \frac{p^{k_1+k_2}(1-p)^{N_1+N_2-k_1-k_2}}{\int p^{k_1+k_2}(1-p)^{N_1+N_2-k_1-k_2}dp} = \frac{p^{k}(1-p)^{N-k}}{\int p^{k}(1-p)^{N-k}dp}$$

→ This is exactly the same as analyzing the whole dataset at once!

Conjugate Priors

In general, if we use a beta distribution as the prior we get out another beta distribution as the posterior

→ The beta distribution is a *conjugate* prior to the Binomial distribution

Other examples:

- Gamma distribution is the conjugate prior for a Poisson Likelihood
- Dirichlet distribution is the conjugate prior for a Multinomial Likelihood
- Normal* distribution is the conjugate prior for a Normal Likelihood

(*) under certain assumptions

→ For more, see: https://en.wikipedia.org/wiki/Conjugate_prior

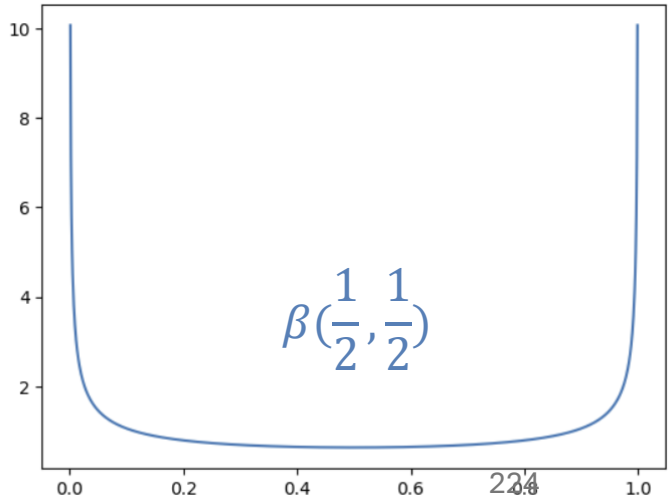
Jeffreys Prior

- The Jeffreys prior is intended to be a non-informative prior distribution
- It is constructed from the Fisher information I (→ see earlier Lecture)
- Jeffreys prior $p(\theta) \sim \sqrt{\det I(\theta)} = \sqrt{\mathbb{E}[(\partial\theta \log p(x|\theta))^2]}$

- Example for our Binomial:
  - $I(p) = -E[\partial^2 p log p] = \frac{Np}{p^2} + \frac{N-Np}{(1-p)^2} = \frac{N}{p(1-p)} \sim p^{-1}(1-p)^{-1}$

  → $P_{Jeffreys}(p) = \sqrt{I(p)} \sim p^{-1/2}(1-p)^{-1/2}$

  Which is a beta distribution $\beta(\frac{1}{2}, \frac{1}{2})$



Invariance

- Jeffreys prior is constructed in such a way that it is invariant under reparameterizations $\theta \to \xi$ since

- $$p(\xi) = p(\theta)\left|\frac{\partial\theta}{\partial\xi}\right| \sim \sqrt{\det I(\theta)}\left|\frac{\partial\theta}{\partial\xi}\right| = \sqrt{\mathbb{E}[(\partial\theta \log p(x|\theta))^2]\left(\frac{\partial\theta}{\partial\xi}\right)^2}$$

$$= \sqrt{\mathbb{E}[(\partial\xi \log p(x|\theta))^2]} = \sqrt{\det I(\xi)}$$

Example 2: Gaussian mean

- Assume we have data x distributed according to a normal distribution

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- We also assume that σ is fixed (= known)

Let's see what we can learn about μ from a single measurement x!

- We will use a flat prior for μ that extends well beyond the measured value x

$$p(\mu|x,\sigma) = \frac{p(x|\mu,\sigma)p(\mu)}{\int p(x|\mu,\sigma)p(\mu)d\mu}$$

$$\int p(x|\mu,\sigma)p(\mu)d\mu = \int_{\mu_{min}}^{\mu_{max}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \frac{1}{\mu_{max}-\mu_{min}} \mathrm{d}\mu$$

$$\approx \frac{1}{\mu_{max}-\mu_{min}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}\mu = \frac{1}{\mu_{max}-\mu_{min}} = p(\mu)$$

→ $p(\mu|x,\sigma) = \frac{p(x|\mu,\sigma)p(\mu)}{p(\mu)} = p(x|\mu,\sigma)$ which is the same Normal distribution, but as a function of μ instead of x

Multiple observations

Let's go back to our standard example of measuring n independent and identically distributed ($i.i.d.$) samples from a normal distribution

Using the "update of knowledge" procedure

$$p_{2(\mu|x_2)} = \frac{p(x_2|\mu,\sigma_2)p_1(\mu)}{\int p(x_2|\mu,\sigma_2)p_{1(\mu)}d\mu}$$

Prior (= Posterior from first measurement):

$$p_{1(\mu)} = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma_1}\right)^2}$$

Likelihood for second measurement:

$$p_{(x_2|\mu,\sigma_2)} = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x_2-\mu}{\sigma_2}\right)^2}$$

... some tedious calculus later ... (by completion of squares)

$$p(\mu|x1, x2, \sigma1, \sigma2) = \frac{1}{\sqrt{2\pi}\sigma_A} e^{-\frac{1}{2}\left(\frac{x_A-\mu}{\sigma_A}\right)^2}$$

With the weighted average $x_A = \frac{x_1/{\sigma_1}^2 + x_2/{\sigma_2}^2}{1/{\sigma_1}^2 + 1/{\sigma_2}^2}$

And $\frac{1}{{\sigma_A}^2} = \frac{1}{{\sigma_1}^2} + \frac{1}{{\sigma_2}^2}$

For n independent measurements x_i this generalizes to a Gaussian with mean $\frac{\sum x_i/{\sigma_i}^2}{\sum 1/{\sigma_i}^2}$ and variance $\left(\sum 1/{\sigma_i}^2\right)^{-1}$

Combining Measurements

- When combining measurements
  - The weight of individual data is proportional to the inverse of the square of the resolution

  → you can win quickly by improving resolution!

  → you can also win by adding more data, but it does not scale as fast $(\sim\sqrt{n})$

- Let's assume all $\sigma_i = \sigma$ are the same, meeting the CLT again

$$\rightarrow x_A = \frac{\sum x_i/\sigma^2}{\sum 1/\sigma^2} = \frac{\sum x_i}{\sum 1} = \frac{1}{n}\sum x_i = \bar{x}$$

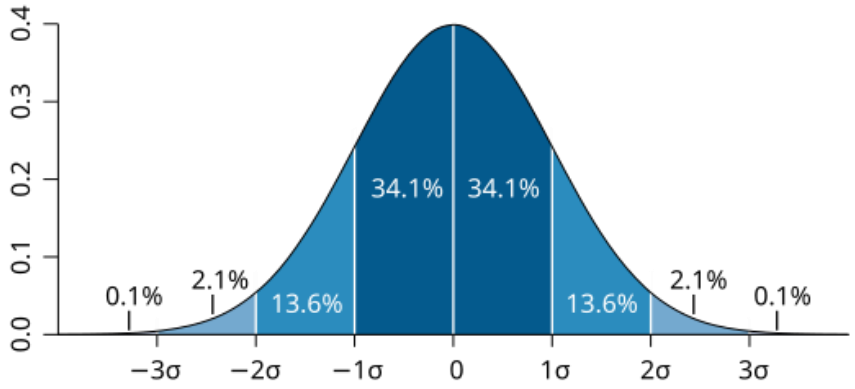
$$\rightarrow {\sigma_A}^2 = \left(\sum 1/\sigma^2\right)^{-1} = \frac{\sigma^2}{n}$$

Summary of posterior

- We can construct “credible intervals” from our posterior to summarize it
  - Calculate intervals in μ that contain a desired amount of probability,

For instance the central or smallest intervals:

Probability Content (in %)	μ range
68.3	$x \pm \sigma$
90.0	$x \pm 1.65\sigma$
95.0	$x \pm 1.96\sigma$
99.0	$x \pm 2.58\sigma$
99.7	$x \pm 3\sigma$



These Bayesian credible intervals are often abbreviated as “C.I.”, as opposed to the Frequentist confidence level intervals “C.L.”

Exercise

Bayes

Let‘s again do the Poisson Example, where we measure x=7

- Use a (conjugate) Gamma Prior $p(\lambda|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$
- What‘s the posterior $p(\lambda|\mathrm{x})$?

- Hint: don‘t solve the integral, just compare shapes
  - Prior $p(\lambda)\sim\lambda^{\alpha-1}e^{-\beta\lambda}$
  - Likelihood $p(\mathrm{x}|\lambda)\sim\lambda^{\mathrm{x}}e^{-\lambda}$
  - Posterior $p(\lambda|\mathrm{x})\sim p(\mathrm{x}|\lambda)p(\lambda)\sim\lambda^{\mathrm{x}}e^{-\lambda}\,\lambda^{\alpha-1}e^{-\beta\lambda} = \lambda^{\mathrm{x}+\alpha-1}e^{-(1+\beta)\lambda}$

  - This is again a gamma with $\alpha \to \mathrm{x}+\alpha$ and $\beta \to \beta+1$

- Plugging in numbers: let‘s choose the prior $\alpha = 2, \beta = 1$ and we observed x=7
- Posterior is a gamma with $\alpha = 9$ and $\beta = 2$

Beyond Simple Models

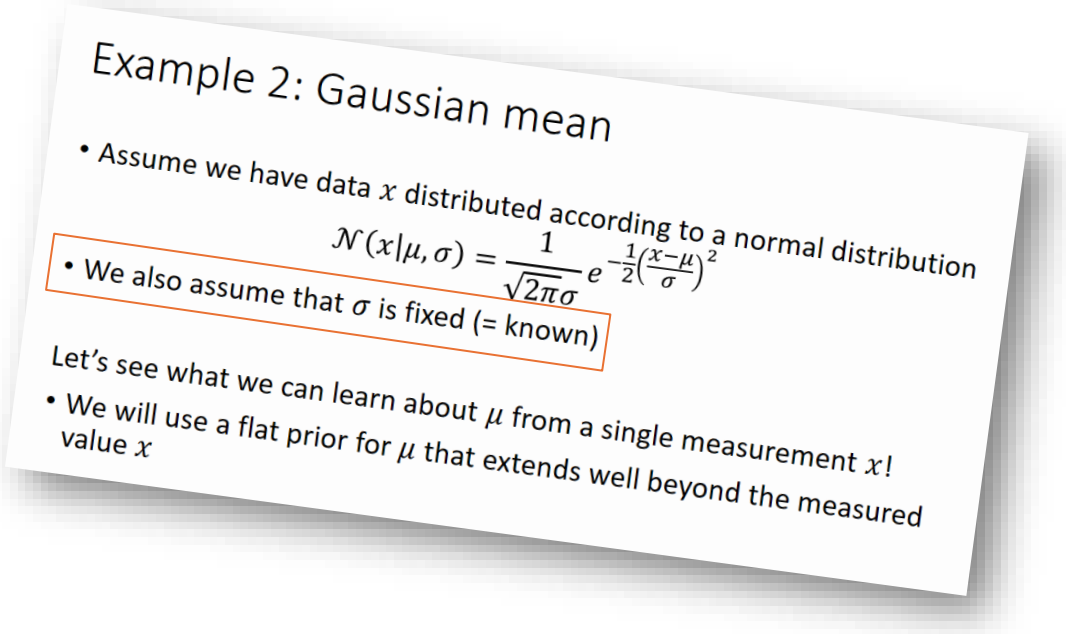
...again!!

