



# Statistics for Experimental Physicists

20th School for Astroparticle Physics October 7 – 15, 2024 in  
Obertrubach-Bärnfels in the countryside near Erlangen

Philipp Eller (TUM)

[philipp.eller@tum.de](mailto:philipp.eller@tum.de)

# A working example

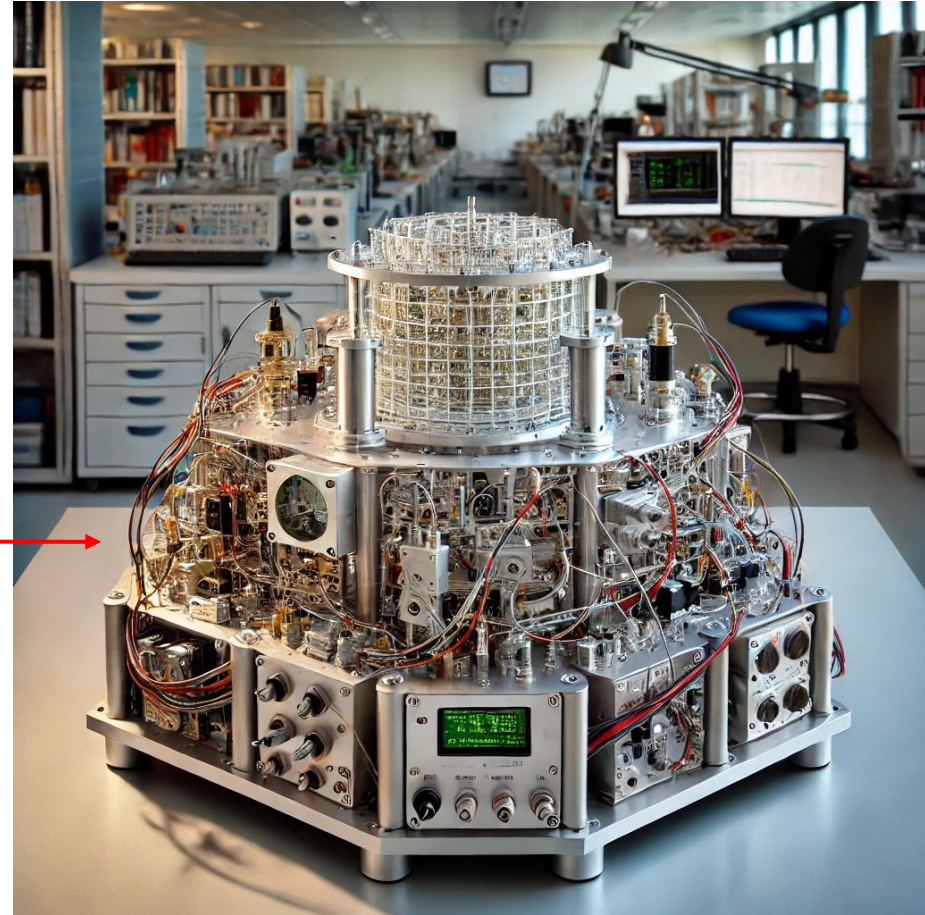
Reactor



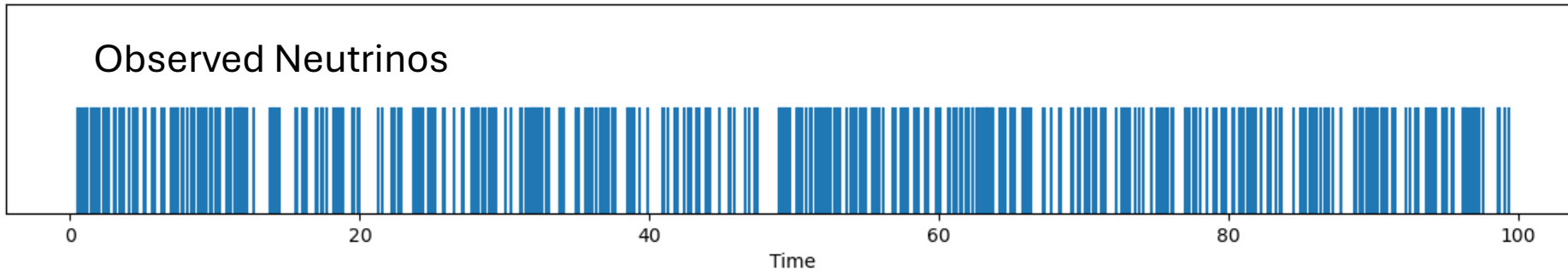
$\nu$



Neutrino Detector (according to GPT)



Observed Neutrinos

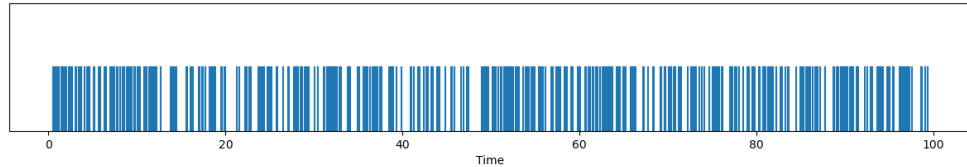


# Your job as the scientist:

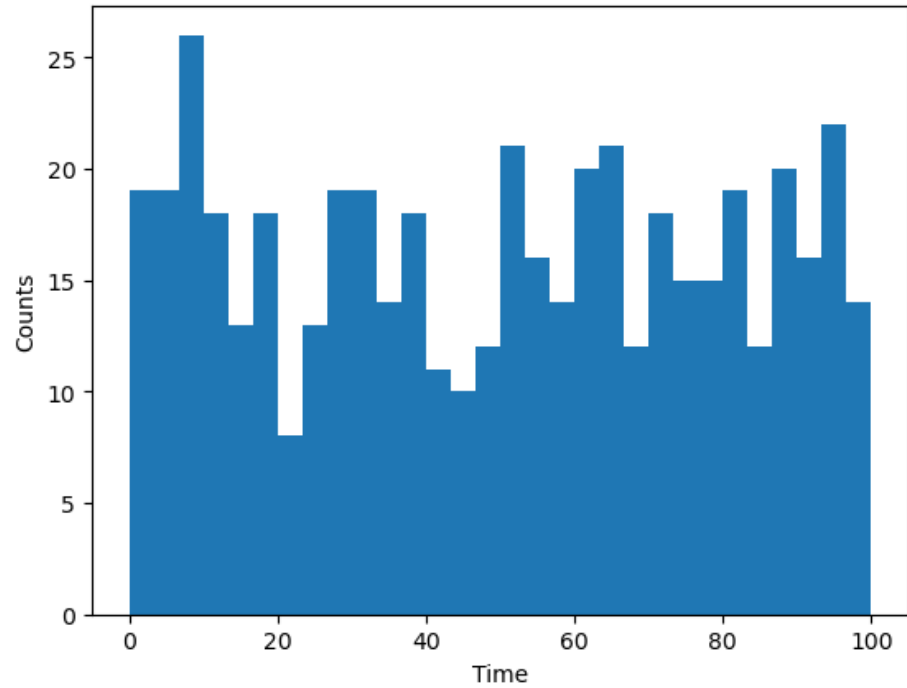
1. Monitor the rate
2. Tell if the rate suddenly changed, i.e. if anyone removed or add nuclear fuel



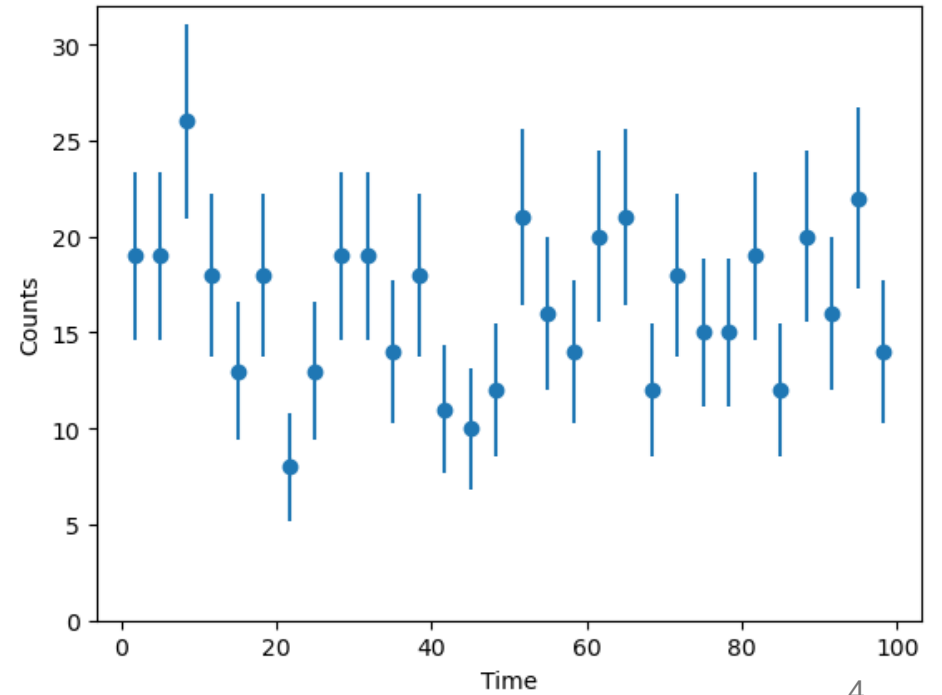
# Naive Analysis



Histogram



Add Errors



# Calculating Averages...

How to correctly average values including their errors?



To correctly average values that have associated errors (uncertainties), you must use **weighted averages**, where each value contributes according to how certain (or uncertain) it is. This approach ensures that values with smaller errors (greater certainty) have more influence on the final average than those with larger errors (less certainty). Here's a step-by-step guide on how to do this:

## 1. Given data

Let's assume you have  $n$  measured values  $x_1, x_2, \dots, x_n$  with their associated uncertainties (errors)  $\sigma_1, \sigma_2, \dots, \sigma_n$ .

## 2. Weighted average formula

The weighted average  $\bar{x}$  is given by:

$$\bar{x} = \frac{\sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}$$

Where:

- $x_i$  are the individual measured values.
- $\sigma_i$  are the uncertainties in the individual measured values.

This formula assigns more weight to values with smaller uncertainties.

## 3. Uncertainty in the weighted average

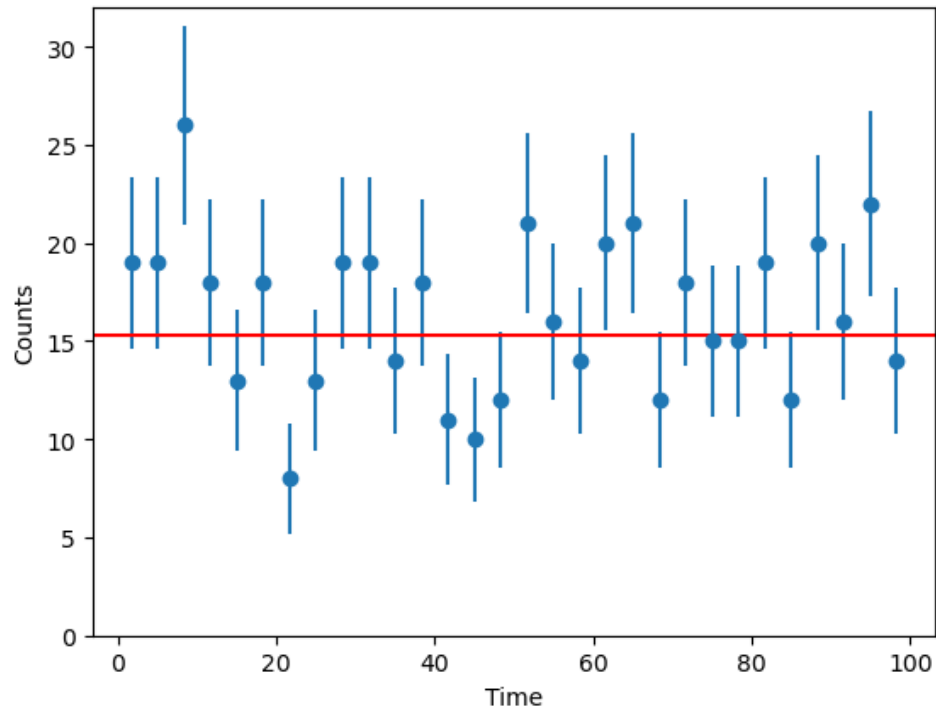
The uncertainty (error)  $\sigma_{\bar{x}}$  in the weighted average is calculated as:

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}}$$

This uncertainty reflects how precisely the weighted average is determined, taking into account the uncertainties of the individual values.

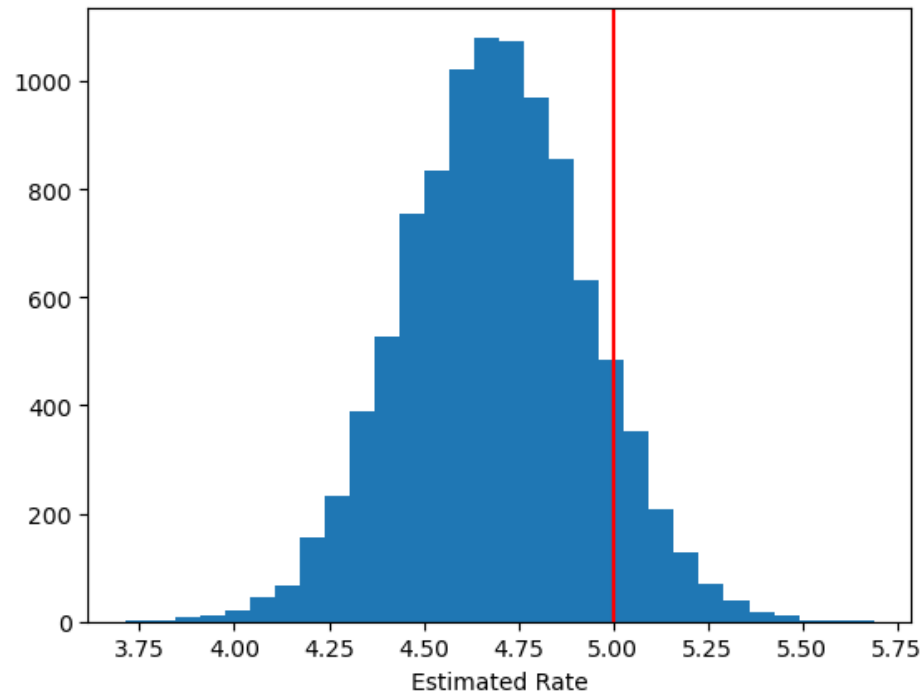
# Result

- Estimated rate:  $4.599 \pm 0.214$
- The rate from the theory calculation should be 5...checks out!



# Problems arise...

When you run this procedure over and over on new data, you find that your predictions are biased:



You are getting slightly puzzled....but it's probably a systematic uncertainty somewhere....or the theory guys are just wrong

# Anyway...

The rate is a little off, but you still carry on to figure out whether it's constant...

You test your constant model with a goodness-of-fit test

- Evaluating this gives a p-value of 25.8 %, which looks pretty good, right?

## 1. Chi-Square Goodness-of-Fit Test

This test is suitable for binned data, like histograms.

Steps:

### 1. Formulate Hypotheses:

- Null Hypothesis ( $H_0$ ): The observed data follow the expected distribution.
- Alternative Hypothesis ( $H_A$ ): The observed data do not follow the expected distribution.

### 2. Calculate the Expected Frequencies:

- Determine the expected frequency  $E_i$  for each bin based on the theoretical distribution.

### 3. Calculate the Chi-Square Statistic:

- For each bin  $i$ , calculate:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- $O_i$  = Observed frequency in bin  $i$ .
- $E_i$  = Expected frequency in bin  $i$ .

### 4. Determine the Degrees of Freedom:

- Degrees of freedom = (Number of bins - 1) - (Number of estimated parameters).

### 5. Compare with Critical Value or p-value:

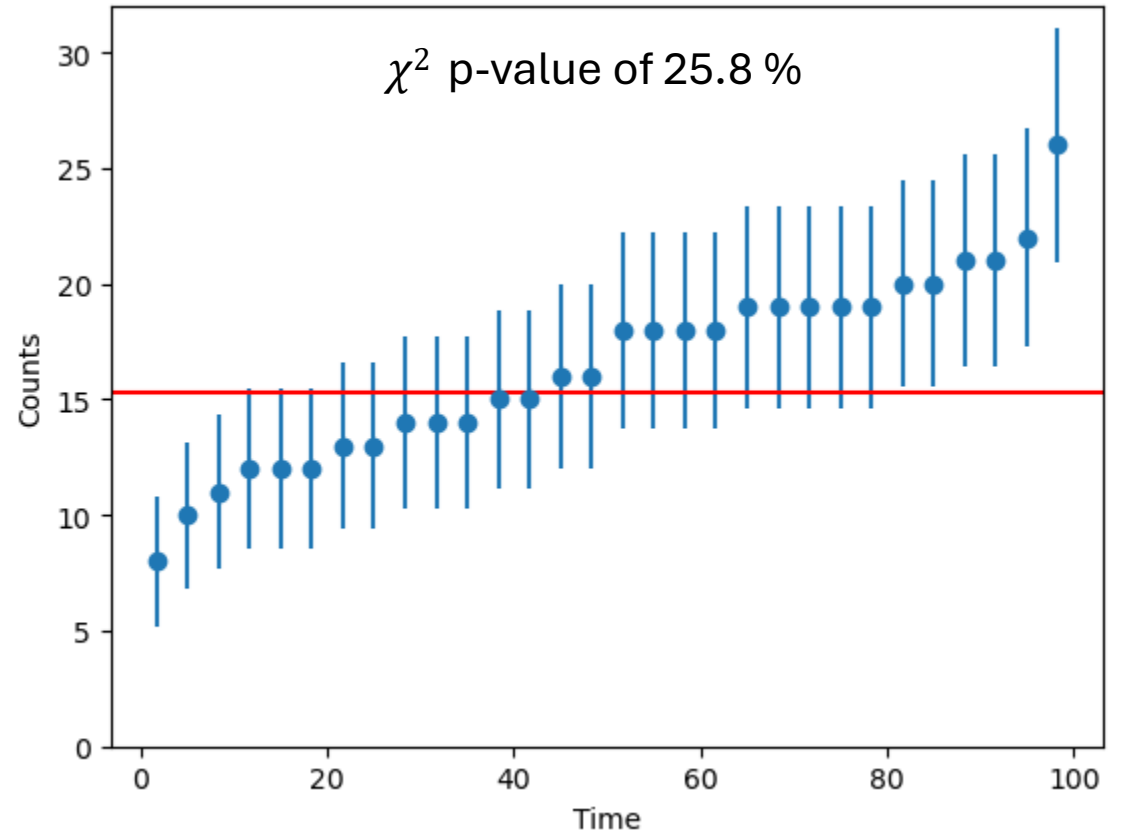
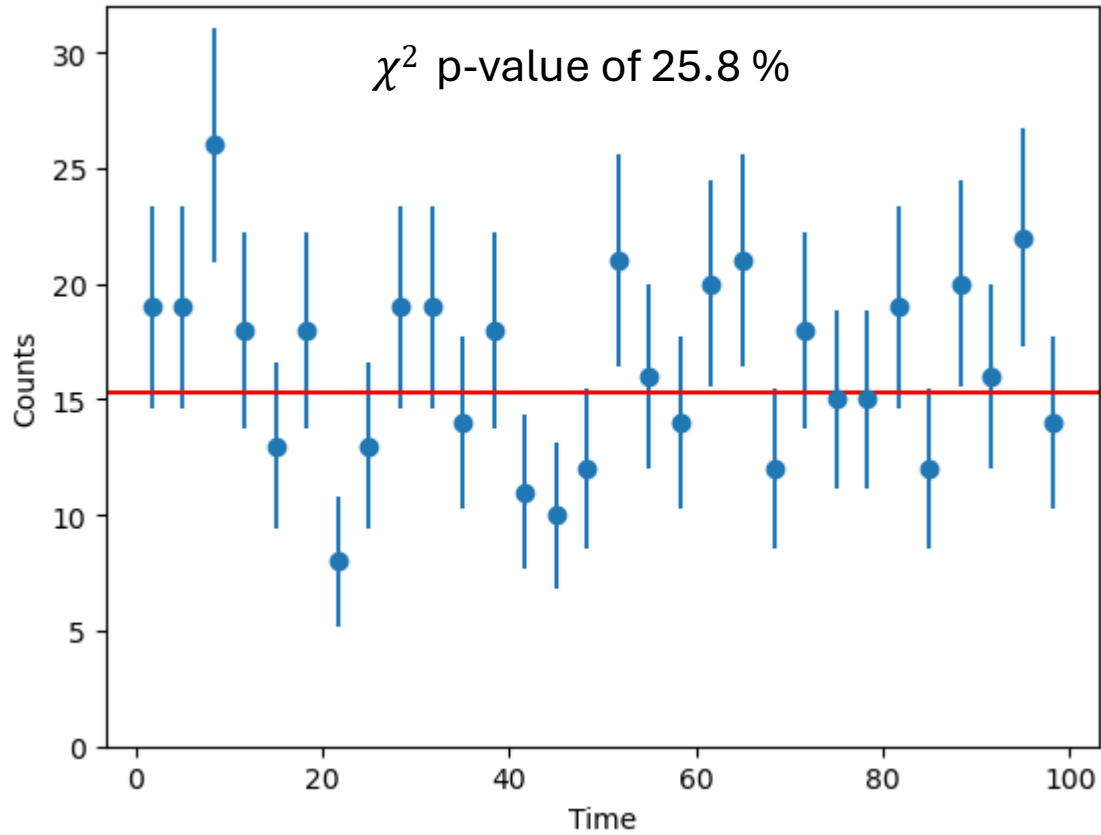
- Use a Chi-Square distribution table to find the critical value at a chosen significance level (e.g., 0.05).
- Alternatively, calculate the p-value and compare it with the significance level.

### 6. Conclusion:

- If  $\chi^2$  is greater than the critical value or p-value is less than the significance level, reject  $H_0$ .



# Identical $\chi^2$ p-value



# ?!?

- Our estimated rate is systematically low 😞
- Our test cannot really tell the difference between a constant rate and something else 😞

→ We're now sufficiently confused, and conclude that we need to take a statistics course



# ~ A Statistics Course ~



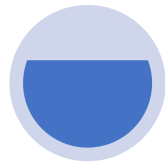
# Overview

---



## Today

Basic Properties  
Probability vs.  
Likelihood  
Estimators Theory  
/ Point Estimators  
Principle of  
maximum  
likelihood



## Tomorrow

Hypothesis  
testing  
Size / Power, p-  
values, etc.  
Constructing  
Confidence  
Intervals



## Thursday

Bayes' Approach  
Closed form  
solution /  
conjugate priors  
etc  
MCMC for  
numerical  
solution

# Probability & Basic Properties

# Probability

Most things in the world are not certain. Yet, we are used to reason about in our everyday lives

**During Games: It's rare (but not impossible) to roll two sixes**

**Forecasting: It's likely we will see a recession**

**Science: The data points towards the existence of a Higgs**

# 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 dice) in, 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

# Conditional Probability

Most commonly, we encounter probabilities of an outcome  $x$  that depend on certain parameters  $\theta$

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

We say „p of  $x$  given  $\theta$ “, and it is also called „conditional probability“

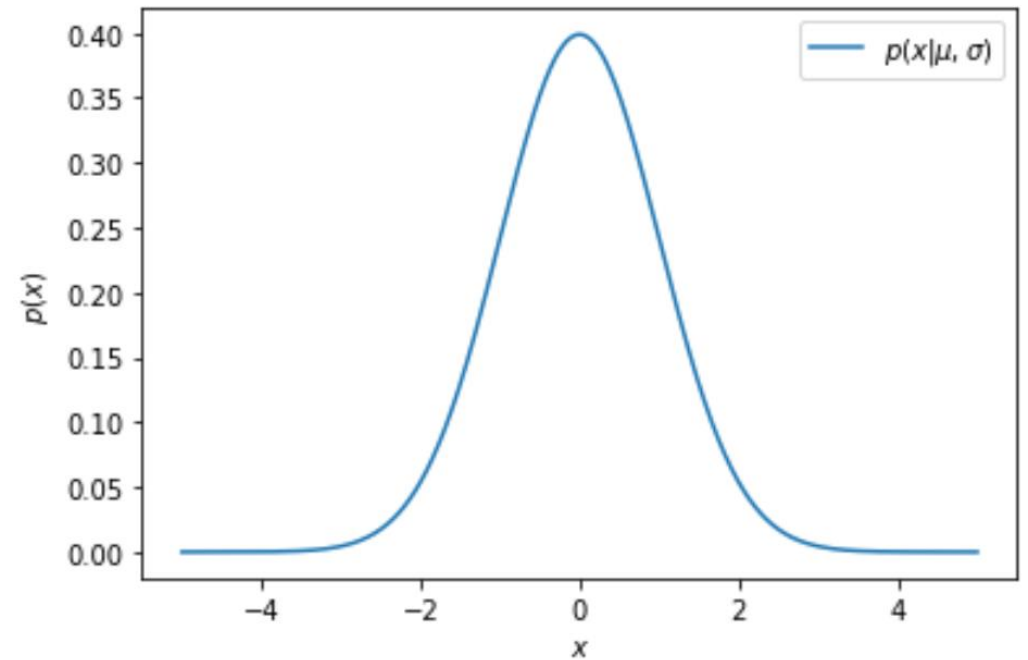
We speak of a „parametric model“ if we can express this in the form of distributions, or at least a data-generating process (forward model)

- Sometimes this is a simple distribution with one, two parameters
- ... sometimes a very complicated model with many 100s of parameters

# Examples of Parametric Models

**Normal Model:**  $\theta = (\mu, \sigma)$

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



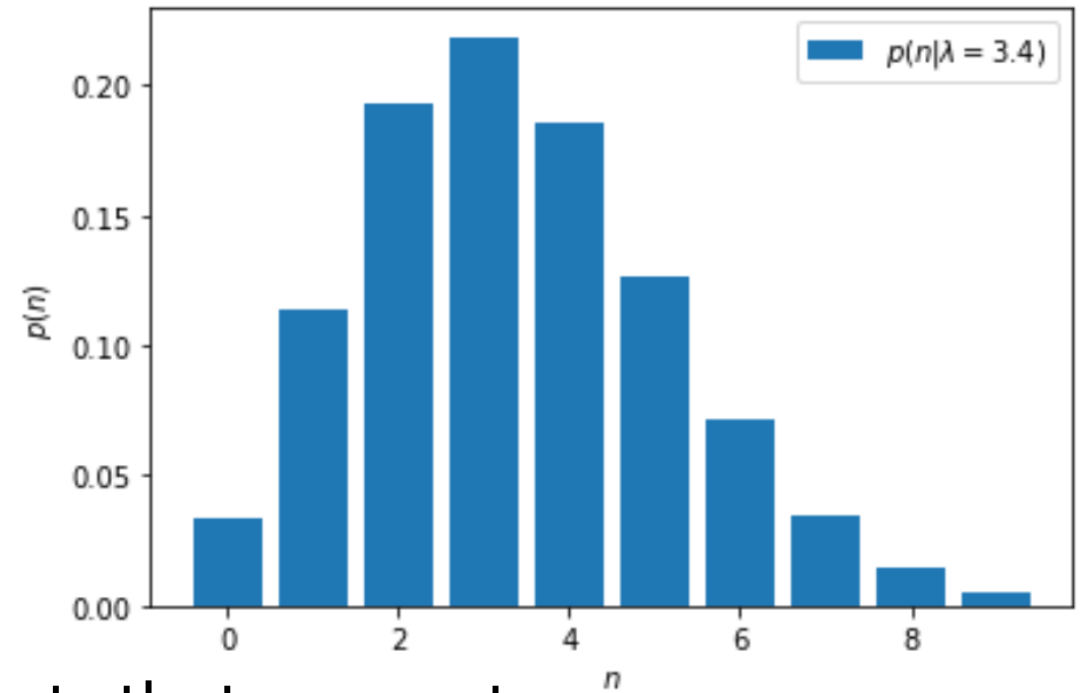
Occurs naturally if the measured quantity is the sum of many individual random processes (e.g. measurements on a detector module)

→ **"Central Limit Theorem"**

# Examples of Parametric Models

**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.

# Other Distributions

- What distributions do you know / have you heard of?
  
- See for example:  
[https://en.wikipedia.org/wiki/List\\_of\\_probability\\_distributions](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

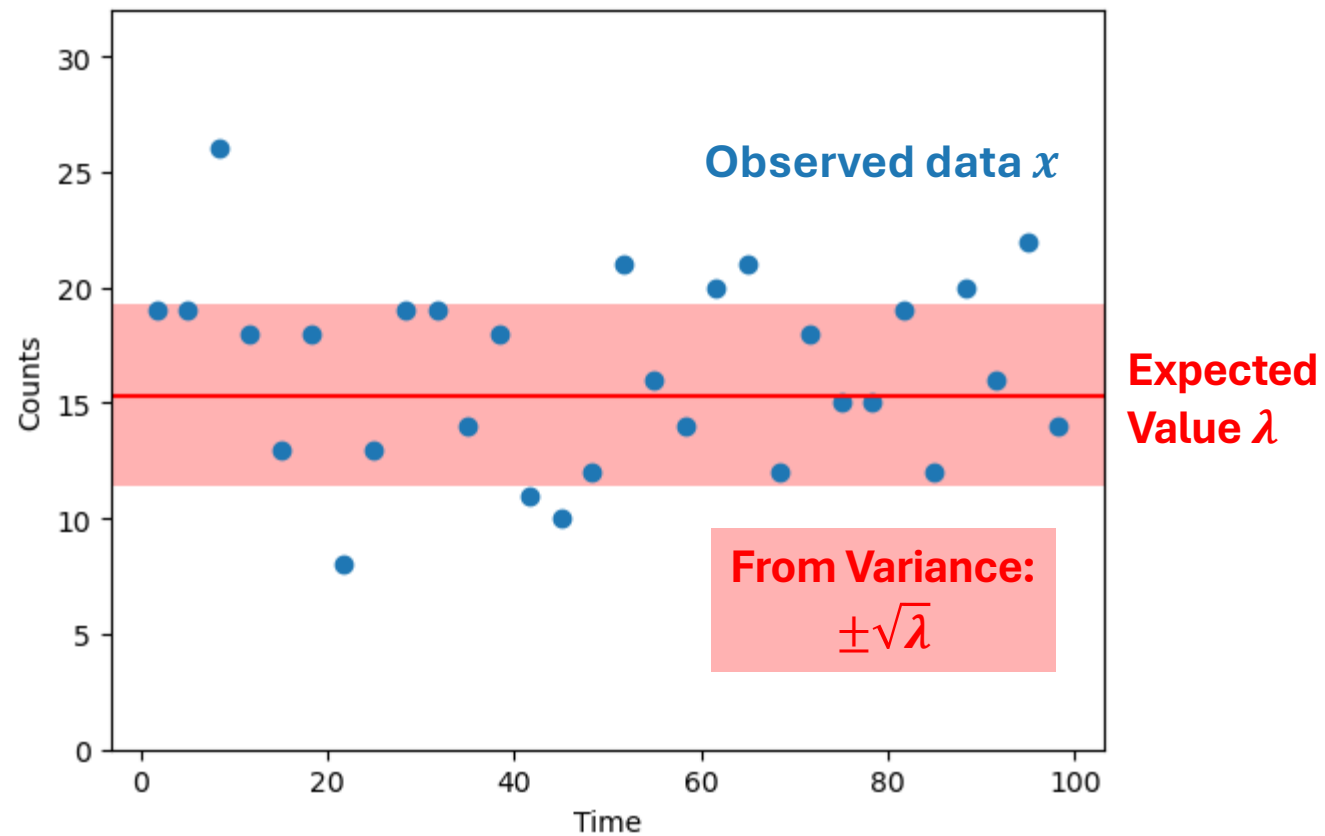
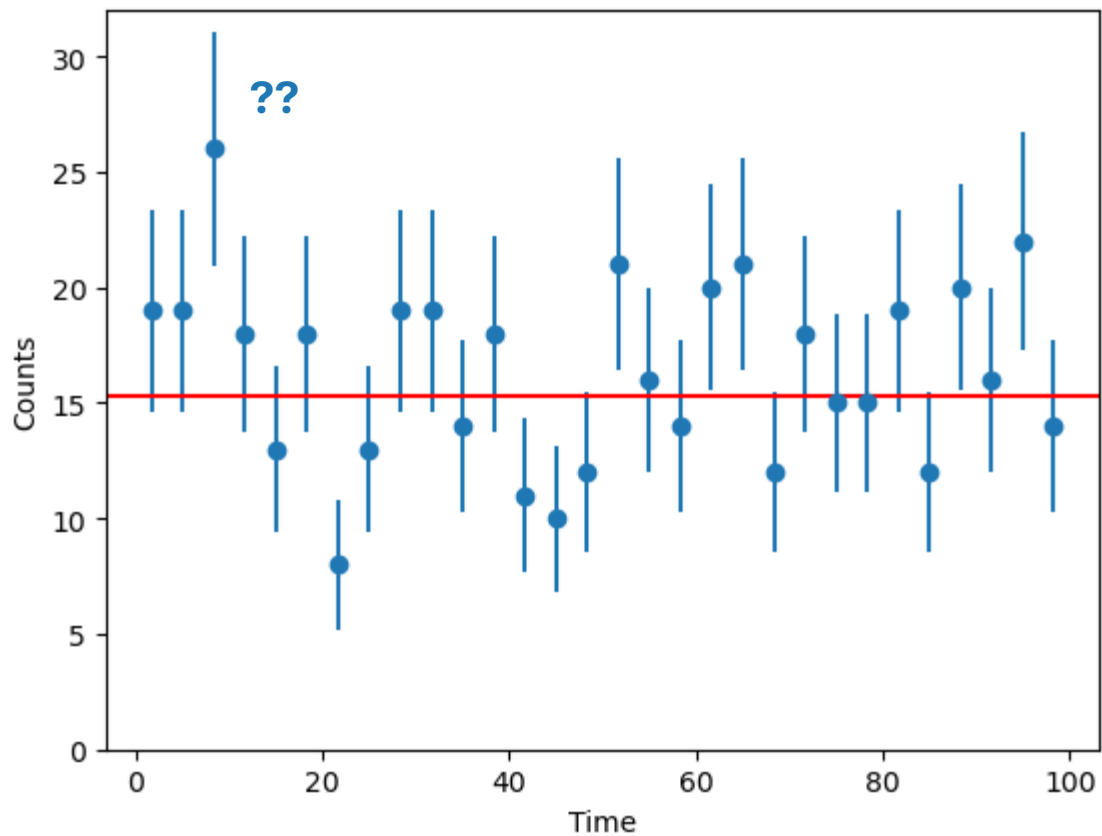
$$Var(x) = E[(x - E[x])^2]$$

→ For Poisson:  $Var(x) = \lambda$

# 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  $p$  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  $x = F^{-1}(y)$

# Fixing the 1<sup>st</sup> problem: our plot





# Likelihood Function

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

**The likelihood function:**

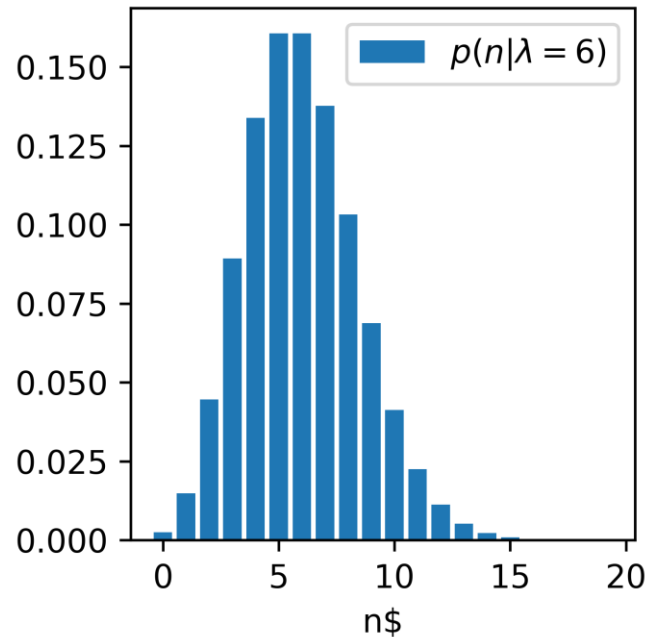
$$L_x(\theta) = p(x|\theta)$$

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

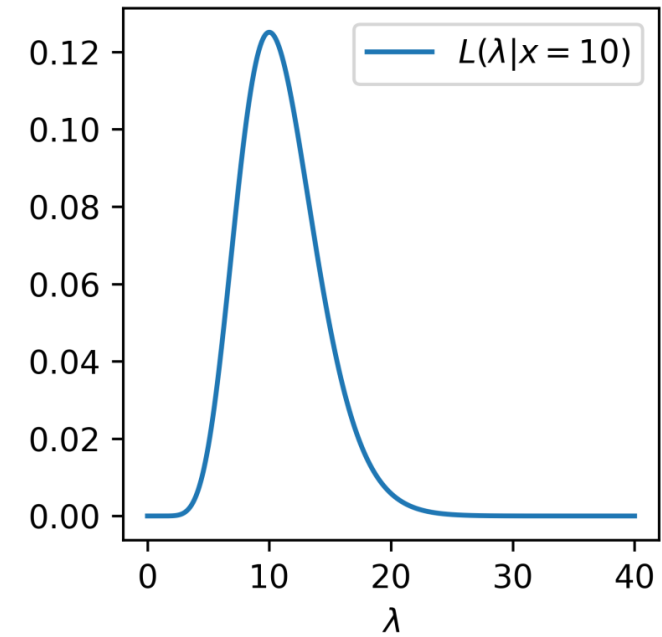
This is **very different from a probability!** We did not specify what  $\theta$  is, and it cannot be assumed these are random variables

# Example: Poisson Likelihood

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



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

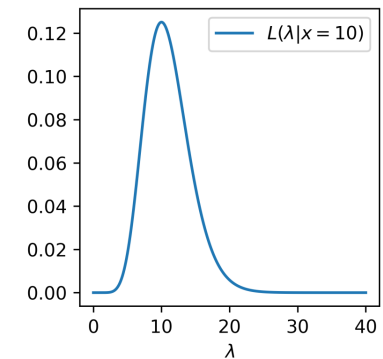
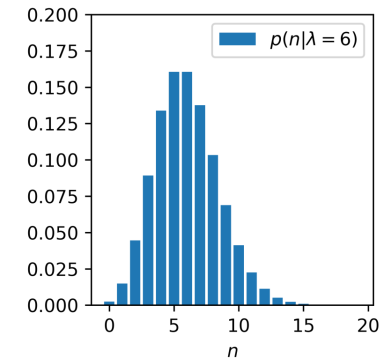


likelihoods for  
observations for  
a fixed  $n = 10$

# Likelihood vs. Probability functions

Important to remember that likelihood and probability functions are different

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



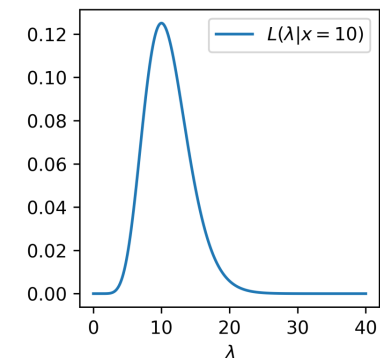
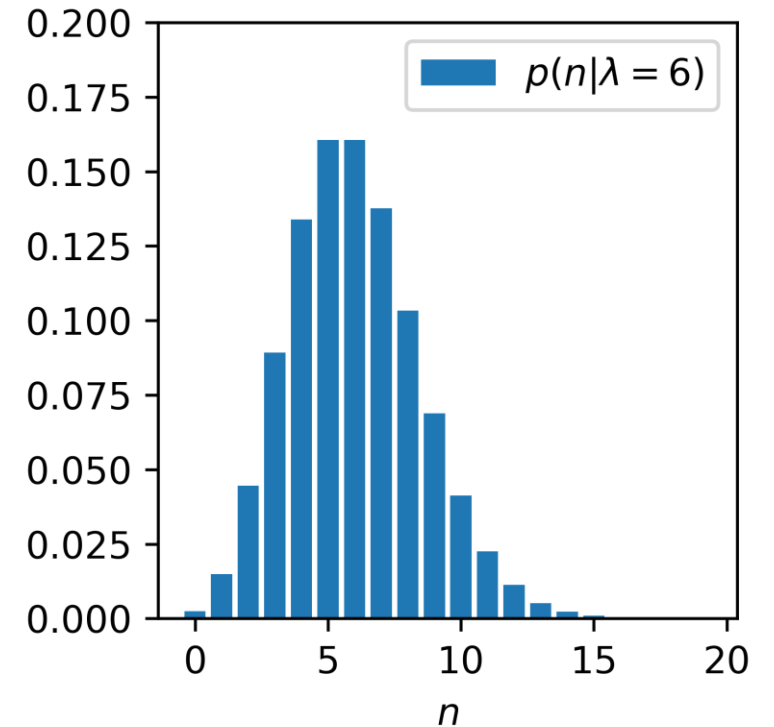
# Likelihood vs. Probability functions

Important to remember that likelihood and probability functions are different

## Probability:

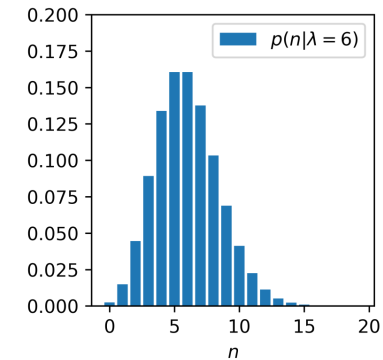
fixed parameters  $\theta$ , variable data  $x$

normalized  $\int dx p(x|\theta) = 1$



# Likelihood vs. Probability functions

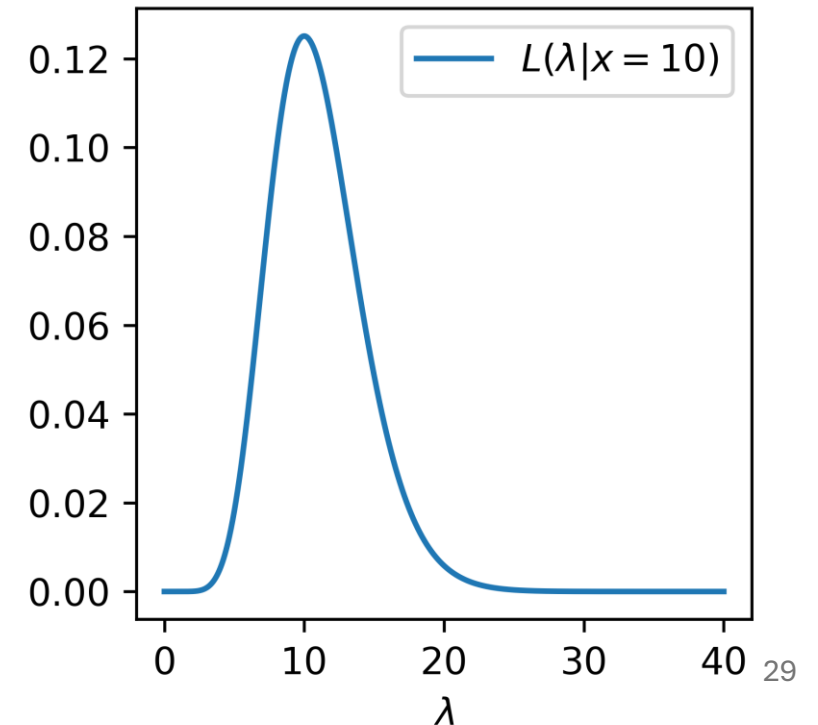
Important to remember that likelihood and probability functions are different



## Likelihood:

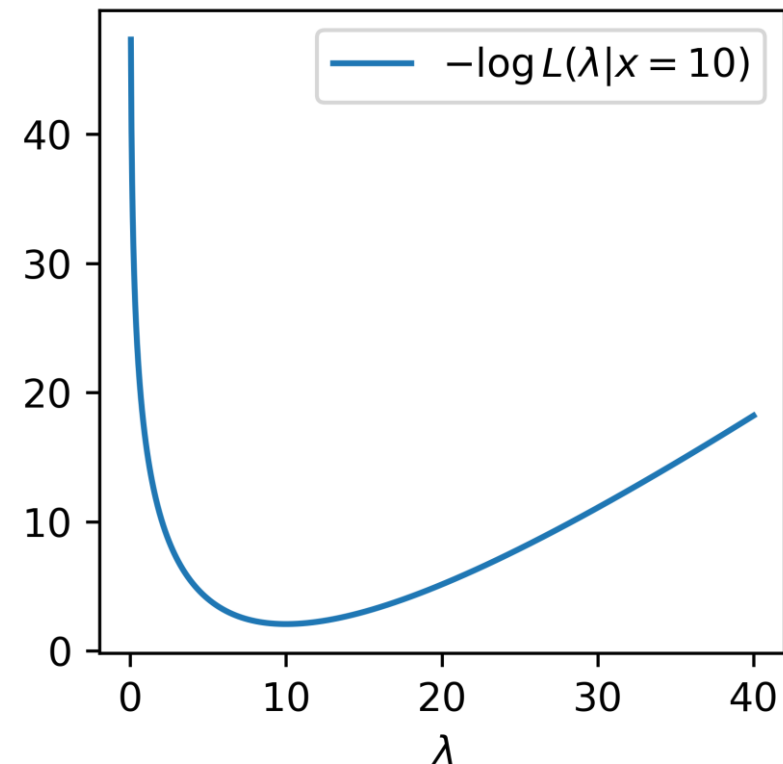
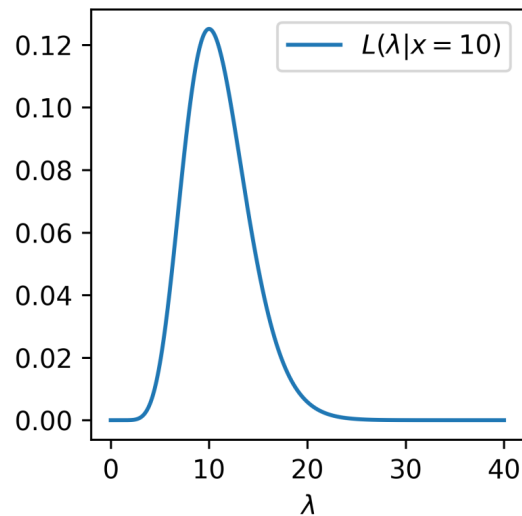
fixed data  $x$ , variable parameters  $\theta$

not normalized  $\int d\theta p(x|\theta) \neq 1$

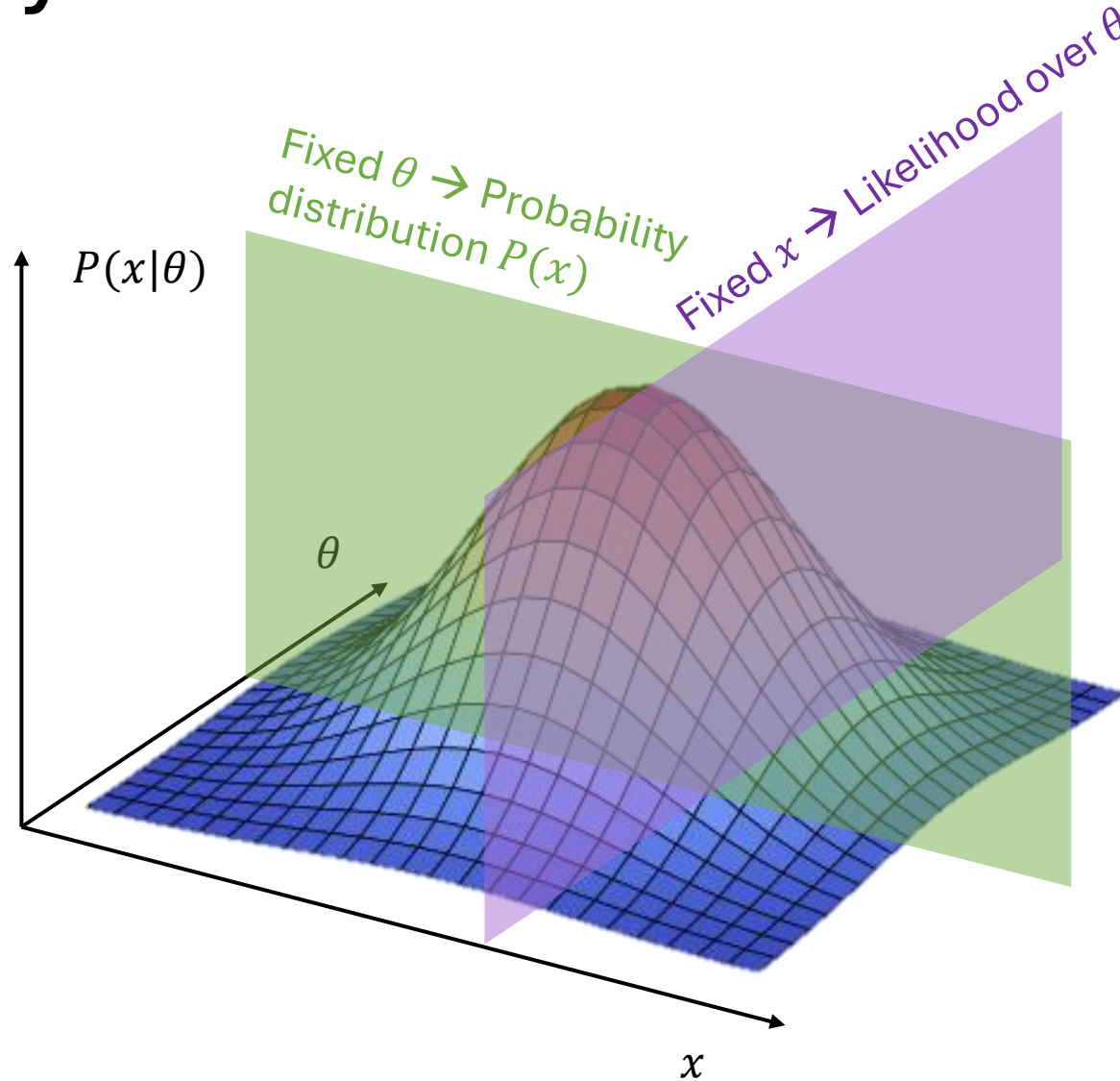


# Negative Log Likelihood

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



# 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

# 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



**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

$$p_{\text{sig}}(x) = p(x|\text{sig}) \quad p_{\text{bkg}}(x) = p(x|\text{bkg})$$

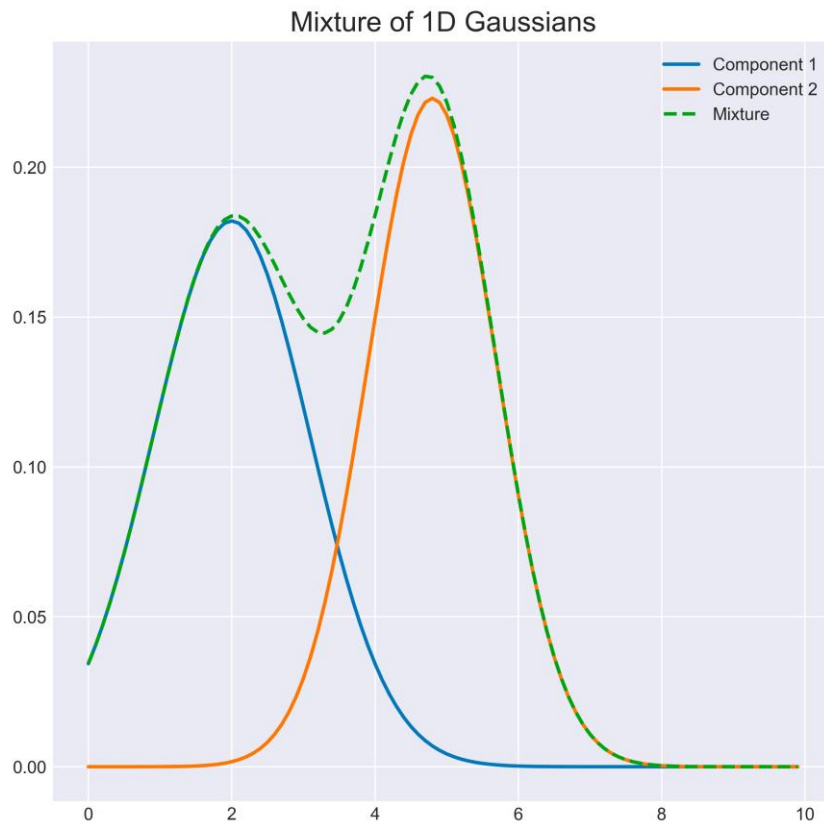
The data density can then be modelled as "mixture"

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

With  $p(\text{sig}) + p(\text{bkg}) = 1$

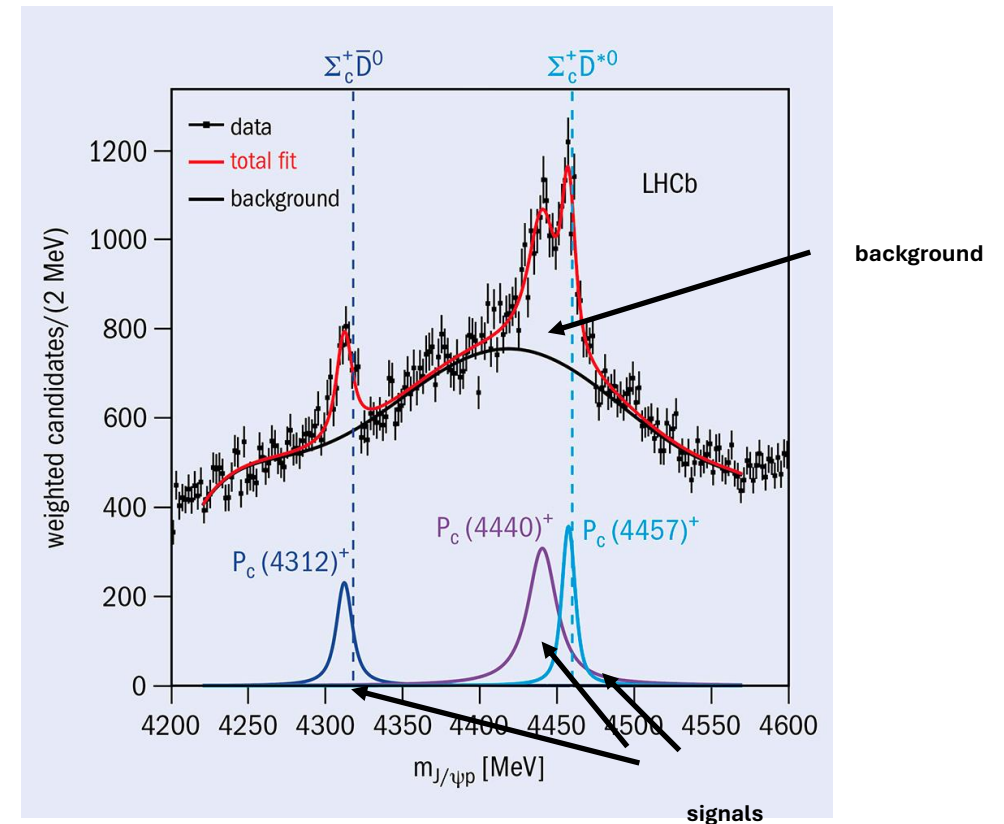
# Mixture Models

## Example:



**Gaussian Mixture Model**

[\[source\]](#)



**Particle Resonances**

# Simultaneous Measurements

Sometimes a 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

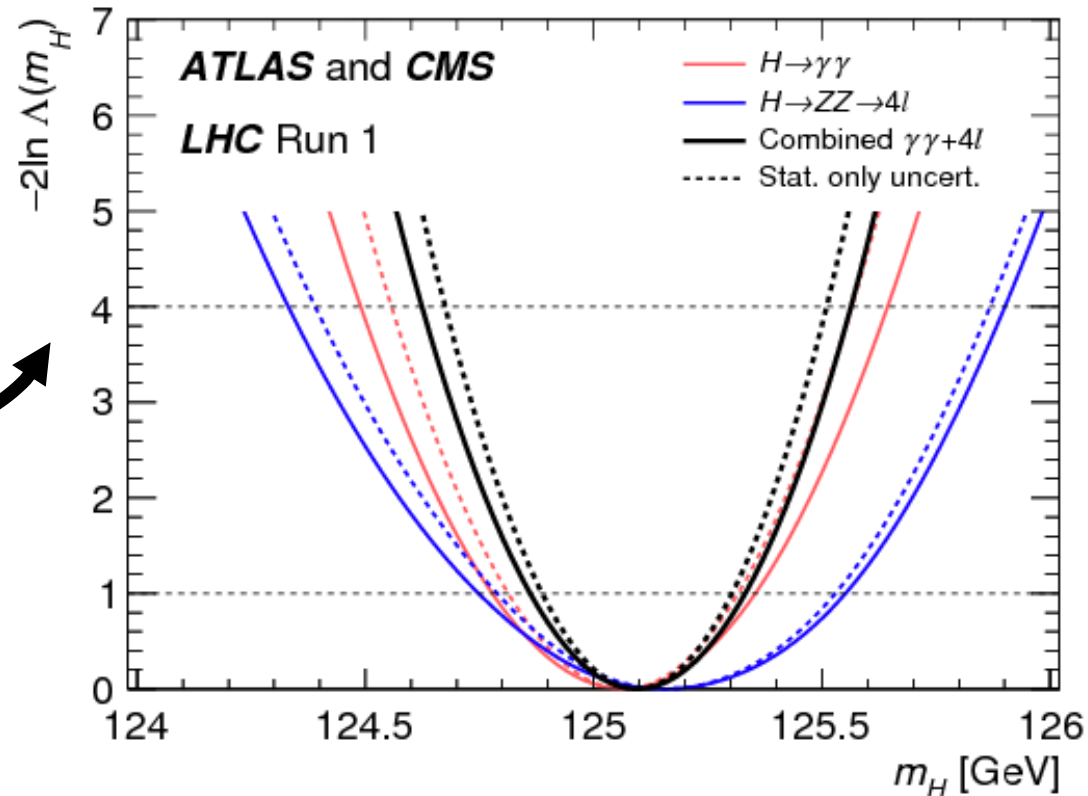
$$\log p_{\text{global}} = \log p_{E1} + \log p_{E2}$$

# 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

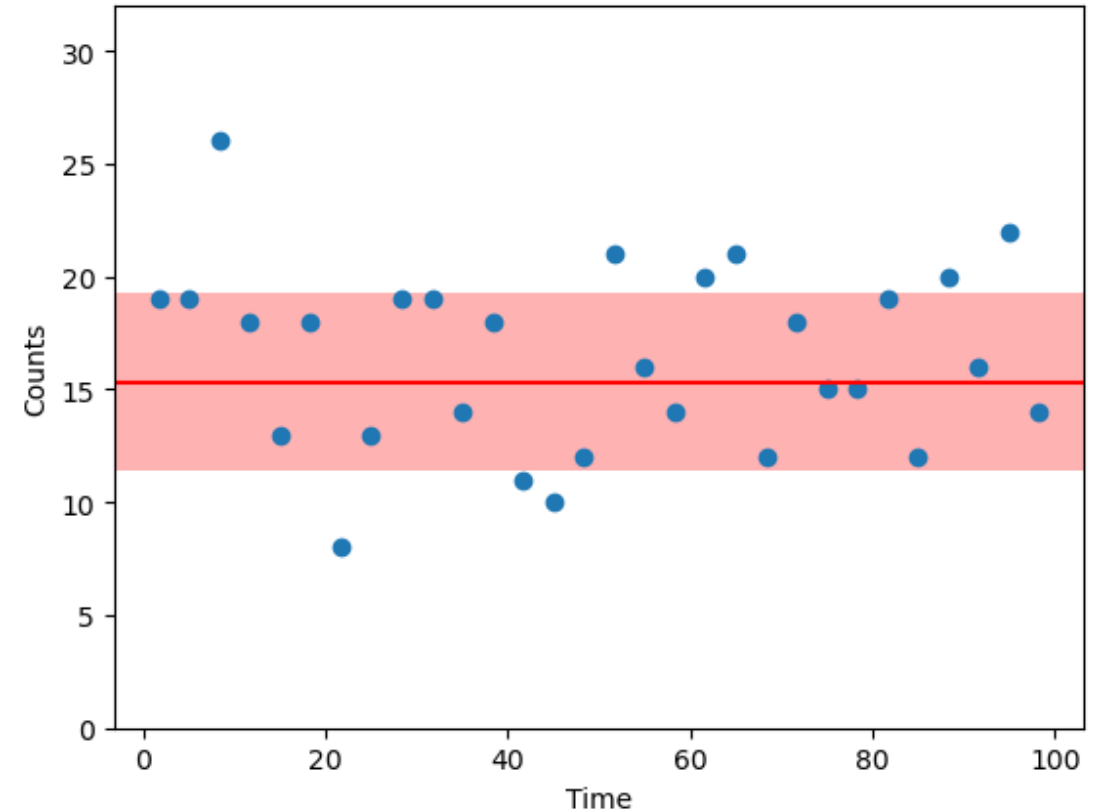


# Simultaneous Measurements

- 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:

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



# Estimator Theory

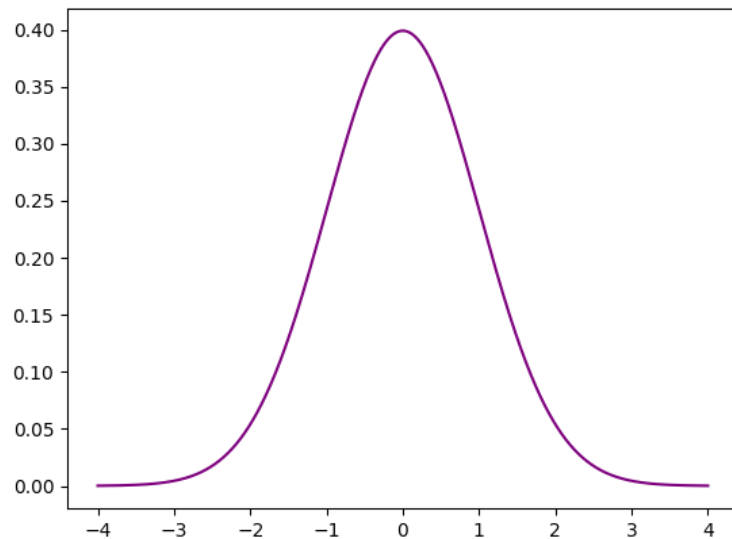


# Recap: Model vs. Observation

## Abstract Space

i.e. a Model, often containing a number of parameters  $\theta$

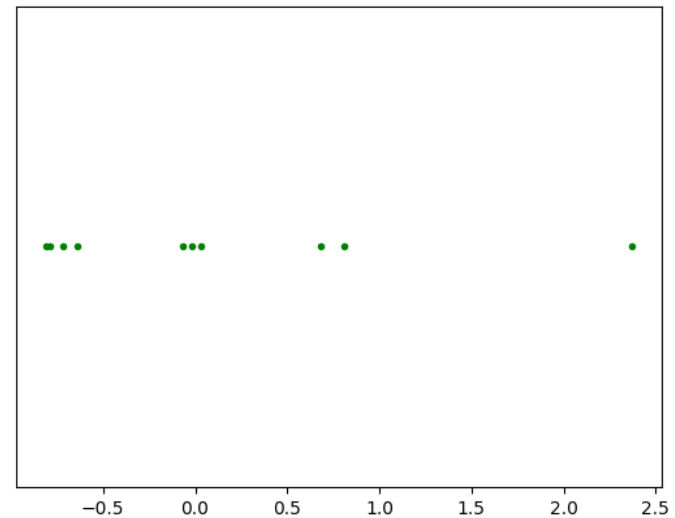
Example: Gaussian with parameters  $\mu$  and  $\sigma$



## Observable Space

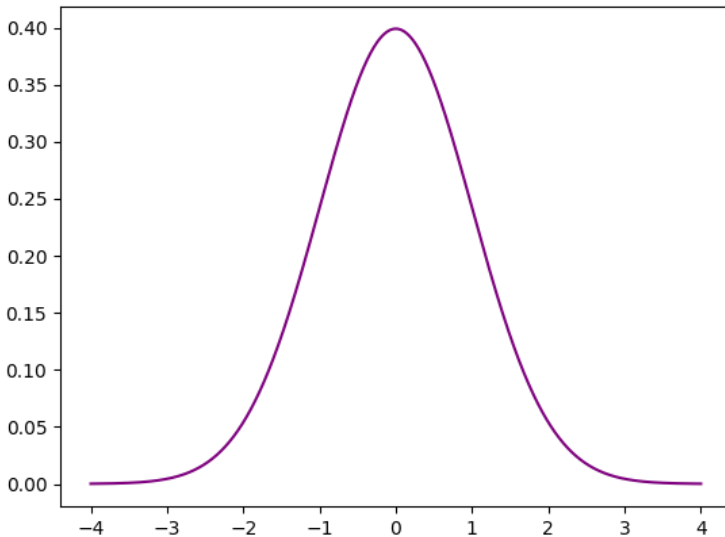
i.e. our “Data”  $x$

Example: Values  $x = \{1.2, -0.7, 0.3, \dots\}$



# From Model to Observation

Abstract Space



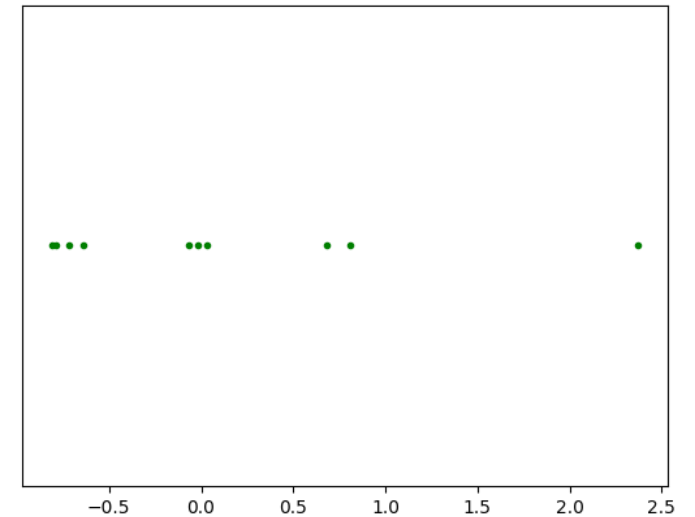
Choose model parameters

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

Observable Space

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

Gives Probability  $P(x)$   
 $x$  is a random variable

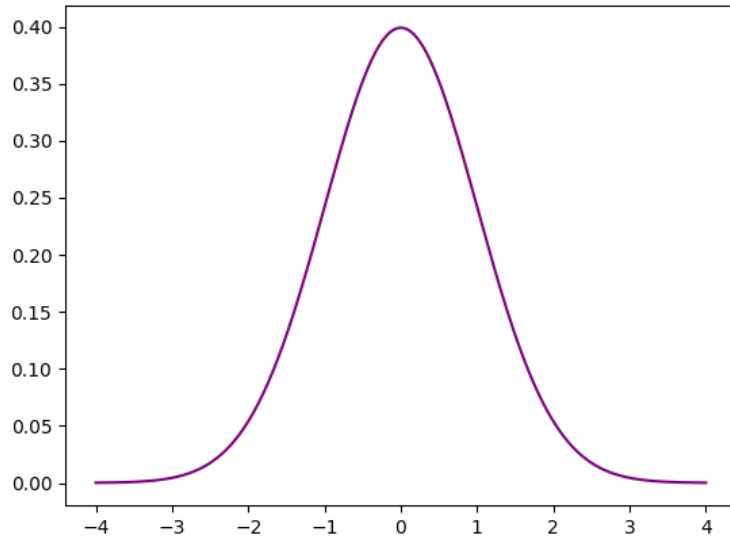


Joint probability of observing  $n$  data:  $\prod_i p(x_i)$

# And back...?

Abstract Space

Estimate model parameters

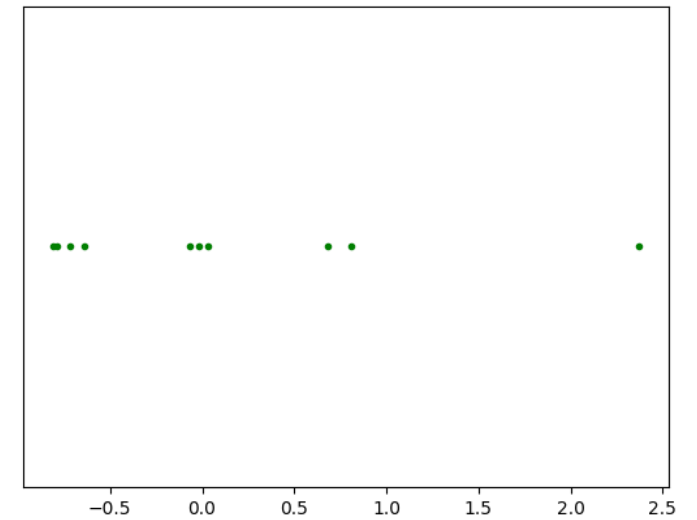


Observable Space

Given a set of observations  $x$

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

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

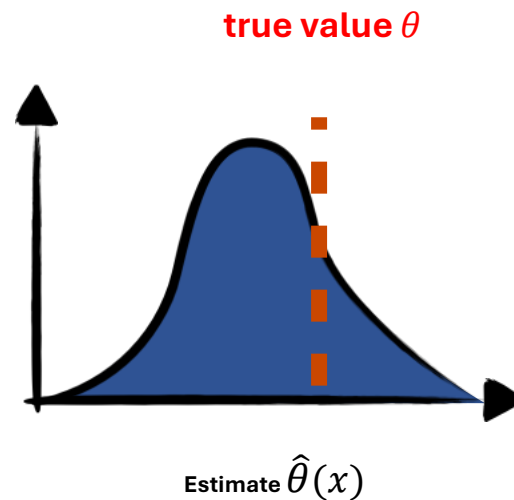
Estimators are **functions of the data** (i.e. "statistics")  $\hat{\theta}(x)$  that give an **estimate** of a parameter of the underlying model  $p(x|\theta_0)$ .

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

- no guarantees that *any particular* estimate is close to  $\theta_0$
- but we can make statements w.r.t repeated experimentation  
i.e. long-run properties of estimators

# Estimate Distribution

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



# Example: Gaussian Mean

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

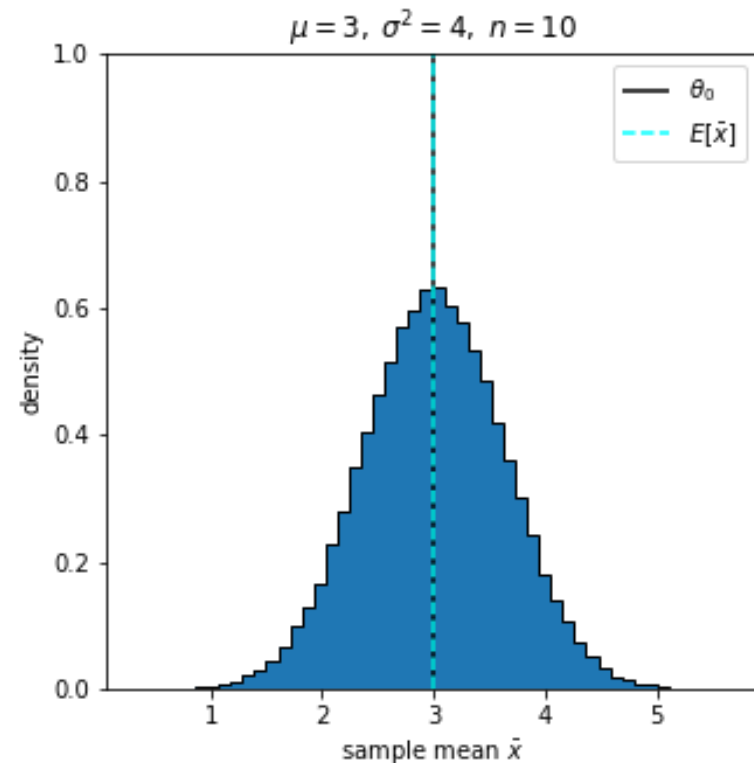
A simple estimator is the "sample mean"

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

Distribution of the estimator for

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

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

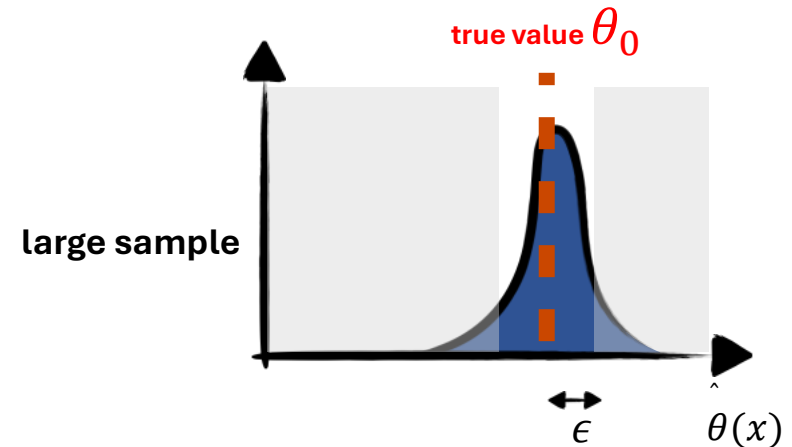
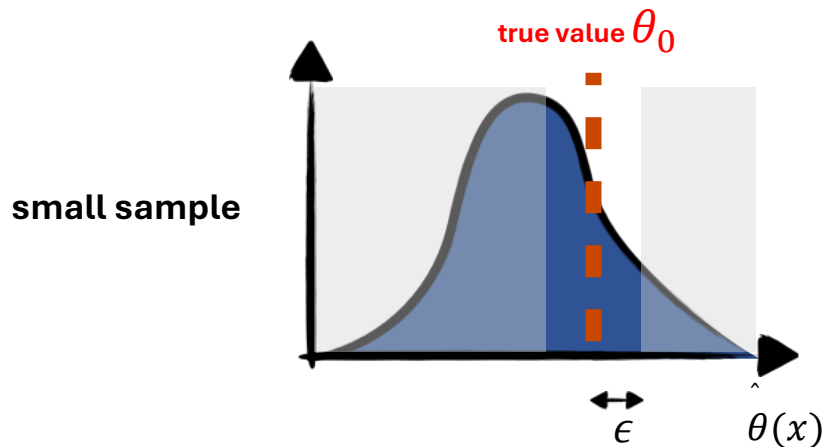


# Consistency

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

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



# Example: Gaussian Mean

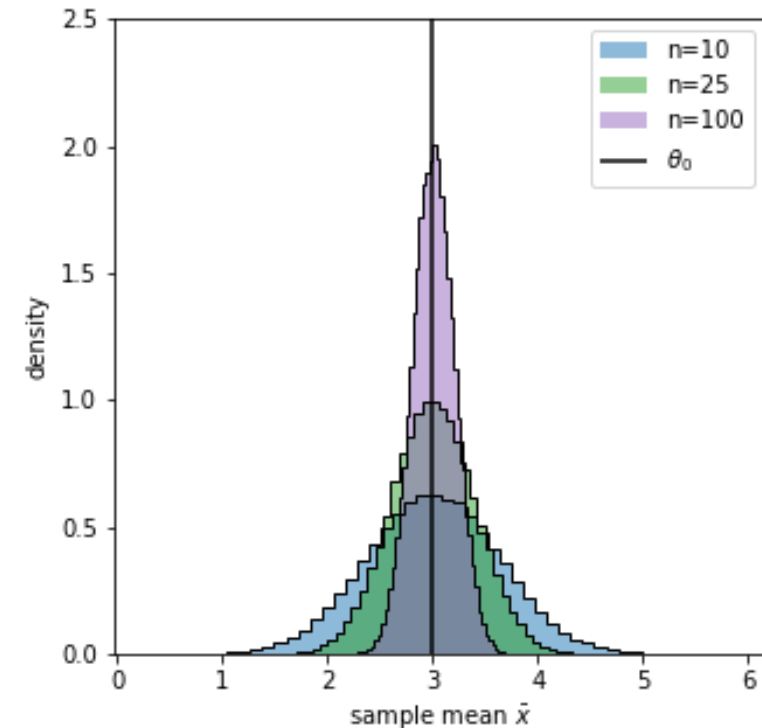
Consistency at play in our Gaussian Example

- sharpening of the distribution around the true value  $\mu$

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

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

$\bar{x}$  is a consistent estimator of  $\mu$



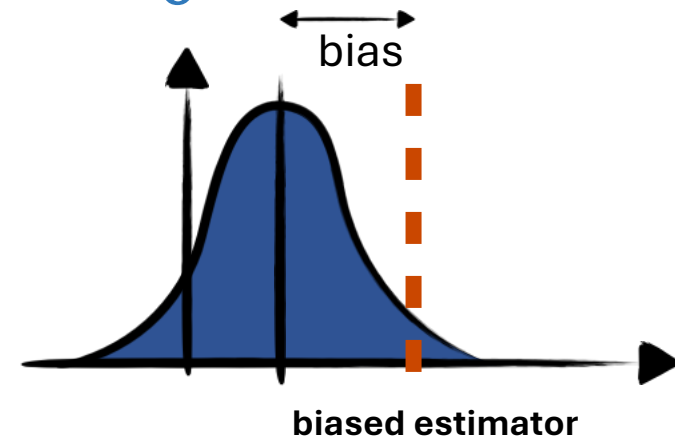
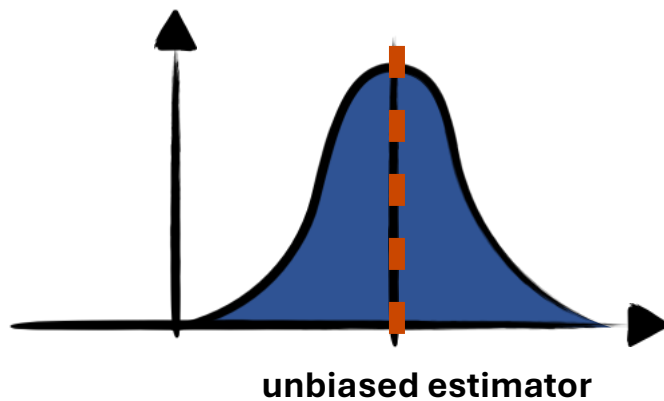


# Estimator Bias

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

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



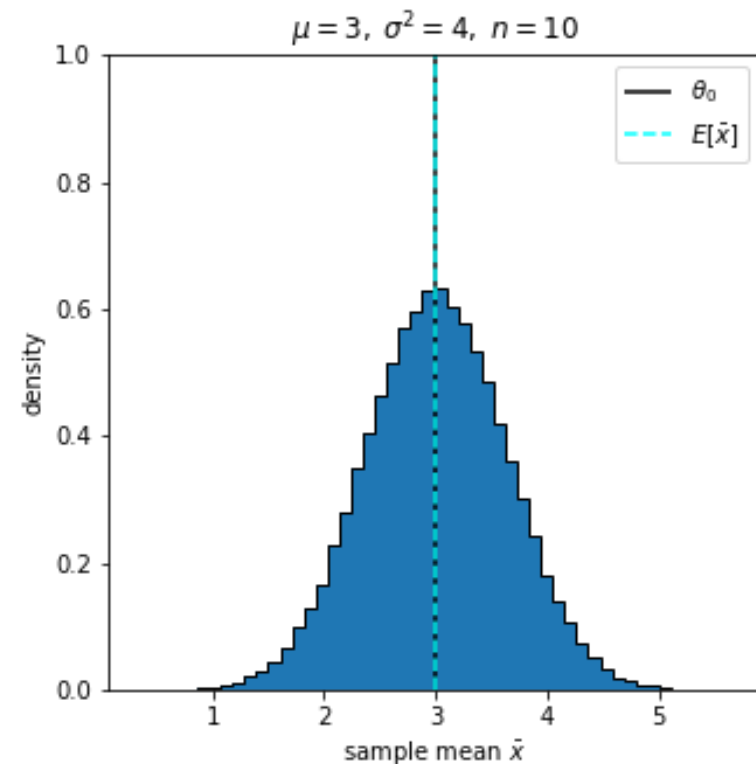
# Example: Gaussian Mean

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

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

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

$\bar{x}$  is a consistent and unbiased estimator of  $\mu$



# Example: Gaussian Variance

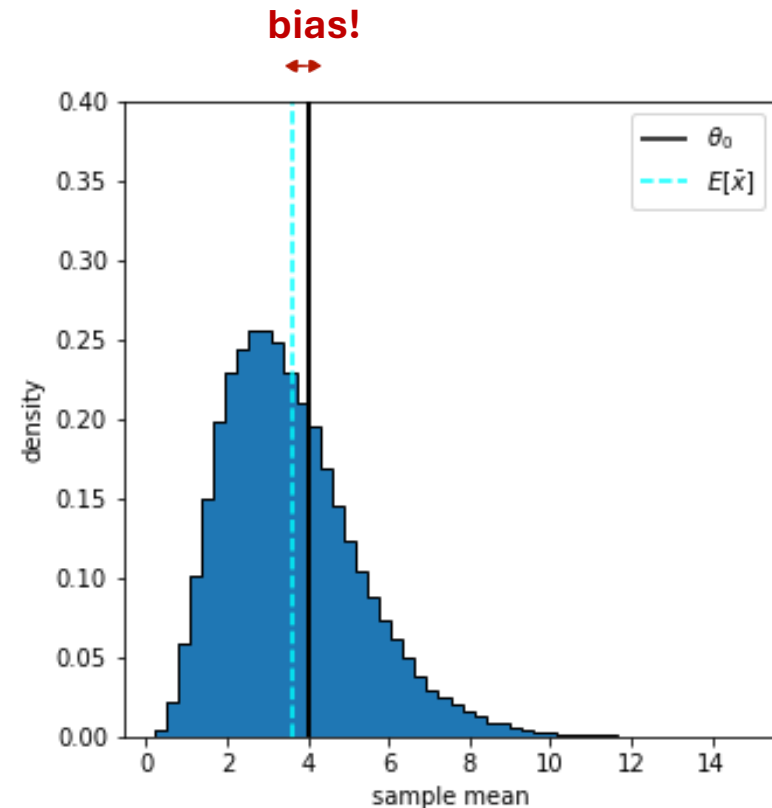
Same Gaussian Setup: but with the **sample variance** as a statistic

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

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

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

$s^2$  is a biased estimator of  $\sigma^2$



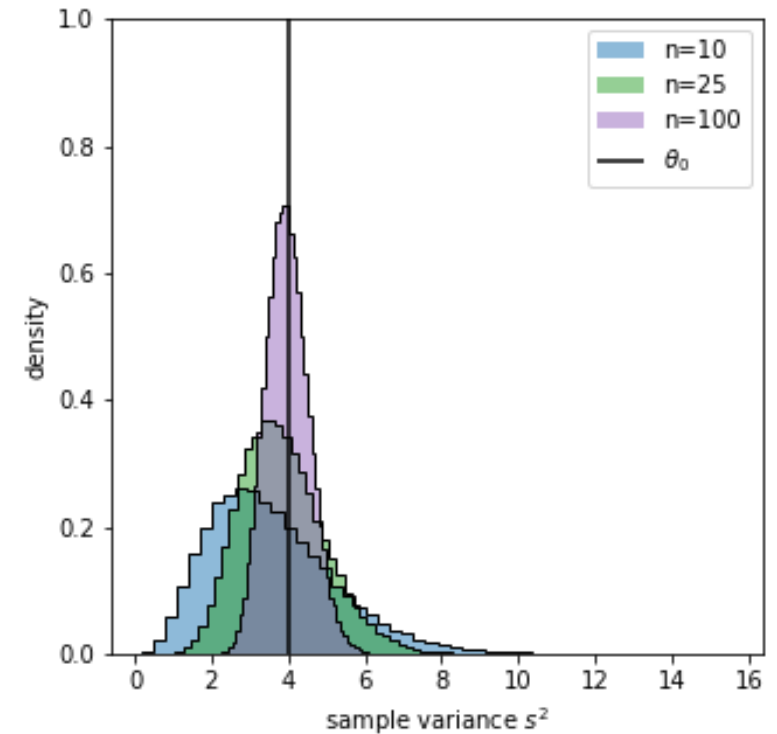
# Example: Gaussian Variance

The sample variance is still consistent though:

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

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

$s^2$  is a consistent but biased estimator of  $\sigma^2$

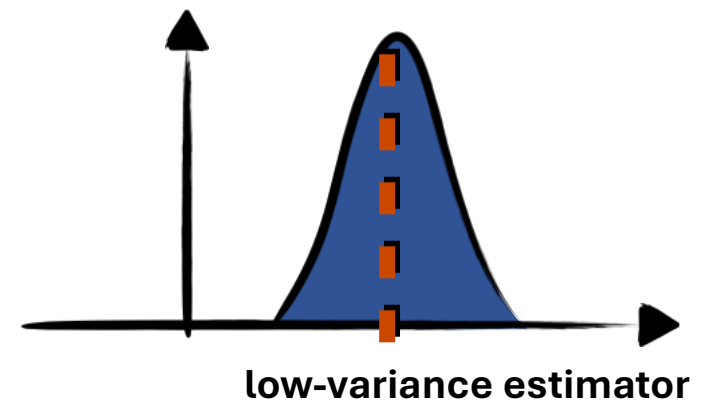
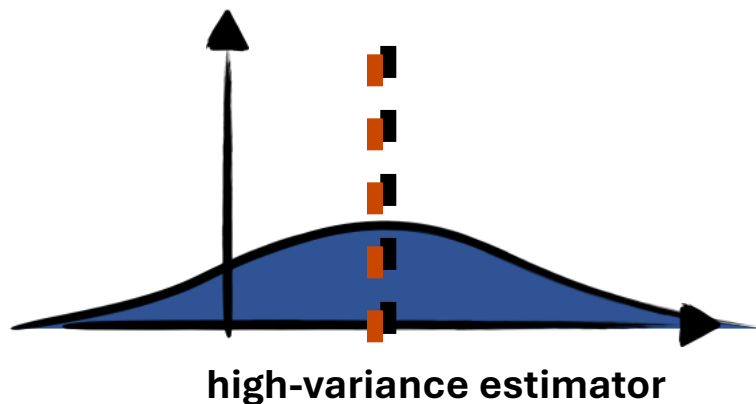