Gaussian Noise revisited

- Before, we considered a Gaussian process with fixed variance
- What if we do not know the variance, but still want to infer μ?
  - The likelihood is now a function of both, σ and μ

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

  - Still, we can use Bayes' theorem in the same way



Prior

The prior now has to be a joint probability function over σ and μ

i.e. of the form $p(\mu, \sigma)$

We can again choose a “flat” prior

$$p(\mu, \sigma) = \begin{cases} const.\ if\ \sigma > 0 \\ 0\ \ elsewhere \end{cases}$$

Let’s assume the boundaries μ_{min}, μ_{max} and σ_{max} are far enough away

Posterior

Then we know the posterior is proportional to:

$$p(\mu, \sigma | \{x\}) \sim p(\{x\} | \mu, \sigma) p(\mu, \sigma)$$

$$\sim \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

Of course this posterior is now also a joint probability over σ and μ!

But we may not be interested in σ…

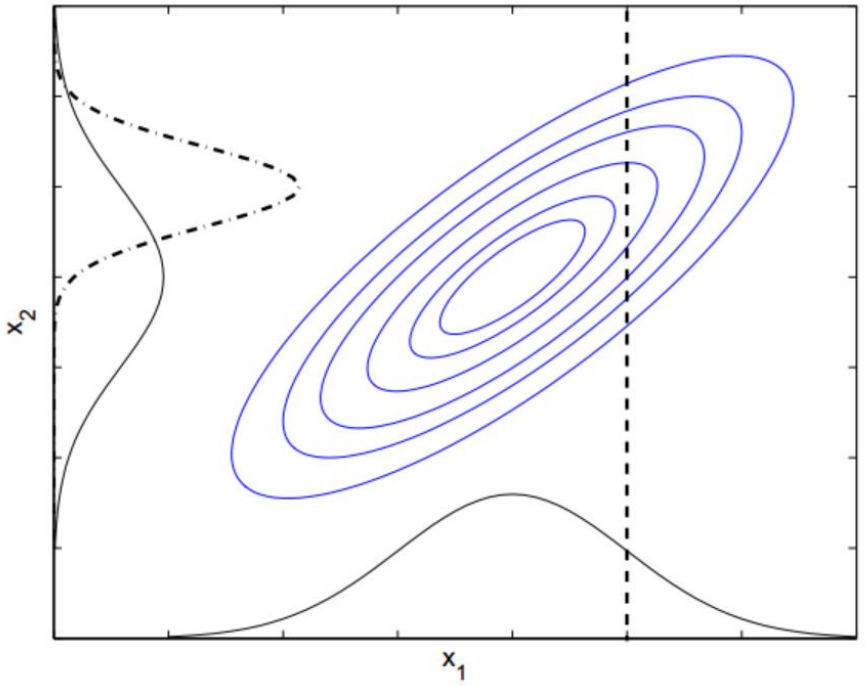
Marginal distribution

- We can integrate over an "unwanted" parameter
  - $p(\mu|\{x\}) = \int p(\mu, \sigma|\{x\})d\sigma$
- (and vice versa we could integrate over μ to have the marginal posterior of σ)

Note that what we did last lecture was to calculate $p(\mu|\{x\}, \sigma)$, which is also a posterior for μ, but it is conditional on a particular choice of σ

→We speak of the "**marginal**" distribution if nuisance parameters take into account our prior ignorance

→We speak of the "**conditional**" distribution if nuisance parameters are set to fixed values



Marginal Posterior

$$p(\mu|\{x\}) \sim \int \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} d\sigma$$

$$= \int (\sqrt{2\pi}\sigma)^{-N} e^{-\frac{1}{2\sigma^2}\Sigma_i (x_i-\mu)^2} d\sigma$$

We'll make a substitution $t = \frac{1}{\sigma}$

$$\sim \int t^{N-2} e^{-\frac{t^2}{2}\Sigma_i (x_i-\mu)^2} dt$$

$$\int t^{N-2} e^{-\frac{t^2}{2}\sum_i (x_i-\mu)^2} dt$$

Using another substitution of $\tau = t\sqrt{\sum_i (x_i - \mu)^2}$ makes the integral independent of μ, so it can be absorbed into the proportionality.

This reduces our posterior to:

$$p(\mu|\{x\}) \sim \left(\sum_i (x_i - \mu)^2\right)^{-(N-1)/2}$$

Properties of Posterior

We can find the MAP (maximum a posteriori probability estimate) easily by finding the root of the derivative (of the log)

$$\left.\frac{dlogp}{d\mu}\right|_{\mu_0} = 0$$

$\rightarrow \mu_0 = \frac{1}{N}\sum_i x_i = \bar{x}$

Which is still our usual sample mean!

Properties of Posterior

What about the shape of the posterior?

We can get an idea about the shape close to the MAP by doing a Taylor expansion:

$$logp(\mu) = logp(\mu_0) + \frac{dlogp}{d\mu}\bigg|_{\mu_0}(\mu - \mu_0) + \frac{1}{2}\frac{d^2logp}{d\mu^2}\bigg|_{\mu_0}(\mu - \mu_0)^2 + \cdots$$

- The first term is constant → doesn't tell us anything about the shape
- Second term is 0 because we are at the maximum
- The quadratic term is the dominant one for the shape

Properties of Posterior

So this leads us to:

$$p(\mu) \approx const \times e^{(\frac{1}{2}\frac{d^2 logp}{d\mu^2}\Big|_{\mu_0}(\mu-\mu_0)^2)}$$

Which has the form of a Gaussian!!

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

→This means we have approximated our posterior with a Gaussian with:

- mean μ_0
- width $1/\sqrt{-\frac{d^2 logp}{d\mu^2}}$

Properties of Posterior

- Plugging in the numbers then leaves us with:

$$\mu = \bar{x} \pm \frac{S}{\sqrt{n}}$$

where $S = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ is simply the sample standard deviation

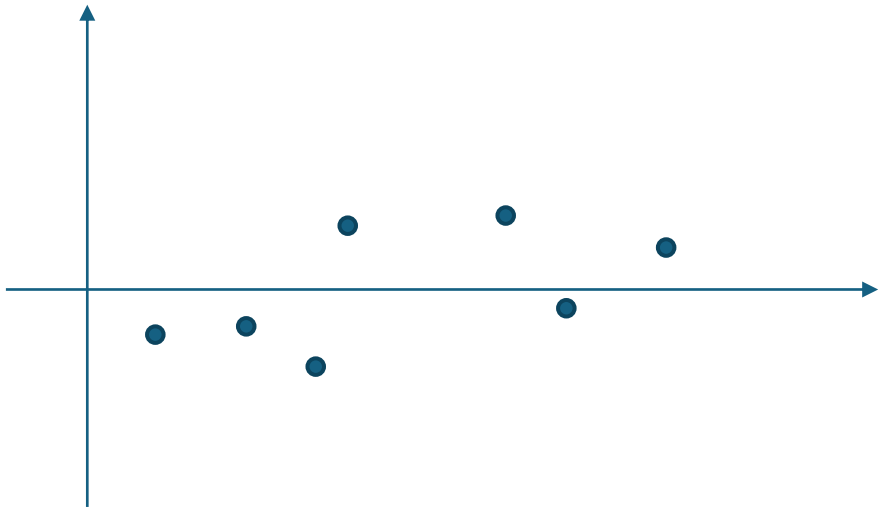
The above has the same form as the case with fixed σ, except that σ has been replaced by an estimate from the data!

Bayesian Model Selection

Model selection

- So far, we talked about parameter estimation
- Sometimes, however, we're more interested to test and compare different models

- Example:



Are these points better described by:

- Model A: $y = 0$
- Model B: $y = a$
- Model C: $y = ax + b$
- Model D: …

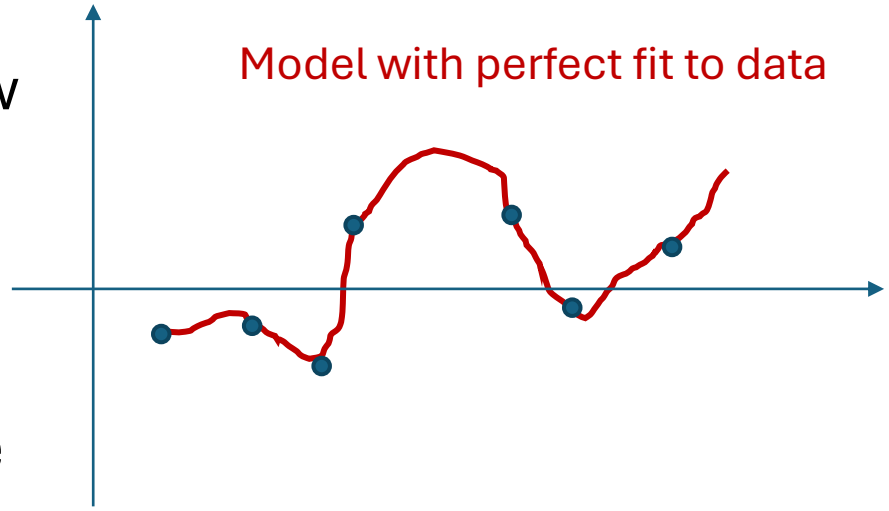
?

Goodness of fit

- We could be tempted to base our decision simply on how well the model can explain the data
  - For example quantifying the residuals, i.e. how close does the model get to the data points

The problem with that:

- An n-dimensional polynomial, for instance, will be able to perfectly describe the data
- But does that make it a better model…?