# Estimator Variance

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

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



# Variance Example: Gaussian Mean

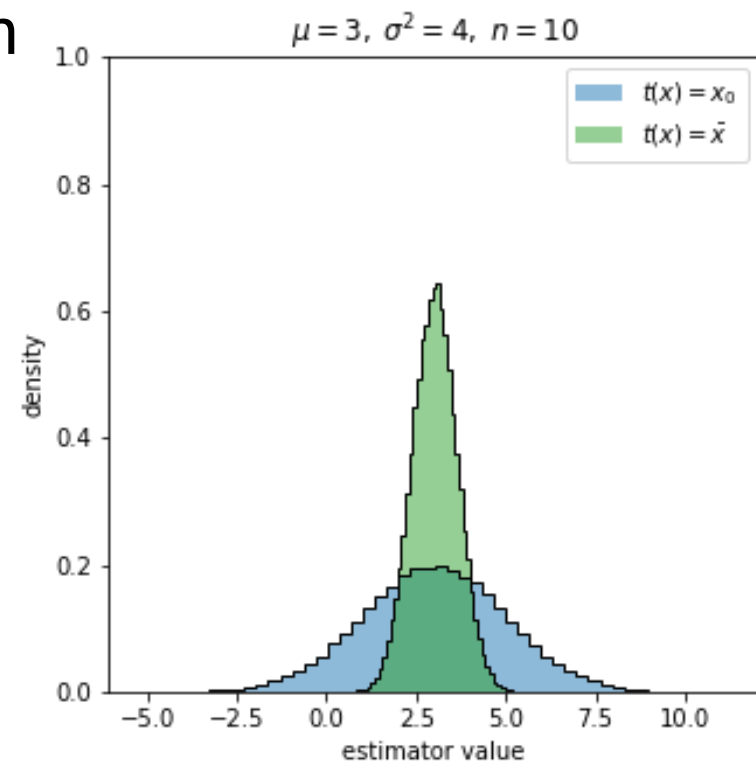
Same Gaussian Model with  $n$  i.i.d. samples  $x_i \sim \mathcal{N}(x_i | \mu, \sigma^2)$

In a sample  $\mathcal{x} = (x_1, \dots, x_n)$  each  $x_i$  is an unbiased, consistent estimator of  $\mu$ .

e.g.  $f(\mathcal{x}) = x_1$

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

**It has much lower variance!**



# Example

- $n$  samples from a unit normal, which has

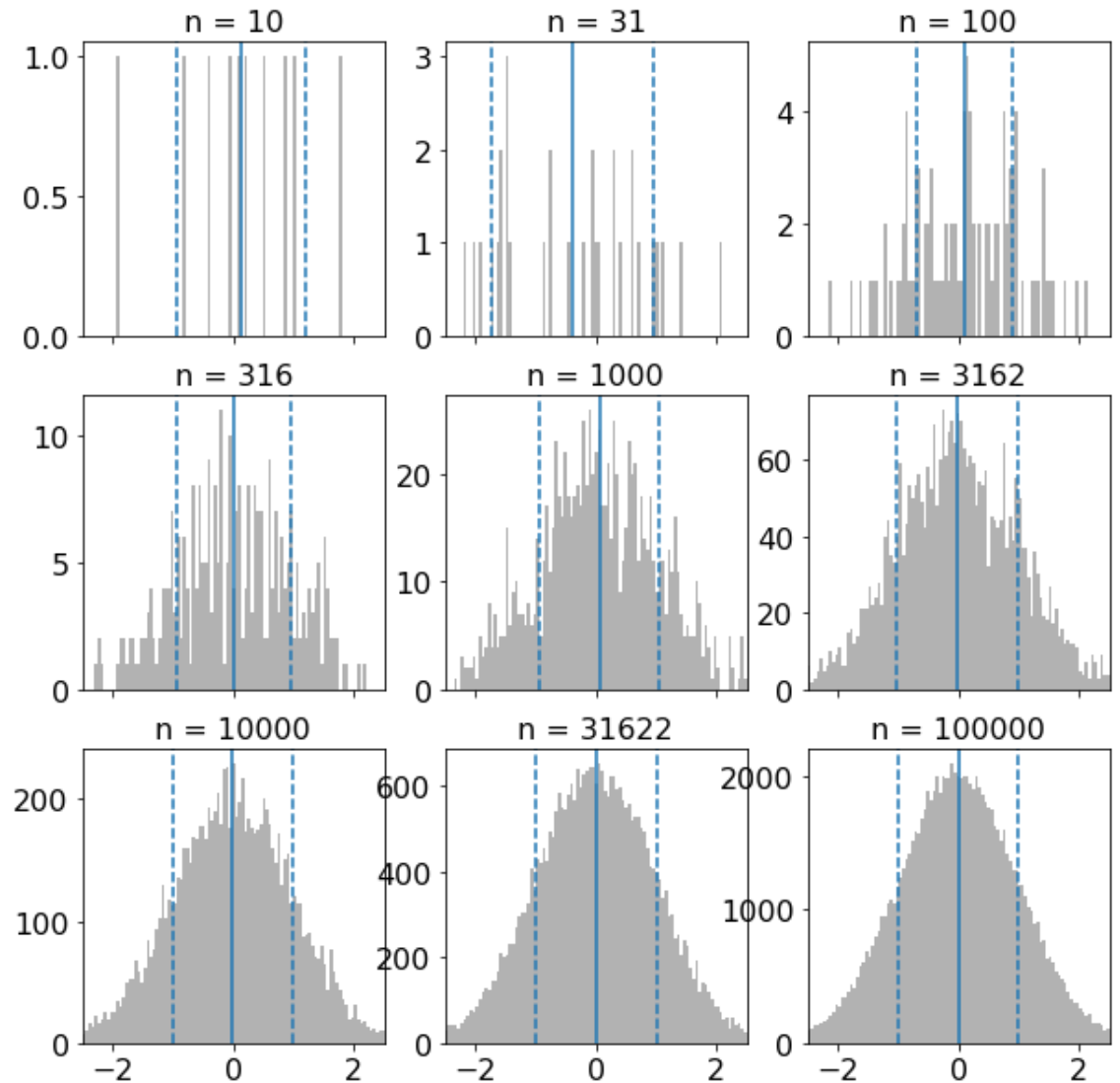
$$\mu = 0$$

$$\sigma = 1$$

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

- What is the variance of the sample mean?

$$\text{Var}(\bar{x}) = ?$$



# Variance of sample mean

- Exercise: calculate

$$\text{Var}(\bar{x}) = \dots$$

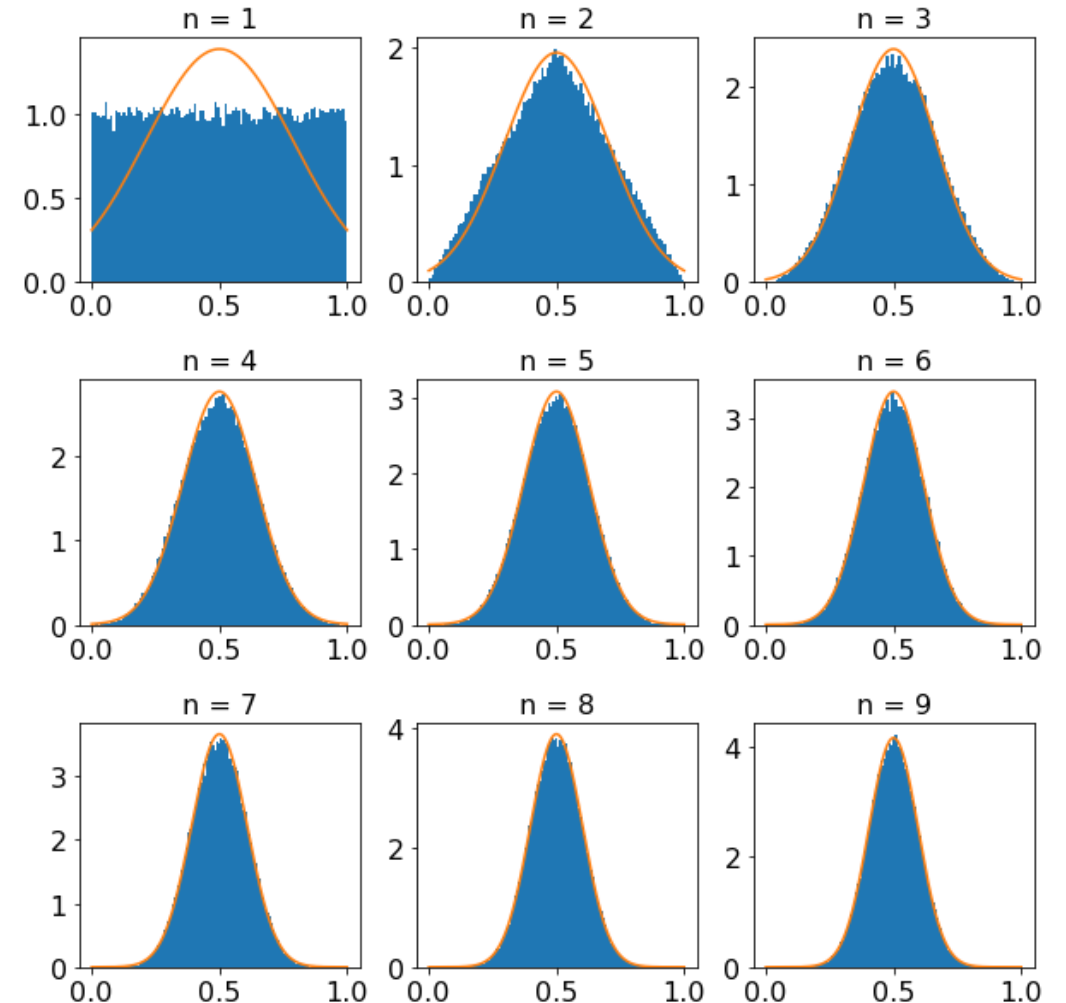
$$= \frac{\sigma^2}{n}$$



# CLT revisited

- What does this remind you of?  
→ Central Limit Theorem of course!
- Suppose we make repeated, independent draws  $x_i$  of a 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 distributed according to a  $\mathcal{N}(\mu, \sigma/\sqrt{n})$ !

CLT using uniform distribution



# Cauchy

What about a Cauchy distribution?

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

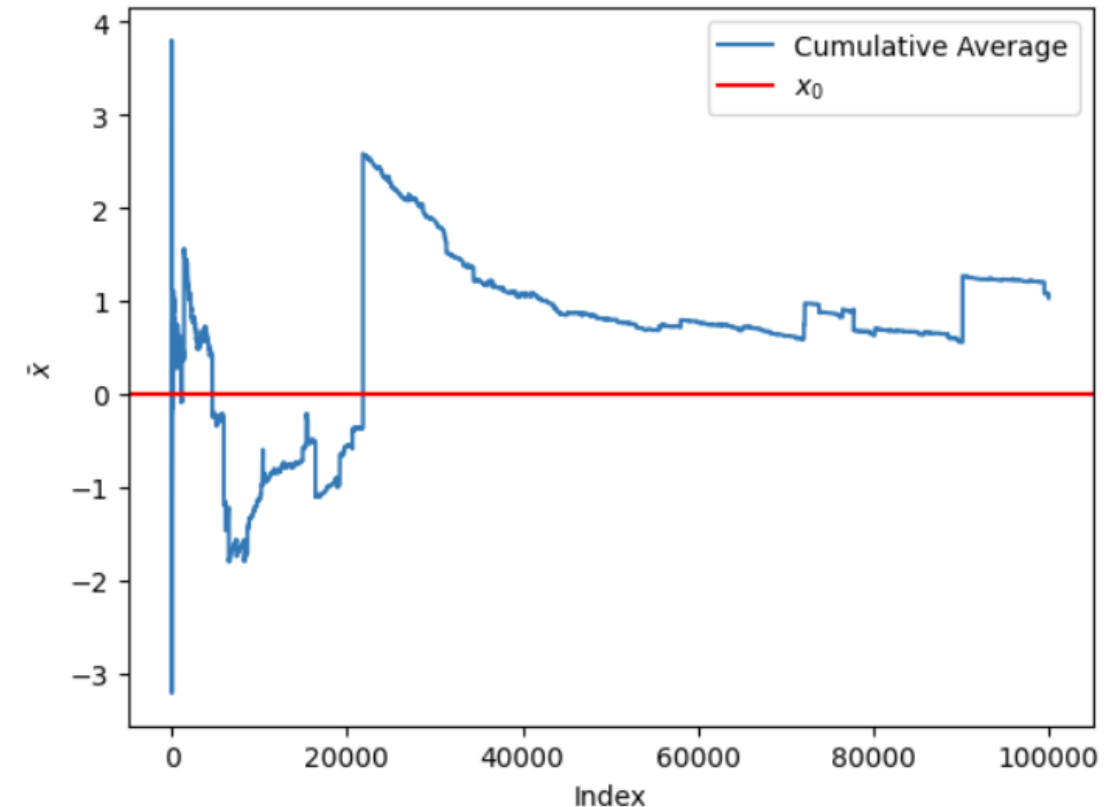
- The distribution does not have finite moments!!
  - $\bar{x}$  is **not** a good estimator for  $x_0$ ! It is not consistent

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

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

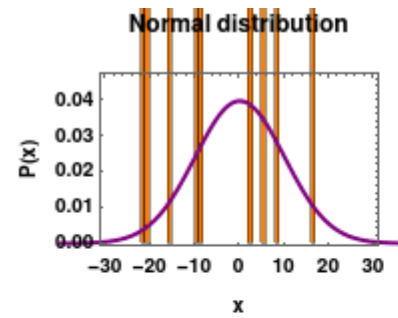
```
fig, ax = plt.subplots()
ax.plot(xbar, label="Cumulative Average")
ax.set_ylabel(r"$\bar{x}$")
ax.set_xlabel("Index")
ax.axhline(0, c='r', label=r"$x_0$")
ax.legend()
```

<matplotlib.legend.Legend at 0x7f8d226ba5e0>

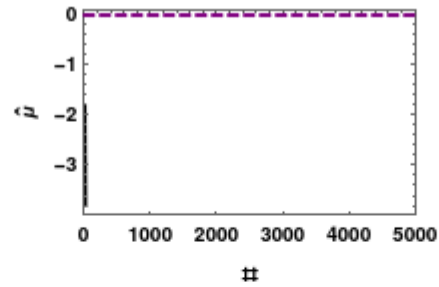


# Normal vs. Cauchy

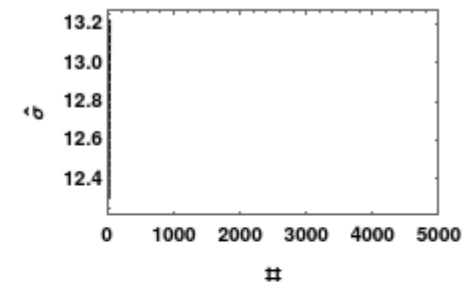
Gaussian



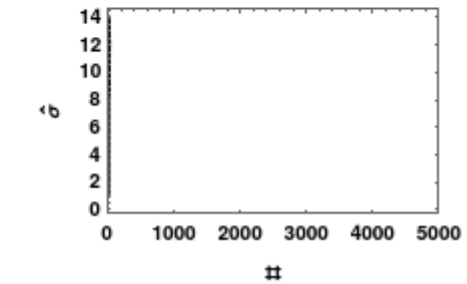
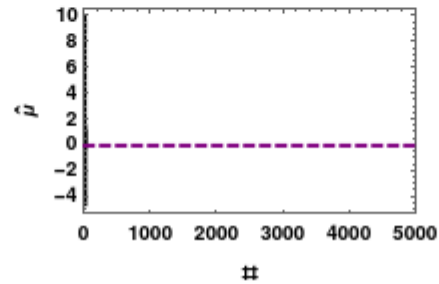
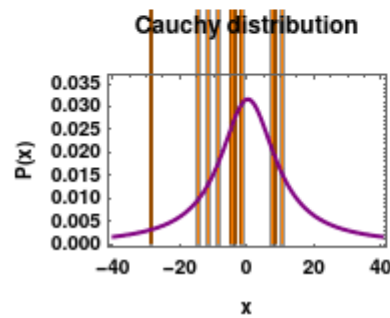
Sample mean



Sample standard deviation



Cauchy



# Bias - Variance Decomposition

Both bias and variance contribute to the overall expected deviation, i.e. mean squared error (MSE), from the true value  $\theta_0$

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

# 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  $\rightarrow$  small variance and vice versa
- Known as the **Cramér-Rao Bound**

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

# 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  $I(\theta)$

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

# Fisher Information

**Exercise:** For a Gaussian,  $x \sim p(x | \mu, \sigma^2)$ , calculate

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

...

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

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  $\bar{x}$  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:**

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

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

# Example

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

- Normal distribution:

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

(Exercise)

$$\frac{\partial \log \mathcal{L}}{\partial \hat{\mu}} = 0 \rightarrow \hat{\mu} = \frac{1}{n} \sum x_i$$

# 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**

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

$$\hat{\mu}_{\text{MLE}} = \bar{x} = \frac{1}{n} \sum_i x_i$$

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

# 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**

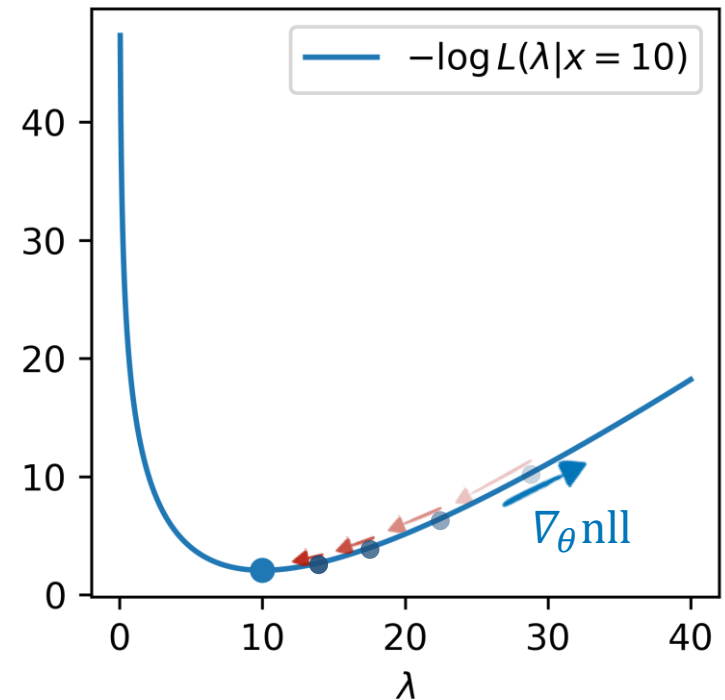
One option via gradient descent

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

while not converged:

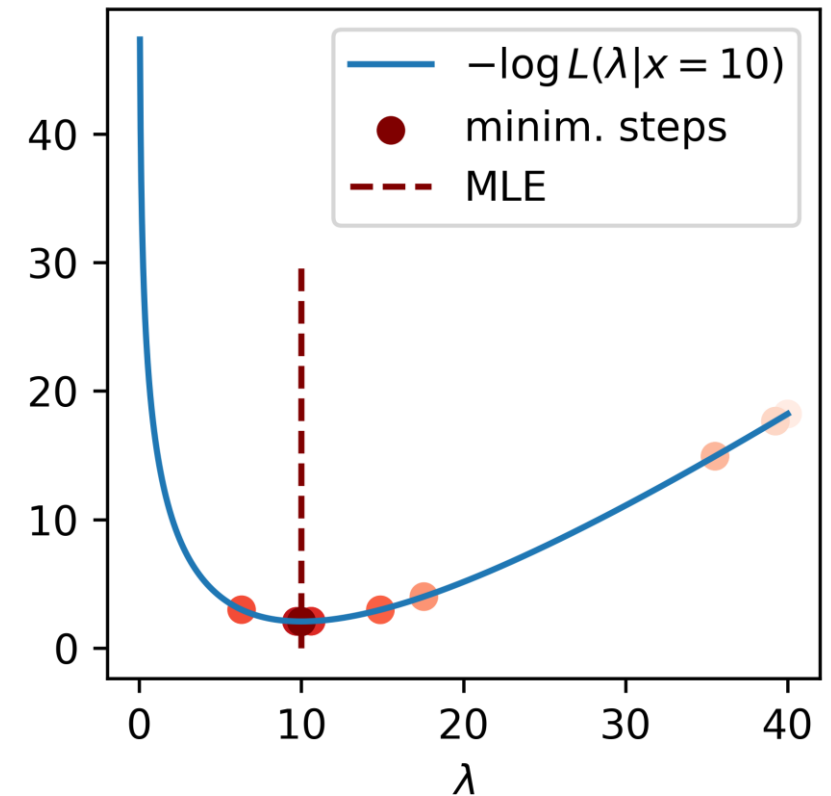
$$g = \nabla_{\theta} f(\theta)$$

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



# Python Example

```
1 def logpdf(p):  
2     return -sps.poisson.logpmf(10,p)  
3  
4 scipy.optimize.minimize(logpdf, x0=20, method = 'SLSQP')  
5  
[69] ✓ 0.5s  
... fun: array([2.07856165])  
    jac: array([2.3603443e-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  $n \rightarrow \infty$

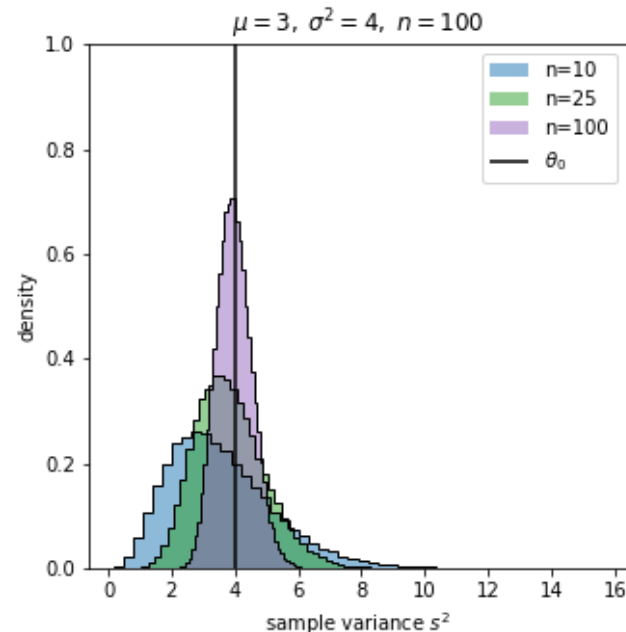
$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow^d \mathcal{N}(0, \text{var}\hat{\theta})$$

= “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**

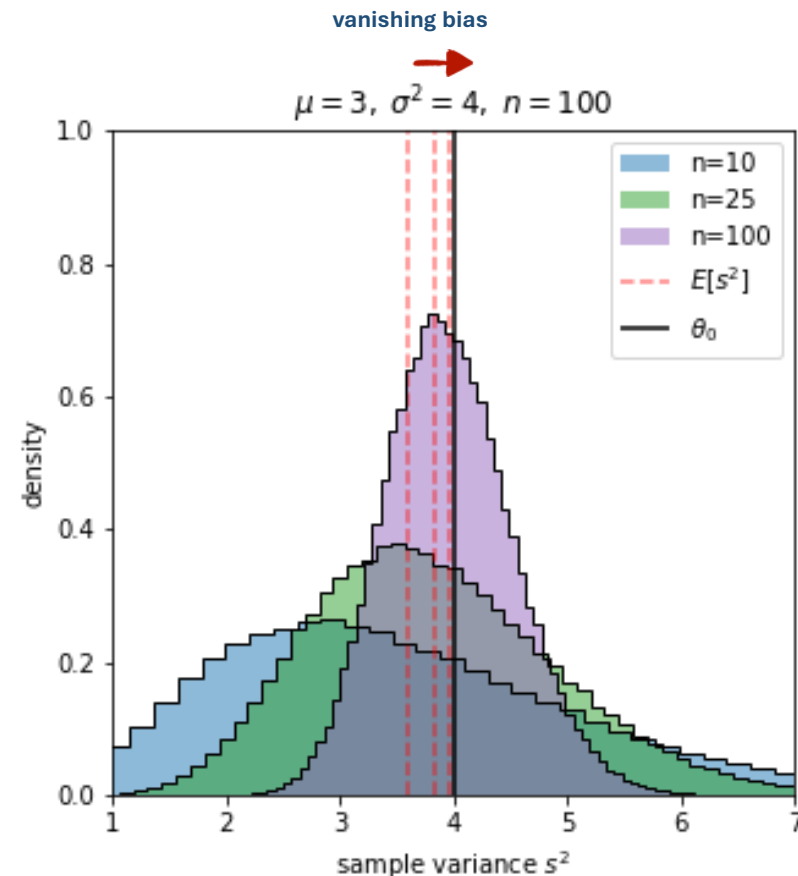
- Note: In general finite-sample MLE are biased

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

  
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 Efficient

MLE estimators saturate the Cramér-Rao bound:

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

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

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

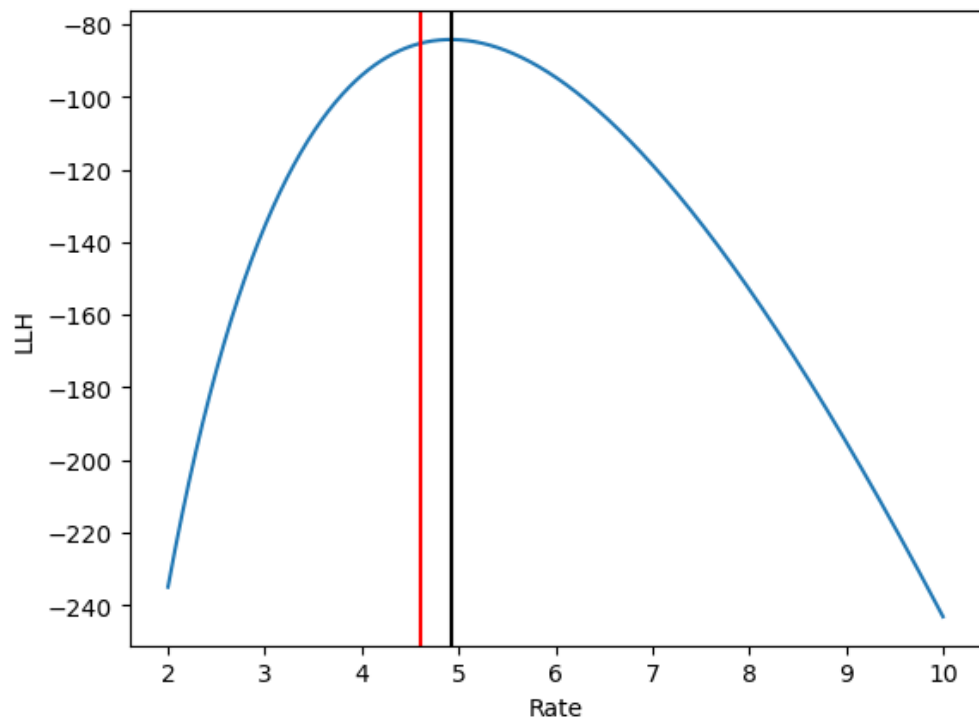


Inverse Fisher Information

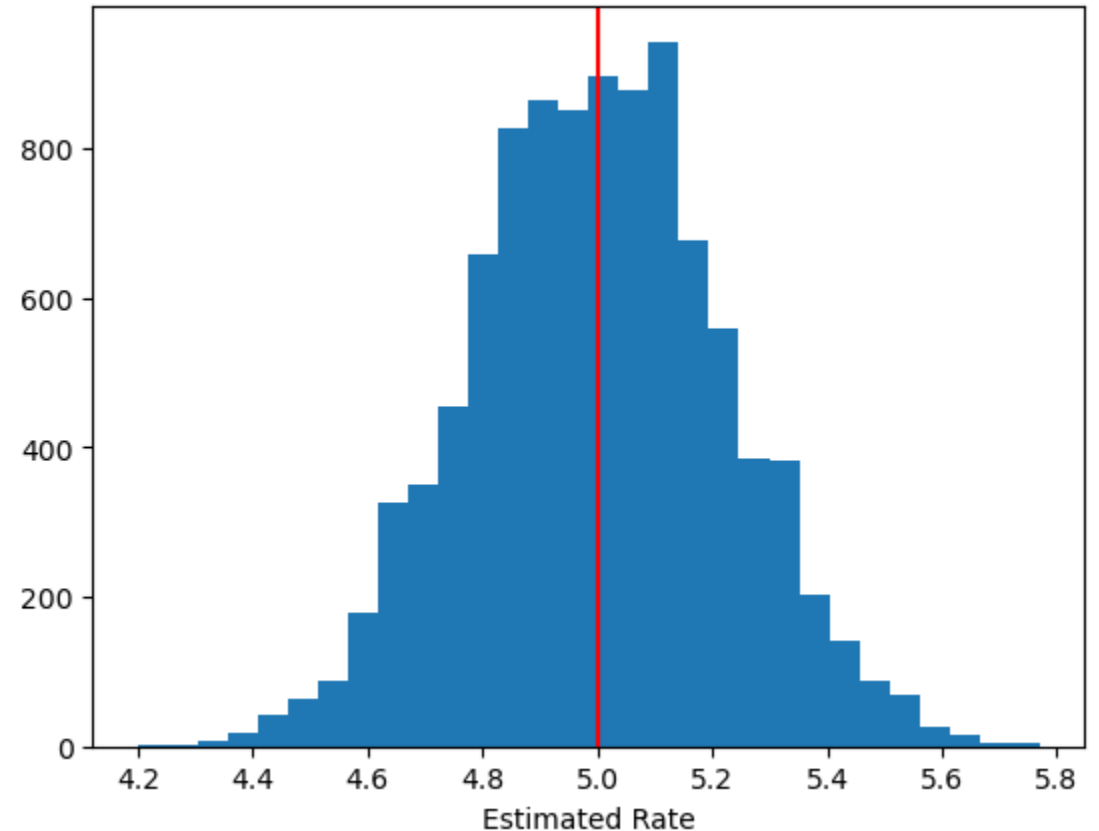
# Fixing the 2<sup>nd</sup> Problem: MLE for our Example

- 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,...)

# Day II



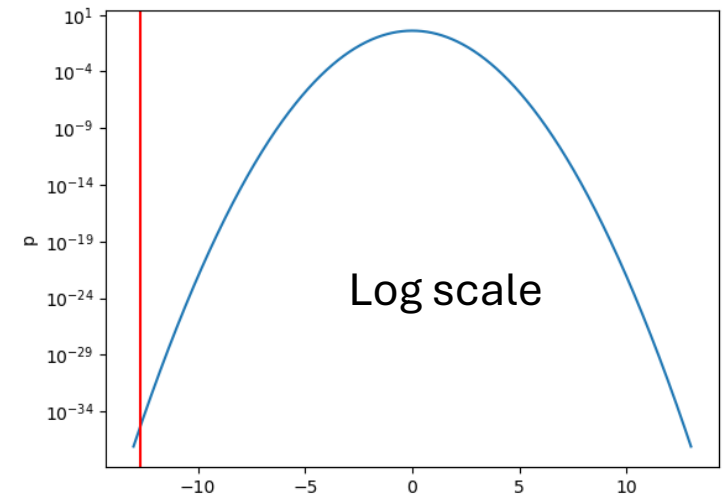
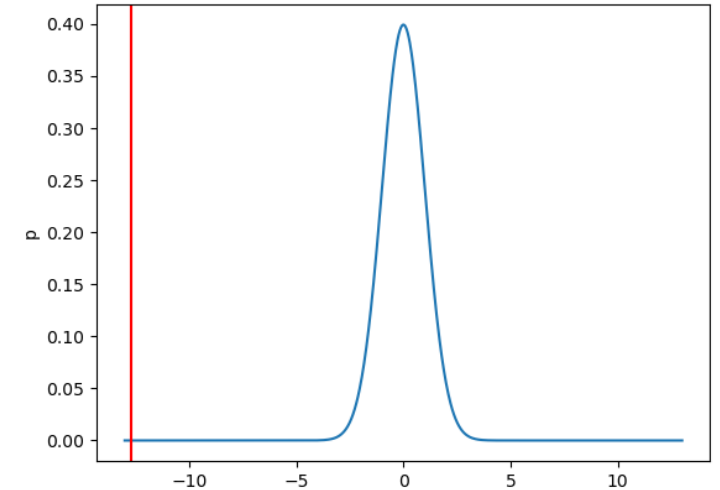
# Hypothesis Testing

# „goodness-of-fit“ tests

- Question: Does the model describe the data?

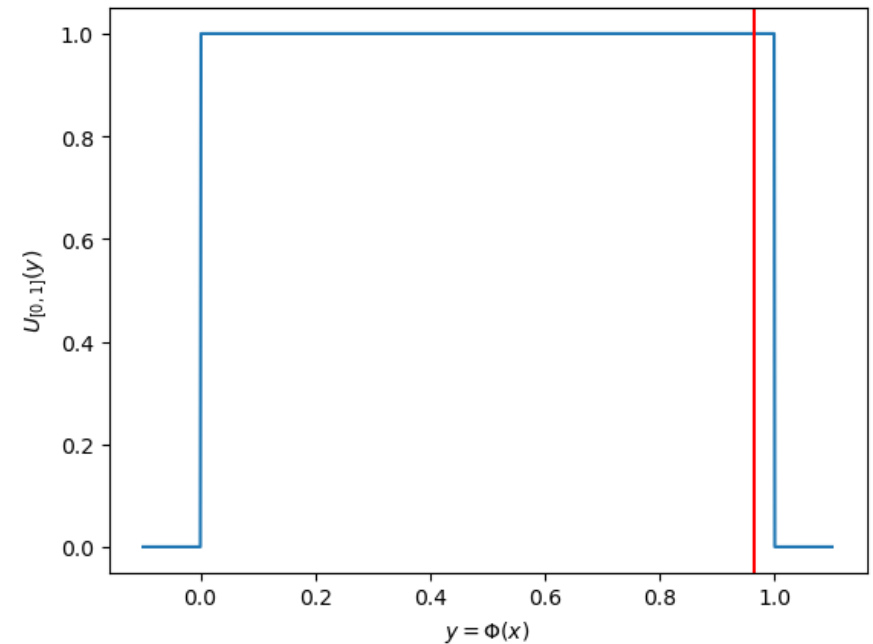
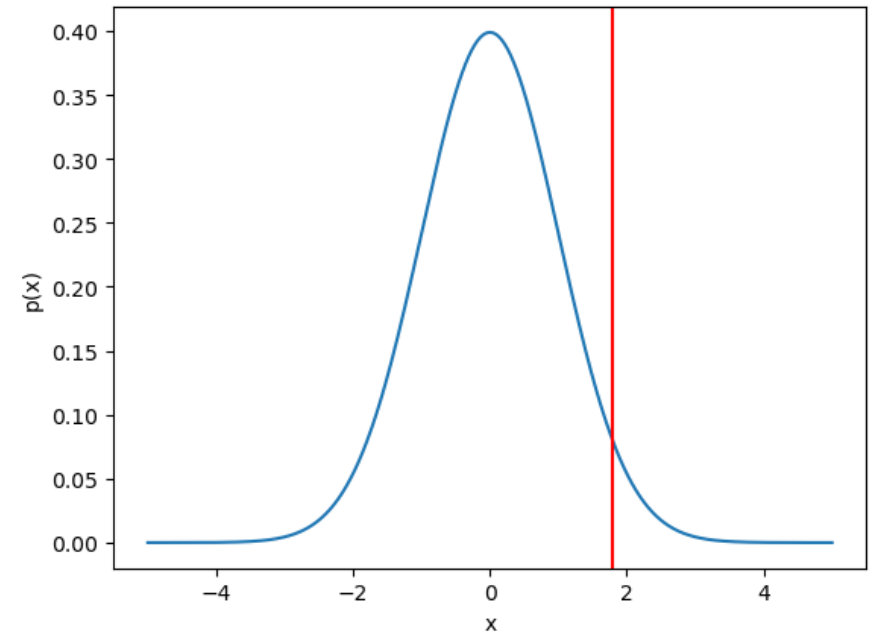
## Example:

- Your model is the unit normal distribution
  - Your observed value is -12.7
  - The probability to get a value at least as extreme, is simply the CDF at -12.7 =  **$2.96 \times 10^{-37}$**
- very poor goodness-of-fit



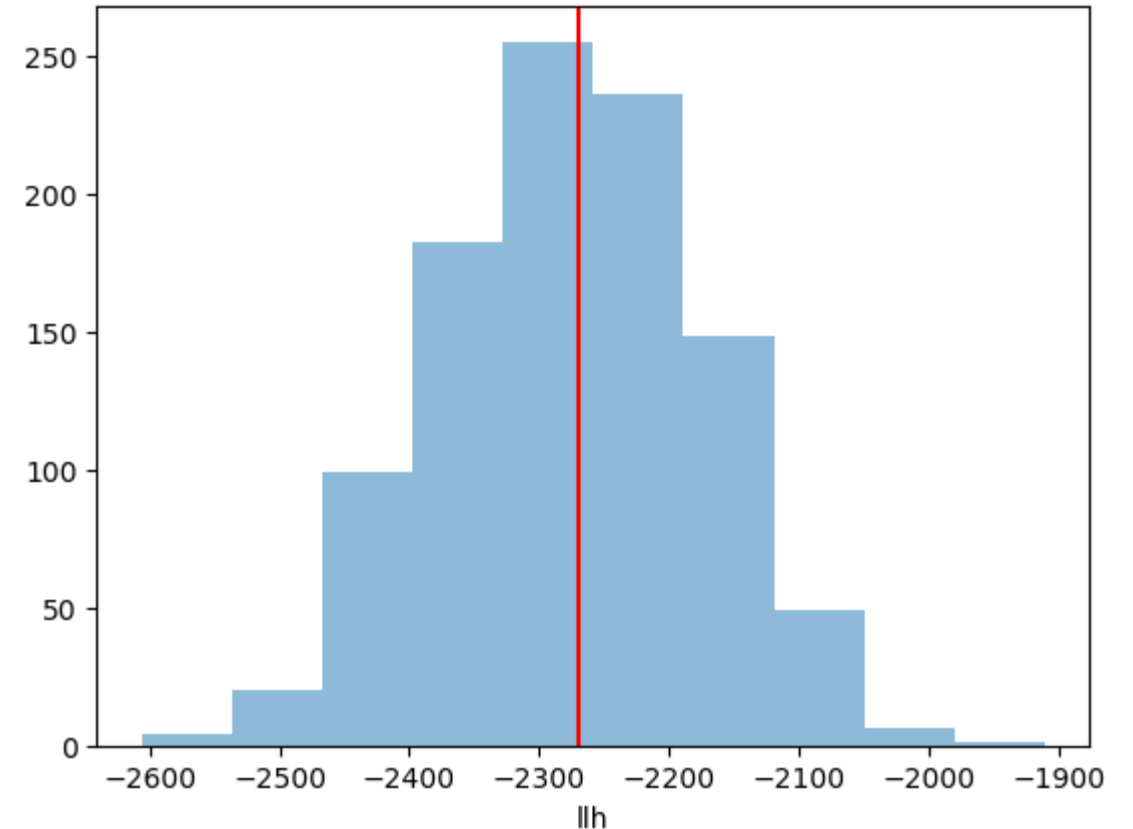
# Equivalence

- Any goodness-of-fit test can be reduced to assessing whether samples come from the standard uniform distribution
  - If you are interested to test whether samples  $x_i$  stem from  $f(x)$ , this is equivalent of asking whether samples  $y_i = F(x_i)$  come from the standard uniform  $U(0,1)$
  - ( $F$  is the cumulative of  $f$ )
  - This can then also directly be interpreted as  $1 - \text{the p-value space}$



# Test Statistic

- In order to work with more complex models (more than a simple observation), we need to define a test statistic
- For example, in our case, we could try to use the likelihood
- And then generate null trials to see how typical data looks like



Why using the likelihood is not a good idea:

<https://www.slac.stanford.edu/econf/C030908/papers/MOCT001.pdf>

# Chi2 approximation

- Let's look again at our likelihood expression:

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

- A Poisson distribution can be approximated by a Gaussian with mean  $\mu = \lambda$  and std. dev.  $\sigma = \sqrt{\lambda}$

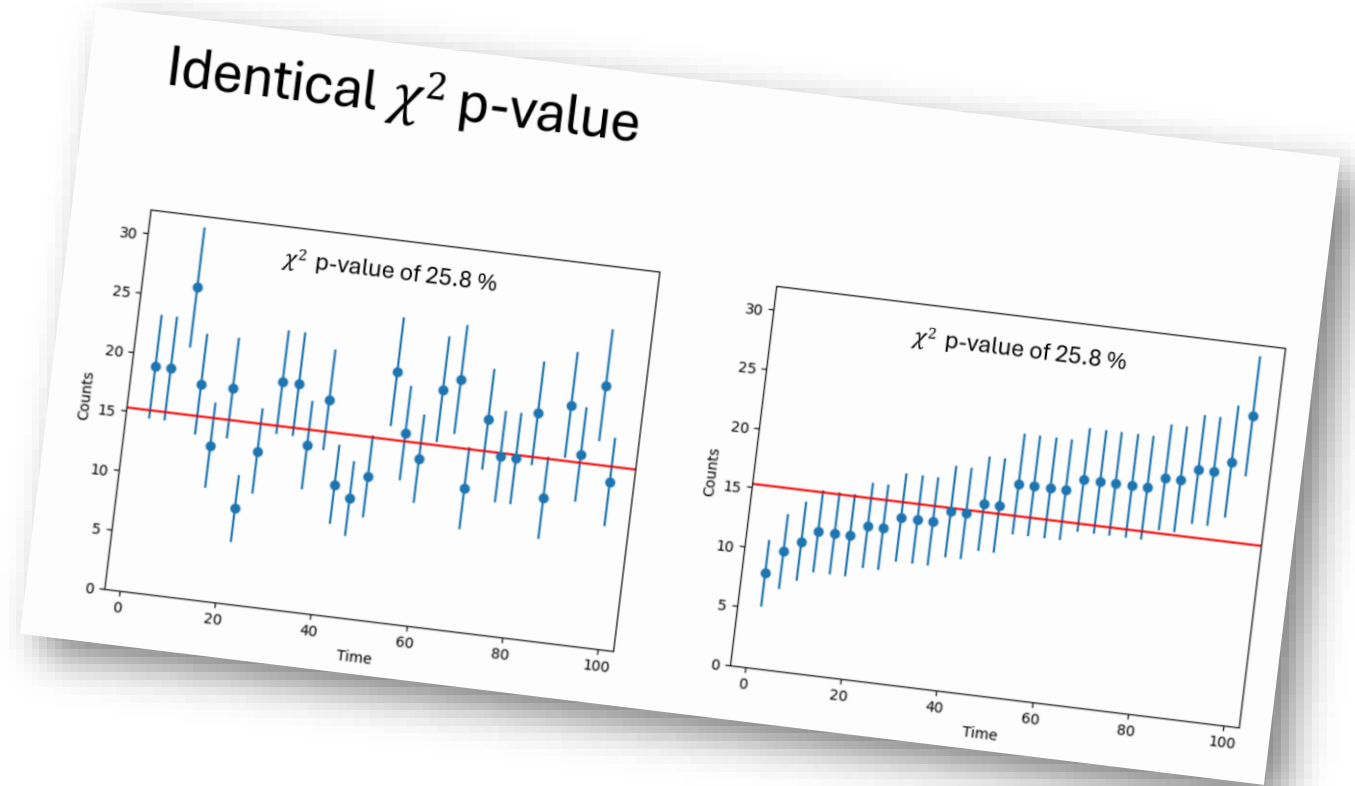
$$L(\lambda) \approx \prod_i \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2} \frac{(x_i - \lambda)^2}{\lambda}}$$

→ A test statistic can be built out of this:  $TS(\lambda) = -2(\ln L(\lambda) - \ln L(\hat{\lambda})) \approx \sum_i \frac{(x_i - \lambda)^2}{\lambda}$

- Distribution of TS under the null hypothesis is analytically known! It follows a chi2 distribution with  $df = n - \#params$ .
- This is known as the Pearson chi2 test
- It is statistically more sound that using the bare likelihood

# Problems

- Order of the bins plays no role in the TS!
- (In the un-binned case, it is even worse 😞)
- Many other test statistics have been proposed
  - Kolmogorov-Smirnov (KS)
  - Anderson-Darling (AD)
  - Cramer-von Mises (CvM)
  - Moran, Greenwood, ...etc.



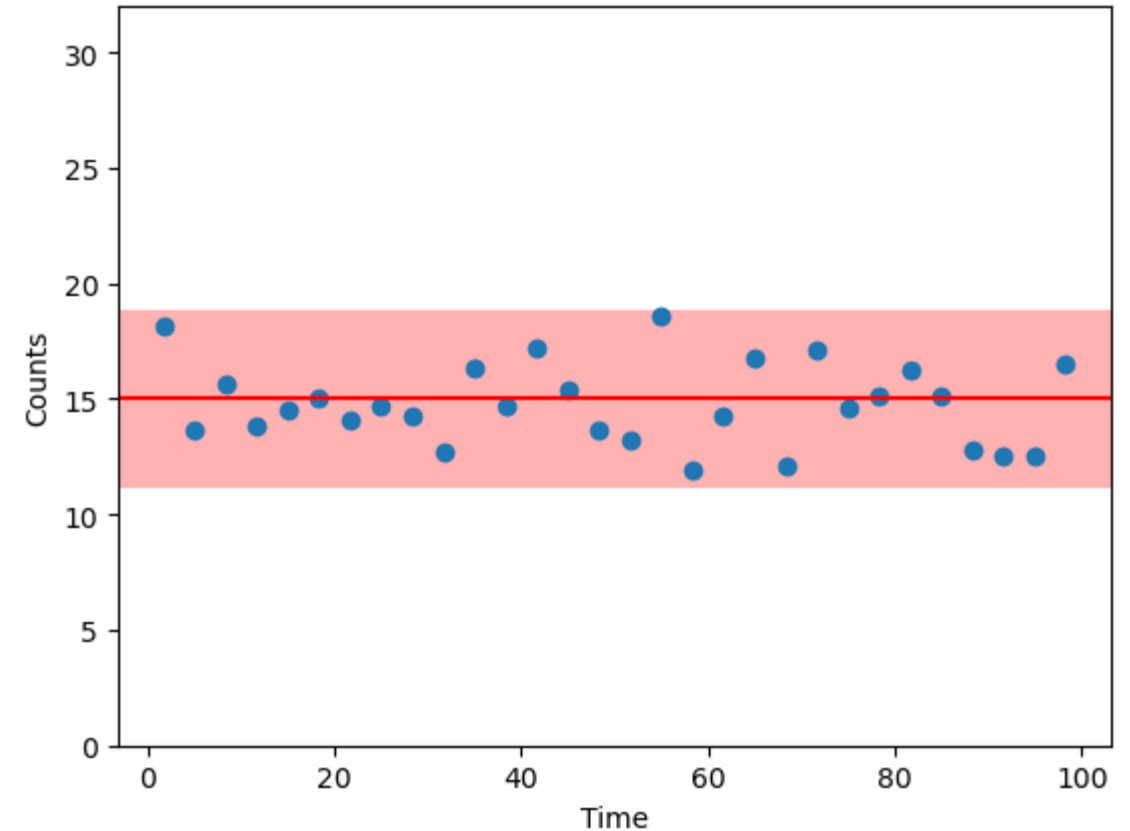
# The real issue here

## Example:

- Generate data from a narrower Gaussian, i.e. from a different model!
- Chi2 G.o.f. is now 99.9999%

## Hard Questions:

- Should this be considered a good g.o.f.?
- The data is following the model so closely, isn't this also very extreme / unlikely!
- Why would we define the rejection region on one tail only? The scenario here I would also reject by eye
- In fact, any (small) region of the same size in the distribution is equally (un)likely...so why should we reject anything in favour of anything else?
- What does our intuition tell us? Is there anything we implicitly assumed?



# Two hypotheses testing

- We can make the situation much better, if we can specify an alternate hypothesis  $H_1$
  - So we can formulate a binary decision on which hypothesis to reject/accept
  - We can formulate our decision based on comparing:
    - The sampling distributions of data under either candidate models
    - with the actually realized data
- See what kind of data the theories produce, and then compare with what we got from our experiment, then make a decision.



# 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

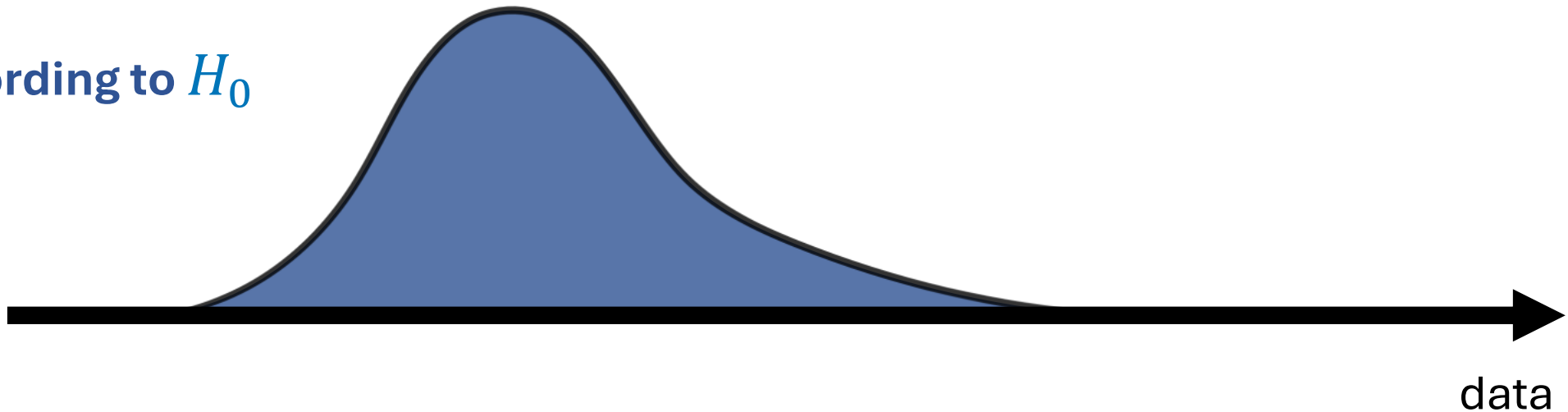