Using Bayes

- We can instead compare the two posteriors
  - Probability of Model A given the data D: $p(A|D)$
  - Probability of Model B given the data D: $p(B|D)$

- By building the *posterior ratio* = $\frac{p(A|D)}{p(B|D)}$

  we would prefer theory A if the posterior ratio is $\gg 1$
  (or theory B if it is much smaller then 1)

Using Bayes

- Pluggin in Bayes' theorem, this can be expressed as:

$$\frac{p(A|D)}{p(B|D)} = \frac{p(D|A)}{p(D|B)}\frac{p(A)}{p(B)}$$

(since $p(D)$ cancels out in the ratio)

$\frac{p(A)}{p(B)}$ is the ratio of the priors for the two models
To be fair it could be set to 1

Models with parameters

- What if model B contains an unknown parameter λ?
- We can marginalize over it:

$$p(D|B) = \int p(D, \lambda|B) d\lambda = \int p(D|\lambda, B) p(\lambda|B) d\lambda$$

Our usual likelihood under Model B

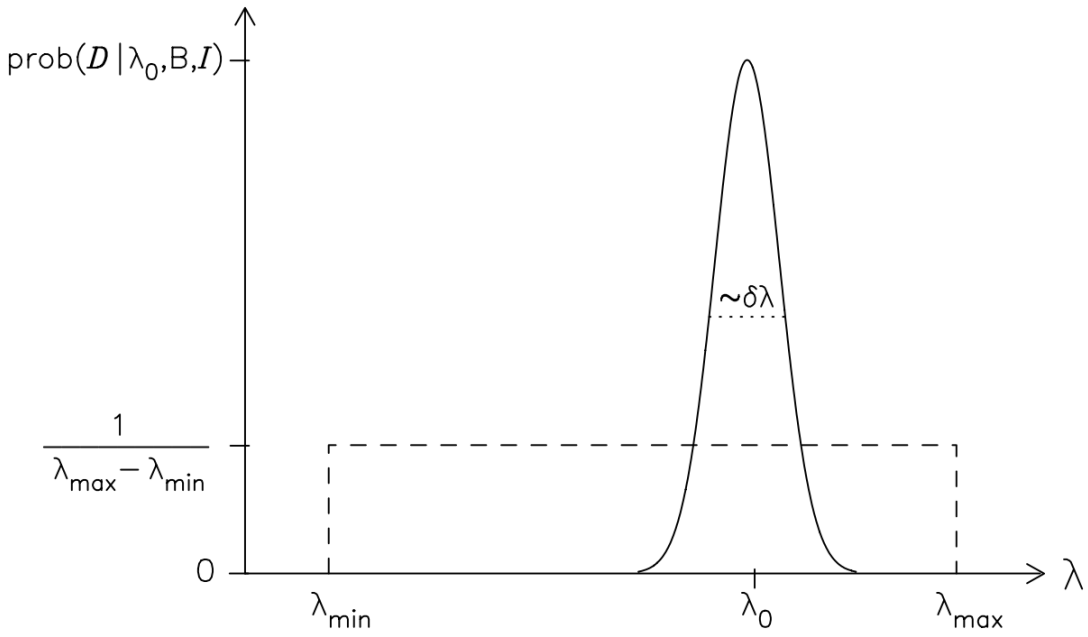
The prior for λ under model B

Example

- Let's use a uniform prior for λ:
  - $p(\lambda|B) = \frac{1}{\lambda_{max} - \lambda_{min}}$

  $\rightarrow p(D|B) = \frac{1}{\lambda_{max} - \lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} p(D|\lambda, B) d\lambda$

- Let's also assume that the likelihood is normal around the MLE λ_0 (i.e. parabolic in LLH) with a width (uncertainty) of $\delta\lambda$
  - $p(D|\lambda, B) = p(D|\lambda_0, B) \times e^{-\frac{(\lambda - \lambda_0)^2}{2\delta\lambda^2}}$



- $\rightarrow p(D|B) = \frac{1}{\lambda_{max} - \lambda_{min}} \int_{\lambda_{min}}^{\lambda_{max}} p(D|\lambda_0, B) \times e^{-\frac{(\lambda - \lambda_0)^2}{2\delta\lambda^2}} \, d\lambda$

$$= \frac{p(D|\lambda_0, B)}{\lambda_{max} - \lambda_{min}} \times \int_{\lambda_{min}}^{\lambda_{max}} e^{-\frac{(\lambda - \lambda_0)^2}{2\delta\lambda^2}} \, d\lambda$$

$$= \frac{p(D|\lambda_0, B) \times \delta\lambda\sqrt{2\pi}}{\lambda_{max} - \lambda_{min}}$$

Posterior ratio

$$\frac{p(A|D)}{p(B|D)} = \frac{p(A)}{p(B)} \times \frac{p(D|A)}{p(D|\lambda_0, B)} \times \frac{\lambda_{max} - \lambda_{min}}{\delta\lambda\sqrt{2\pi}}$$

Prior preference for models
Can be set to 1 for example

Likelihood ratio of the best fit of the models to the data (MLEs) – "goodness of fit"

Ockham factor: penalty for introducing additional degrees of freedom into a theory

- The likelihood ratio alone will always favour the model with the closer fit to data
- The Ockham factor penalizes more complex theories (Ockham's razor)
- Large prior ranges penalize a theory
- Small $\delta\lambda$ penalize a theory (→ only a very narrow range of parameter values are compatible with the data)

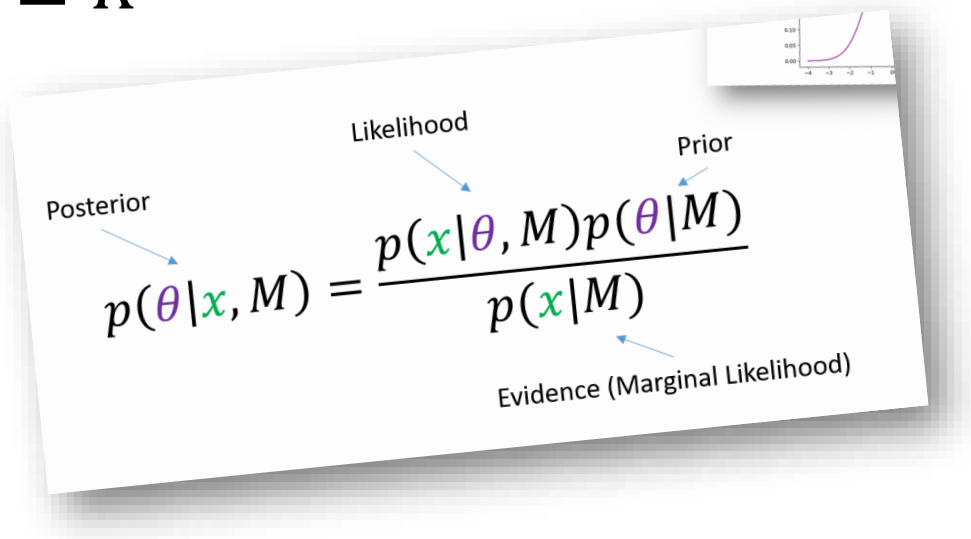
Bayes Factor

When we assign equal prior weight to either model, as discussed the prior ratio cancels out

$$\frac{p(A|D)}{p(B|D)} = \frac{p(D|A)}{p(D|B)} \equiv K$$

Where we define K as the Bayes factor

This is the ratio of the **Evidence** under each model!



$\log_{10} K$	K	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Exercise

Bayes

Model Comparsion

- Two theorists are arguing who‘s model is better
  - Theorist A: has a model with 2 parameters
  - Theorist B: has a model with 12 parameters
- Whose model is better?

- Let‘s agree to use one model
  - Theorist A gives you very narrow priors on the parameters
  - Theorist B gives you very wide priors on the parameters
- Whose model is better?

Solving the integral…

$$p(\theta|x,M) = \frac{p(x|\theta,M)p(\theta|M)}{\int p(x|\theta,M)p(\theta|M)d\theta}$$

- To this end, we have either:
  - Solved integrals analytically
  - Recognized the form of the posterior and then identified the correct pdf
  - Expanded the posterior around its MAP with a Taylor series, i.e. approximated the posterior with a normal distribution
  - Build ratios to avoid integrals by canceling them out
  - ….

- This is somewhat unsatisfactory, and we want methods applicable to the general case → we need to resort to **numerical methods**

MCMC

- In general, highly dimesnional integrals are very difficult to solve
- Instead, we typically resort to **Markov Chain Monte Carlo** (MCMC) methods to get around solving the integral
- An MCMC algorithm (e.g. Metropolis Hastings) allows us to generate samples from the posterior distribution directly, without evaluating the integral

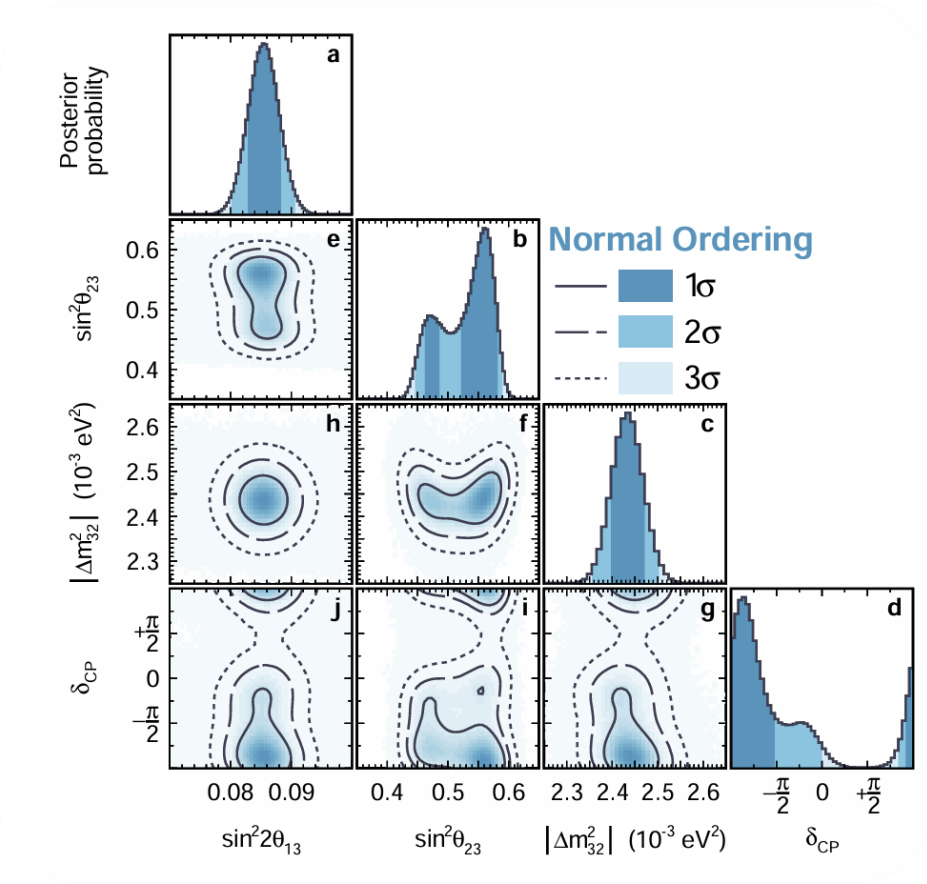
→ **Price to pay:** We will not get an expression for the posterior!

→ Rather, we get samples. With those, we can estimate quantities, e.g. draw density estimates (Historgams and the like)

Example

- Running an MCMC analysis requires:
  - Specifying the priors
  - Specifying the likelihood
  - Running an MCMC chain to produce samples according to the posterior
  - Using the sample to estimate quantities:
    - Maximum a Posterior (MAP), Credible Intervals
    - Densities (e.g. 1d and 2d marginals via Histograms or KDEs…)
    - …

Model comparison requires the actual integral value → needs other tools



Summary of Bayesian Inference

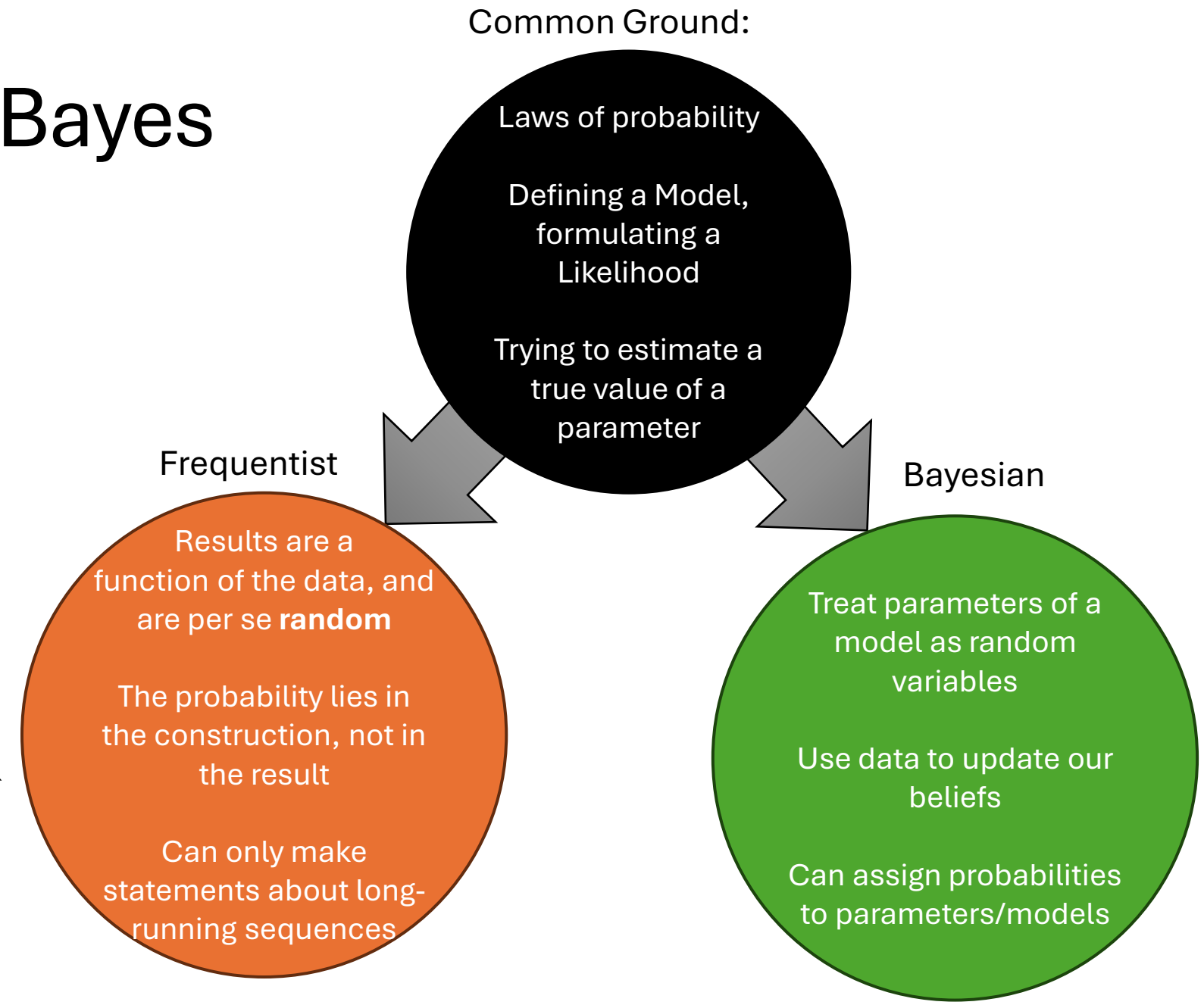
Summary so far

- Bayes Theorem offers a way to turn statements about the outcome of an experiment given its parameters into a statement about the parameters given an outcome
  - This needs the likelihood (just as in the frequentist case)
  - But also a choice of prior distribution for the parameters!
  - Some special priors:
    - Conjugate priors for a given likelihood: Posterior will have same form
    - Jeffreys prior: invariant under reparatemetrization, therefor regarded as “unbiased”
- The result is summarized in the Posterior distribution
  - This is the Prior updated with the knowledge from our data
  - Using the posterior of one measurement as the prior to the next offers a way to continuously “update our knowledge”
  - Credible Intervals can be derived from the posterior

Review of Course

Frequentist vs. Bayes

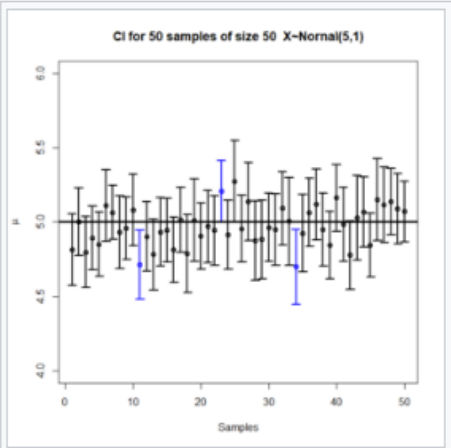
- What should you use?
  - It‘s not that easy, the two answer different questions!
  - **Frequentist C.L.:** “In repeated experiments, 95% of such intervals will contain the true parameter.”
  - **Bayesian C.I:** “Given the model (prior & llh) and the data, there is a 95% probability the parameter lies in this interval.”



Common misunderstandings [edit]

See also: § Counterexamples

Confidence intervals and levels are frequently misunderstood, and published studies have shown that even professional scientists often misinterpret them.[12][13][14][15][16][17]



Plot of 50 confidence intervals from 50 samples generated from a normal distribution.

- A 95% confidence level does not mean that for a given realized interval there is a 95% probability that the population parameter lies within the interval (i.e., a 95% probability that the interval covers the population parameter).[18] According to the frequentist interpretation, once an interval is calculated, this interval either covers the parameter value or it does not; it is no longer a matter of probability. The 95% probability relates to the reliability of the estimation procedure, not to a specific calculated interval.[19] Neyman himself (the original proponent of confidence intervals) made this point in his original paper:[10]

> It will be noticed that in the above description, the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results will tend to α. Consider now the case when a sample is already drawn, and the calculations have given [particular limits]. Can we say that in this particular case the probability of the true value [falling between these limits] is equal to α? The answer is obviously in the negative. The parameter is an unknown constant, and no probability statement concerning its value may be made...

- A 95% confidence level does not mean that 95% of the sample data lie within the confidence interval.
- A 95% confidence level does not mean that there is a 95% probability of the parameter estimate from a repeat of the experiment falling within the confidence interval computed from a given experiment.[16]

https://en.wikipedia.org/wiki/Confidence_interval

Why are C.I. not the same as C.L.??

Pathological Example: Let's suppose you construct intervals for the width σ of a Gaussian centered at 0:

$$\text{Interval} = \begin{cases} \mathbb{R}^{+} \ if \ x \geq 0 \\ \emptyset \ if \ x < 0 \end{cases}$$

The resulting intervals will have a perfect coverage of $\alpha = 50\%$

→The true value of σ will be contained in exactly 50% of cases!

Yet, the intervals are either empty or all real, positive numbers…

C.L. intervals cannot be interpreted the same way as credible intervals!

Why are C.I. not the same as C.L.??

- **Another Pathological Example:** Let‘s suppose you are trying to measure a parameter λ that has a true value of λ_0

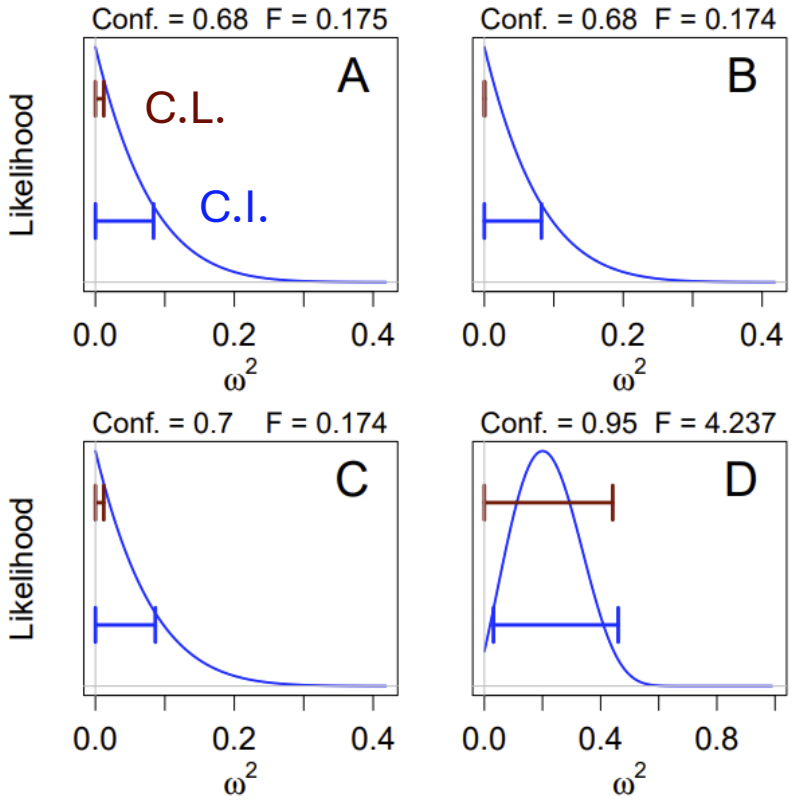
- You perform a Bayesian Analysis, and choose a prior $p(\lambda)$ that happens to have $p(\lambda_0) = 0$

→Any resulting posterior will also have $p(\lambda_0|x) = 0$

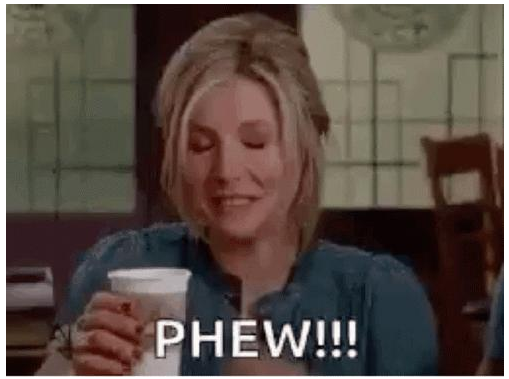
→Any C.I. will never contain the true value (coverage of 0%)

Differences

- The intervals themselves differ
  - Examples can be constructed where they are wildly different
  - Example on the right taken from Morey et al. “The Fallacy of Placing Confidence in Confidence Intervals”
- Their interpretation is different:
  - C.L. only make sense in the context of long running sequences of trials, and there is one true parameter value
  - C.I. express the degree of belief as a probability distribution over possible values of the parameter