Let's see the data that the hypotheses are predicting



# Testing with Sampling Distributions

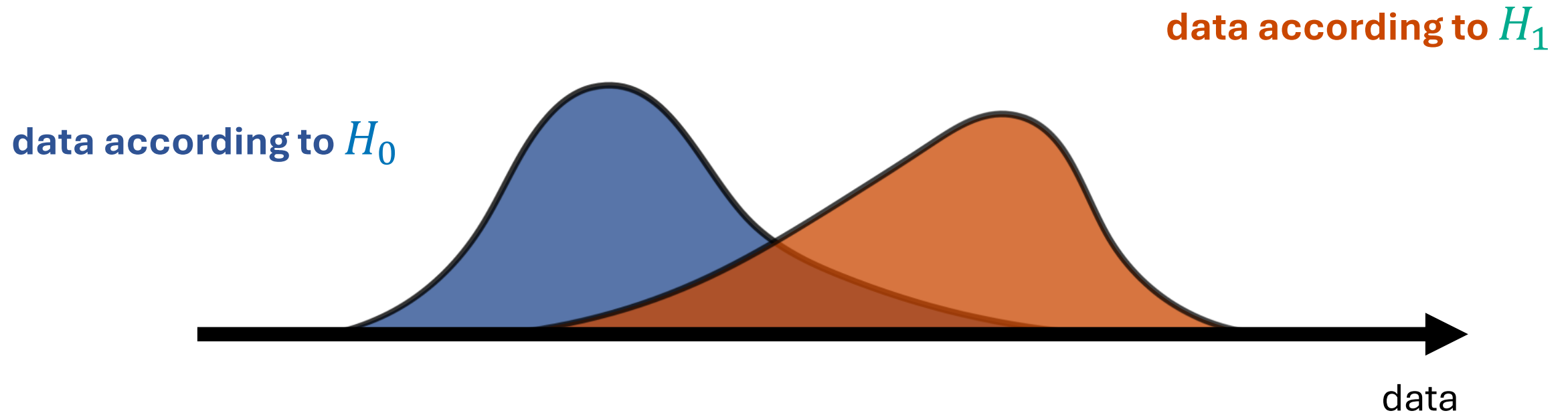
Let's see the data that the hypotheses are predicting

data according to  $H_0$



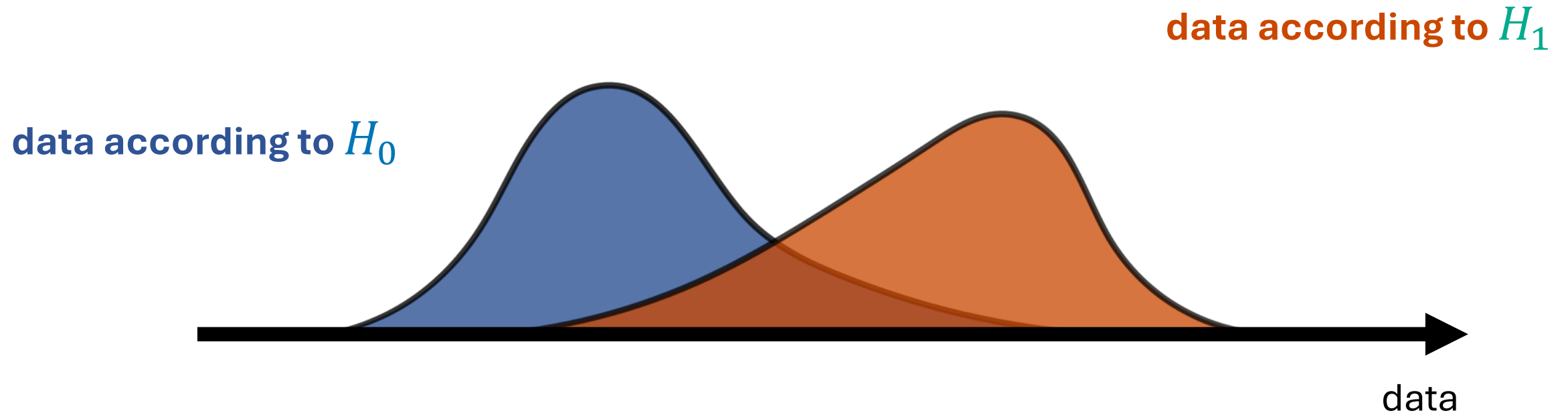
# Testing with Sampling Distributions

Let's see the data that the hypotheses are predicting



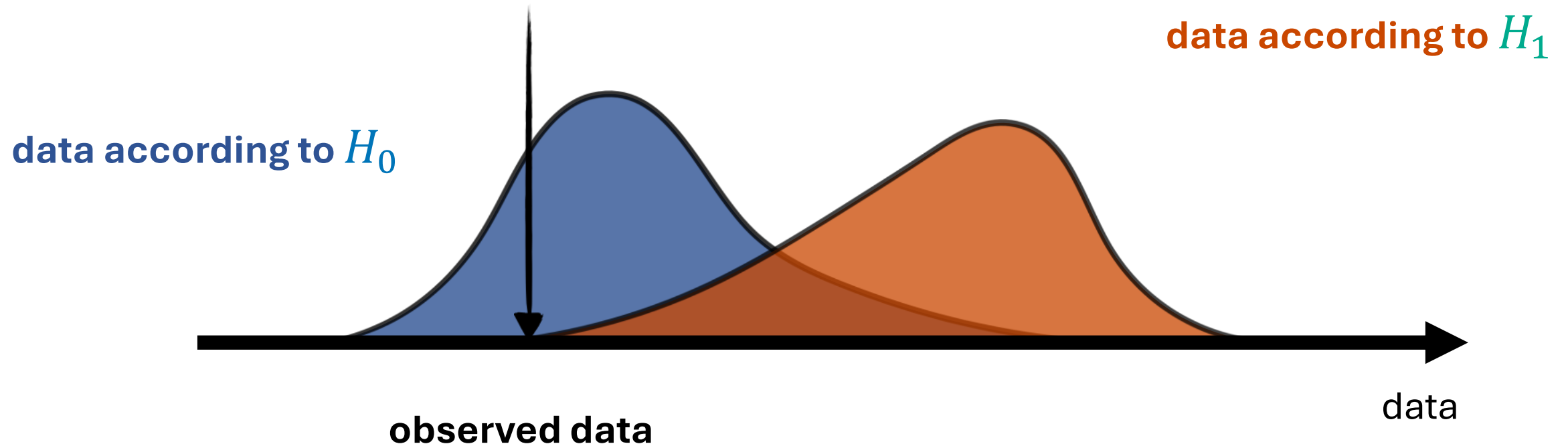
# Testing with Sampling Distributions

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



# Testing with Sampling Distributions

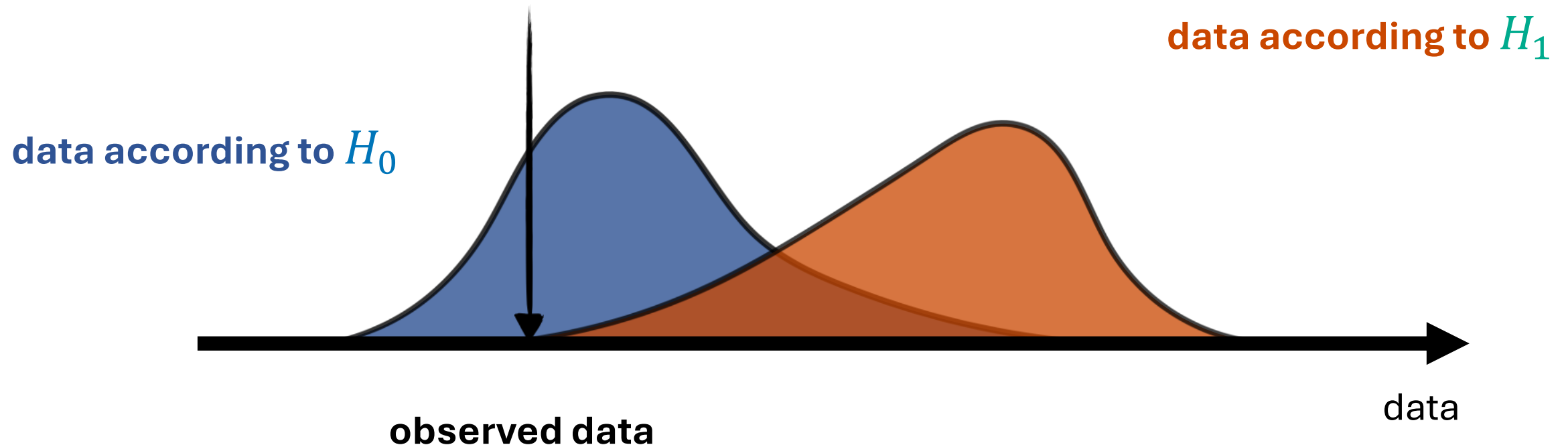
Let's look at the data **observed** in the real world



# Testing with Sampling Distributions

## Question:

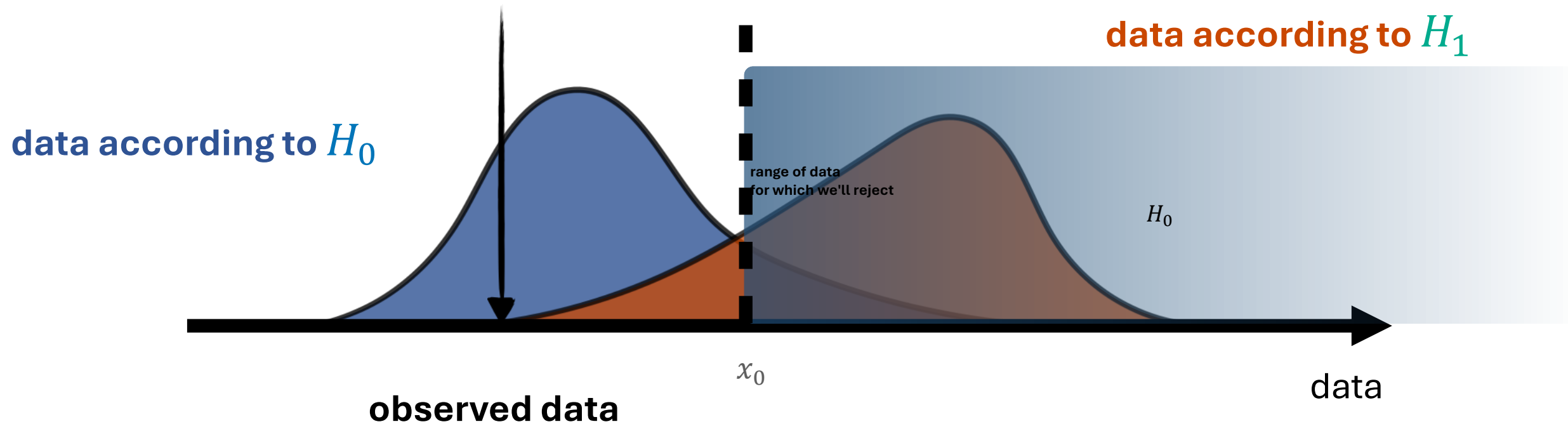
if you had to formulate a rule to reject (or not)  $H_0$  what would it be?



# Testing with Sampling Distributions

A reasonable answer:

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

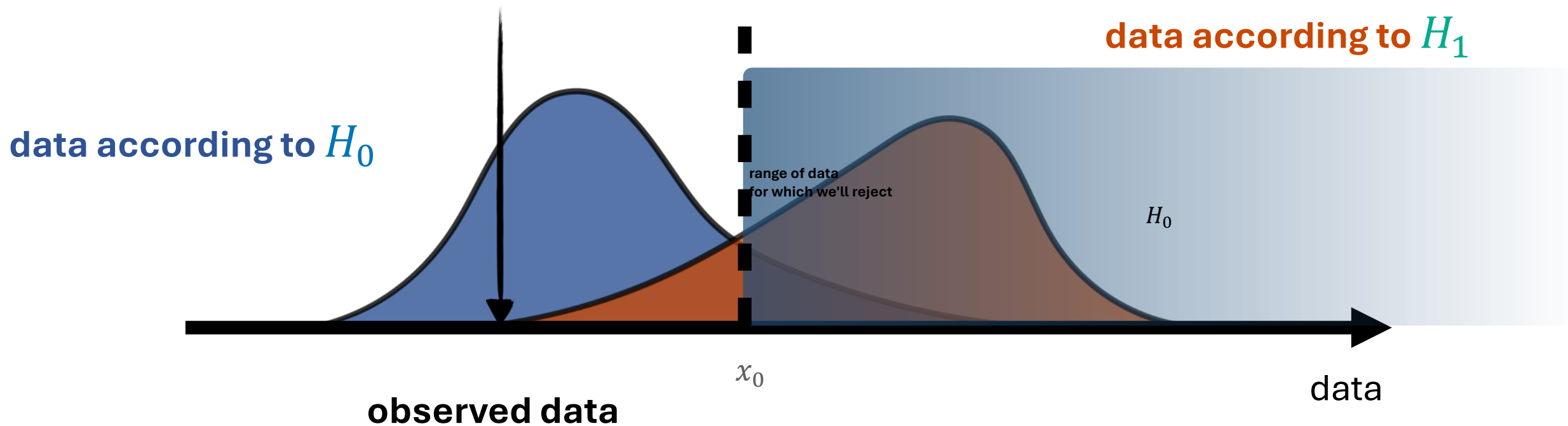




# Testing with Sampling Distributions

A reasonable answer:

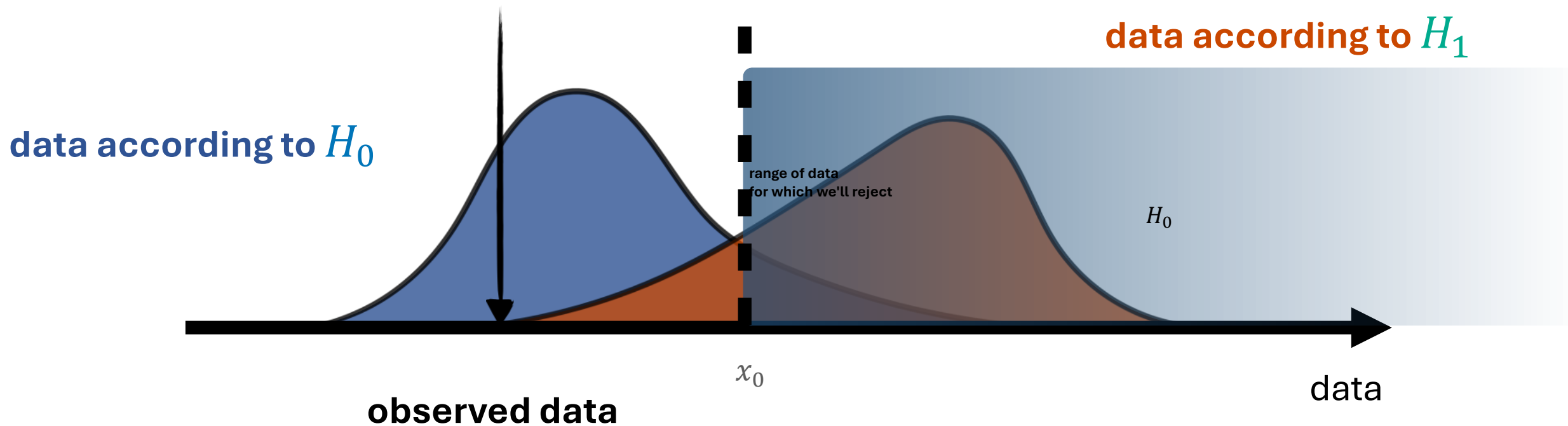
- reject  $H_0$  if data is too far too the right (e.g. data  $> x_0$ )
- follow-up question: how do you choose  $x_0$ ?



# Testing with Sampling Distributions

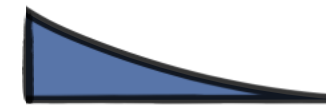
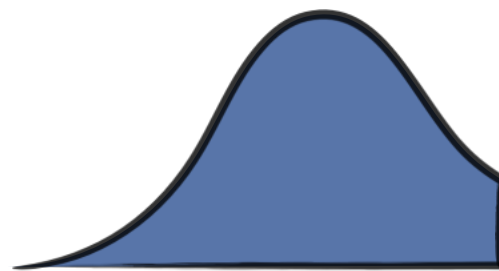
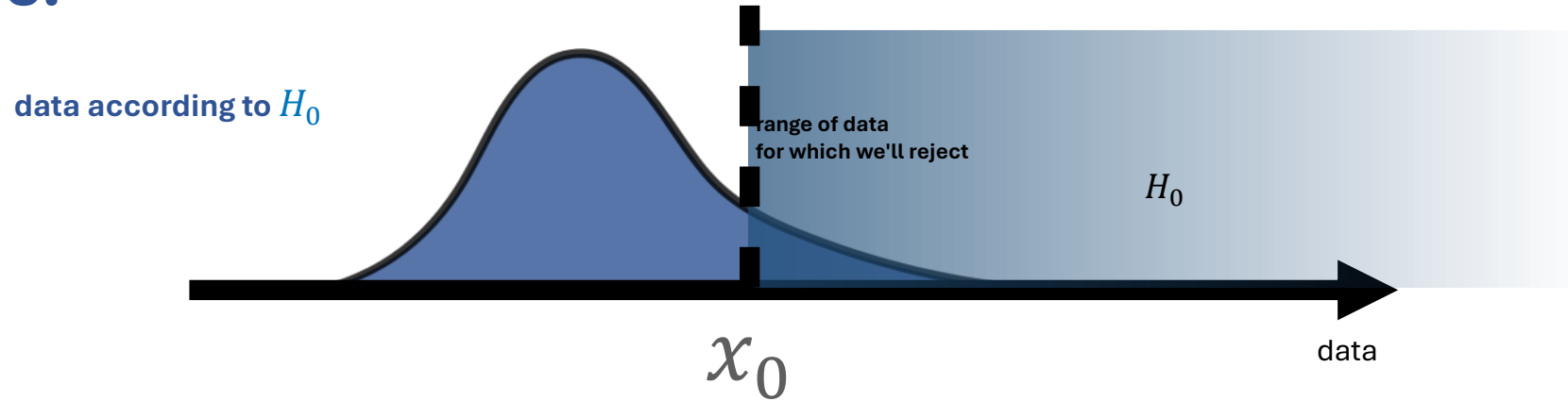
## How good is this rule?

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



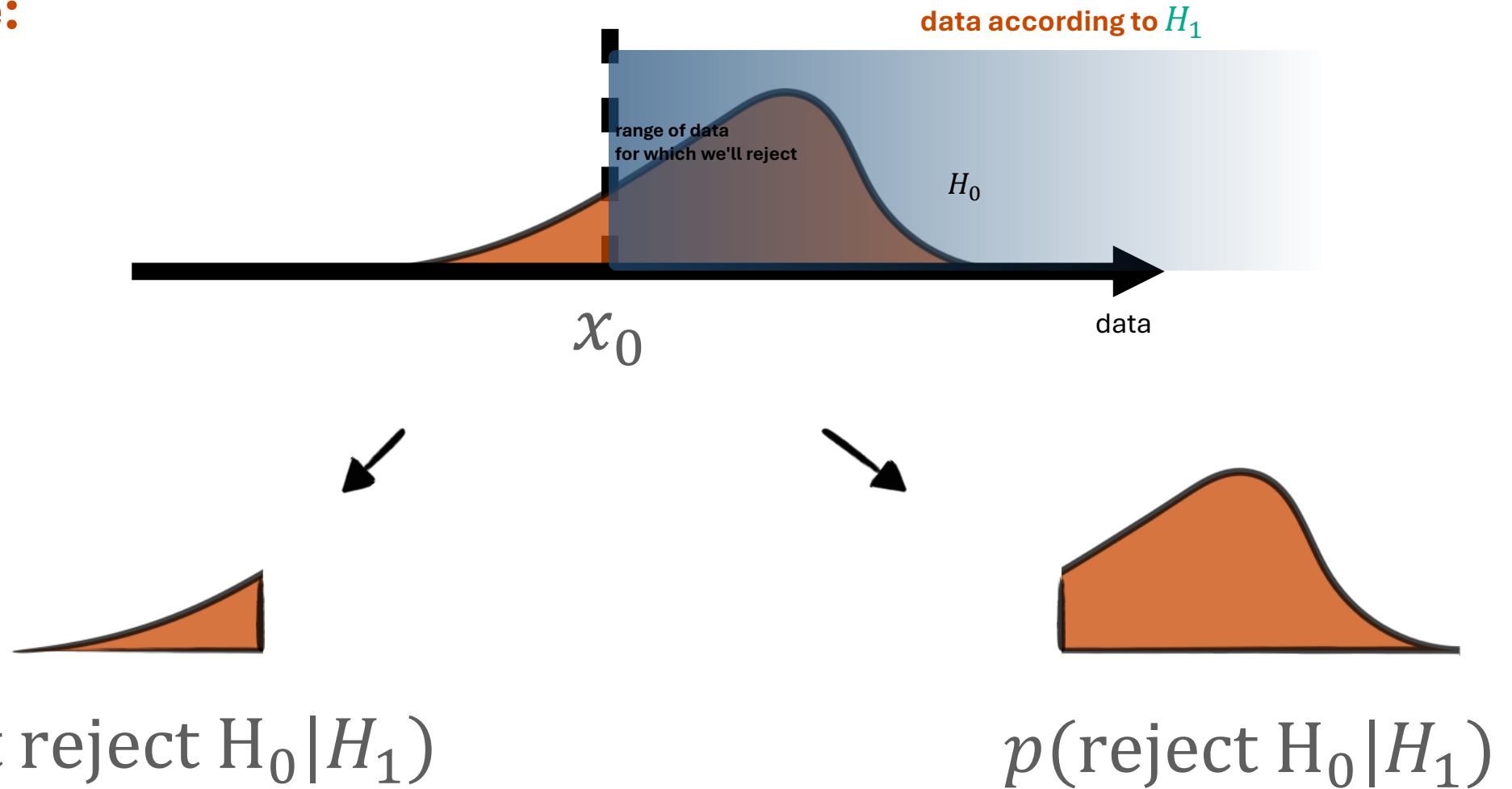
# Testing with Sampling Distributions

If  $H_0$  is true:



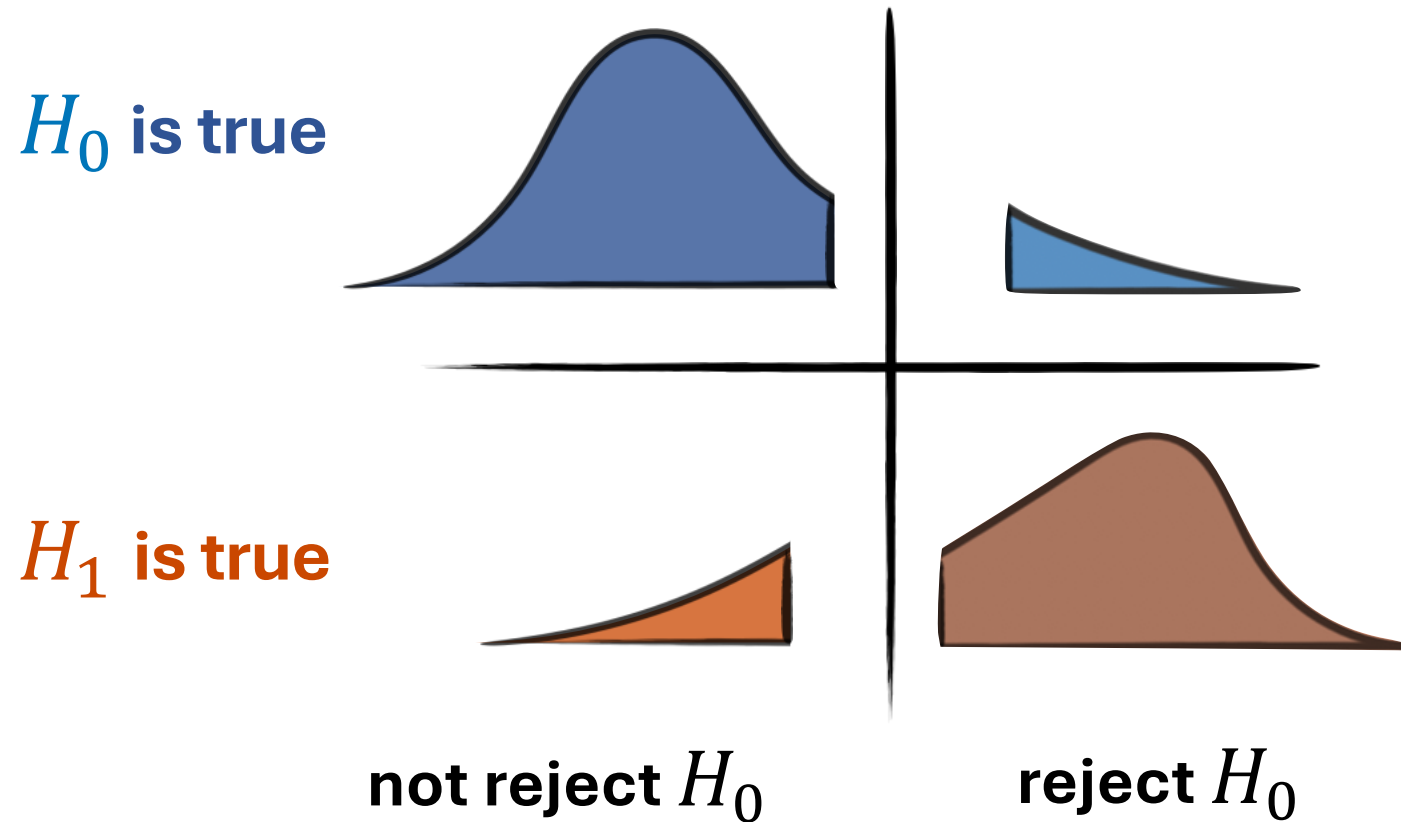
# Testing with Sampling Distributions

If  $H_1$  is true:



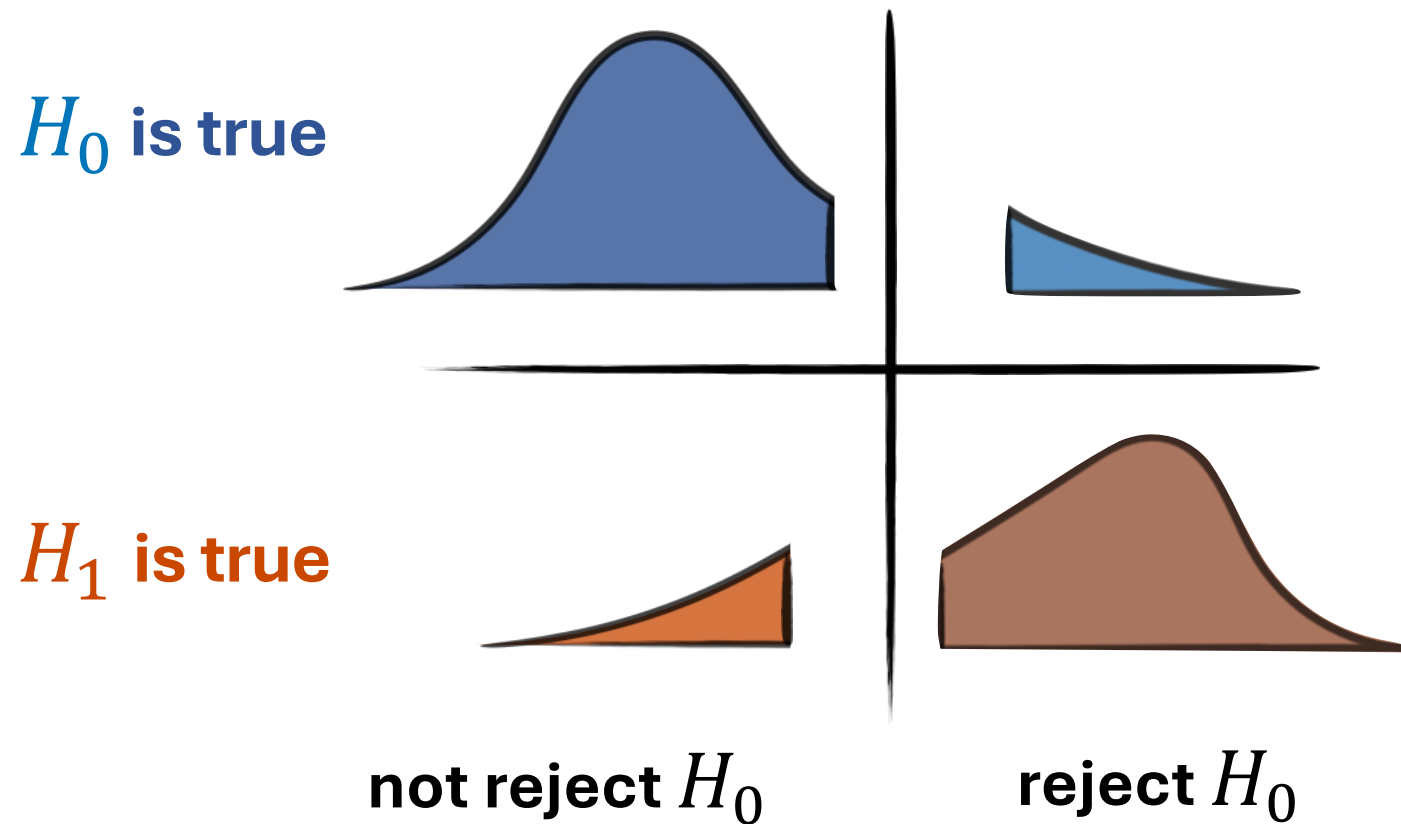
# Type I & Type II Error

## Two Hypotheses & Two Actions



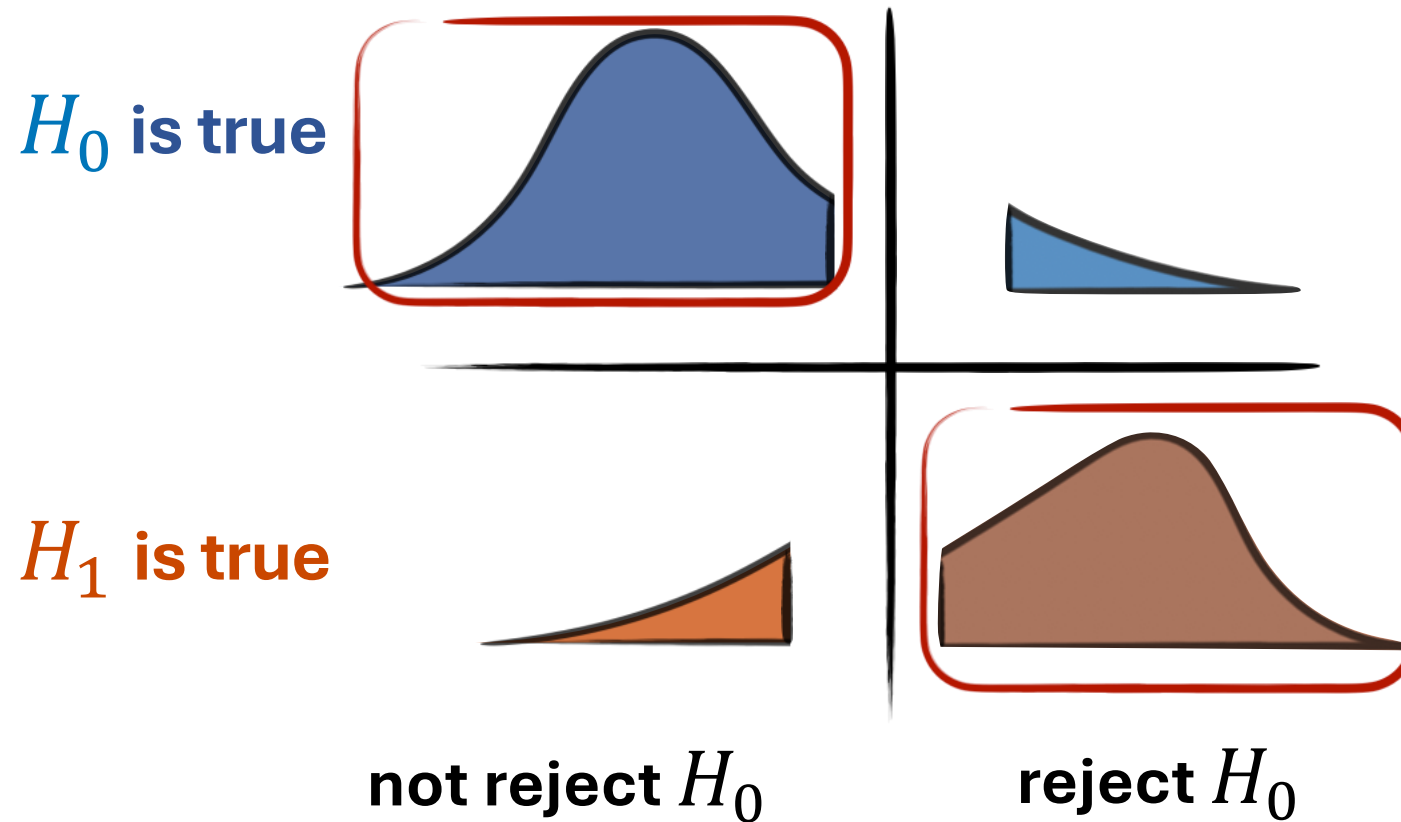
# Type I & Type II Error

What are the favorable decisions?



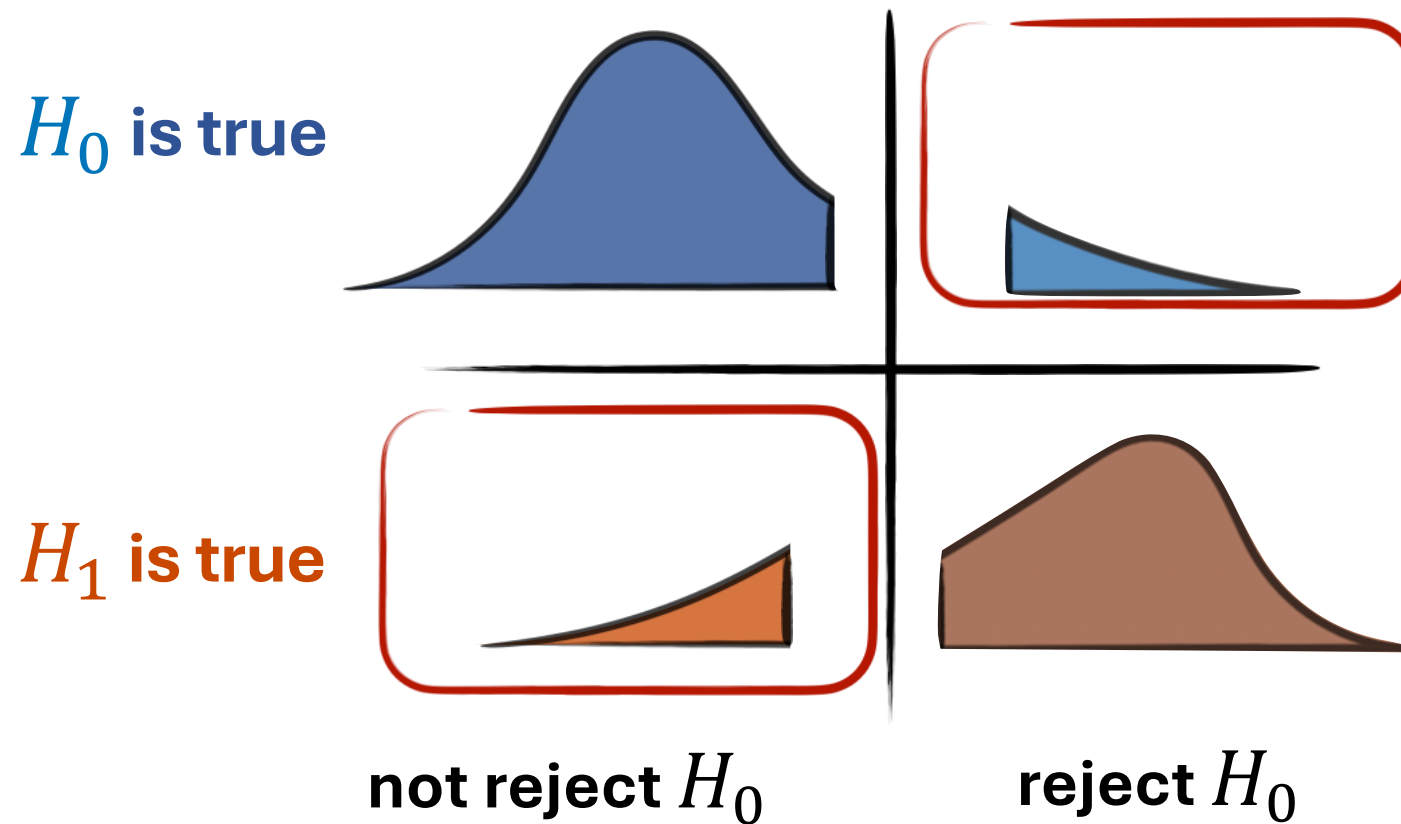
# Type I & Type II Error

What are the favorable decisions?



# Type I & Type II Error

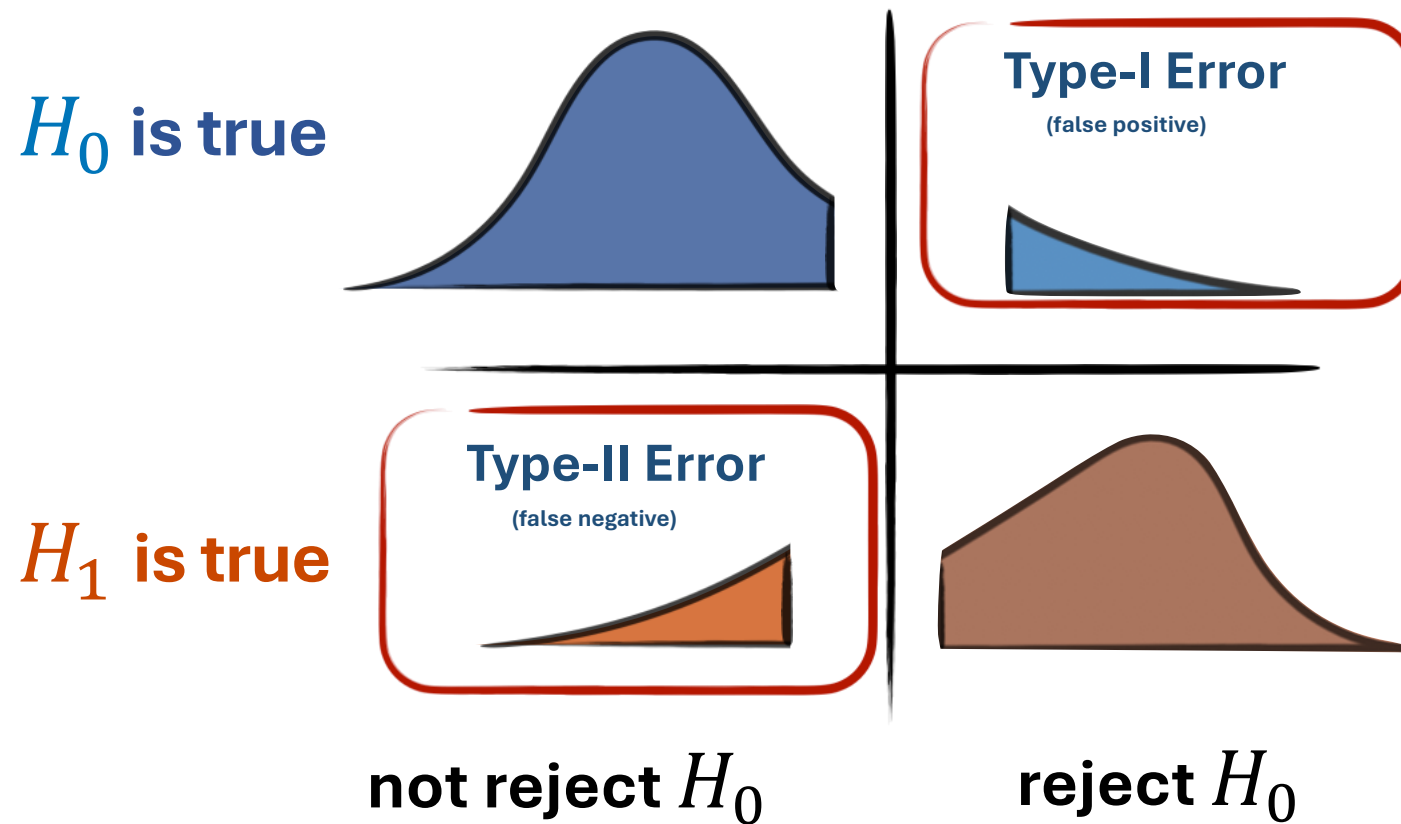
What are the unfavorable decisions?





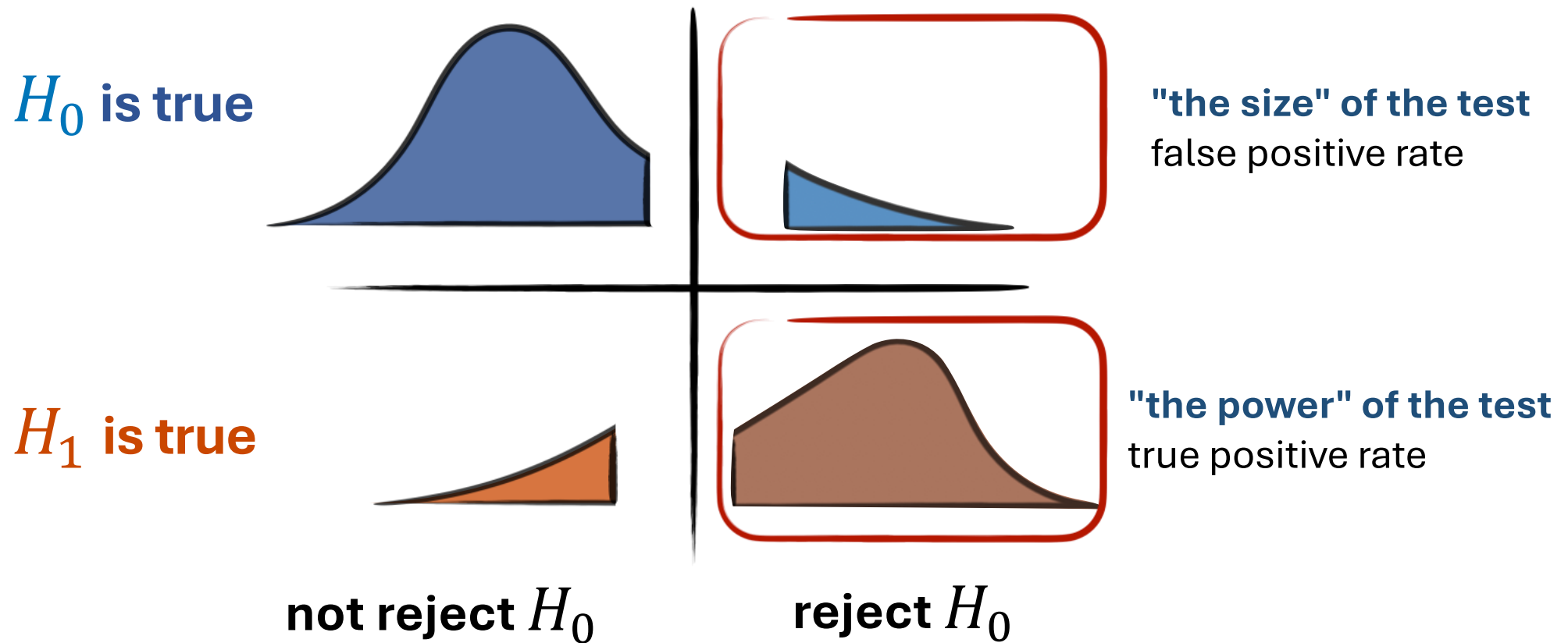
# Type I & Type II Error

What are the bad decisions?



# Type I & Type II Error

The Performance of the test is fully characterized by two numbers



# 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

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

# p-values

- The probability of wrongly rejecting the null hypothesis for the **observed value** obtained from the data
- This is very similar to the size of the test
  - 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

# p-Values: lots of confusion

p-values do not have a good reputation so let's reiterate & understand

- p-values are a useful to report relationship of data with null hypothesis in a portable way independent of test statistic
- In fact, p-values are uniformly distributed under  $H_0$ !
- test stat. value  $t(\text{data})$  is harder to interpret w/o knowing details
- "The p-value" is at **the observed data** expressed in these units  
 $p(\text{data}) \ll \alpha$ : data is **deep in rejection region**  
 $p(\text{data}) \gg \alpha$ : data **not close to rejection region**

# p-Values: lots of confusion

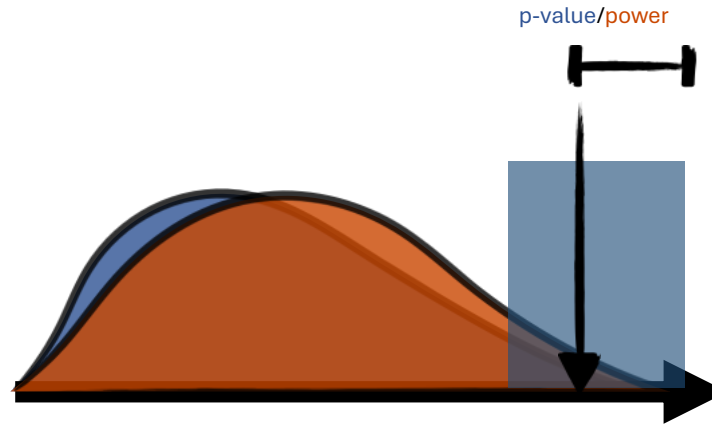
p-values do not have a good reputation so it's reiterate & understand

**p-values are not probability that  $H_0$  is false give the data  $p(\neg H_0 | x)$**

**(frequentist analysis doesn't put probabilities on hypotheses)**

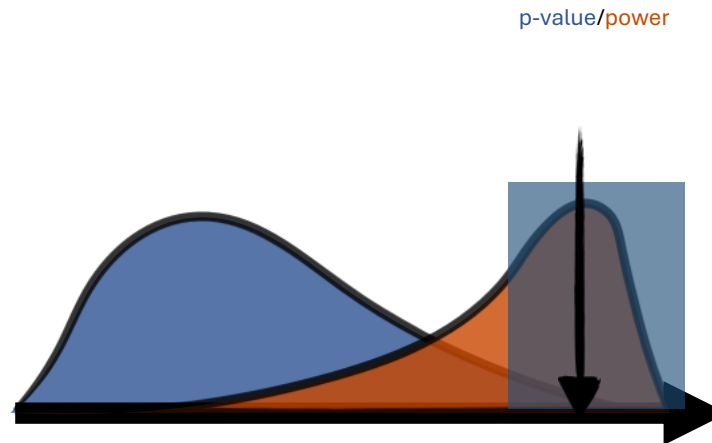
**Frequent source of confusion by many (scientists & not)**

# Two examples of low p-values



**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

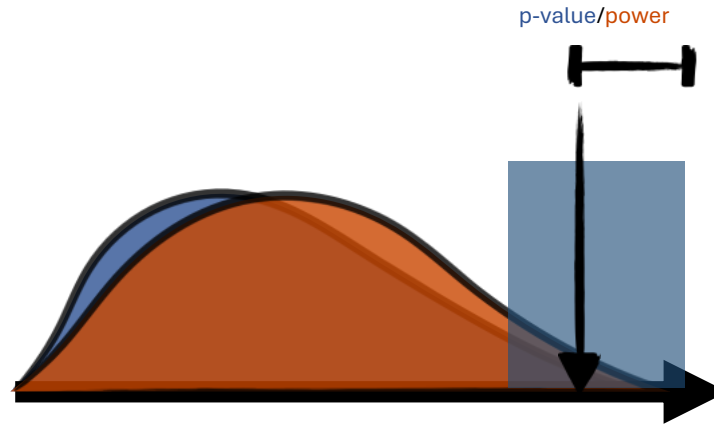
For both experiments  
the observed data has  
low p-value ( $<0.05$ ).

**Should you reject  
the null hypothesis?**

# Two examples of low p-values

## Low-power test:

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

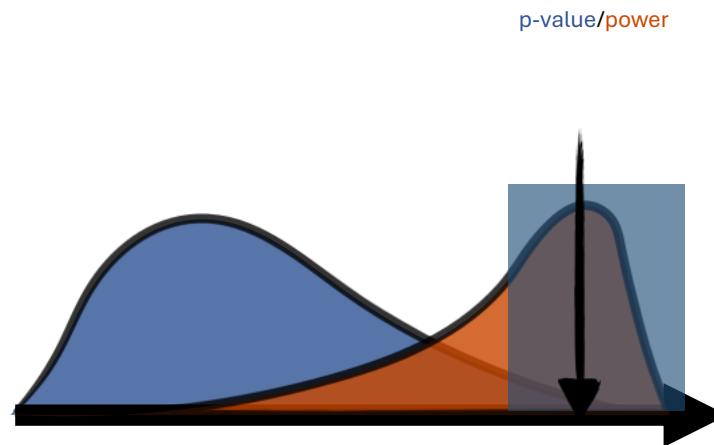


**Experiment 1:  
(low sensitivity)**

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



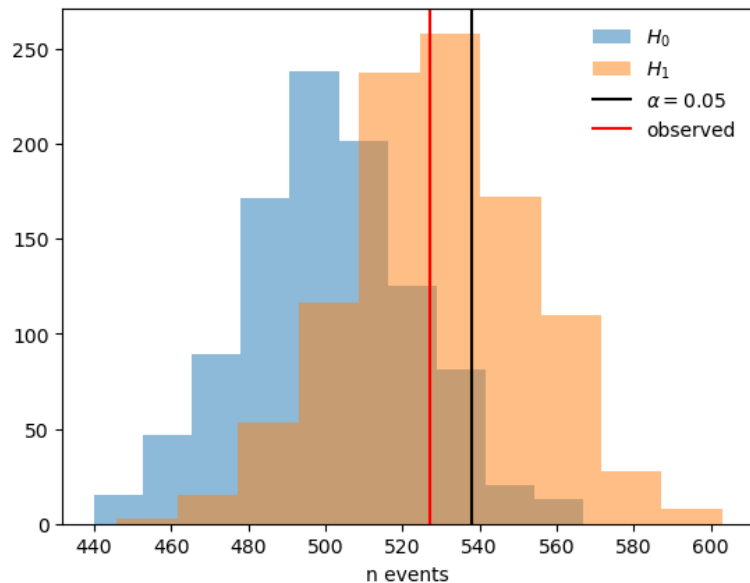
# Back to our Example

- **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:
  - $H_0$ : Rate is constant at 5
  - $H_1$ : Rate increased from 5 to 6 at  $T = 70$

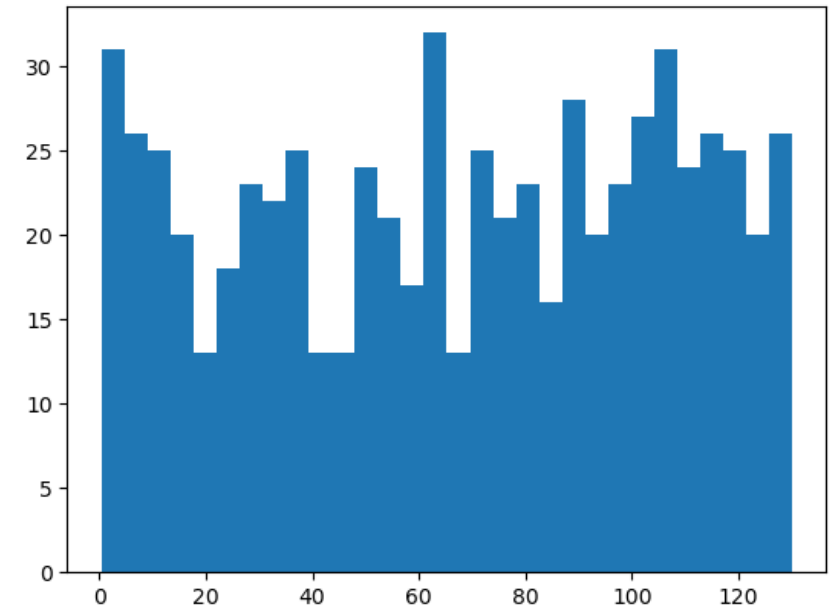


# Test statistic

- We again 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)



The new data you collected:



→ We cannot reject the null Hypothesis at the 5% level

# Neyman-Pearson Lemma

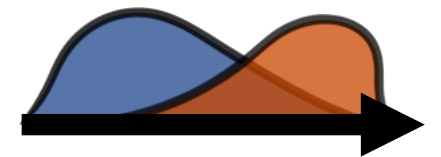
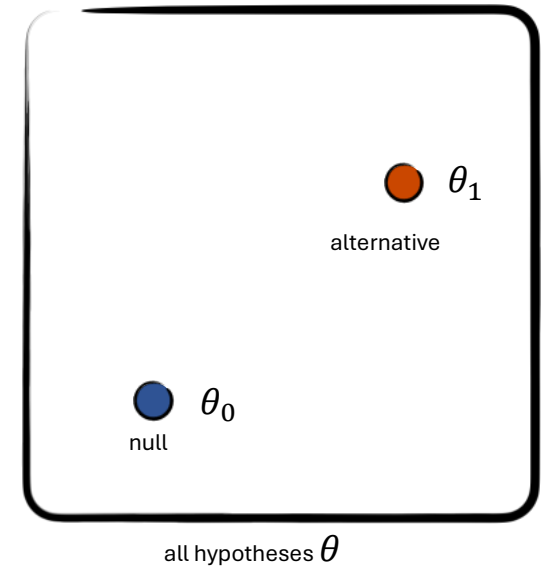
If we want to a null  $p(x|\theta_0)$  vs an alternative  $p(x|\theta_1)$  we have a very compelling answer

The Neyman-Pearson Lemma:

**The Likelihood Ratio:**

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

is the optimal test statistic

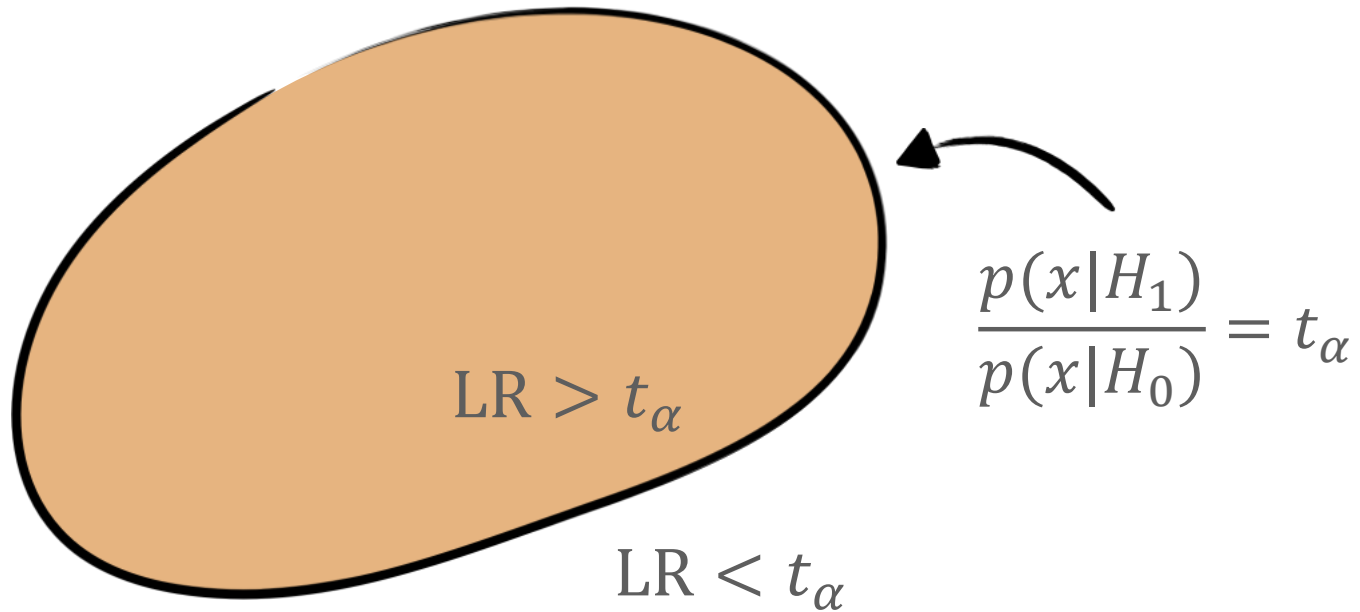


# Neyman-Pearson Lemma

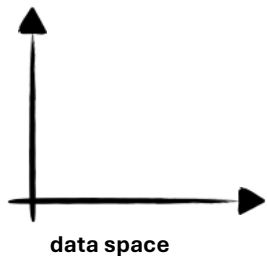
With Likelihood Ratio as test statistic, we reject  $H_0$  when  $t(x)$  indicates that data is too  $H_1$ -like: **more likely to get  $x$  under  $H_1$  than  $H_0$**

$$t(x) = \frac{p(x|\theta_1)}{p(x|\theta_0)} > t_\alpha \quad \text{or equivalently} \quad \lambda(x) = -2\log \frac{p(x|\theta_0)}{p(x|\theta_1)} > \lambda_\alpha$$

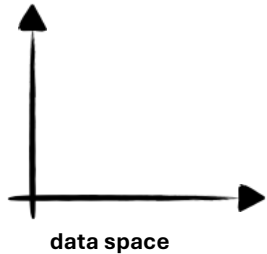
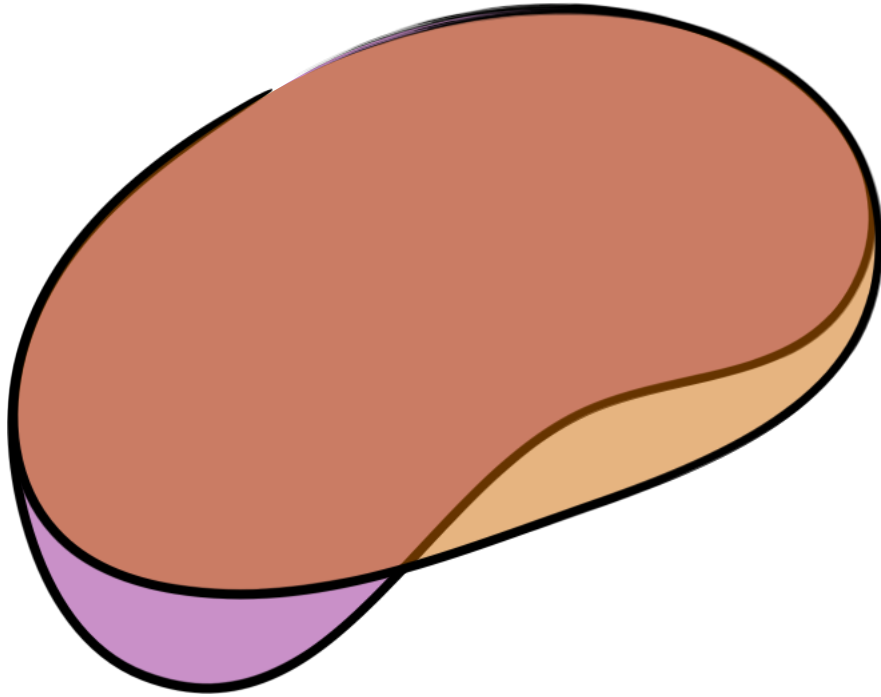
# Visual Proof of NP-Lemma



Start with the **rejection region**  
as defined by Neyman-Pearson

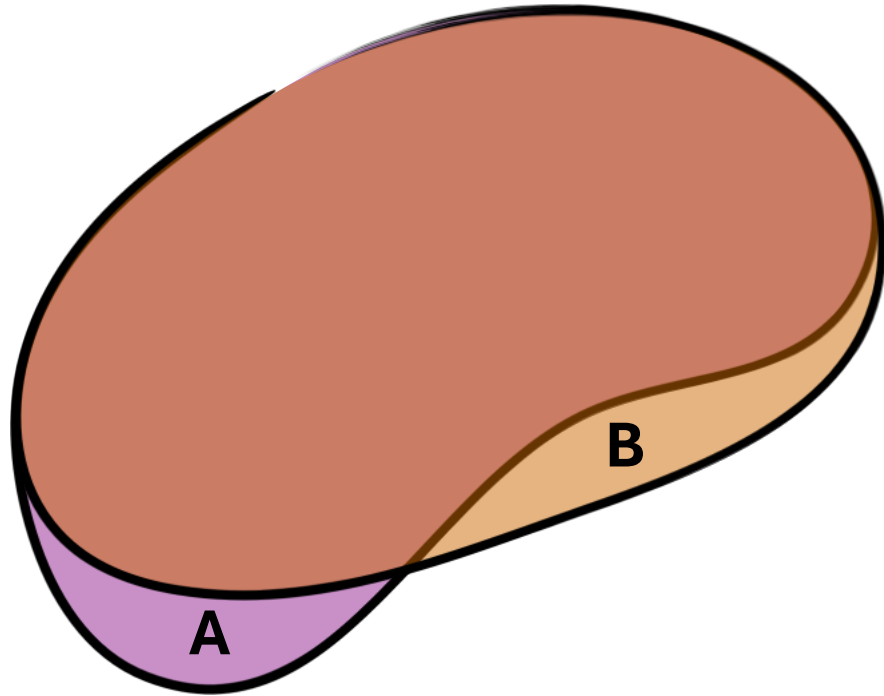


# Visual Proof of NP-Lemma

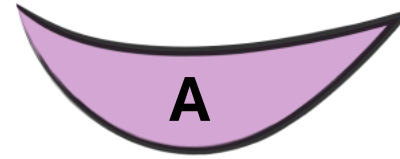


**Consider a potential **alternative region** and see how its power compares**

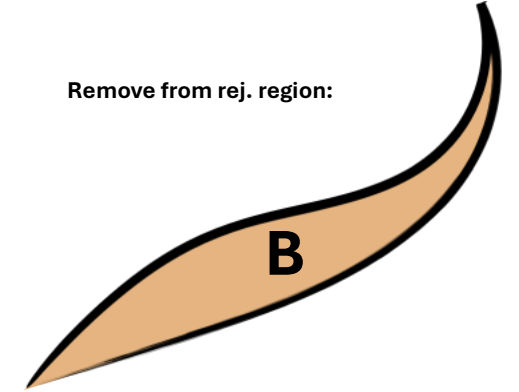
# Visual Proof of NP-Lemma



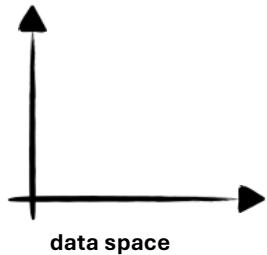
Add to rej. region:



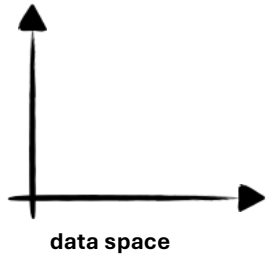
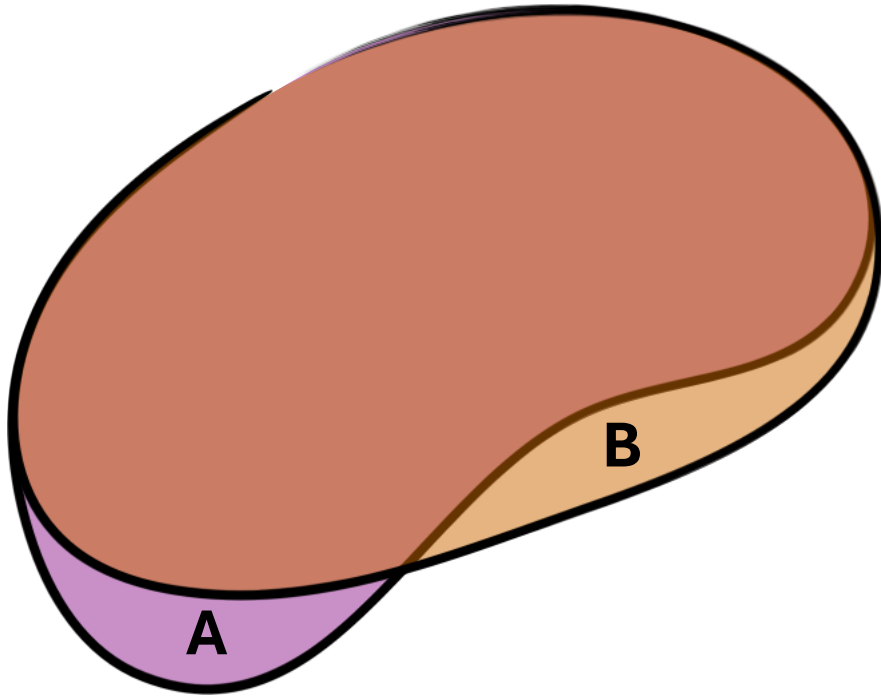
Remove from rej. region:



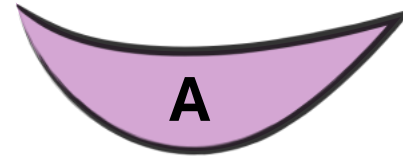
**Focus on the areas that  
are different between the two**



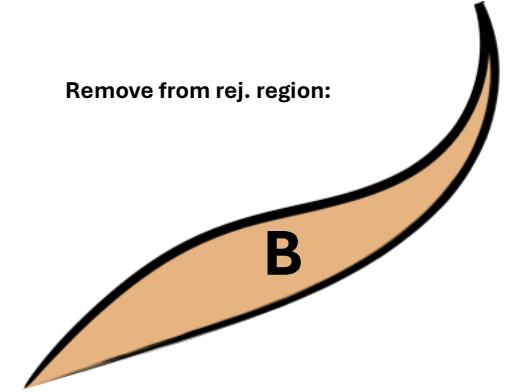
# Visual Proof of NP-Lemma



Add to rej. region:



Remove from rej. region:

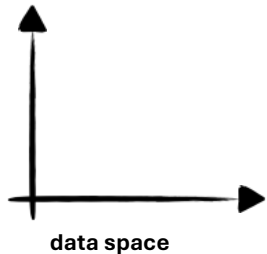
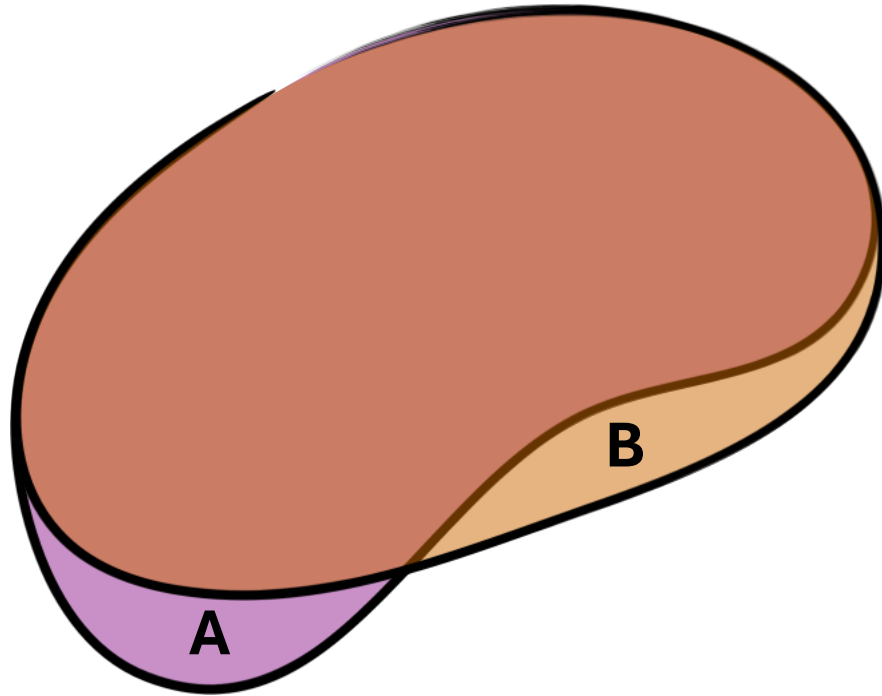


①  $p(A|H_0) = p(B|H_0)$

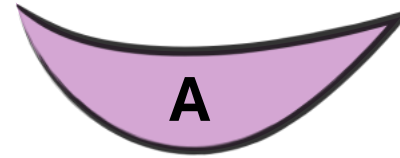
test size should stay constant



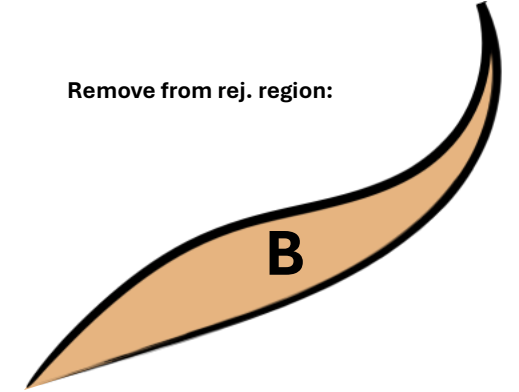
# Visual Proof of NP-Lemma



Add to rej. region:



Remove from rej. region:



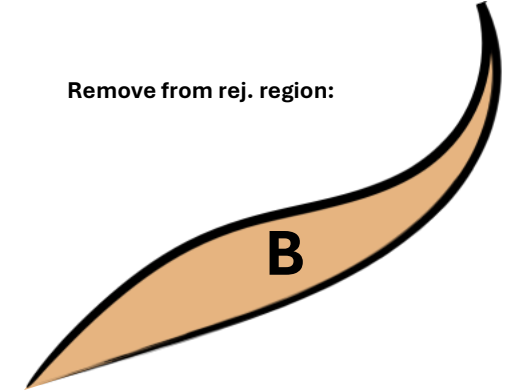
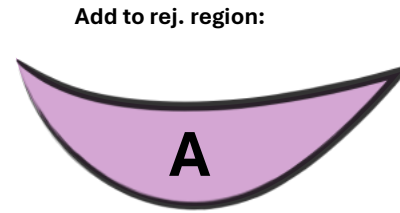
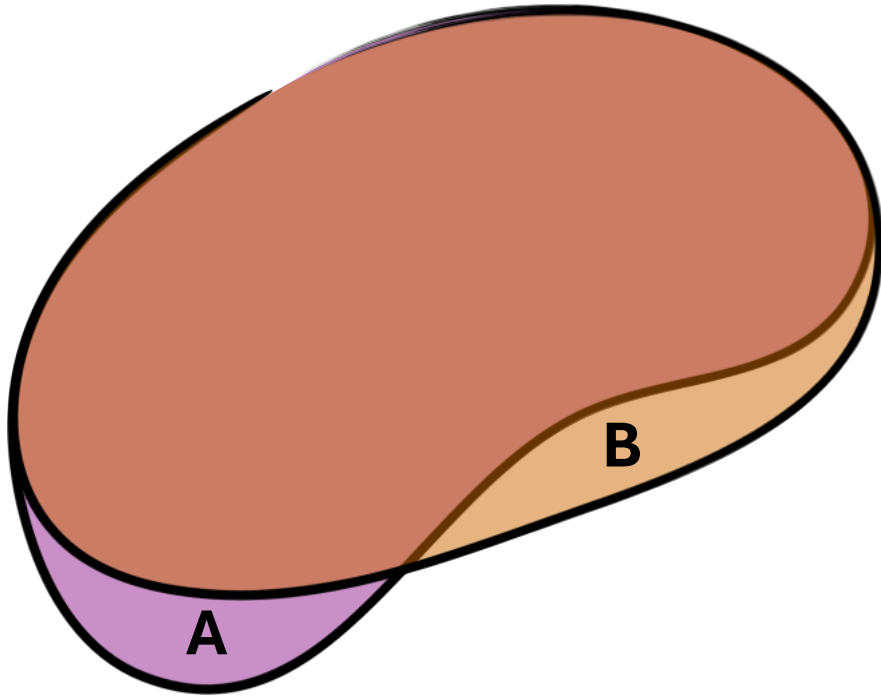
①  $p(A|H_0) = p(B|H_0)$

②  $p(B|H_1) > t_\alpha p(B|H_0)$

test size should stay constant

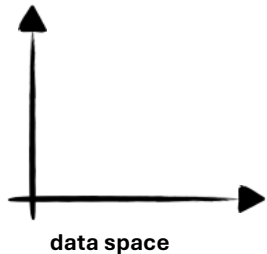
removed region is over threshold  $k_\alpha$

# Visual Proof of NP-Lemma

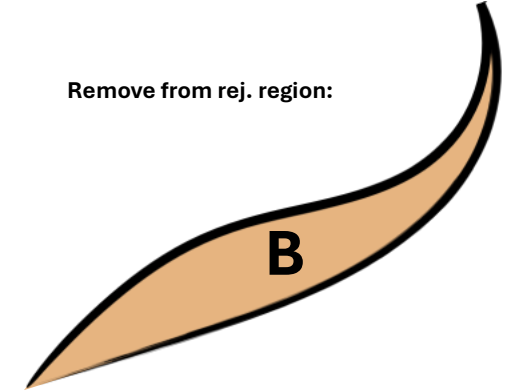
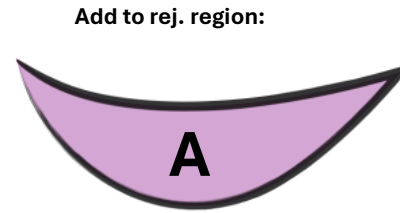
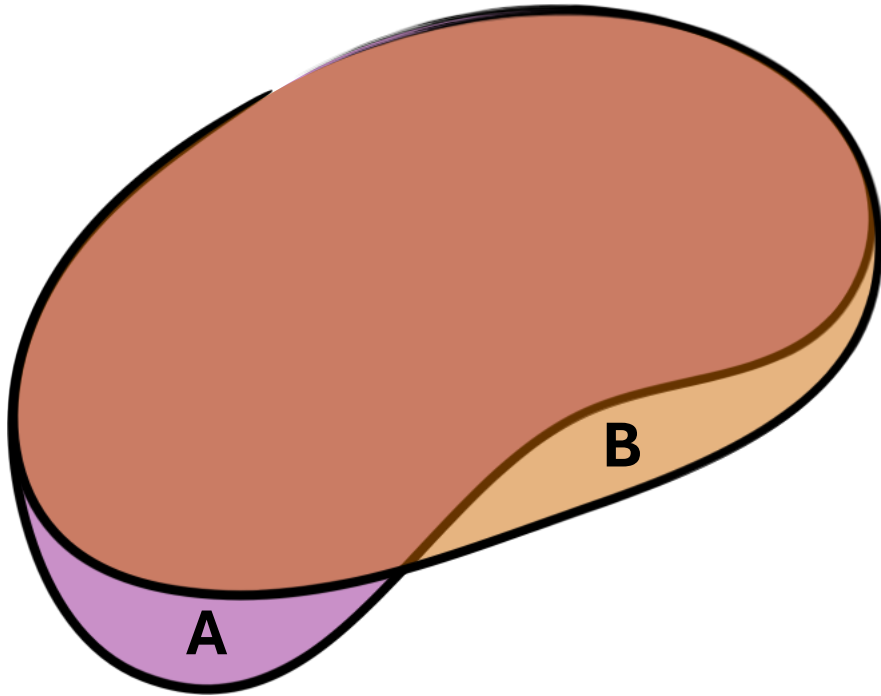


- ①  $p(A|H_0) = p(B|H_0)$
- ②  $p(B|H_1) > t_\alpha p(B|H_0)$
- ③  $p(A|H_1) < t_\alpha p(A|H_0)$

test size should stay constant  
removed region is over threshold  $k_\alpha$   
added region is under threshold  $k_\alpha$



# Visual Proof of NP-Lemma



①  $p(A|H_0) = p(B|H_0)$

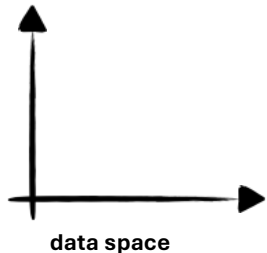
②  $p(B|H_1) > t_\alpha p(B|H_0)$

③  $p(A|H_1) < t_\alpha p(A|H_0)$

test size should stay constant

removed region is over threshold  $k_\alpha$

added region is under threshold  $k_\alpha$

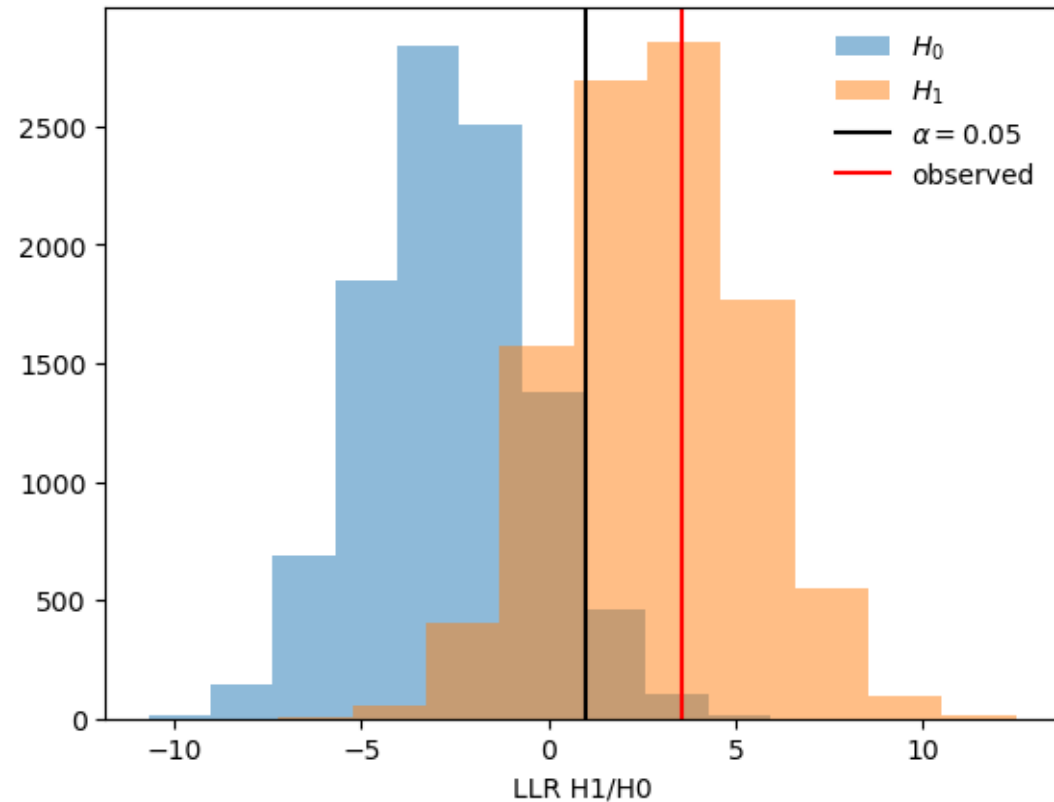


$$p(A|H_1) < t_\alpha p(A|H_0) = t_\alpha p(B|H_0) < p(B|H_1)$$

**new region has less power than NP-region!**

# Using Likelihood Ratio

- Using Neyman-Pearson, i.e. the LHR as TS, we can reject the null at  $> 5\%$ !
- P-value of

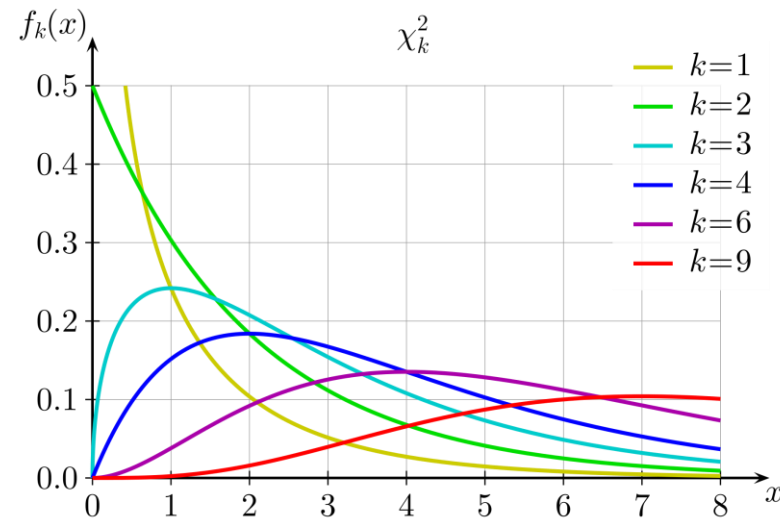


# Nested Hypotheses

- In the example before we were comparing „simple“ hypotheses
  - They were one fully specified model, no parameters to change
- In real life, we usually encounter complex hypotheses.
- In particular, so-called „Nested hypotheses“
  - $H_0$  is part of  $H_1$
- If we do not make assumptions on the rates we observe then our constant model is a strict subset of our model that introduces two separate rate → Nested hypotheses

# Wilk's theorem

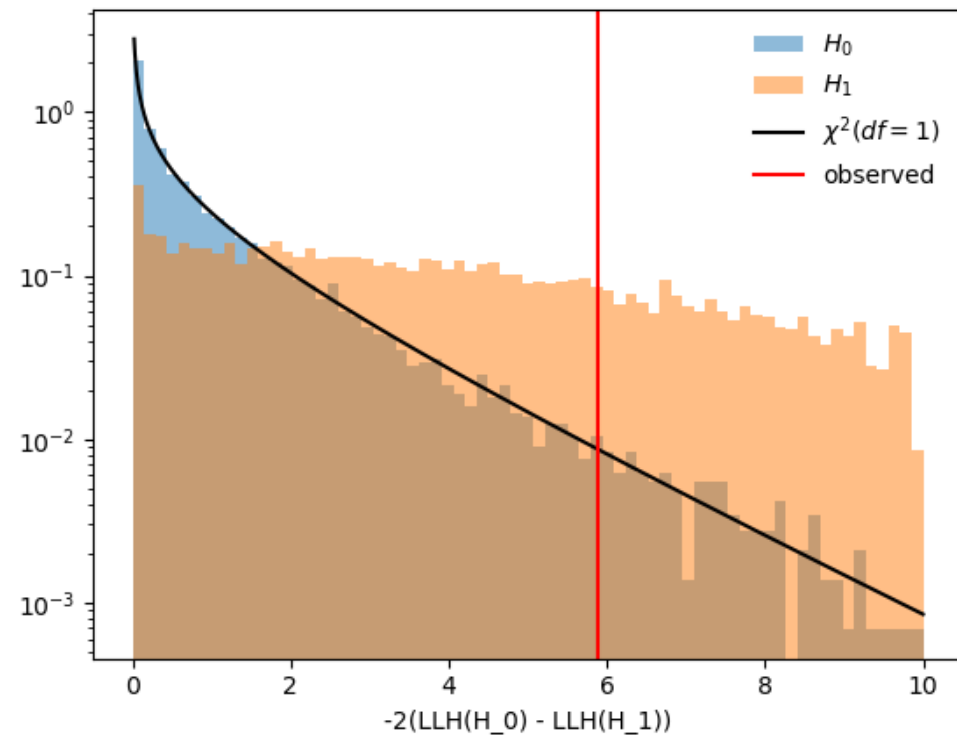
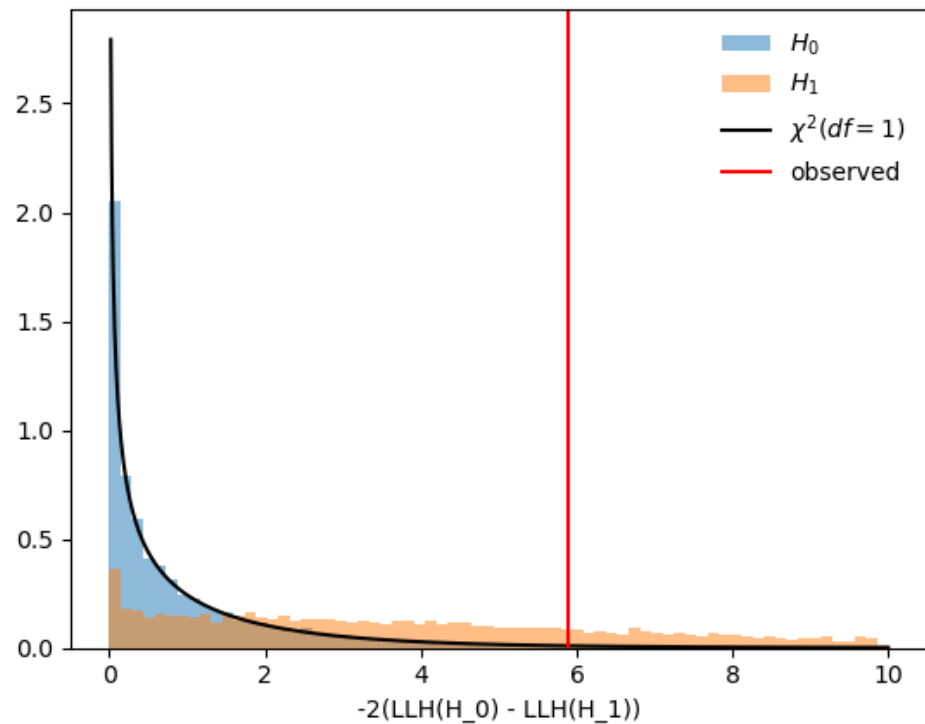
- The LRT test statistic ( $-2\Delta LLH$ ) under the null hypothesis is distributed as the chi2 distribution!



- The degrees of freedom (df) is the difference in free parameters between  $H_0$  and  $H_1$  (In our case  $df=1$ )

# Fixing the 3<sup>rd</sup> Problem: 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  $\chi^2$



# Interval Estimation

# Recap: Point Estimators

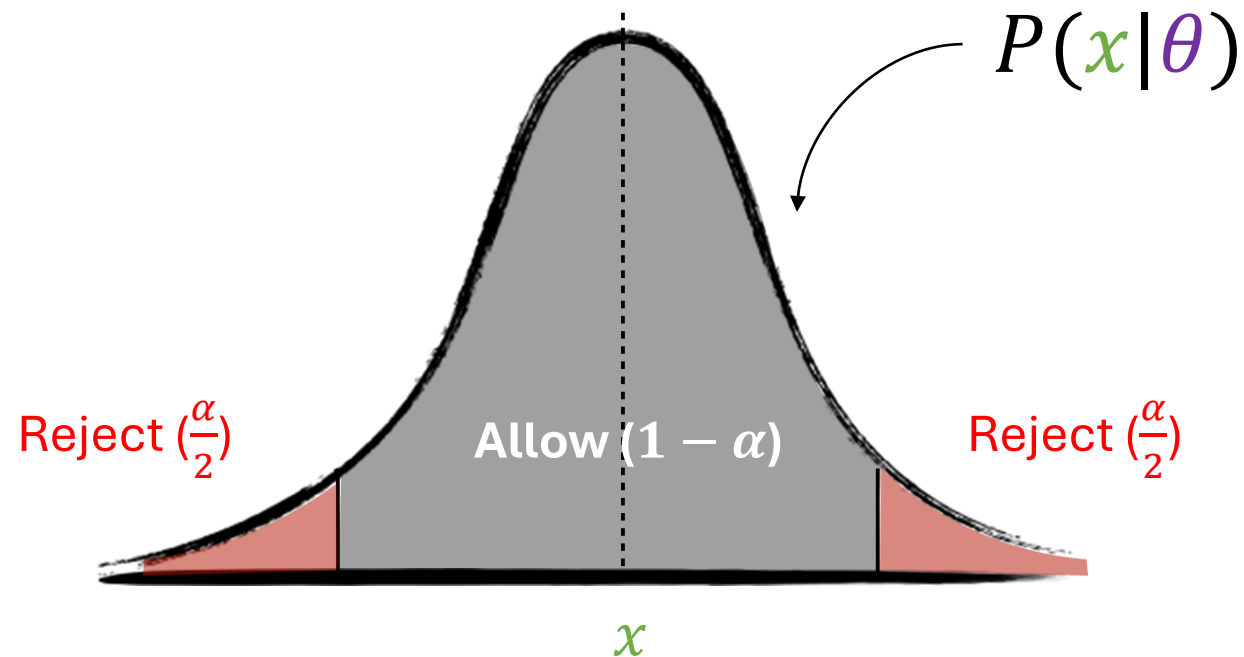
- Remember:
  - The principle of maximum likelihood gives us desirable estimators  
→ Asymptotically efficient and unbiased
- For example in the case of a Gaussian:
  - **Sample mean** as the MLE of the location  $\mu$  of the distribution
  - **Sample standard deviation** as the MLE of the scale  $\sigma$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

→ **How can we then build meaningful intervals?**

# Neyman Construction

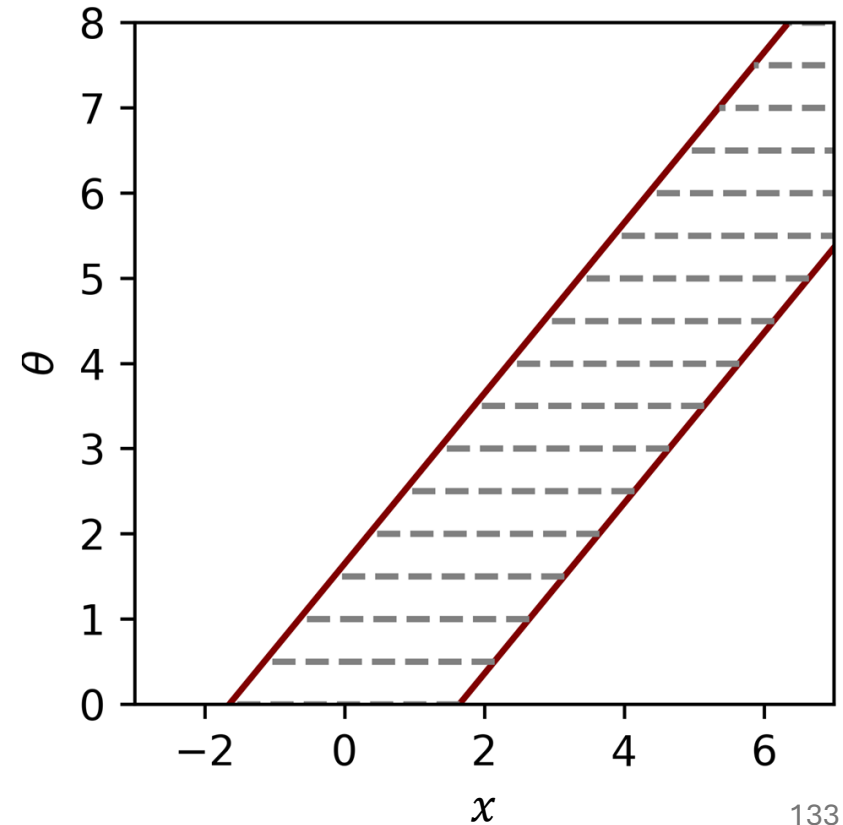
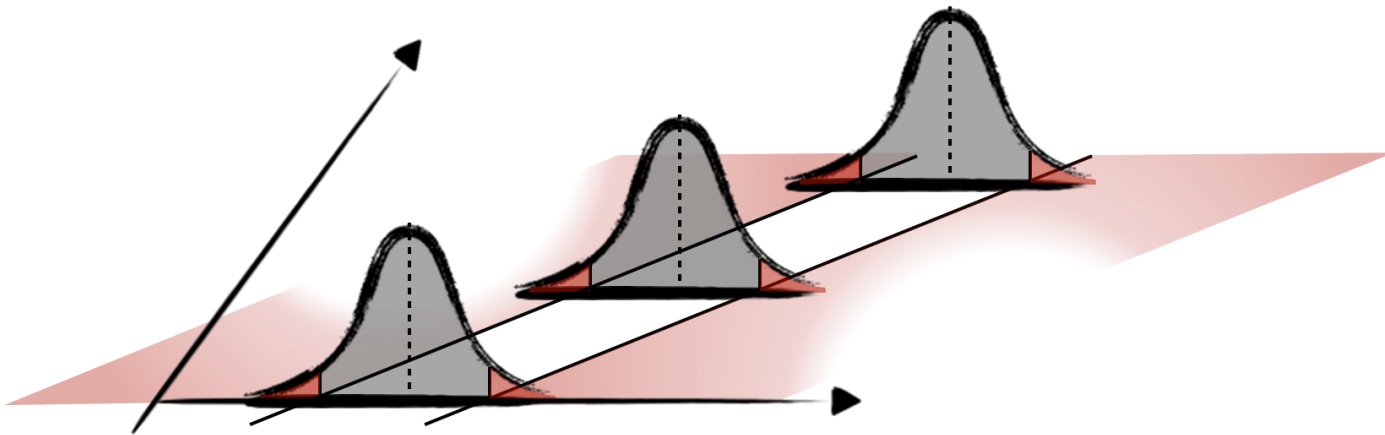
# Start with intervals in observed $x$

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



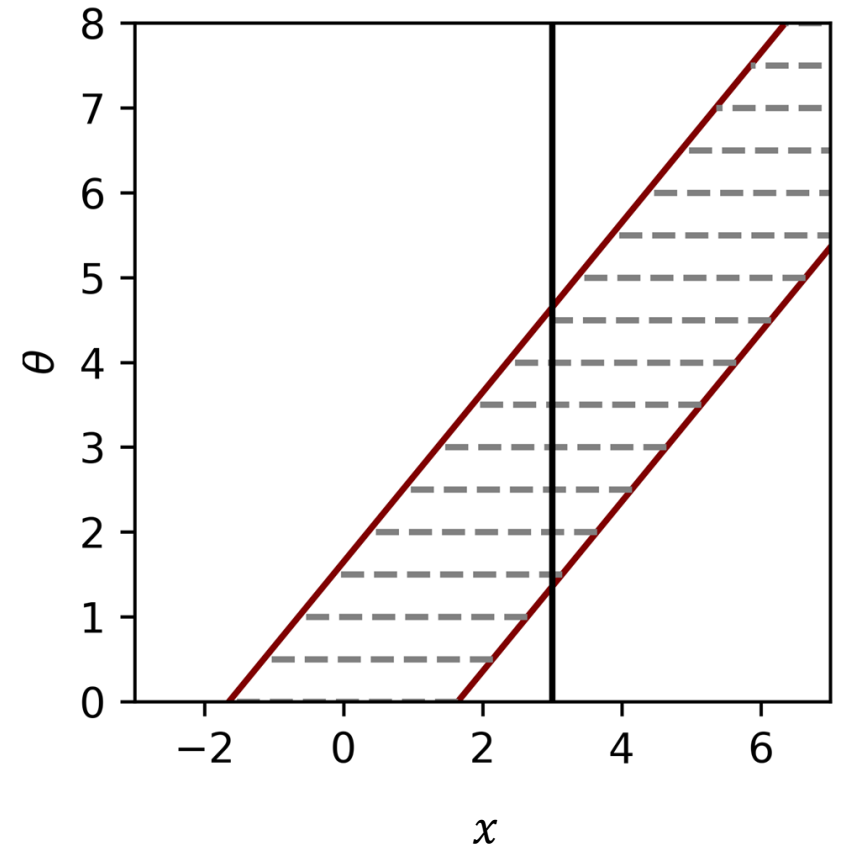
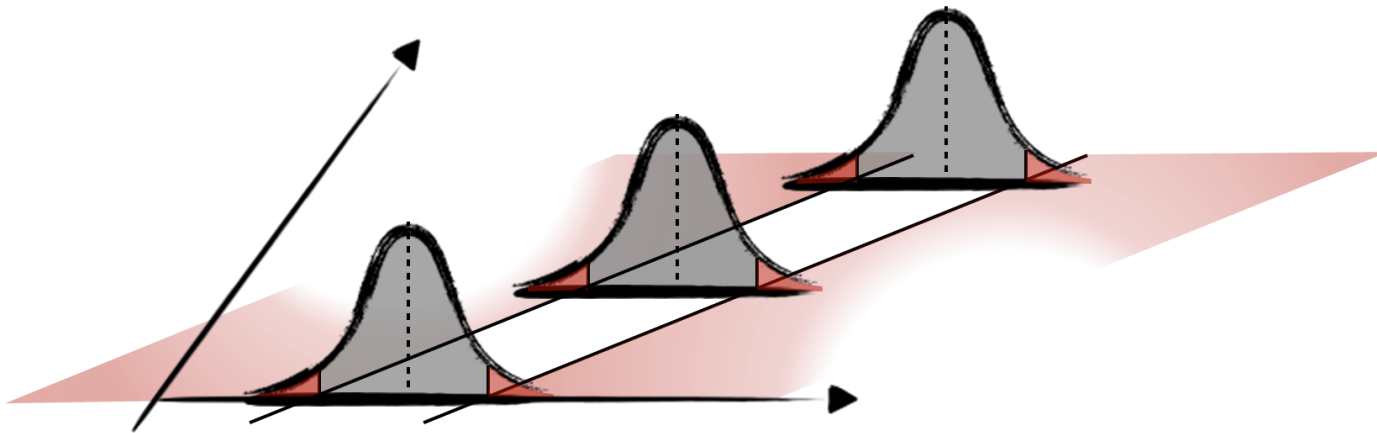
# Band plot walk-through: Step 1

- We can construct such intervals in  $x$  for any choice of  $\theta$ !
- This is called the “Neyman” band plot



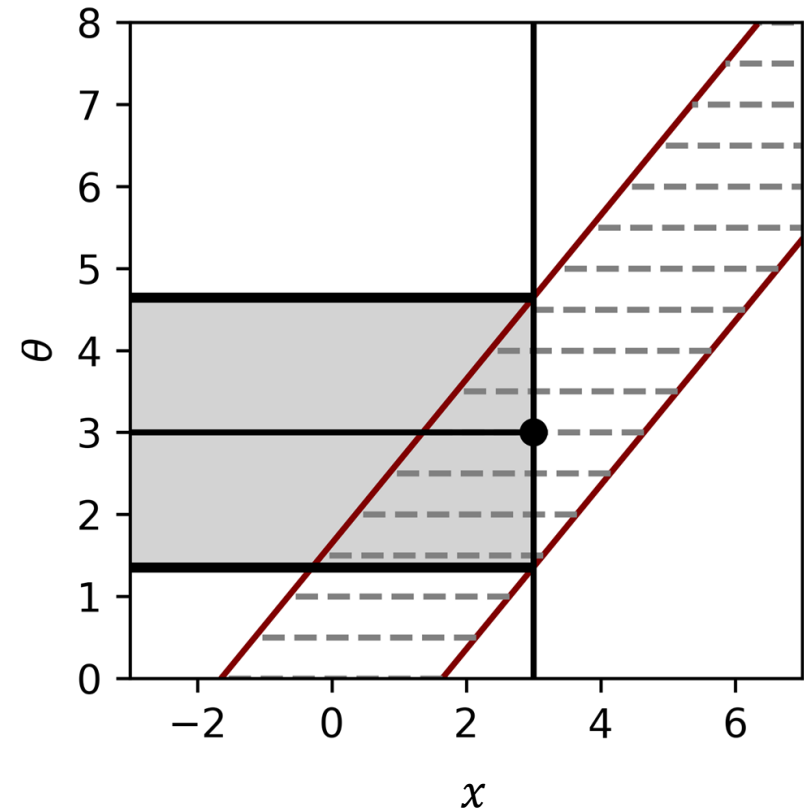
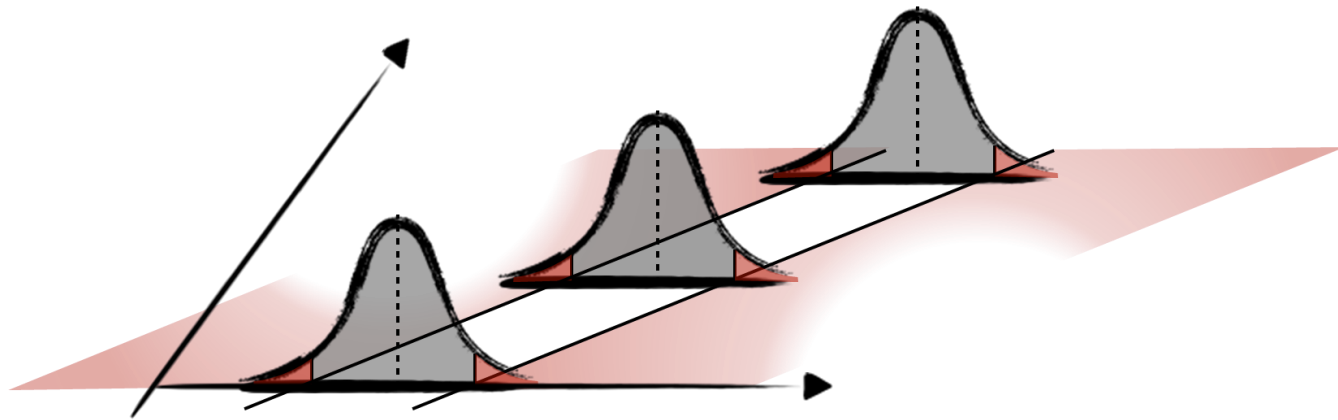
# Band plot walk-through: Step 2

- Now fix the observed value at your measured  $x$



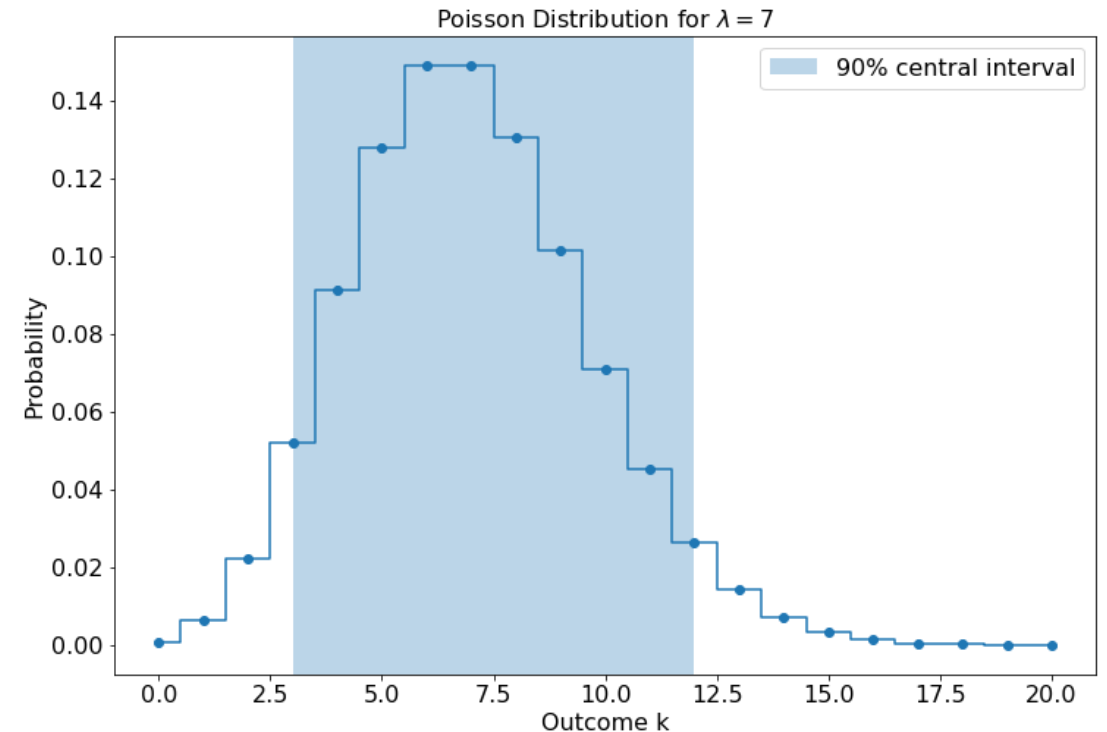
# Band plot walk-through: Step 3

- Read of corresponding interval in  $\theta$



# Example: Poisson

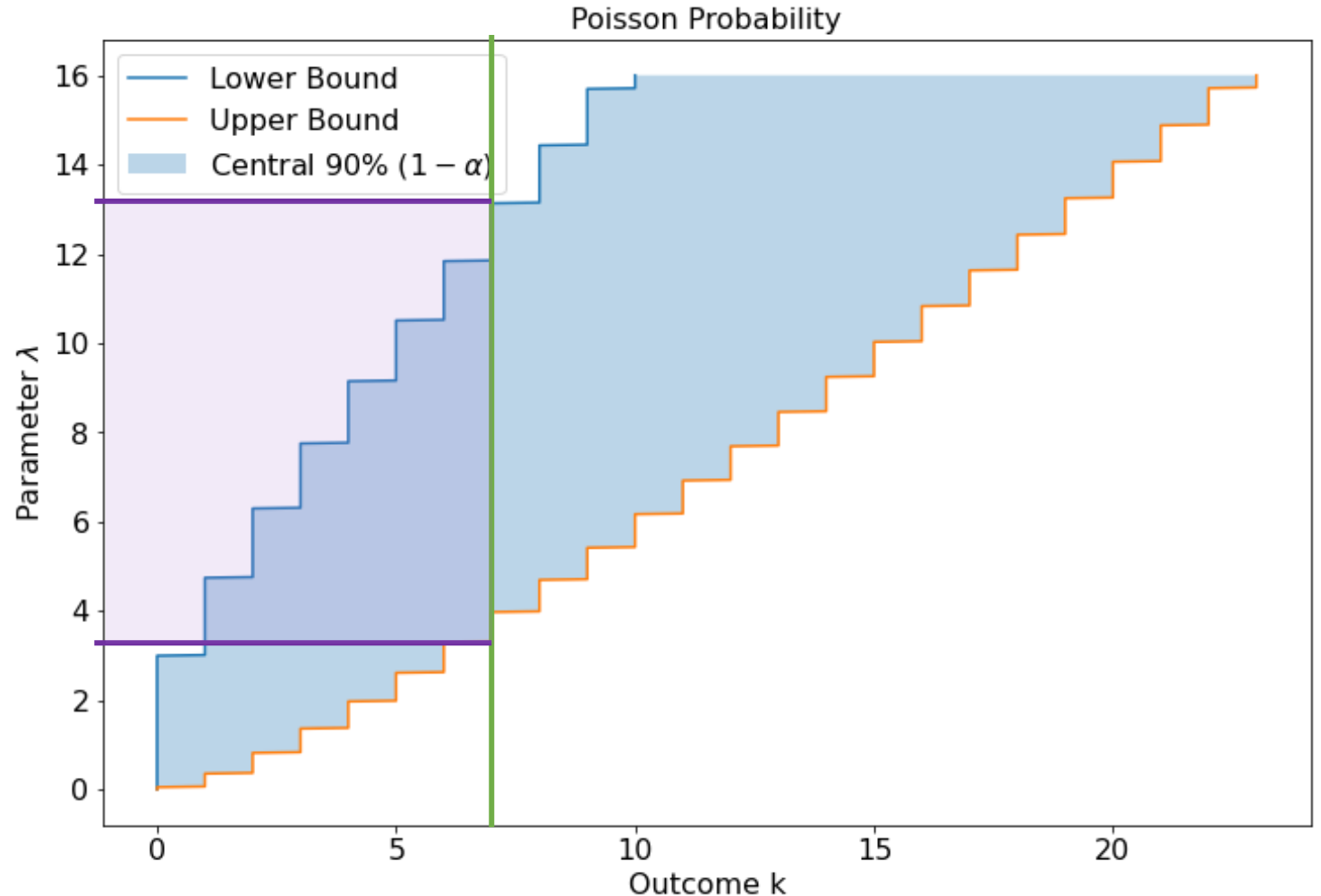