- In the limit of large data (asymptotically), and under some assumptions, C.I.s and C.L.s converge (Bernstein–von Mises theorem)



Handling of Nuisance Parameters

- We are often dealing with unwanted degrees of freedom ν while measuring θ
  - Could for example be a detector uncertainty

- **Conditional**
  - keeping ν fixed
  - Then they are not part of the inference procedure

- **Profiling**
  - Setting ν to their MLE $\hat{\nu}$ under a respective hypothesis
  - e.g. Profile LHR for nested hypothesis

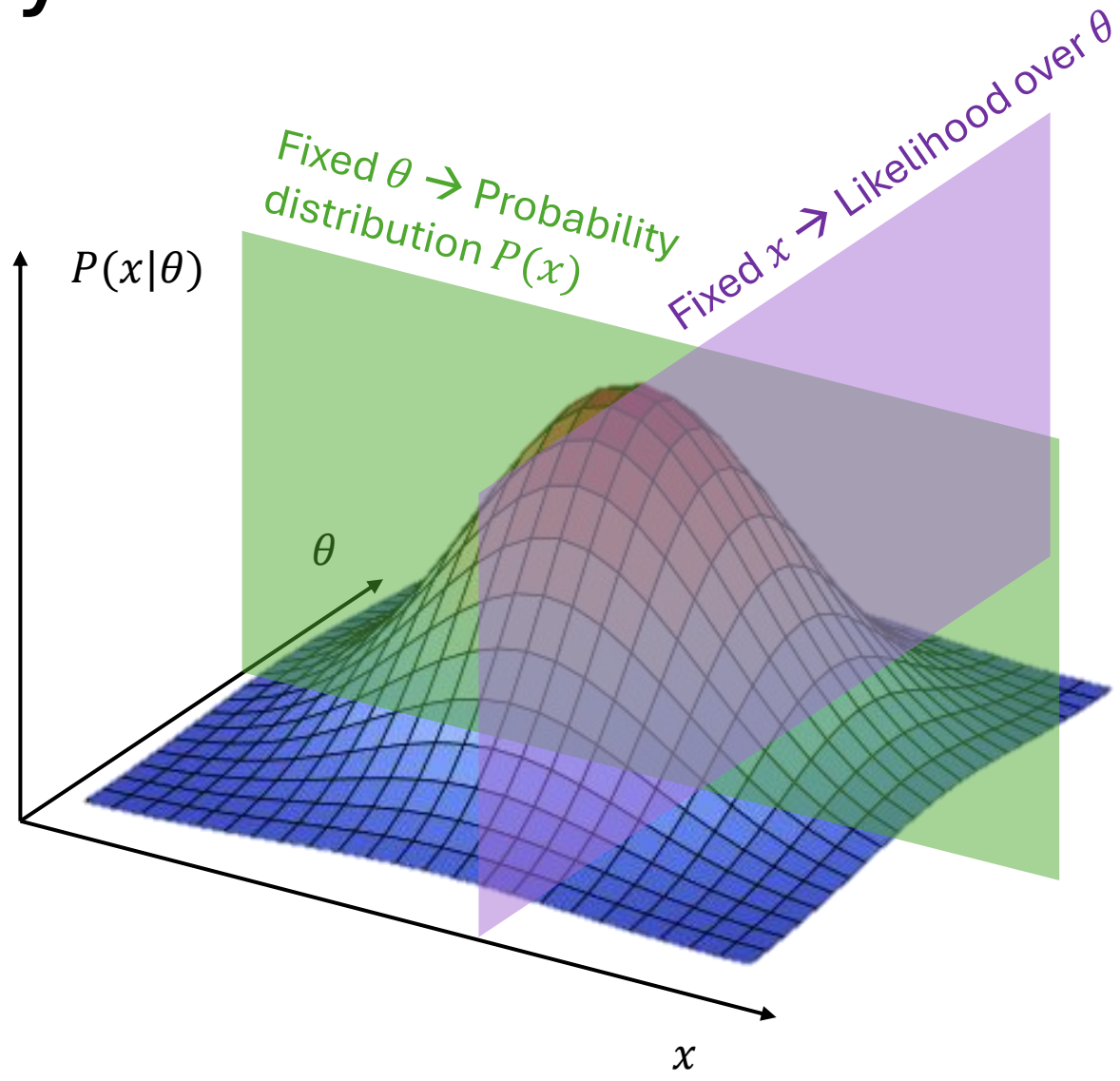
- **Marginalizing**
  - Integrating over all possible values of ν, i.e. $\mathrm{p}(\theta) = \int \mathrm{p}(\theta, \nu)\mathrm{d}\nu$
  - Works only in the Bayesian context, as we need ν to have a pdf $p(\nu)$

Hypo Testing vs. Model Comparison

- **Frequentist Hypothesis Testing:**
  - Does not give a preference for a hypothesis
  - You are testing a null vs. an alternate, trying to reject the null in favour of the alternate
  - It does not give you any support for accepting the alternate!
  - It gives p-values, meaning how unlikely the data are under the null

- **Bayesian Model Comparison:**
  - Given the data, which model has a higher posterior probability
  - It can directly assign probabilities to models
  - Can compare any (non-nested) model against each other

Probability vs. Likelihood

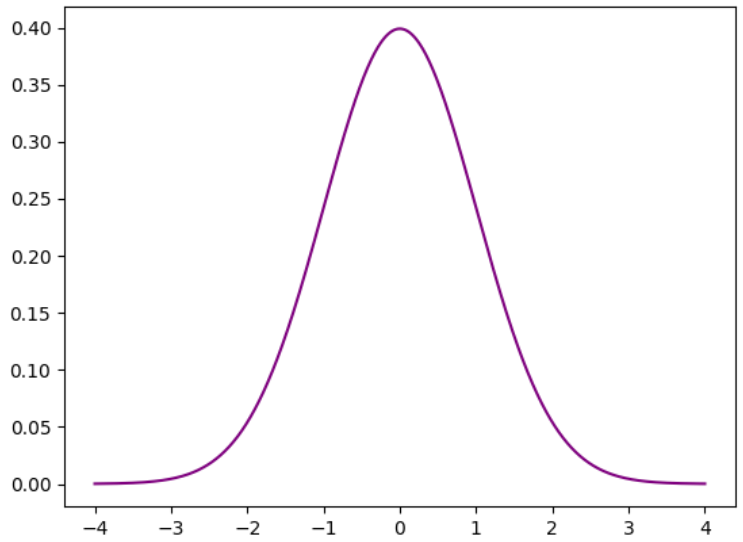


Model vs. Observation

Abstract Space

i.e. a Model, often containing a number of parameters θ

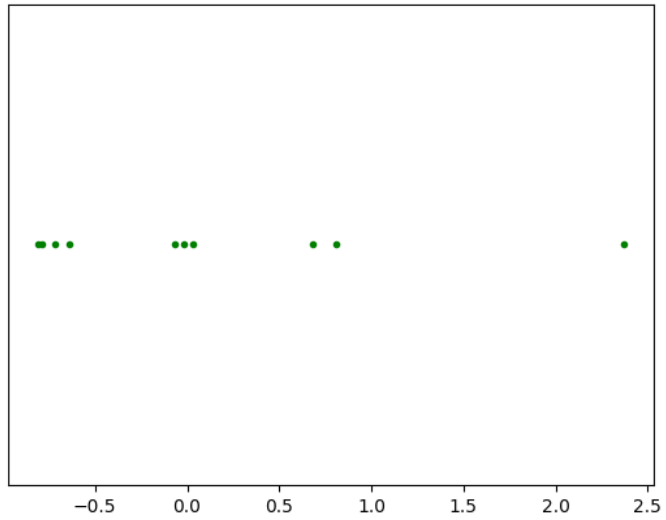
Example: Gaussian with parameters μ and σ



Observable Space

i.e. our "Data" x

Example: Values x = {1.2, -0.7, 0.3, ...}



From Model to Observation

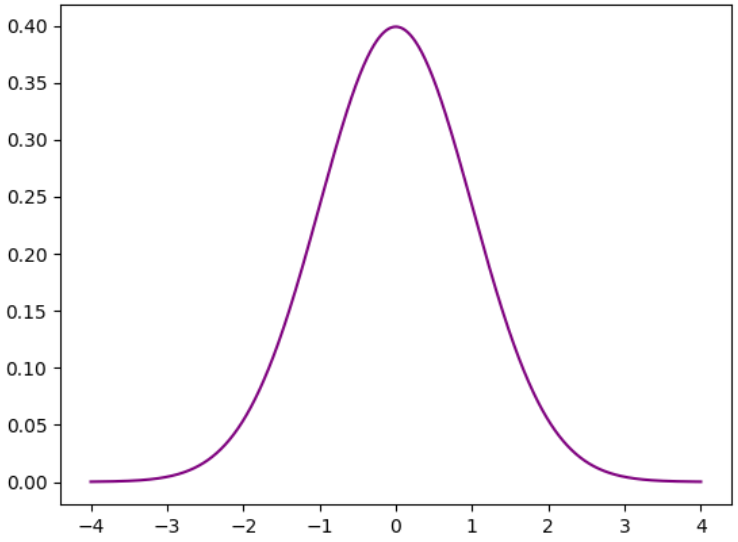
Abstract Space

The **probability** (density) function expresses the probability of observing the data given the model

Observable Space

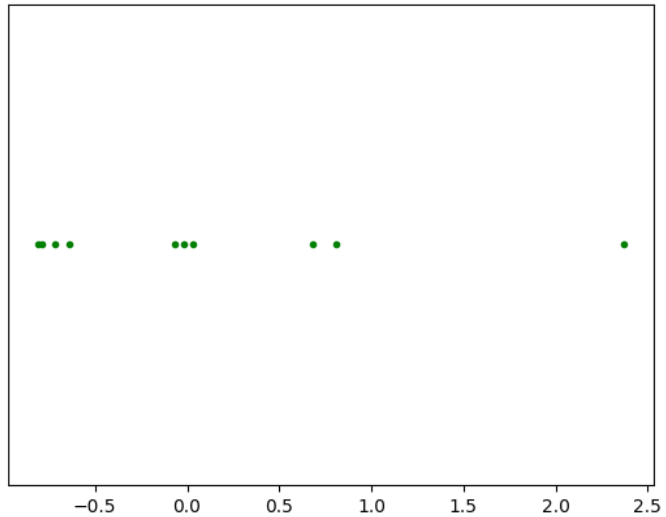
Choose model parameters

$$P(x|\mu = 0, \sigma = 1)$$



Gives Probability P(x)

x is a random variable



Joint probability of observing n data: $\prod_i p(xi)$

And back...?