- Let's suppose we perform a counting experiment and we measure an outcome of  $k = 7$
- We know for a fixed  $\lambda$  the range of possible outcomes  $k$ 
  - $\rightarrow$  i.e., the probability distribution
  - And we know how to construct intervals that contain  $1 - \alpha$  of the probability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





# Neyman Construction

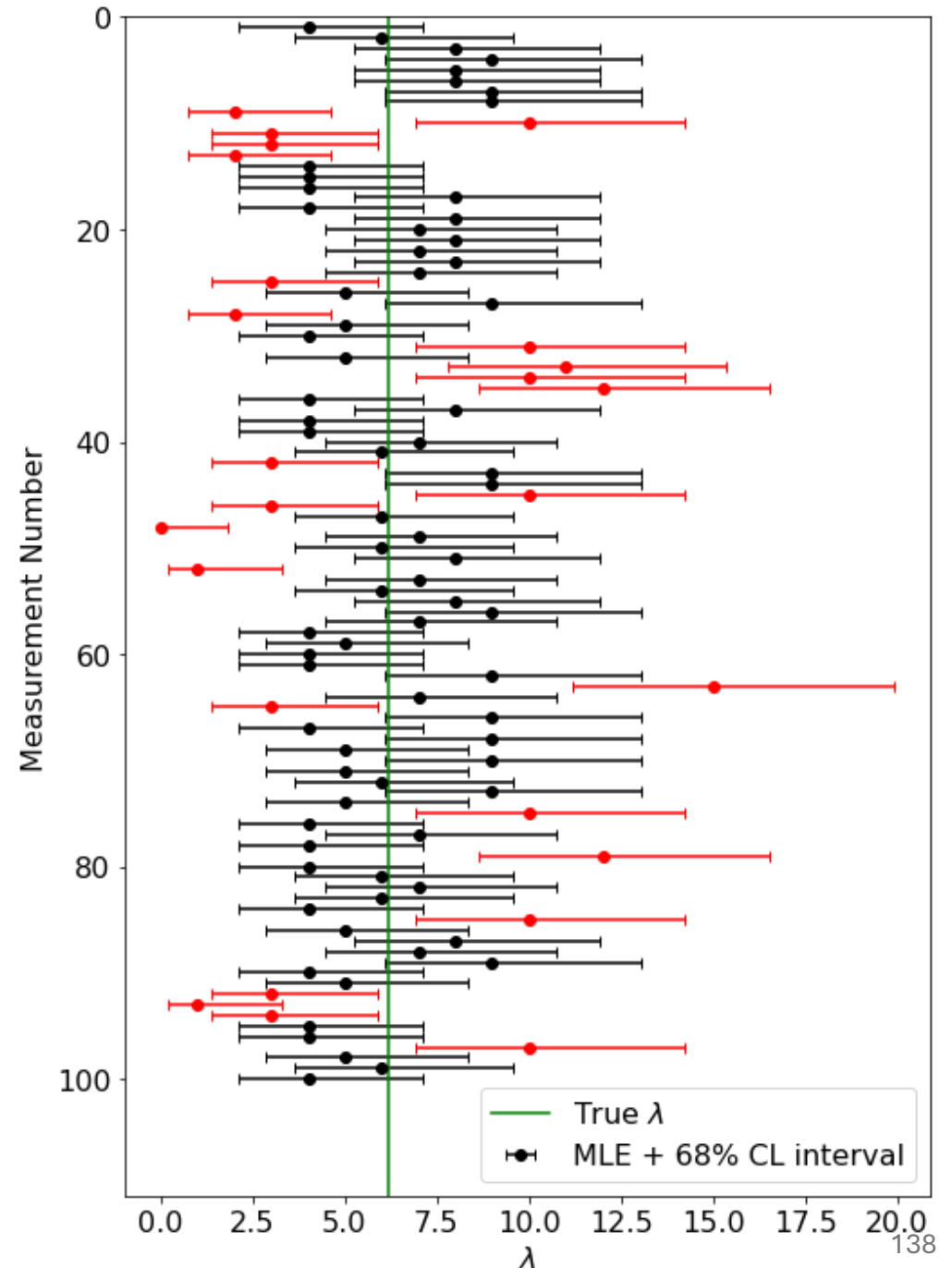
- Poisson band plot:
  - Construct the interval in  $k$  for all possible  $\lambda$   
Here we chose  $\alpha = 0.1$
  - Fix the random variable to our measured value  
 $k = 7$
  - Read off the interval  
 $\lambda \in [3.3, 13.1]$  @ 90% C.L.



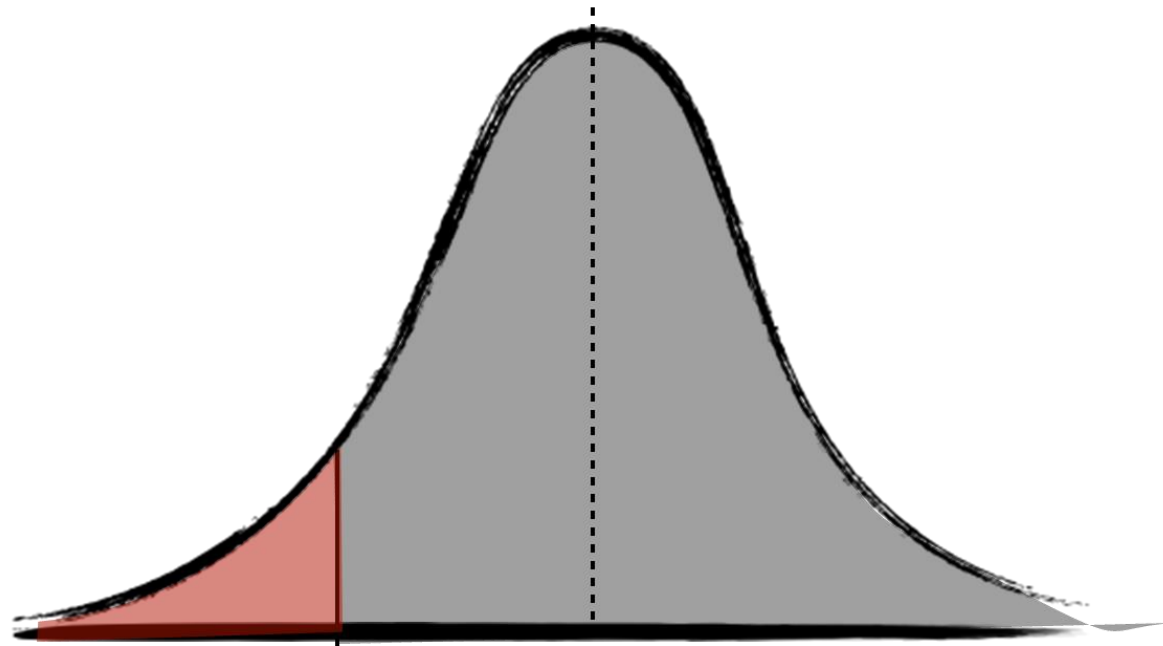
# Interpretation:

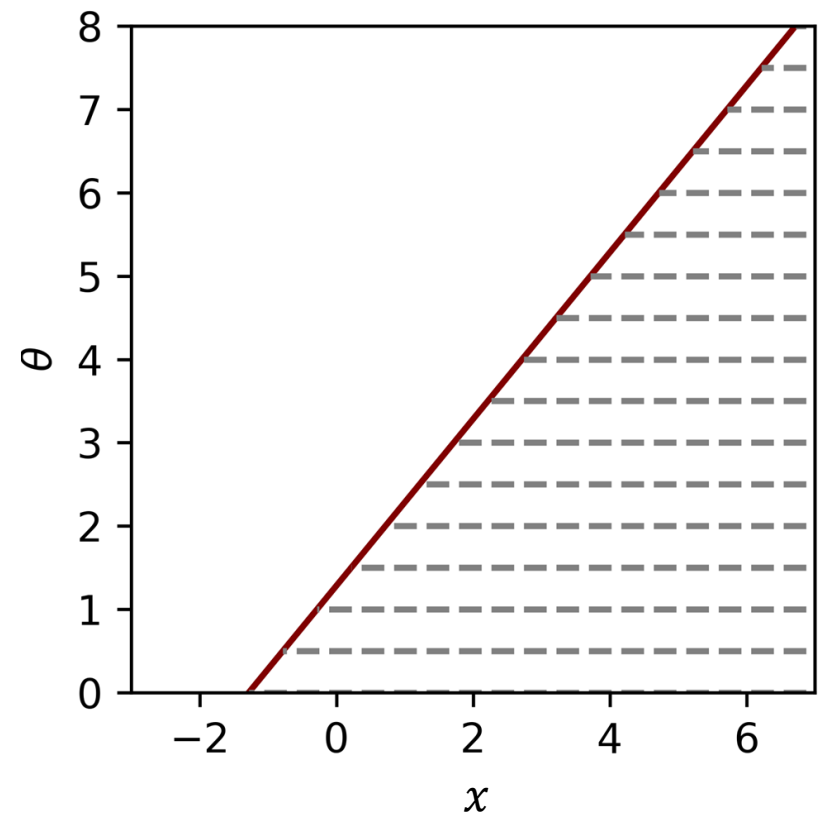
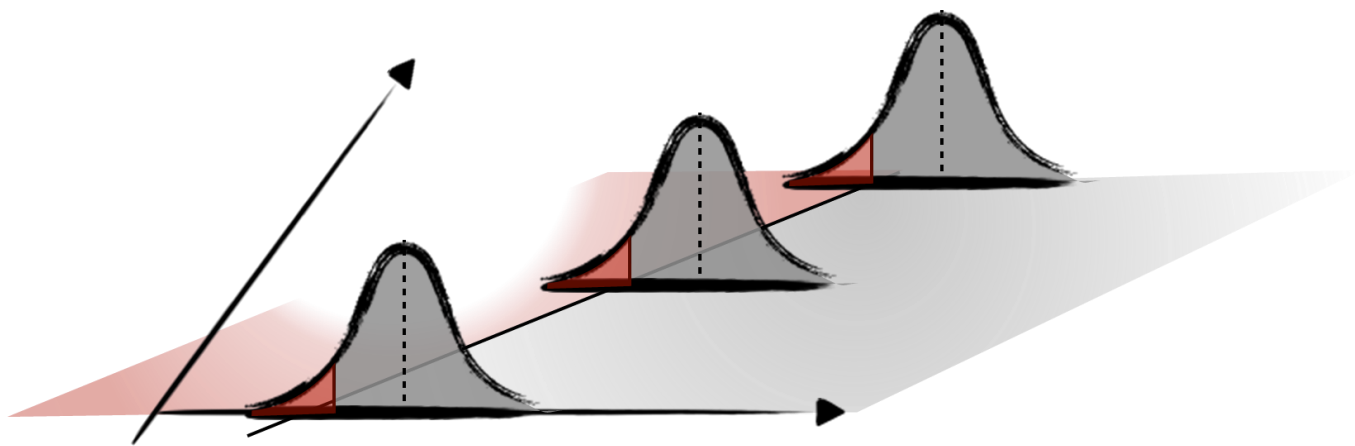
In repeated experiments, the CL interval contains the true value at least  $1 - \alpha$  of times

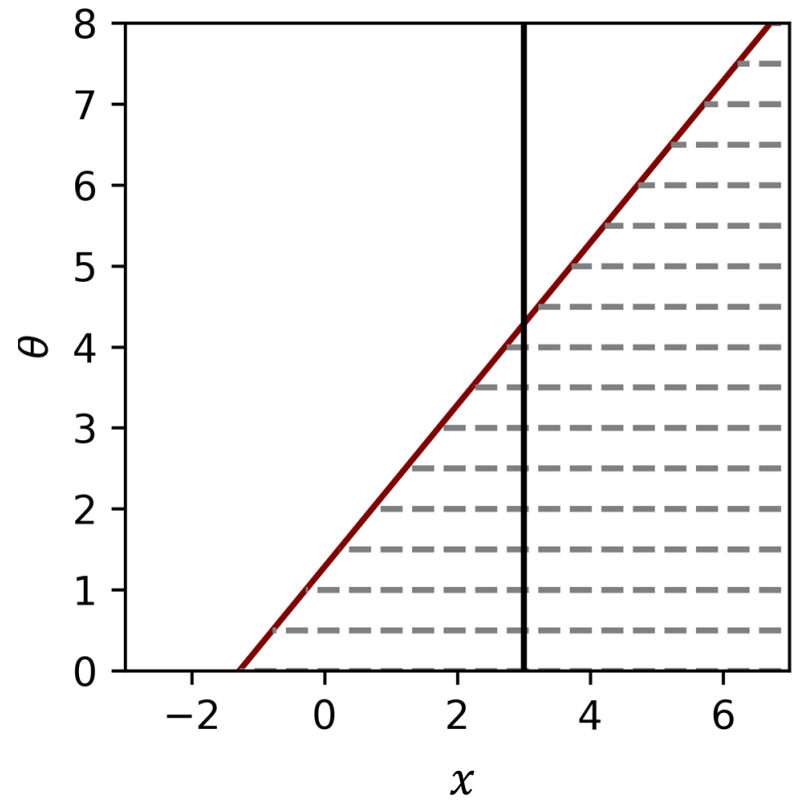
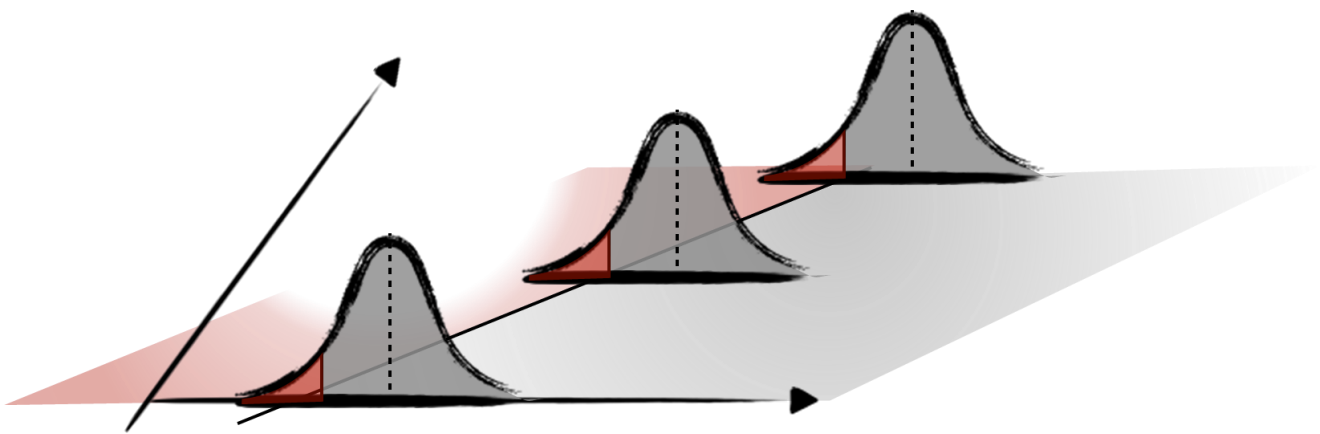
- Here: 100 Poisson experiments with  $\lambda = 6.2$
- Slight *overcoverage*
  - Actually only 25% measurements fall outside
  - Typical problem for discrete distributions

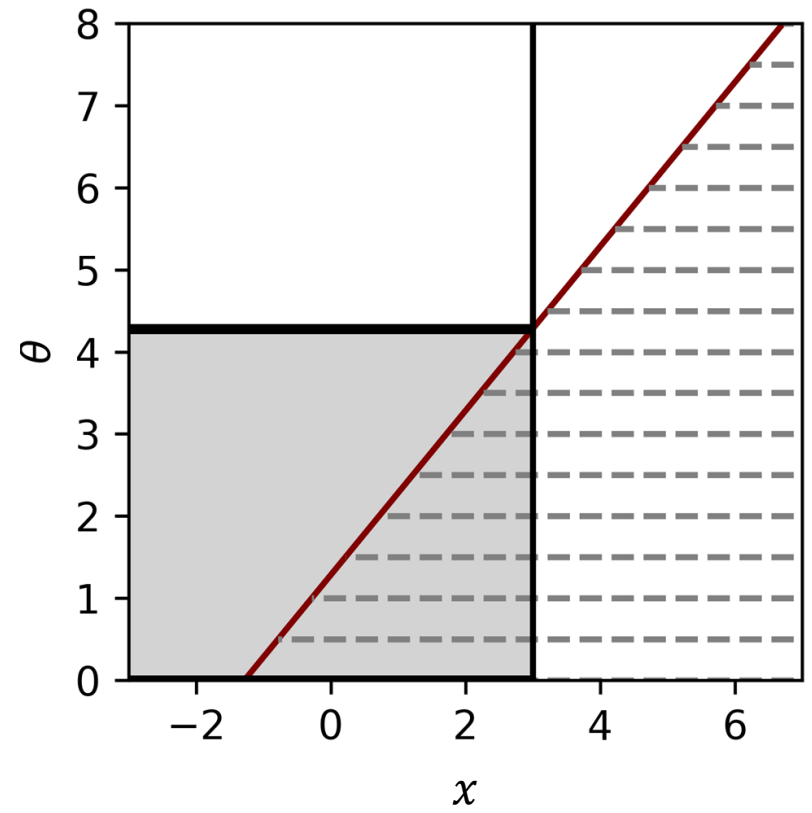
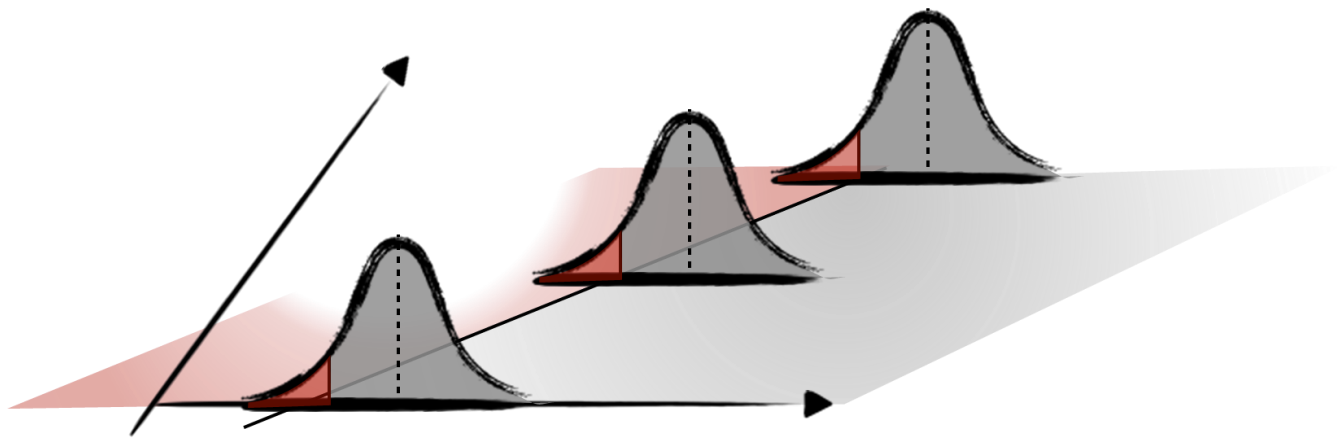


# One-sided Intervals



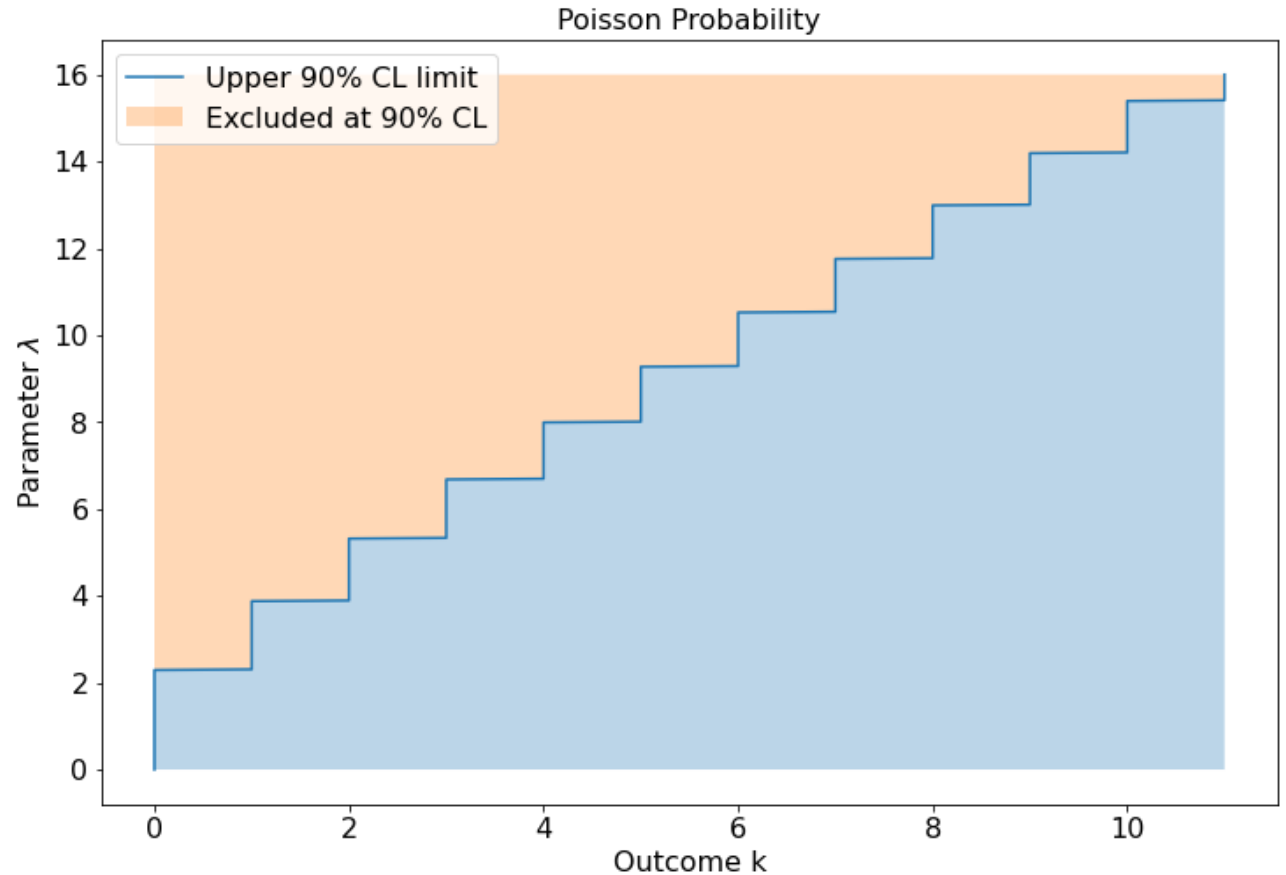






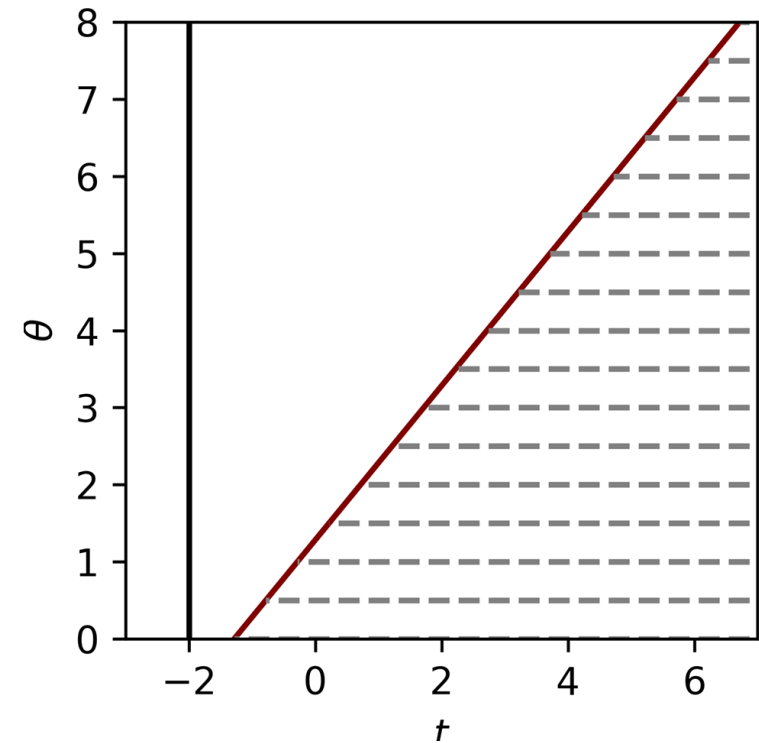
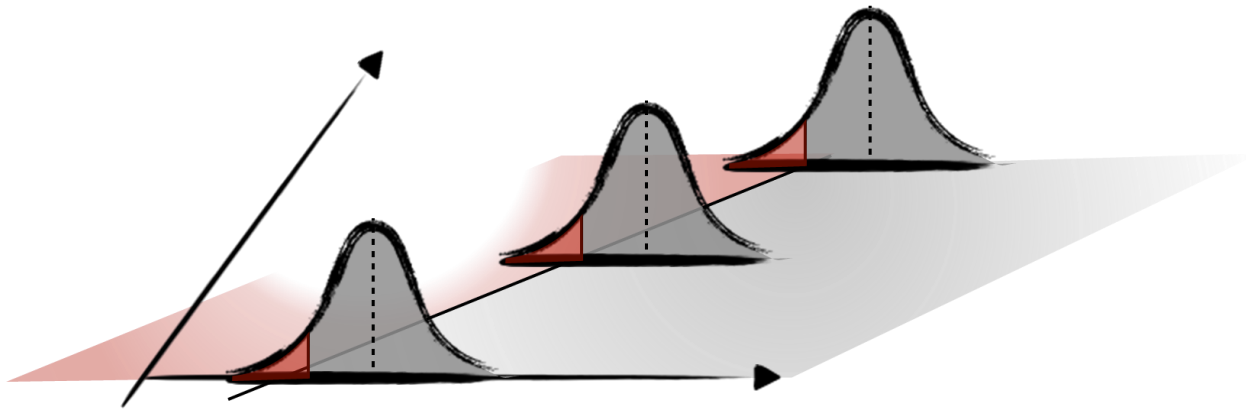
# Example: Poisson

- Upper limit: maximum value of  $\lambda$  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  $\theta$  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 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 Frequentist 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  $\theta$  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  $\theta$

# 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

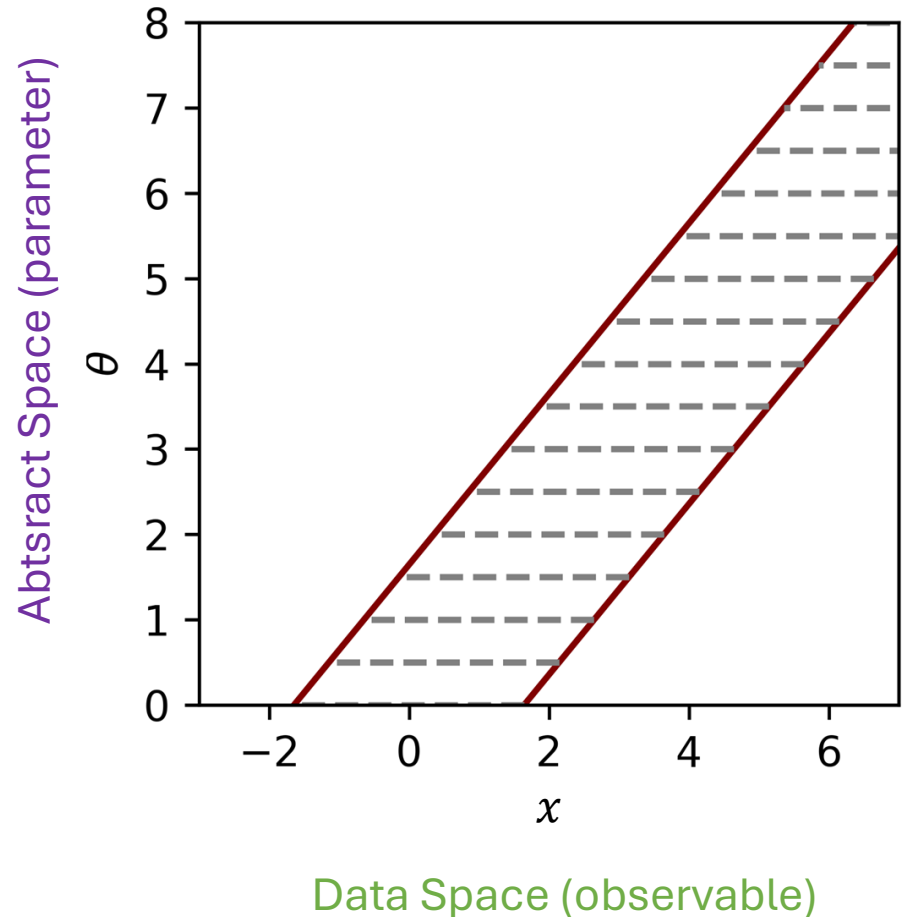
## Frequentist Intervals (confidence intervals):

"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  $\theta$  95% of the time

# Recap

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



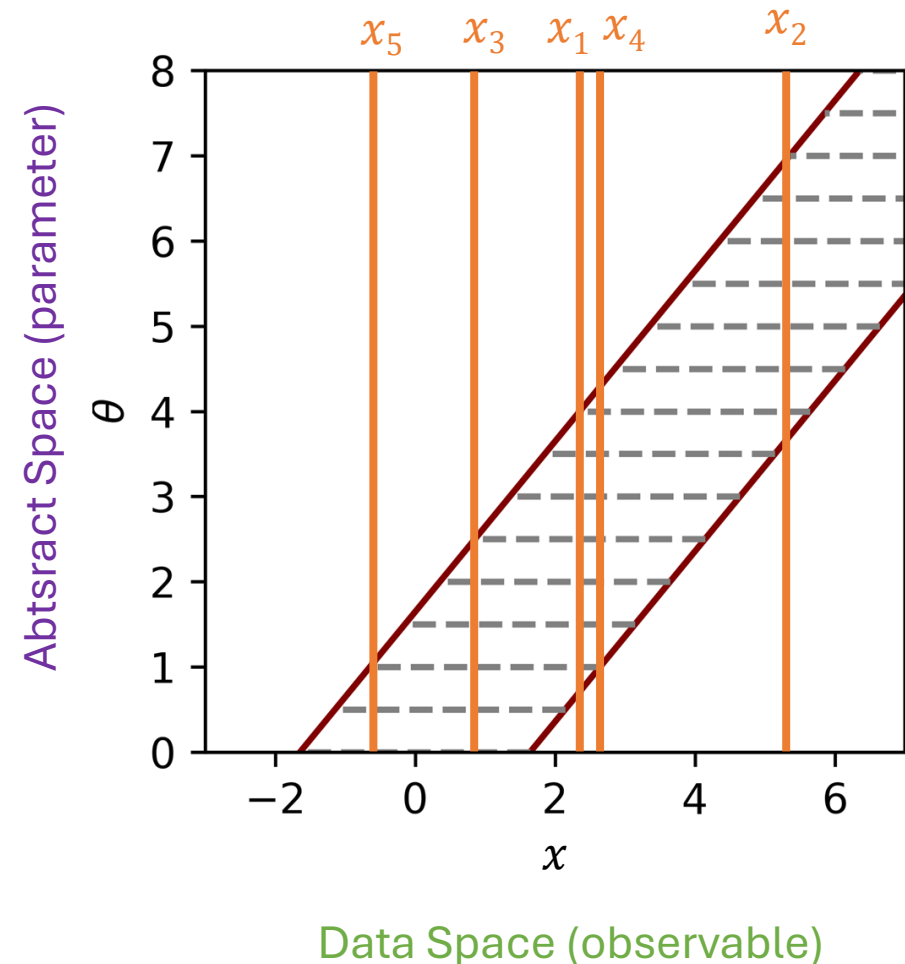


# Beyond Simple Models

...again

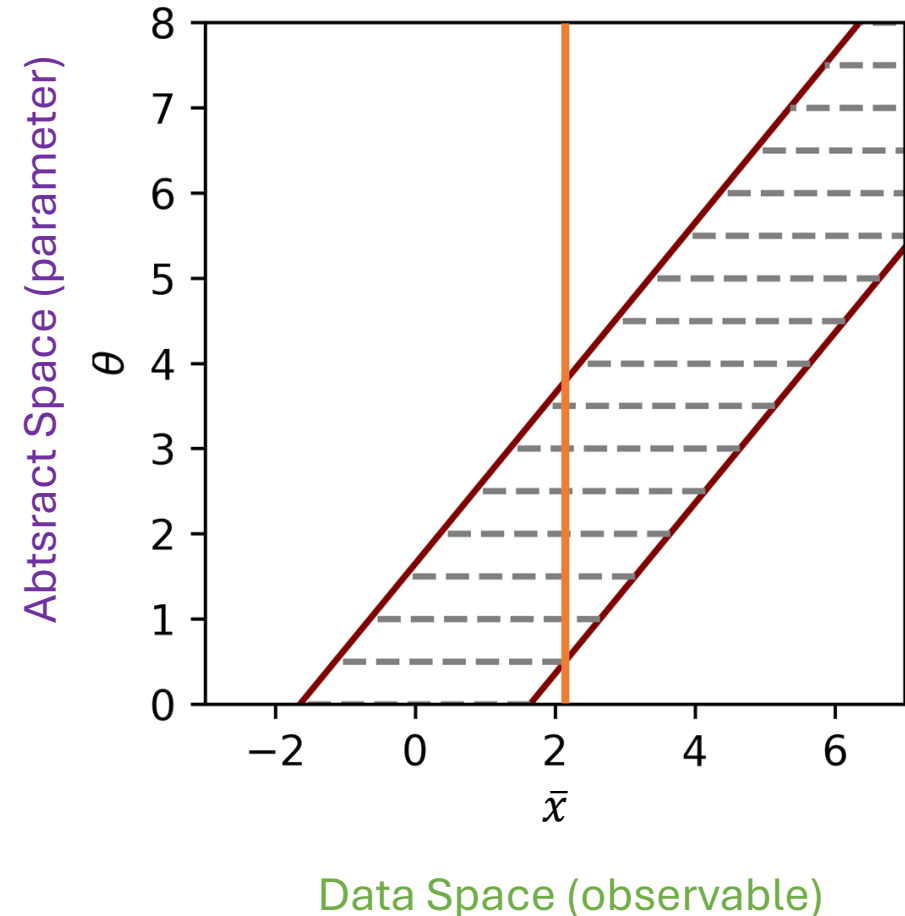
# What if I have more data...?

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

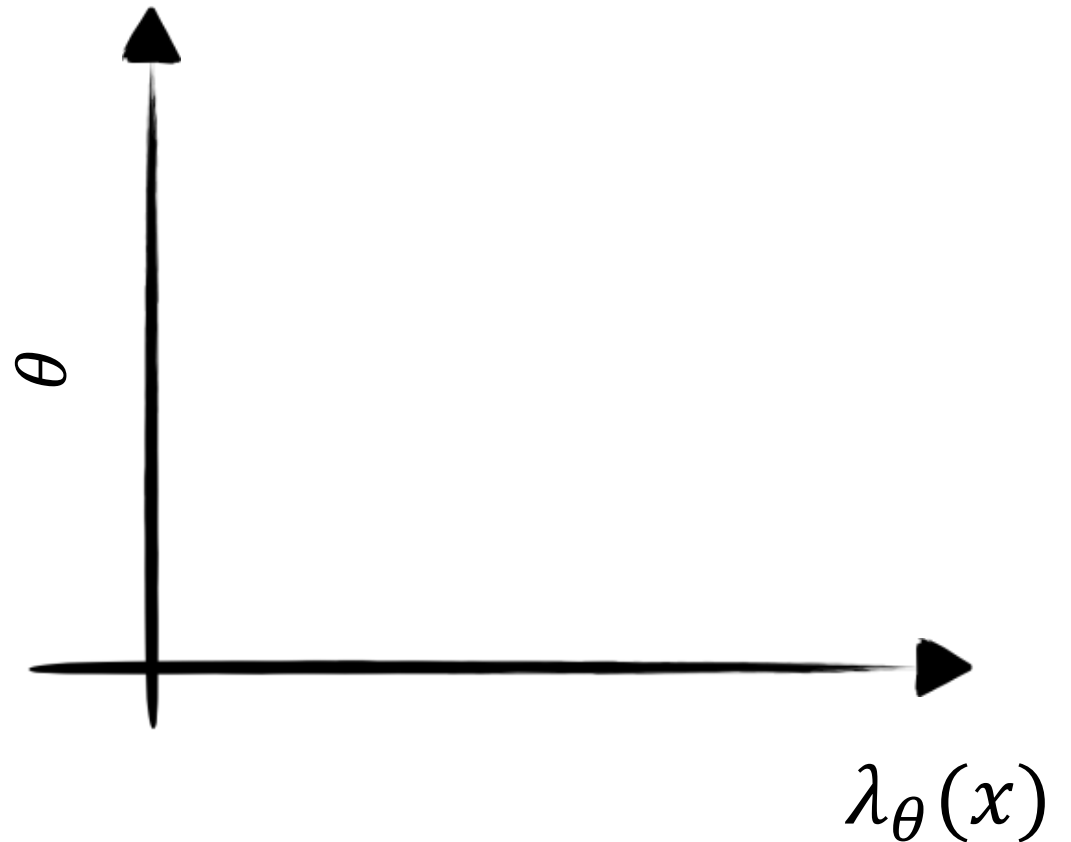
- In the discussion of point estimators, we found that the maximum **likelihood** estimator (MLE) is a desirable way how to summarize our observations
- In the discussion of hypothesis tests, we found that **likelihood**-ratio tests (LRT) have several desirable properties (e.g., Neyman-Pearson Lemma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

# LRT as TS for Neyman Band

- LRT test statistic:

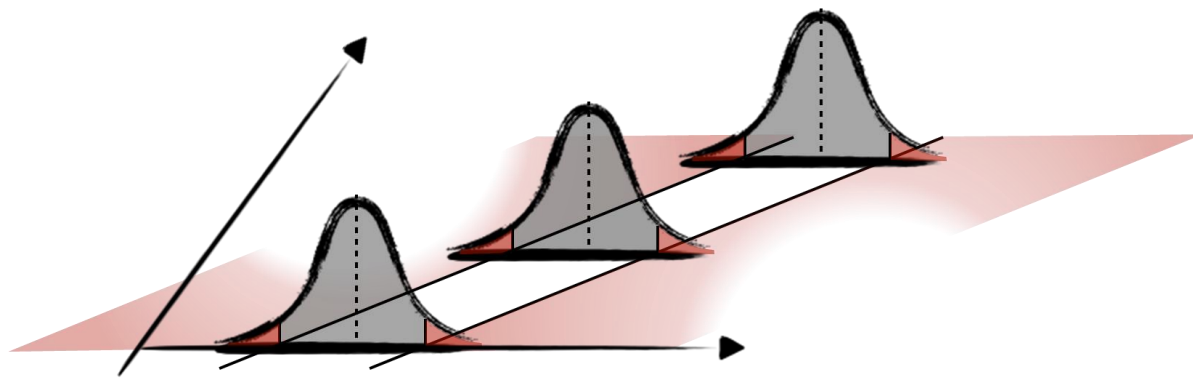
$$t_{\theta} = \lambda_{\theta}(x) = -2 \log \frac{p(x|\theta)}{p(x|\hat{\theta})}$$

- Instead of drawing intervals in  $x$  for every  $\theta$  we construct intervals in  $t_{\theta}$  for every  $\theta$

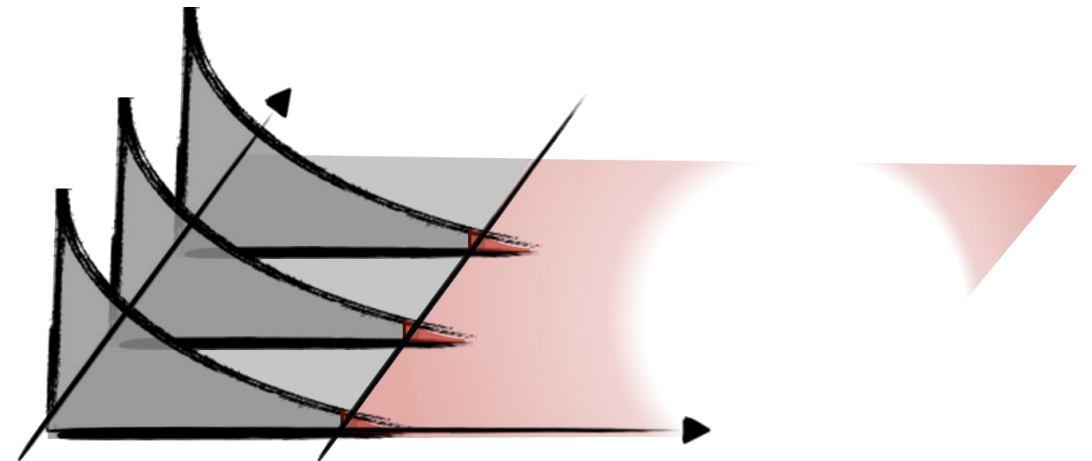


# What do the Rejection Regions look like?

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



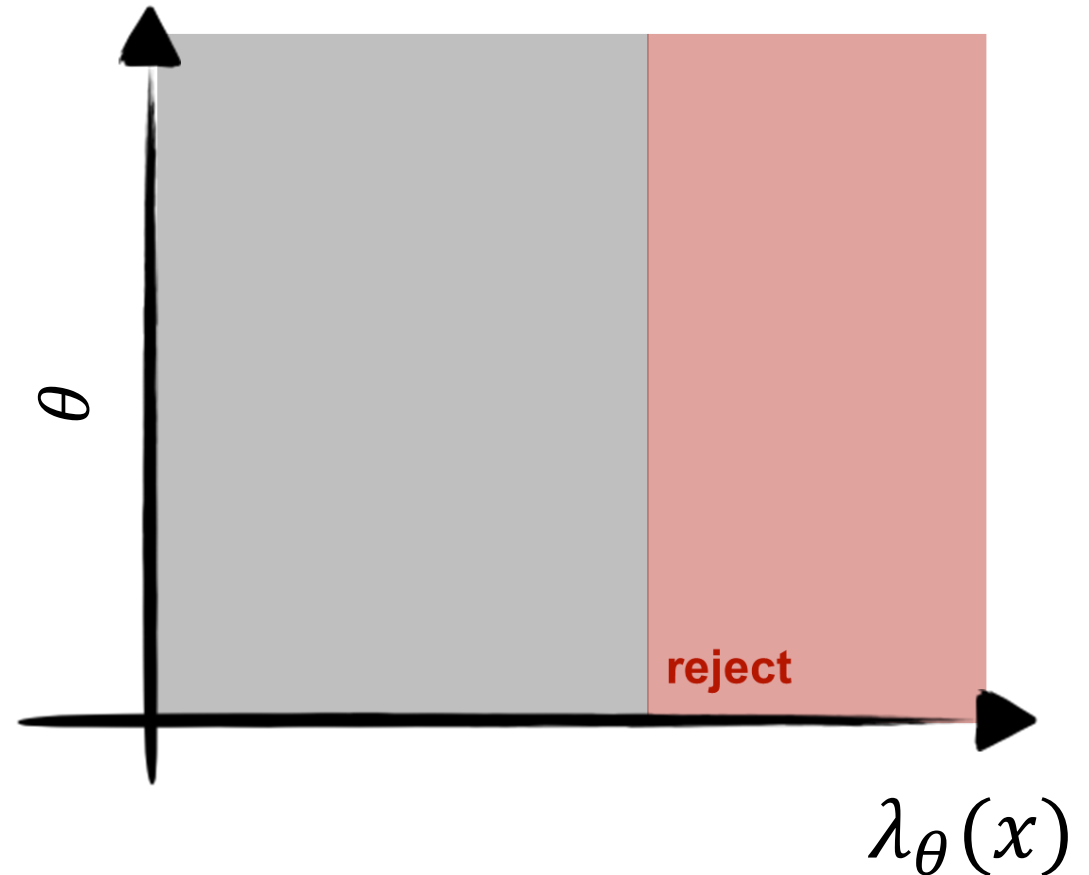
Intervals in  $x$



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  $\chi^2$

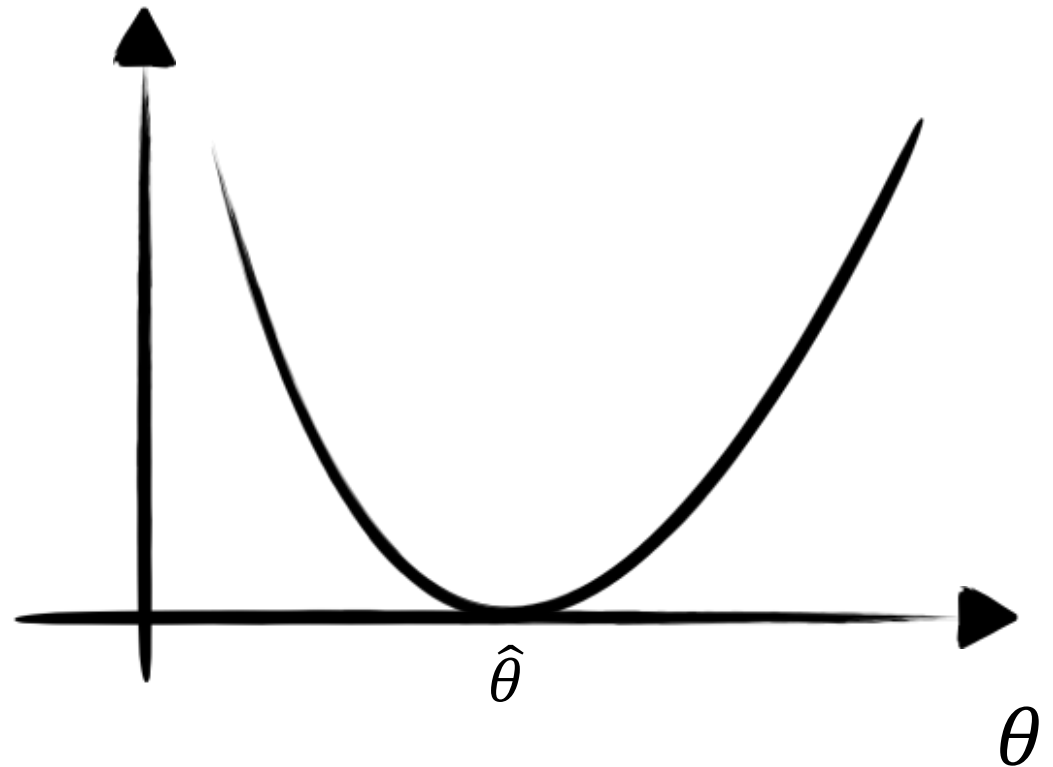


# How does the observed look like?

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

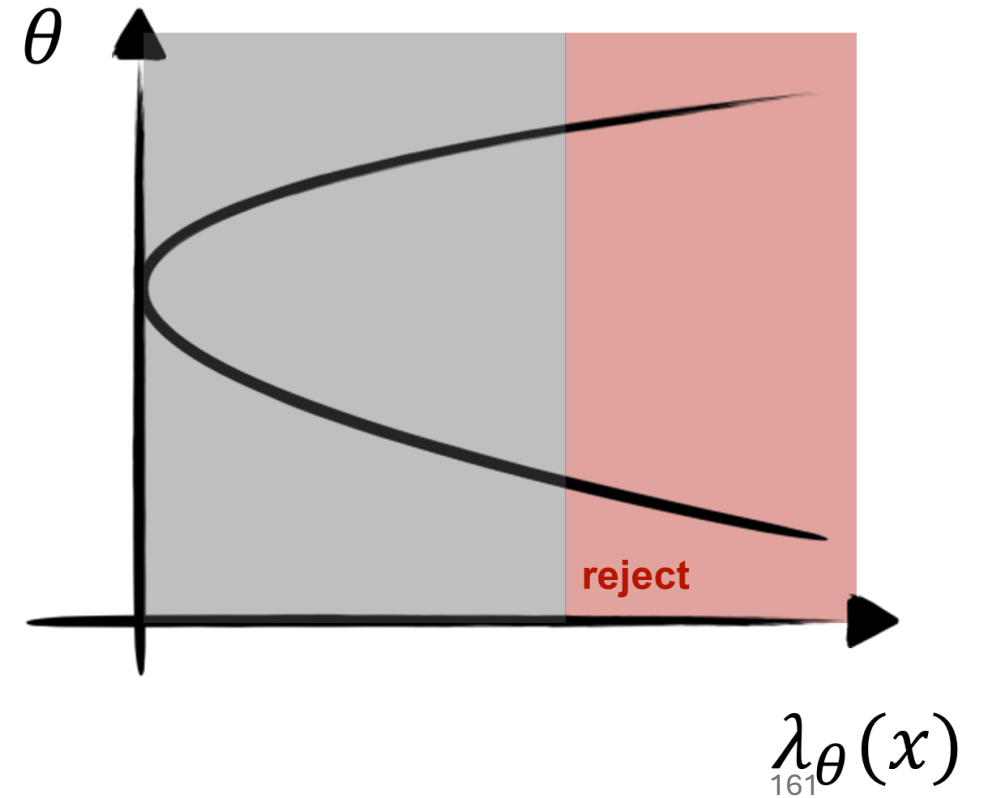
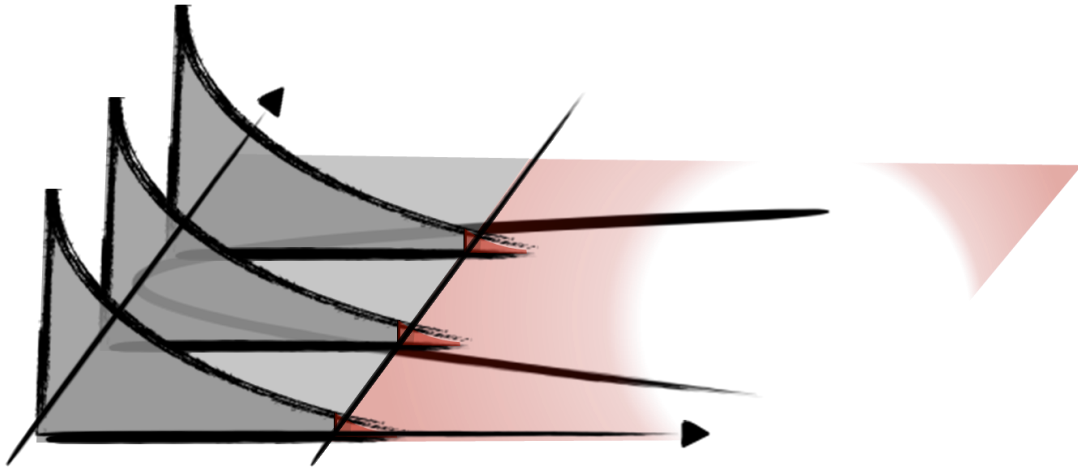
$$t_{\theta} = \lambda_{\theta}(x) = \frac{(\theta - \hat{\theta})^2}{\sigma_{\hat{\theta}}}$$

which is a parabola around the MLE



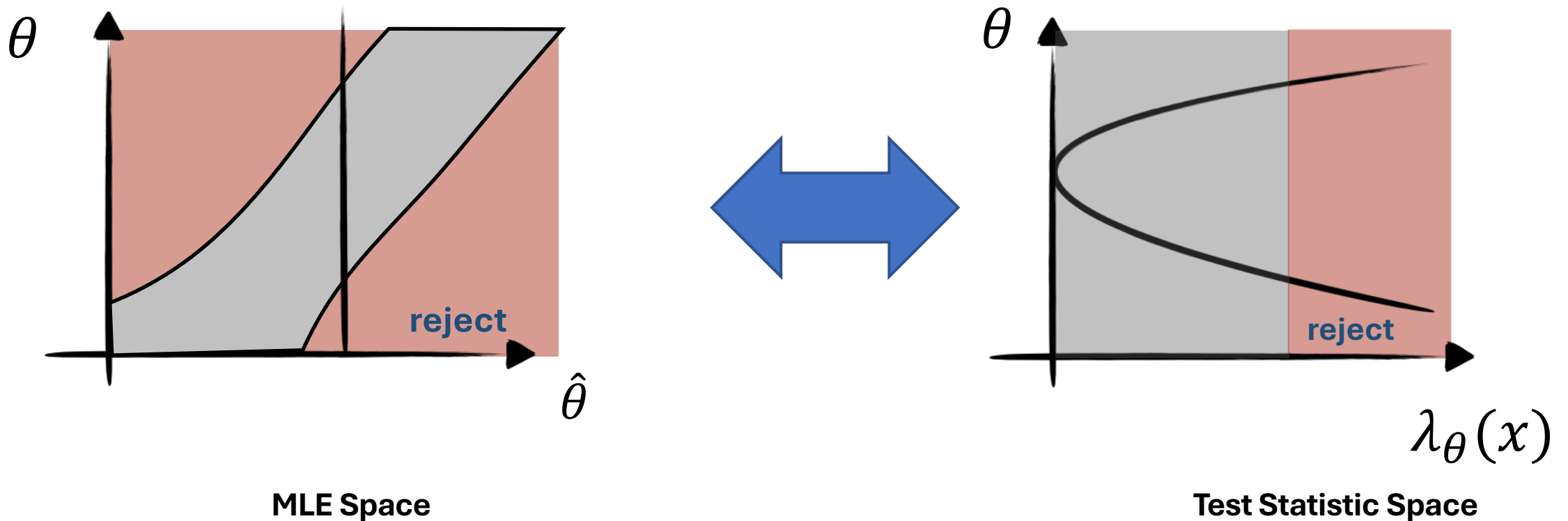


# Putting it together



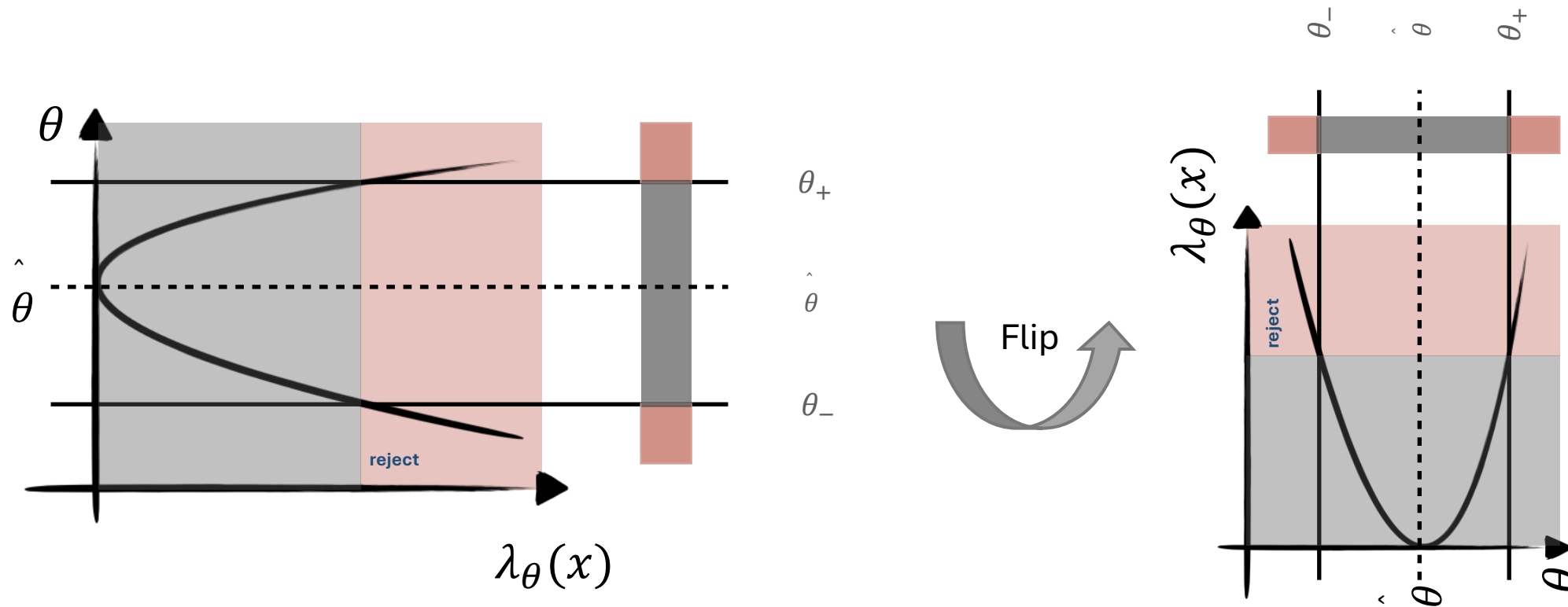
# Equivalence

- Defining a rejection region in **test statistic space** induces a corresponding one in **maximum likelihood space**



# Constructing the Interval

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



# Standard $1\sigma$ intervals

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

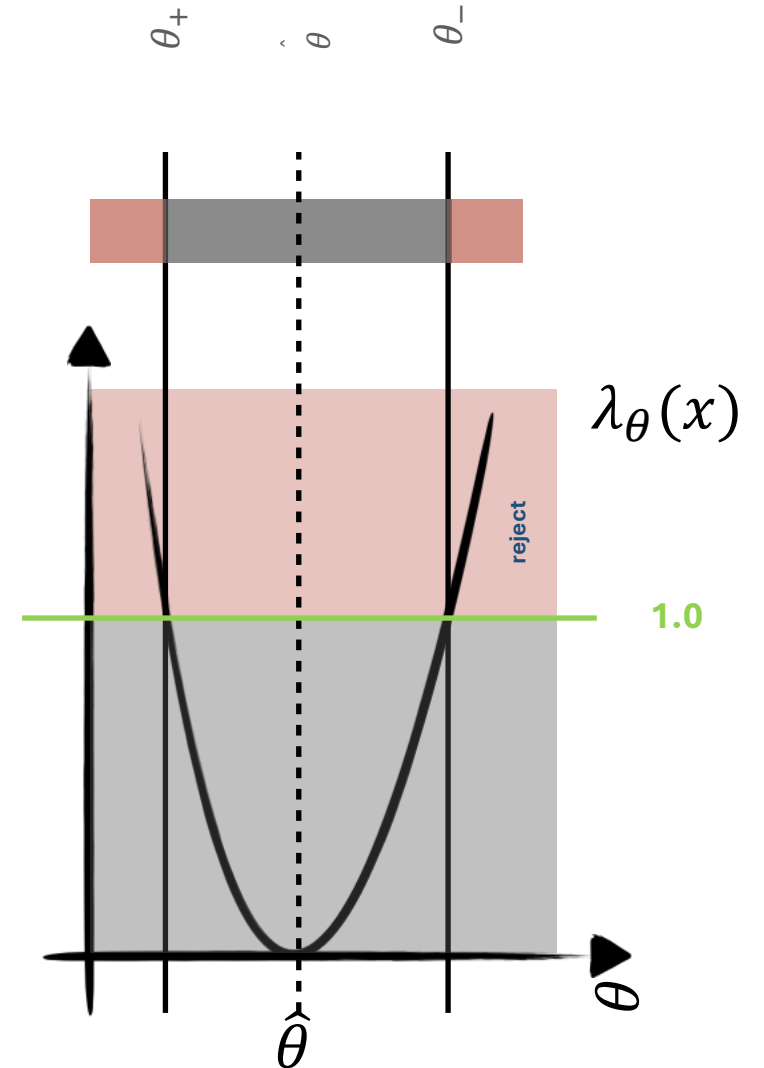
$$\lambda_{\theta}(x) = -2(\text{LL}(\theta_{\text{lim}}) - \text{LL}(\hat{\theta})) = 1$$

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

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

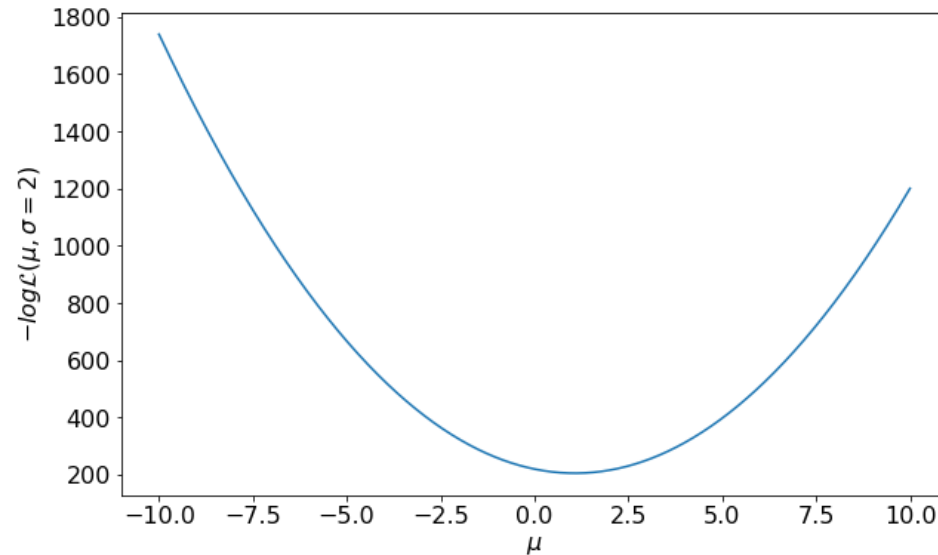
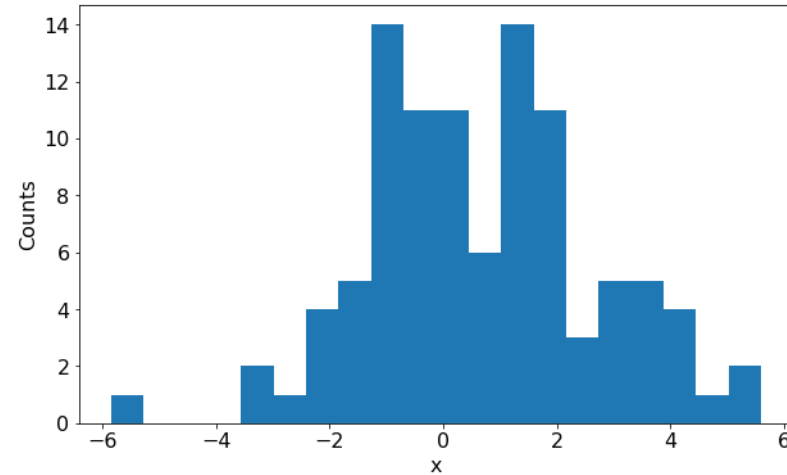
Table 9.5 The values of the quantile  $Q_{\gamma}$  for different values of the confidence level  $1 - \gamma$  for  $n = 1, 2, 3, 4, 5$  fitted parameters.

$1 - \gamma$	$Q_{\gamma}$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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



# Putting it to work

- Let's suppose we draw 100 samples from a normal distribution with  $\mu=1$  and  $\sigma=2$
- Likelihood is obtained from the joint probability
$$\mathcal{L} = \prod_i p(x_i|\mu, \sigma)$$
- Let's fix  $\sigma = 2$  and look at  $\mu$

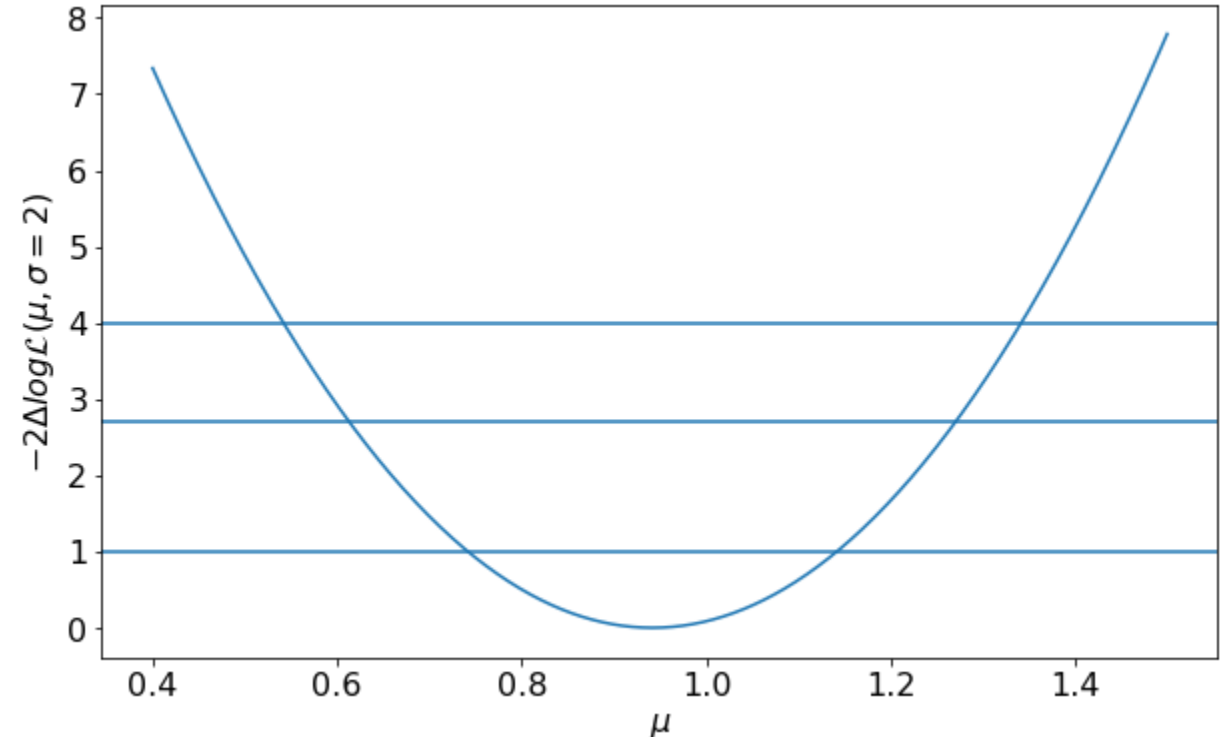