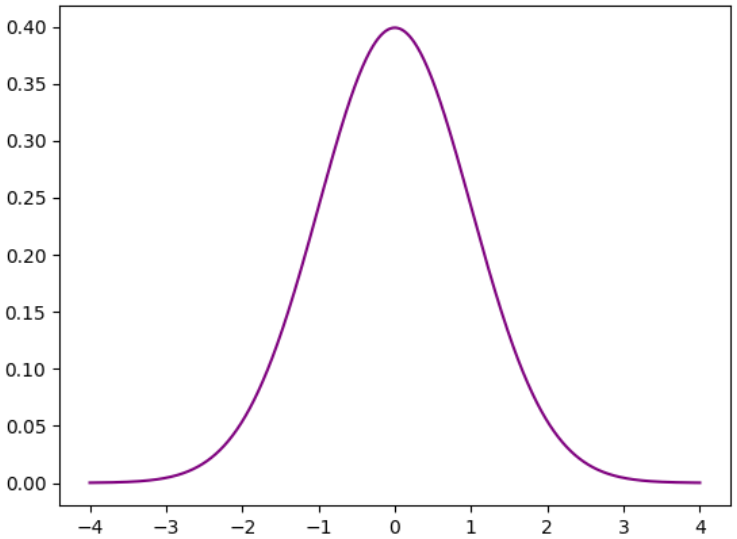
Abstract Space

Observable Space

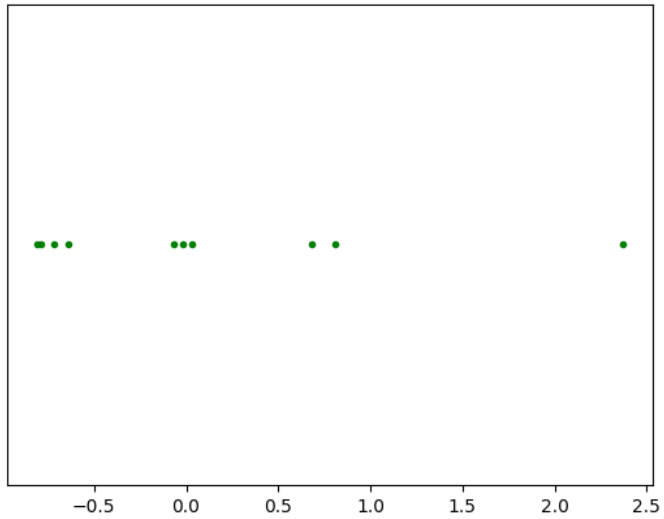
Given a set of observations x

Estimate model parameters

$$P(x = \{1.2, -0.7, 0.3, ...\} | \mu, \sigma)$$



Here P has taken the role of a **likelihood**!
i.e. the probability viewed as a function of its parameters



→ The likelihood allows us to make statements about the model given data

Estimator Variance

A second metric is the **variance of the estimator:**

- spread of the estimator around its expectation value
- generally lower-variance is preferred over high variance

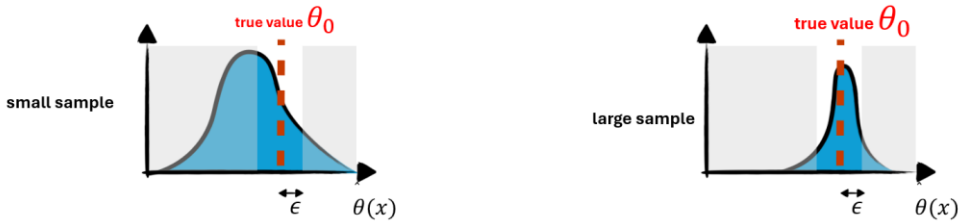
$$\sigma_\theta = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

Consistency

A desirable property is that estimators are **"consistent"**

- more data provides you a better estimate on average
- estimation value probability accumulates close to the true value

$$\lim_{n\to\infty} p(|\hat{\theta}(x) - \theta_0| > \epsilon) = 0; \forall\epsilon$$

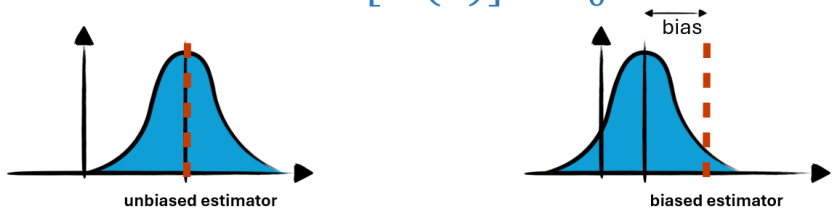


Estimator Bias

A key metric is the **bias of the estimator:**

- deviation of the expectation value of $\hat{\theta}(x)$ from the true value
- generally people prefer unbiased estimators

$$b = \mathbb{E}[\hat{\theta}(x)] - \theta_0$$



Maximum Likelihood

An intuitive way to find a good point is to find the parameter $\hat{\theta}\,(x)$ that **maximizes the probability to observe the data we got:**

$$\hat{\theta}_{MLE}(x) = \text{argmax}_{\theta} p(x|\theta)$$

$\hat{\theta}_{MLE}(x)$ **is called the Maximum-Likelihood Estimator of** θ

Asymptotically Efficient

MLE estimators asymptotically **saturate the Cramér-Rao bound**:

- i.e. achieve the minimum possible variance of all unbiased estimators

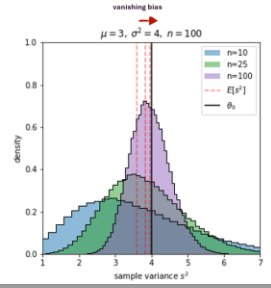
$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d \mathcal{N}(0, \text{var}\hat{\theta})$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d \mathcal{N}(0, I^{-1}(\theta))$$

Inverse Fisher Information

Asymptotically Unbiased

Sample variance is a **biased** MLE estimator, but $\mathbb{E}[s^2]$ moves towards the true value for large samples and the **bias vanishes = asymptotically unbiased**

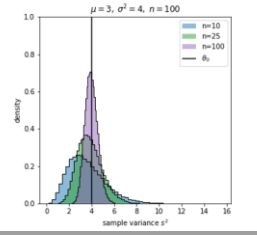


77

Asymptotic Consistency & Normality

We've seen this already for the Gaussian sample variance

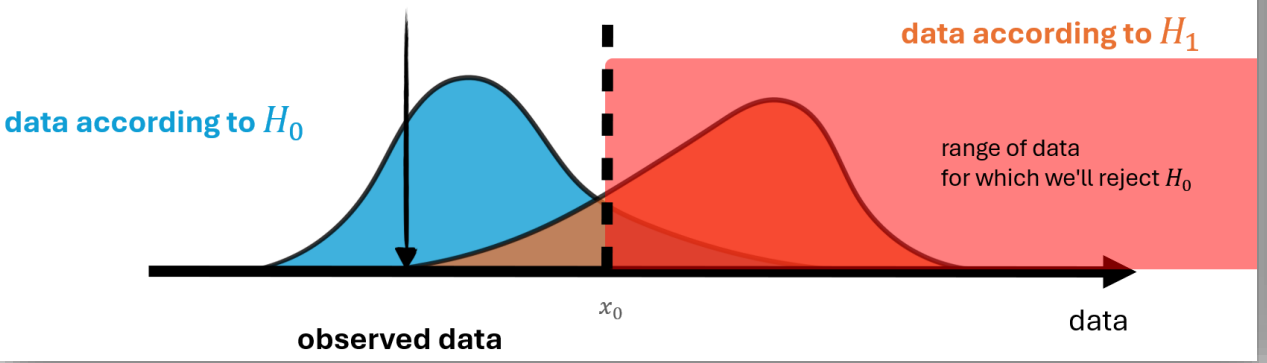
- while the finite-sample $\hat{\theta}$ distributions may not be Gaussian it will progressively be "normalized" (again, CLT)



Testing with Sampling Distributions

A reasonable answer:

- reject H_0 if data is too far too the right (e.g. data > x_0)



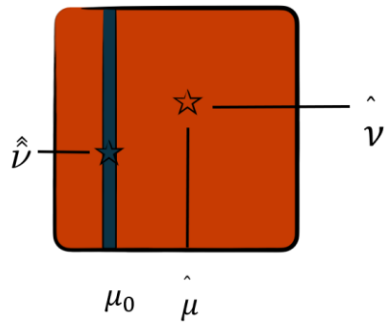
L'hood-Ratio Tests or nested Hypotheses

For the "nested hypotheses" case this is also called the
profile-likelihood ratio statistic

optimal $\mathcal{V}$-values at fixed value of μ

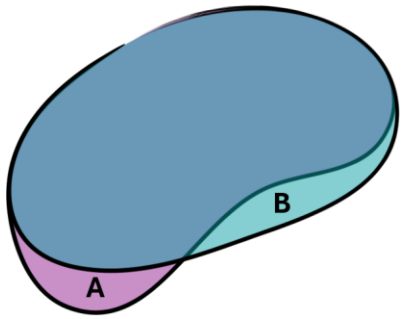
$$\lambda_{\mu_0}(x) = -2\log\frac{L(\mu_0, \hat{\hat{\nu}})}{L(\hat{\mu}, \hat{\nu})}$$

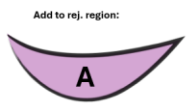
globally optimal values

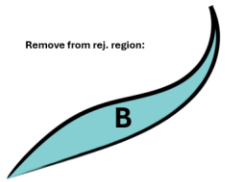


151

Visual Proof of NP-Lemma







① $p(A|H_0) = p(B|H_0)$ test size should stay constant

② $p(B|H_1) > t_\alpha p(B|H_0)$ removed region is over threshold k_α

③ $p(A|H_1) < t_\alpha p(A|H_0)$ added region is under threshold k_α

③ ① ②

$$p(A|H_1) < t_\alpha p(A|H_0) = t_\alpha p(B|H_0) < p(B|H_1)$$

new region has less power than NP-region!

132

Sampling Distribution of LRT

One of the biggest advatages of the LRT: it's not only intuitive & optimal but also its sampling distribution are asymptotically known for nested hypotheses!

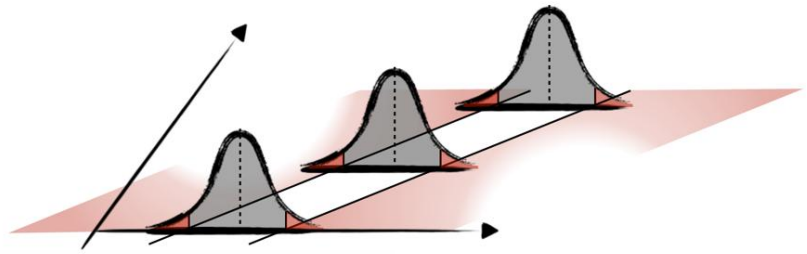
The LRT statistic for r parameters of interest is follows asymptotically

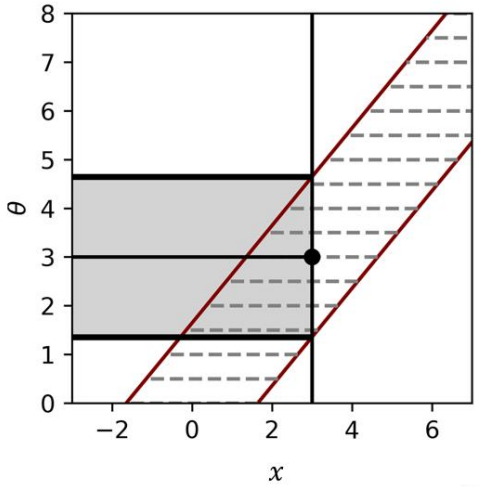
the non-central χ^2 distribution with r degrees of freedom

$$p(t_\mu|\mu') = \chi^2_{\text{nc}}(n, \Lambda^2_{\mu'})$$

Band plot walk-through: Step 3

- Read of corresponding interval in θ

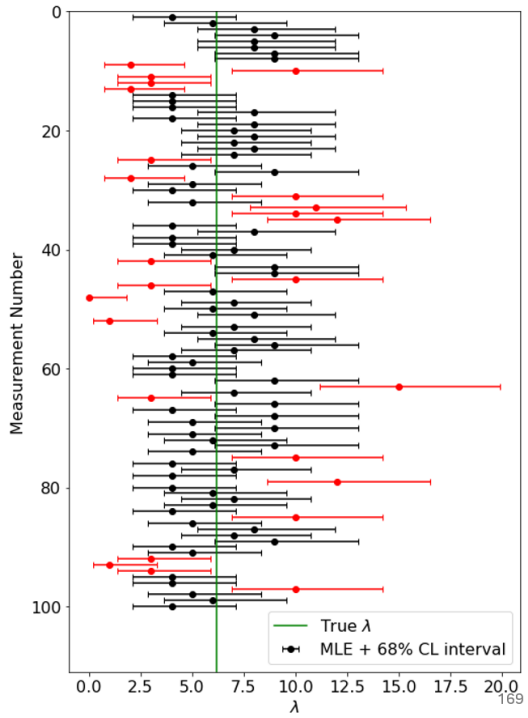




Interpretation:

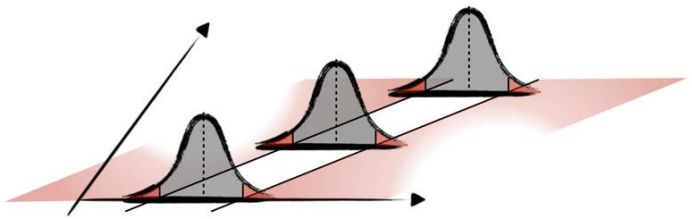
In repeated experiments, the CL interval contains the true value at least $1 - \alpha$ of times

- Here: 100 Poisson experiments with $\lambda = 6.2$
- Slight *overcoverage*
  - Actually only 25% measurements fall outside
  - Typical problem for discrete distributions

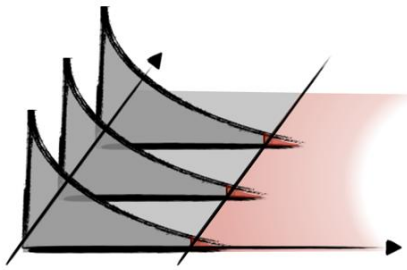


What do the Rejection Regions look like?

- When taking the LRT ratio test, the null hypothesis distribution is always the same, regardless of θ:
  → the χ^2 distribution (Wilk's Theorem)



Intervals in x



Intervals in $\lambda_\theta(x)$

Standard 1σ intervals

Common Interval (with 68% coverage) achieved for the following rejection region

$$\lambda_\theta(x) = -2(\mathrm{LL}(\theta_{\mathrm{lim}}) - \mathrm{LL}(\hat{\theta})) = 1$$

$$\mathrm{NLL}(\theta_{\mathrm{lim}}) - \mathrm{NLL}(\hat{\theta}) = \frac{1}{2}$$

→ Based on a chi2 with d.o.f.=1 (asymptotic)

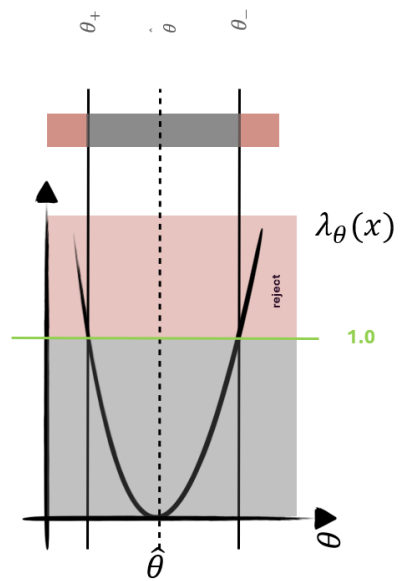


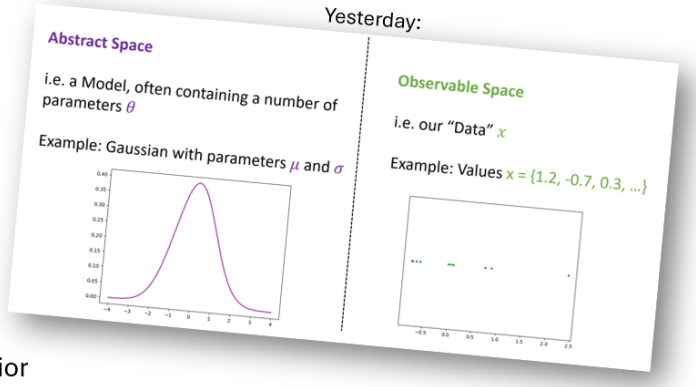
Table 9.5 The values of the quantile Q_γ for different values of the confidence level $1-\gamma$ for $n = 1, 2, 3, 4, 5$ fitted parameters.

$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Cowan, p. 135

In the context of Models and Data

- A = our abstract model M with parameters θ
- B = our data x



Likelihood

Posterior

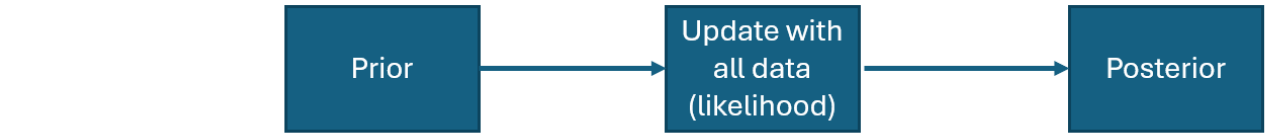
Prior

$$p(\theta|x,M) = \frac{p(x|\theta,M)p(\theta|M)}{p(x|M)}$$

Evidence (Marginal Likelihood)

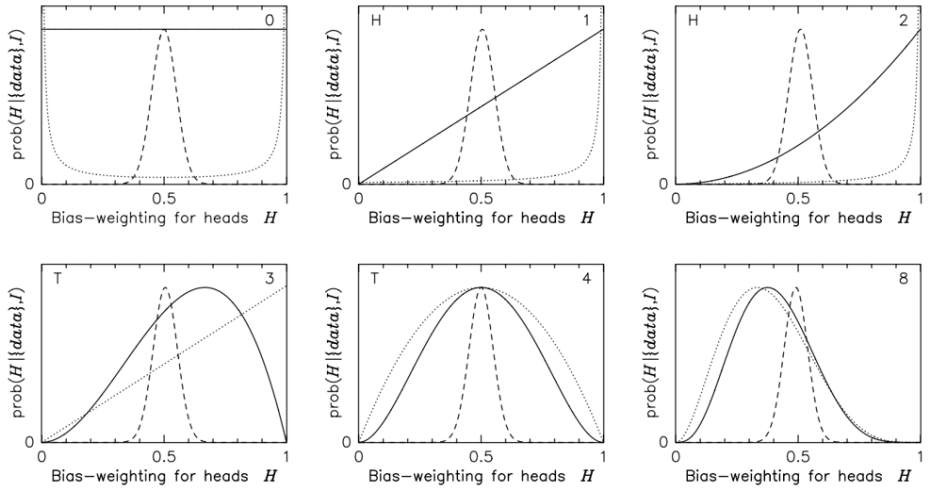
Update of knowledge

- What if instead of analyzing all data at once, we successively performed one coin toss after the other, and use the posterior from one as the prior of the next?



Effect of priors

- If we only analyze a few coin tosses, the effect of the prior is large



Conjugate Priors

In general, if we use a beta distribution as the prior we get out another beta distribution as the posterior

→ The beta distribution is a *conjugate* prior to the Binomial distribution

Other examples:

- Gamma distribution is the conjugate prior for a Poisson Likelihood
- Dirichlet distribution is the conjugate prior for a Multinomial Likelihood
- Normal* distribution is the conjugate prior for a Normal Likelihood

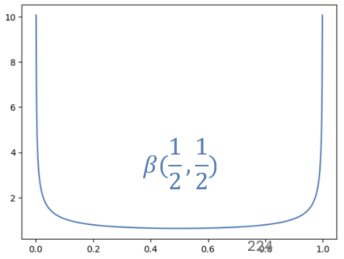
Jeffreys Prior

- The Jeffreys prior is intended to be a non-informative prior distribution
- It is constructed from the Fisher information I (→ see earlier Lecture)
- Jeffreys prior $p(\theta) \sim \sqrt{\det I(\theta)} = \sqrt{\mathbb{E}[(\partial\theta \log p(x|\theta))^2]}$

- Example for our Binomial:
  - $I(p) = -E[\partial^2 plogp] = \frac{Np}{p^2} + \frac{N-Np}{(1-p)^2} = \frac{N}{p(1-p)} \sim p^{-1}(1-p)^{-1}$

  → $P_{Jeffreys}(p) = \sqrt{I(p)} \sim p^{-1/2}(1-p)^{-1/2}$

  Which is a beta distribution $\beta(\frac{1}{2}, \frac{1}{2})$

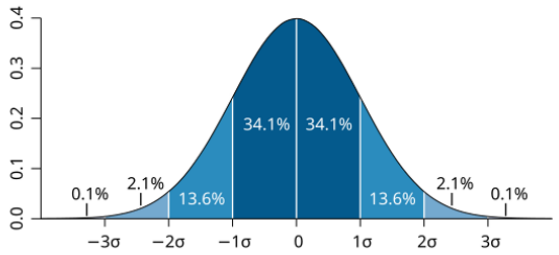


Summary of posterior

- We can construct “credible intervals” from our posterior to summarize it
  - Calculate intervals in μ that contain a desired amount of probability,

  For instance the central or smallest intervals:

Probability Content (in %)	μ range
68.3	$x \pm \sigma$
90.0	$x \pm 1.65\sigma$
95.0	$x \pm 1.96\sigma$
99.0	$x \pm 2.58\sigma$
99.7	$x \pm 3\sigma$



These Bayesian credible intervals are often abbreviated as “C.I.”, as opposed to the Frequentist confidence level intervals “C.L.”

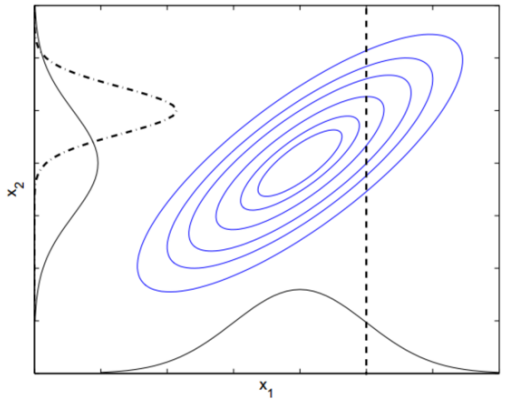
Marginal distribution

- We can integrate over an “unwanted” parameter
  - $p(\mu|\{x\}) = \int p(\mu, \sigma|\{x\})d\sigma$
- (and vice versa we could integrate over μ to have the marginal posterior of σ)

Note that what we did last lecture was to calculate $p(\mu|\{x\}, \sigma)$, which is also a posterior for μ, but it is conditional on a particular choice of σ

→We speak of the “**marginal**” distribution if nuisance parameters take into account our prior ignorance

→We speak of the “**conditional**” distribution if nuisance parameters are set to fixed values



Posterior ratio

$$\frac{p(A|D)}{p(B|D)} = \frac{p(A)}{p(B)} \times \frac{p(D|A)}{p(D|\lambda_0, B)} \times \frac{\lambda_{max} - \lambda_{min}}{\delta\lambda\sqrt{2\pi}}$$

Prior preference for models
Can be set to 1 for example

Likelihood ratio of the best fit of the models to the data (MLEs) – "goodness of fit"

Ockham factor: penalty for introducing additional degrees of freedom into a theory

- The likelihood ratio alone will always favour the model with the closer fit to data
- The Ockham factor penalizes more complex theories (Ockham's razor)
- Large prior ranges penalize a theory
- Small $\delta\lambda$ penalize a theory (→ only a very narrow range of parameter values are compatible with the data)

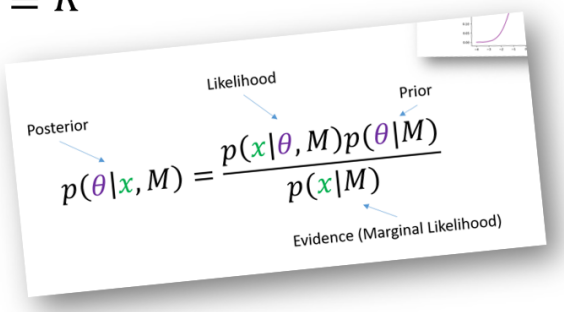
Bayes Factor

When we assign equal prior weight to either model, as discussed the prior ratio cancels out

$$\frac{p(A|D)}{p(B|D)} = \frac{p(D|A)}{p(D|B)} \equiv K$$

Where we define K as the Bayes factor

This is the ratio of the **Evidence** under each model!



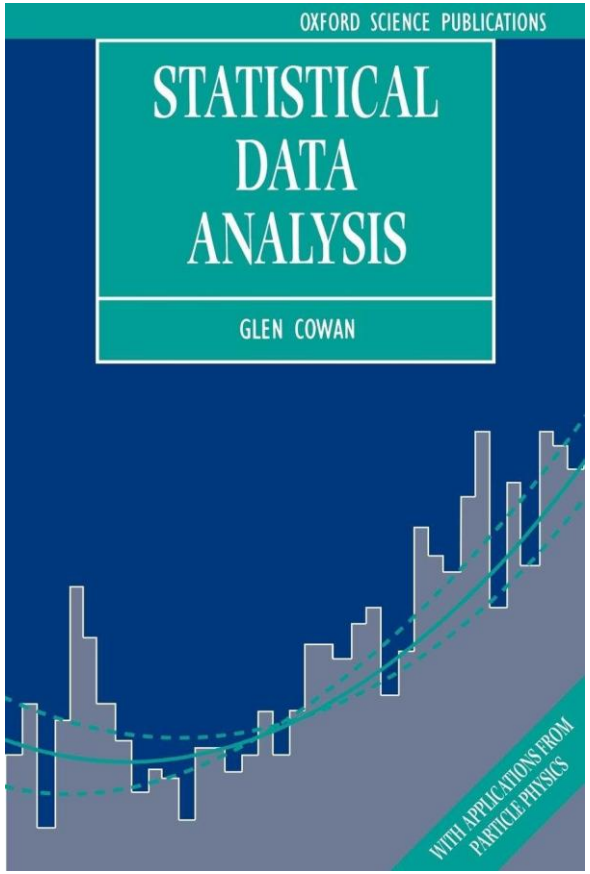
[1]

$\log_{10} K$	K	Strength of evidence
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Concluding Remarks

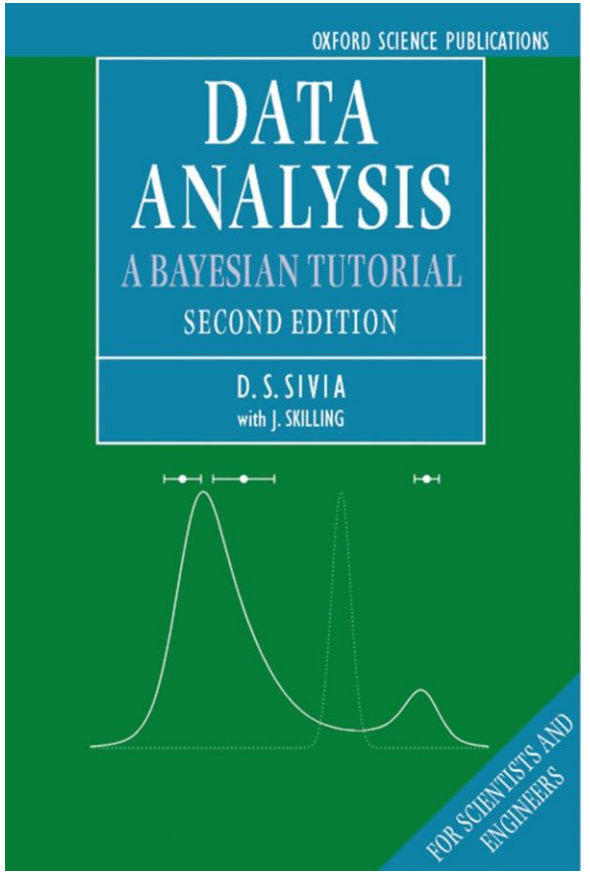
- Whether you prefer the Frequentist or Bayesian school is mostly personal taste
  - Both ways are statistically sound and correct at the same time!
  - They simply differ in the questions they address
- Most importantly, try to understand what you are doing, instead of blindly following some pre-digested recipes
- Often, the crux is in the correct interpretation of the results
  - An „error bar“ or a „probability“ can have many different meaning
- Everything is probably best learned hands-on, applied to an actual data analysis problem (→ your own research!)

Further Reading



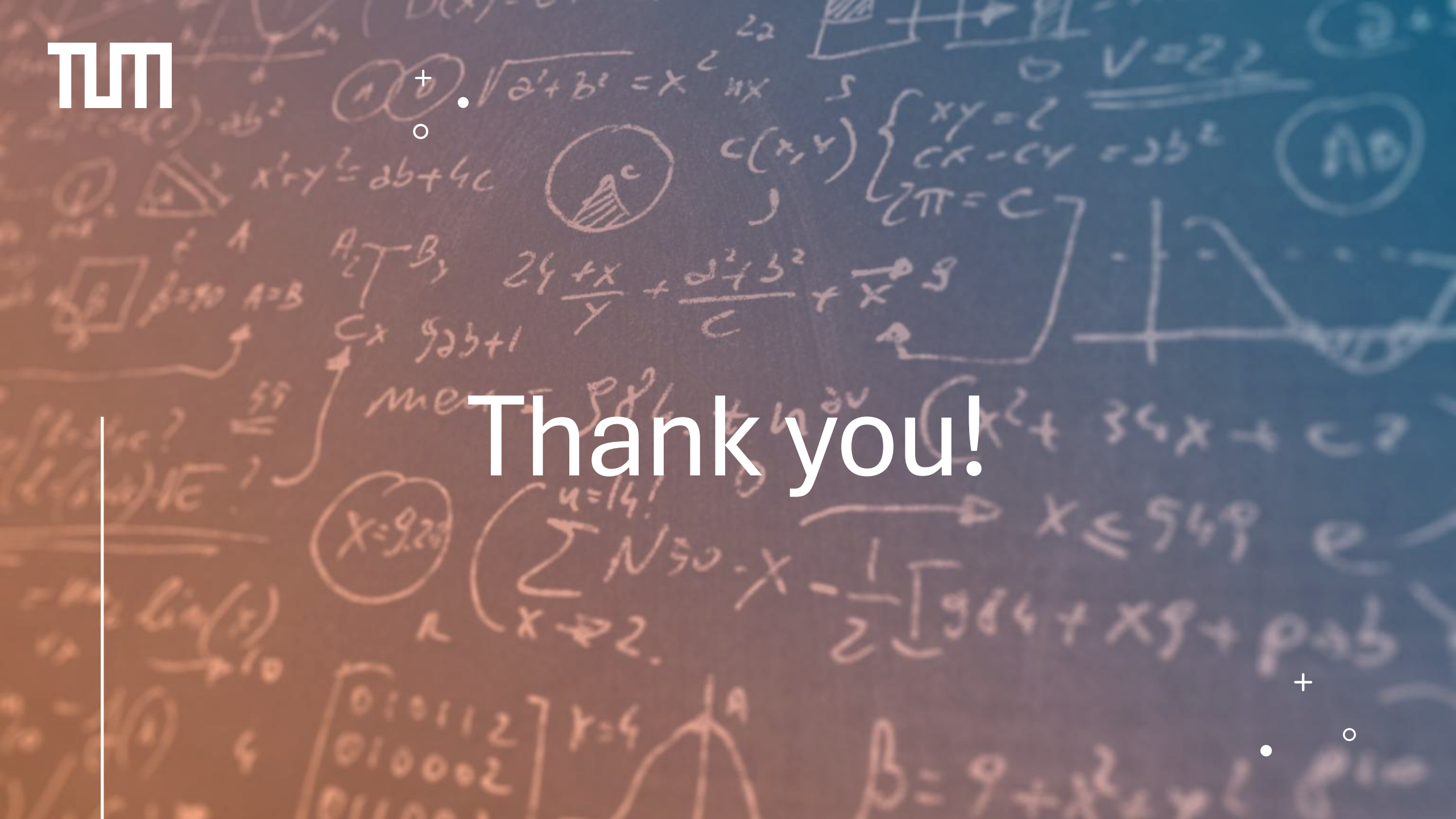
Statistical Data Analysis
by Glen Cowan

- Focused on Frequentist Inference
- Written by and for particle physicists
- Contains many examples connected to particle physics



Data Analysis
by D. S. Sivia

- Focused on Bayesian Inference
- Written by a particle physicist for particle physicists
- Some examples in this lecture were taken from this book
- Contains very didactical examples


Thank you!