# Intervals

- 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  $\mu$  of  $\frac{\sigma}{\sqrt{n}} \rightarrow$  Checks out!!

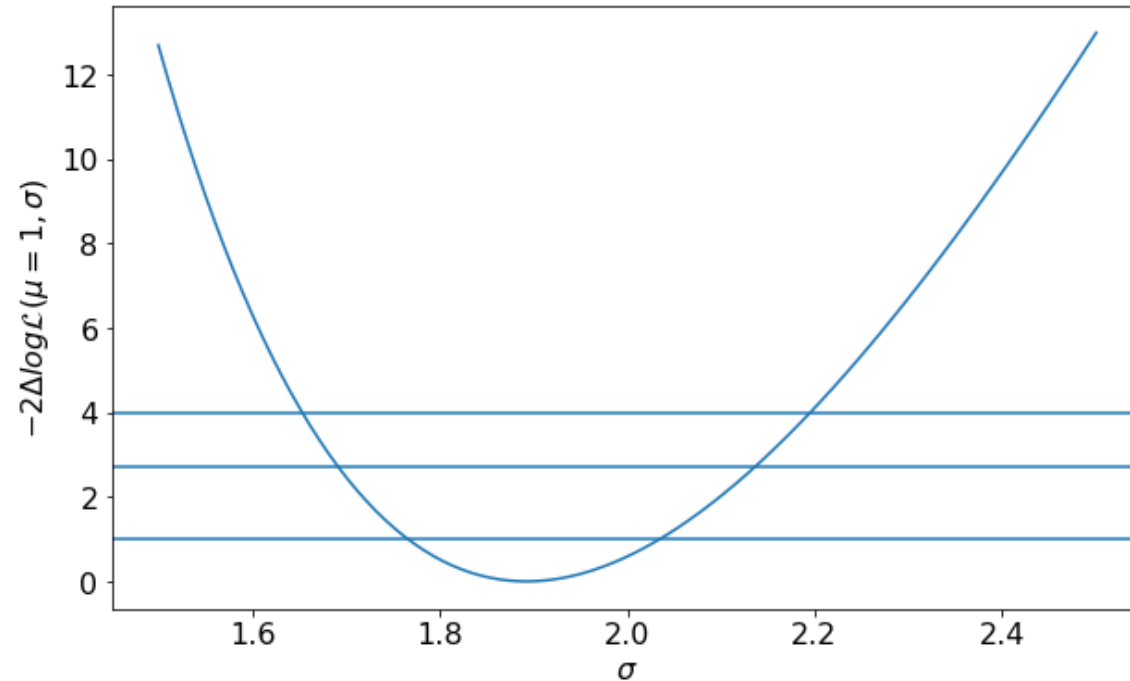
**Table 9.5** The values of the quantile  $Q_\gamma$  for different values of the confidence level  $1 - \gamma$  for  $n = 1, 2, 3, 4, 5$  fitted parameters.

$1 - \gamma$	$Q_\gamma$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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

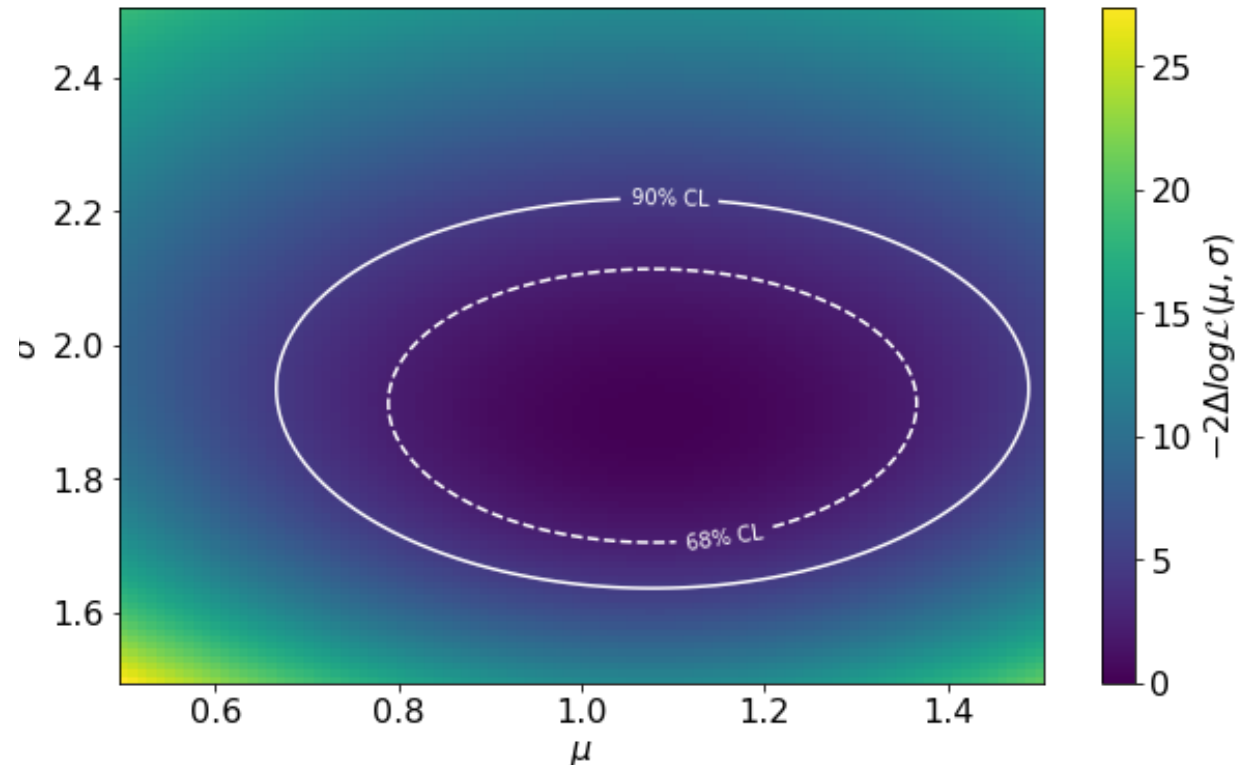


# Both parameters!

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

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

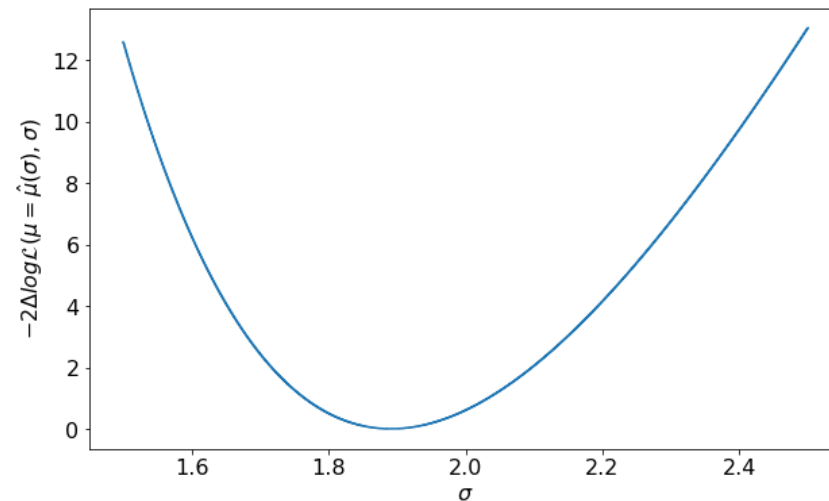
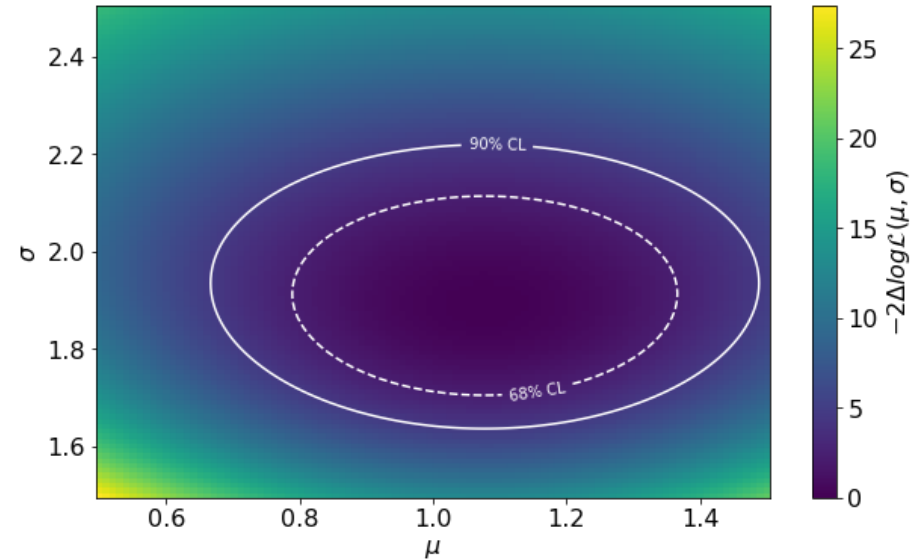
$1 - \gamma$	$Q_\gamma$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1





# Profile LLH

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

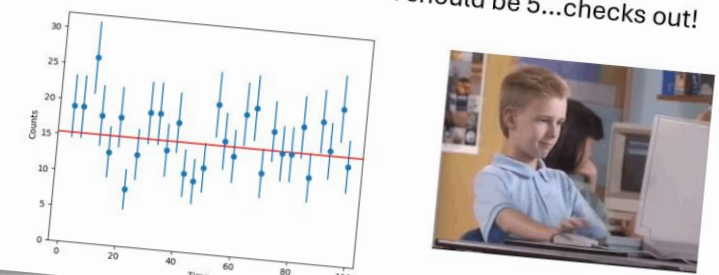


# Fixing the 4<sup>th</sup> Problem

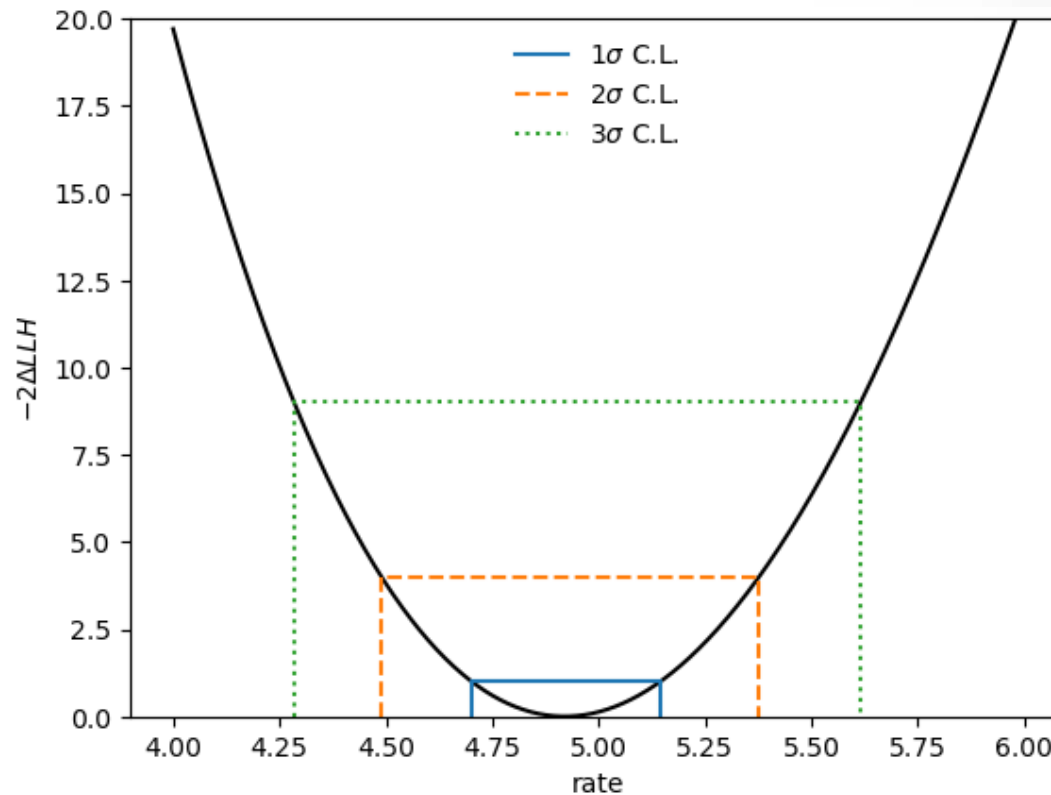
- Finally, we can now quote our estimate with a correct interval:
- Rate =  $4.92 \pm 0.224$  –  $0.218$  @ 68% C.L.

### Result

- Estimated rate:  $4.599 \pm 0.214$
- The rate from the theory calculation should be 5...checks out!



The scatter plot shows 'Counts' on the y-axis (0 to 30) and 'Time' on the x-axis (0 to 100). Blue dots with vertical error bars represent data points, and a red line shows a downward trend. To the right is a small photo of a young boy sitting at a desk with a computer monitor.



# 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  $\theta_0$  the interval will include  $\theta_0$  95% of the time
    - does not mean: 95% belief that  $\theta_0$  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

# Day 3

# Bayesian Inference

# Short recap

- We discussed parameter inference in the Frequentist case:
  - Point estimators
    - Principle of Maximum Likelihood
  - Hypothesis Testing
    - Likelihood ratio tests as most powerful test statistic
  - Confidence Level Intervals
    - Neyman Band Construction

## 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

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

$$\hat{\mu}_{\text{MLE}} = \bar{x} = \frac{1}{n} \sum_i x_i$$

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

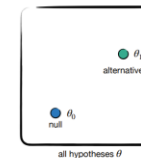
## Neyman-Pearson Lemma

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

The Neyman-Pearson Lemma:

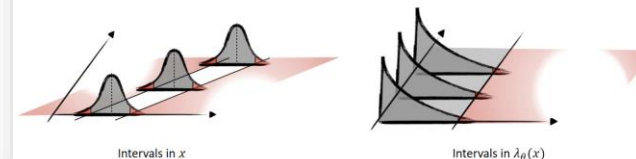
$$\text{The Likelihood Ratio: } l(x) = \frac{p(x|\theta_1)}{p(x|\theta_0)}$$

is the optimal test statistic



## What do the Rejection Regions look like?

- When taking the LRT ratio test, the null hypothesis distribution is always the same, regardless of  $\theta$ :
  - the  $\chi^2$  distribution (Wilk's Theorem)



# Bayesian Probability

- Let's make the stark assumption, that we can upgrade  $\theta$  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  $\theta$  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  $p(\theta)$

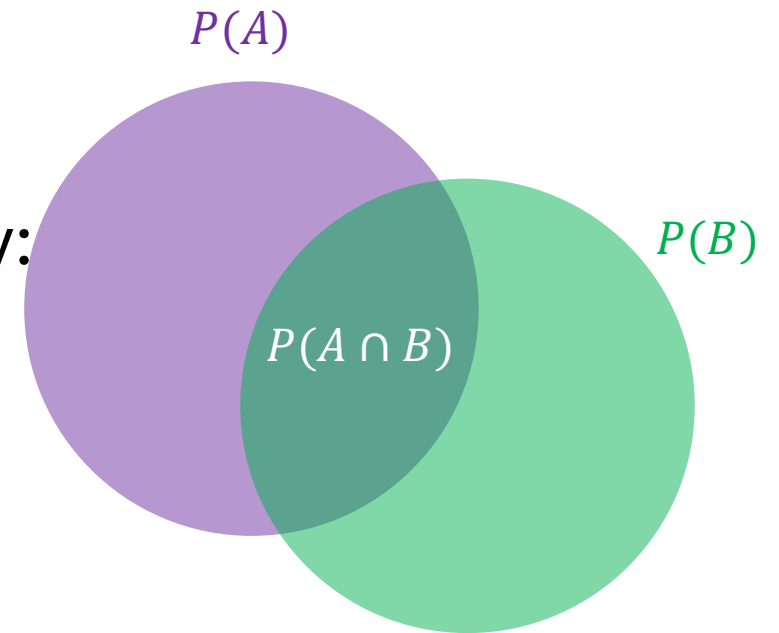
# Bayes' Theorem

Starting from the law of conditional probability:

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

since  $p(A \cap B) \equiv p(B \cap A)$ ,  
we immediately get Bayes' theorem:

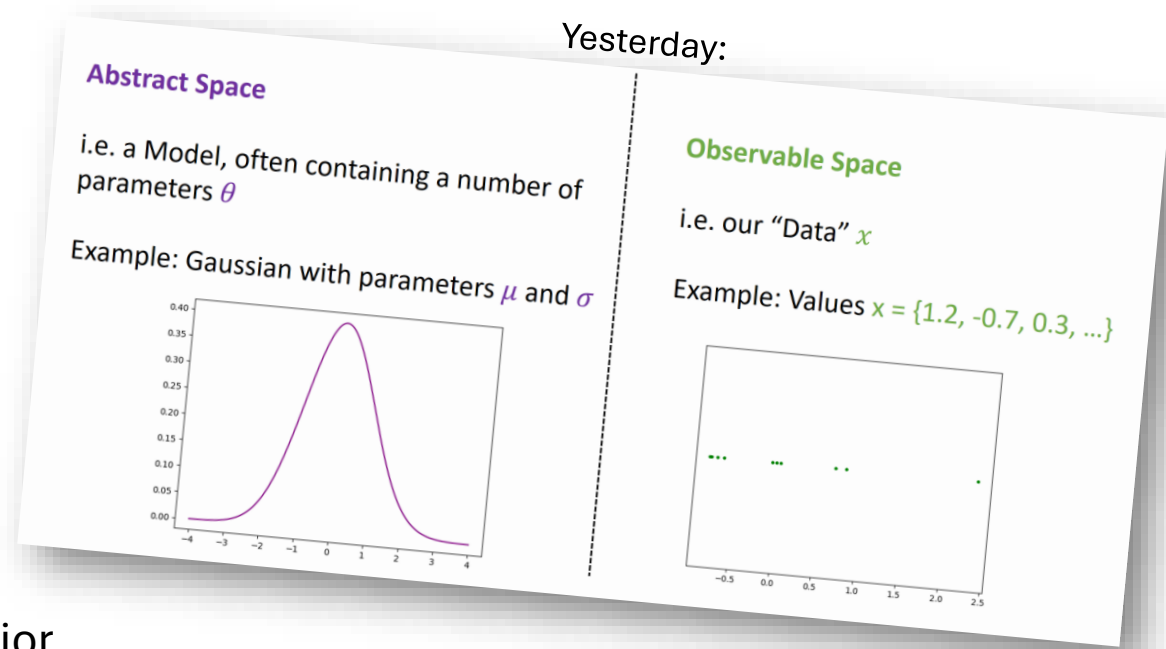
$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$





# In the context of Models and Data

- A = our abstract model  $M$  with parameters  $\theta$
- B = our data  $x$



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

Posterior

Likelihood

Prior

Evidence (Marginal Likelihood)

# 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 ( $\rightarrow$  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)

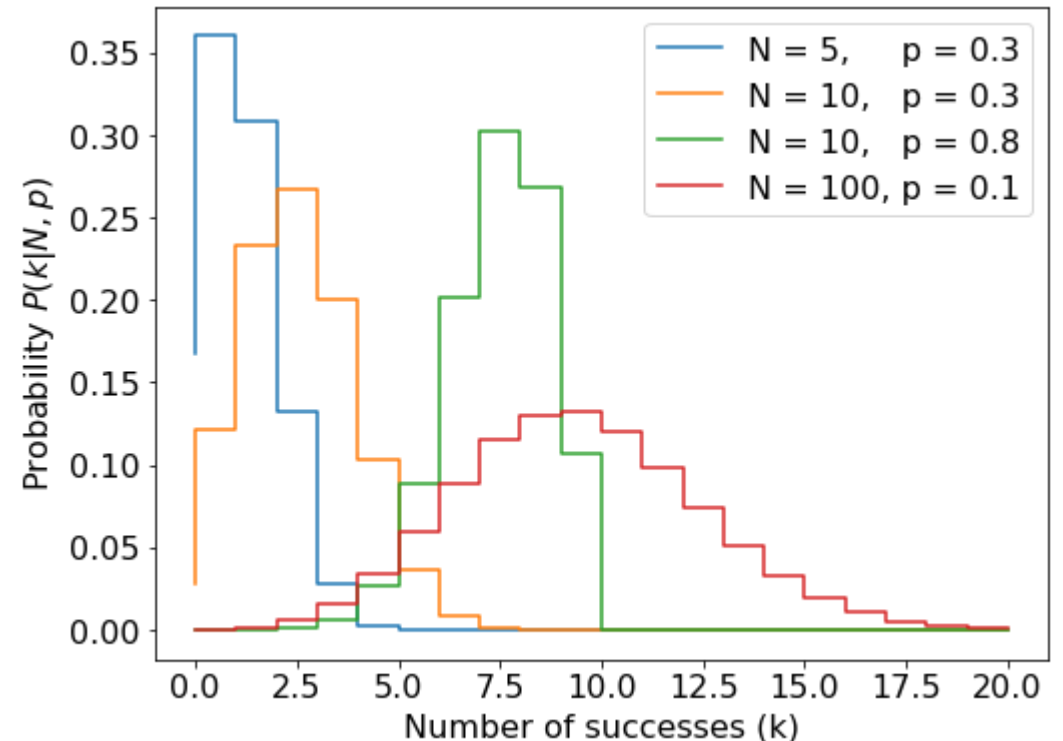


No country for old men  
(<https://www.youtube.com/watch?v=OLCL6OYbSTw>)

# 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 ( $\rightarrow p(p) = 1$  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:

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

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

Variance:

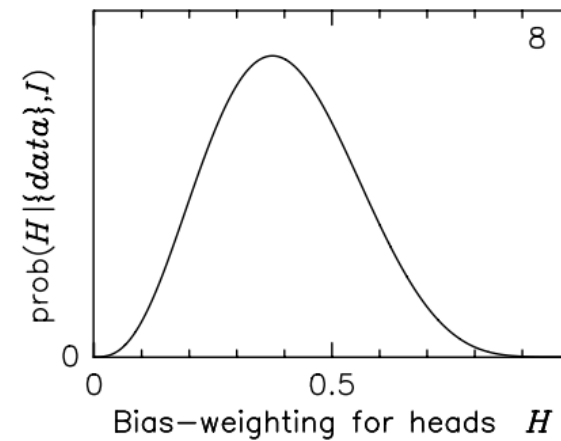
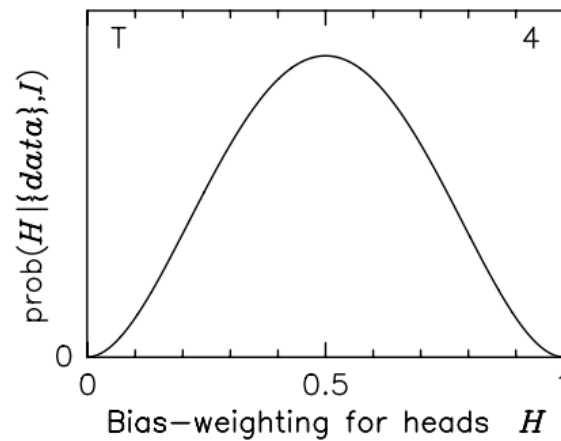
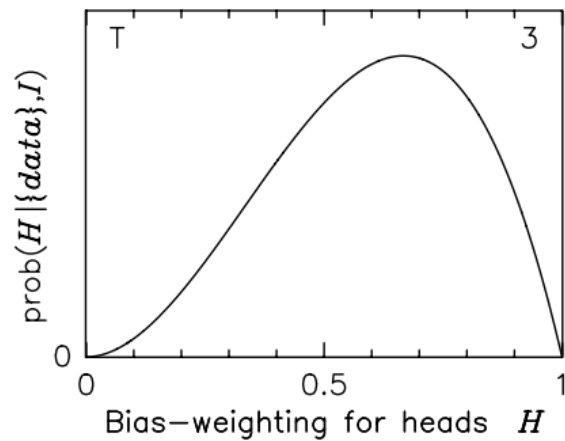
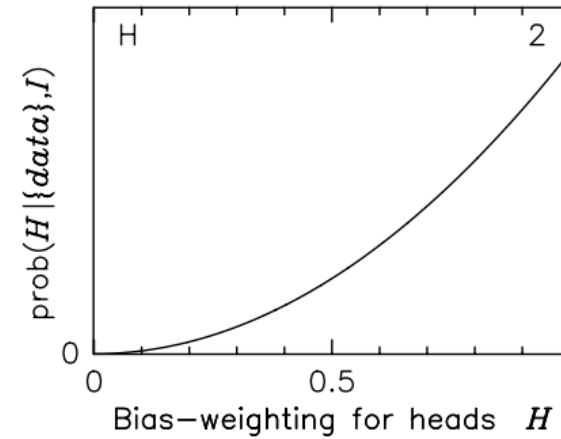
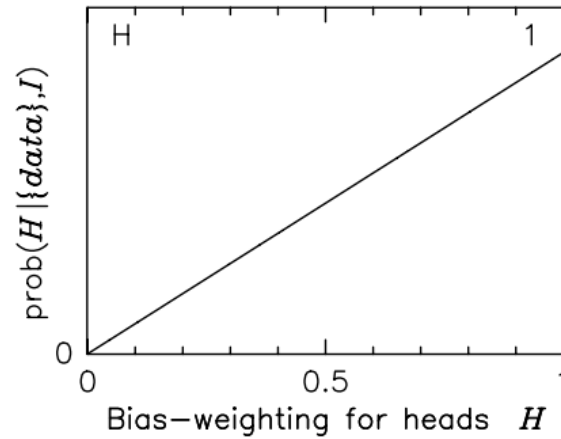
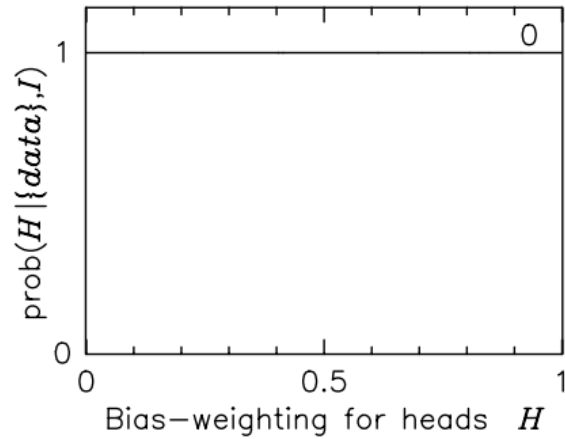
$$V[p] = \frac{(k+1)(N-k+1)}{(N+2)^2(N+3)}$$

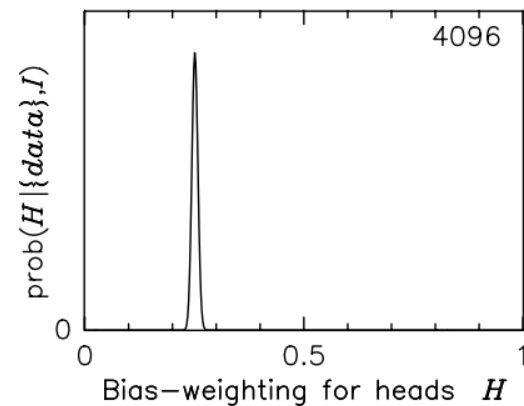
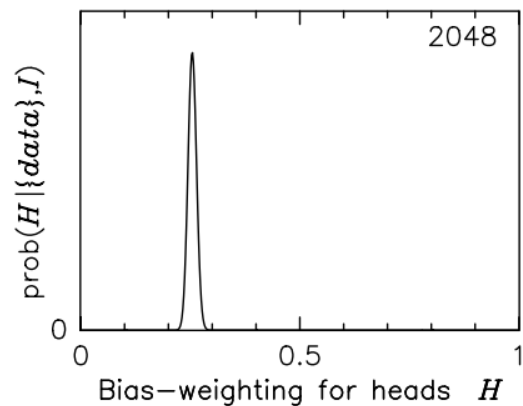
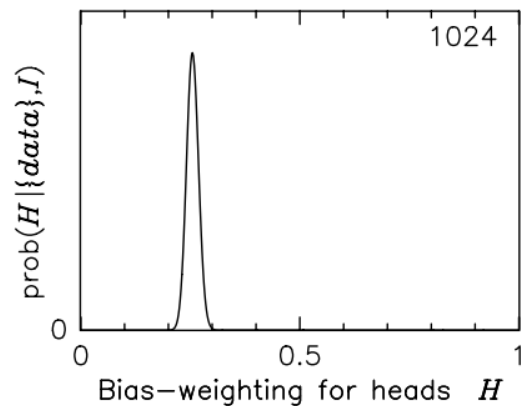
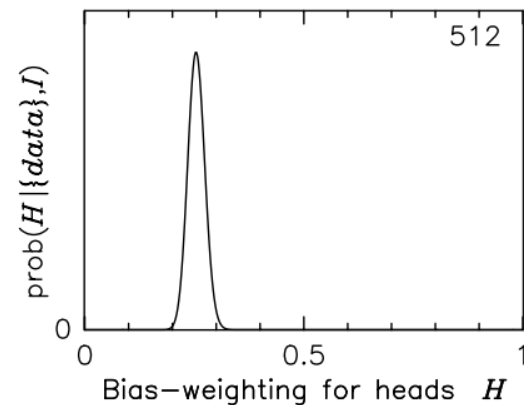
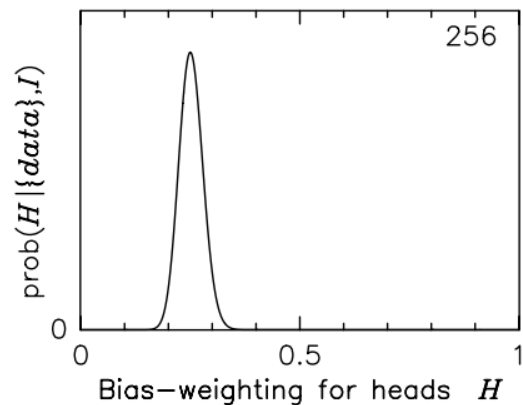
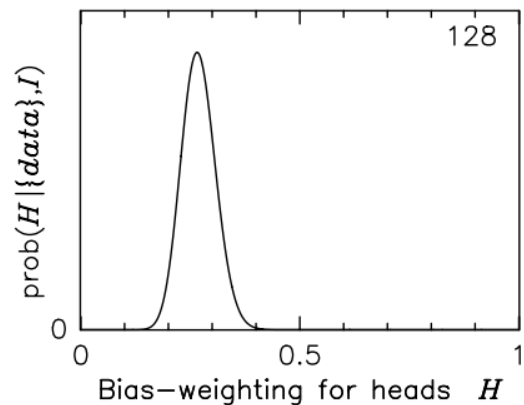
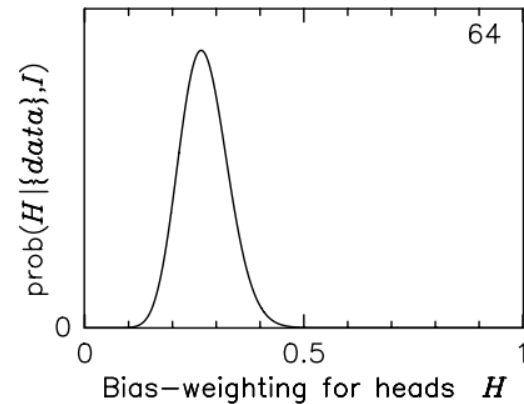
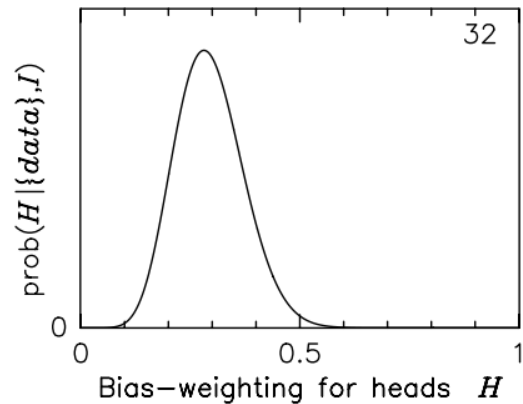
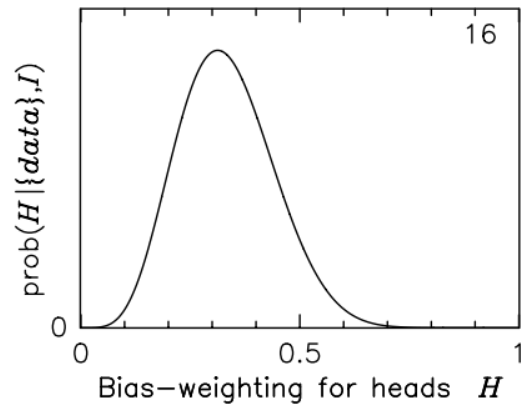
→ These are also valid for  $N = k = 0$ , in which case we get the  $E = \frac{1}{2}$  and  $V = \frac{1}{12}$

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



# Visualization of Posterior

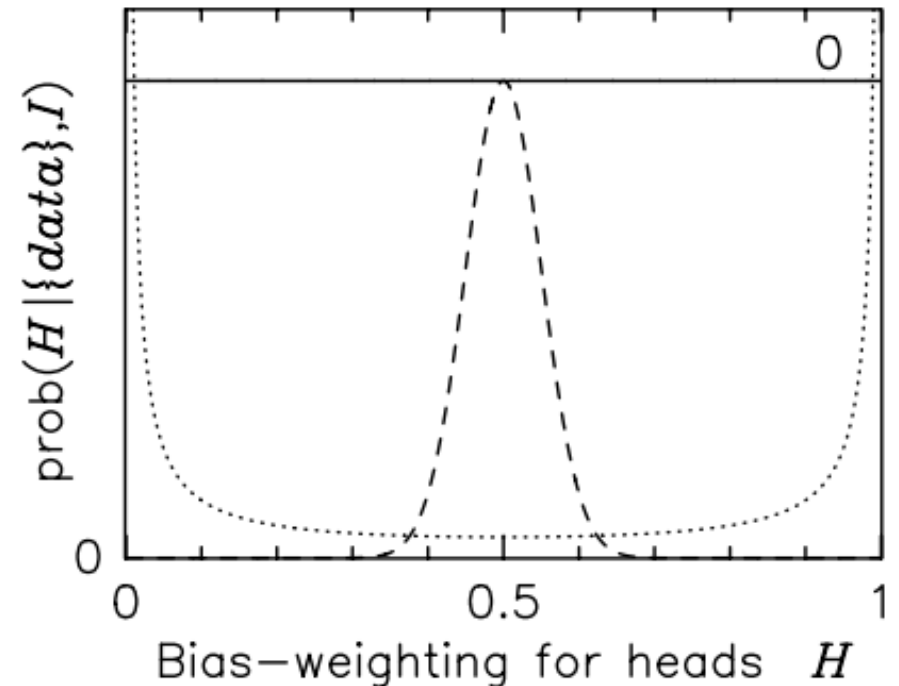




# Priors

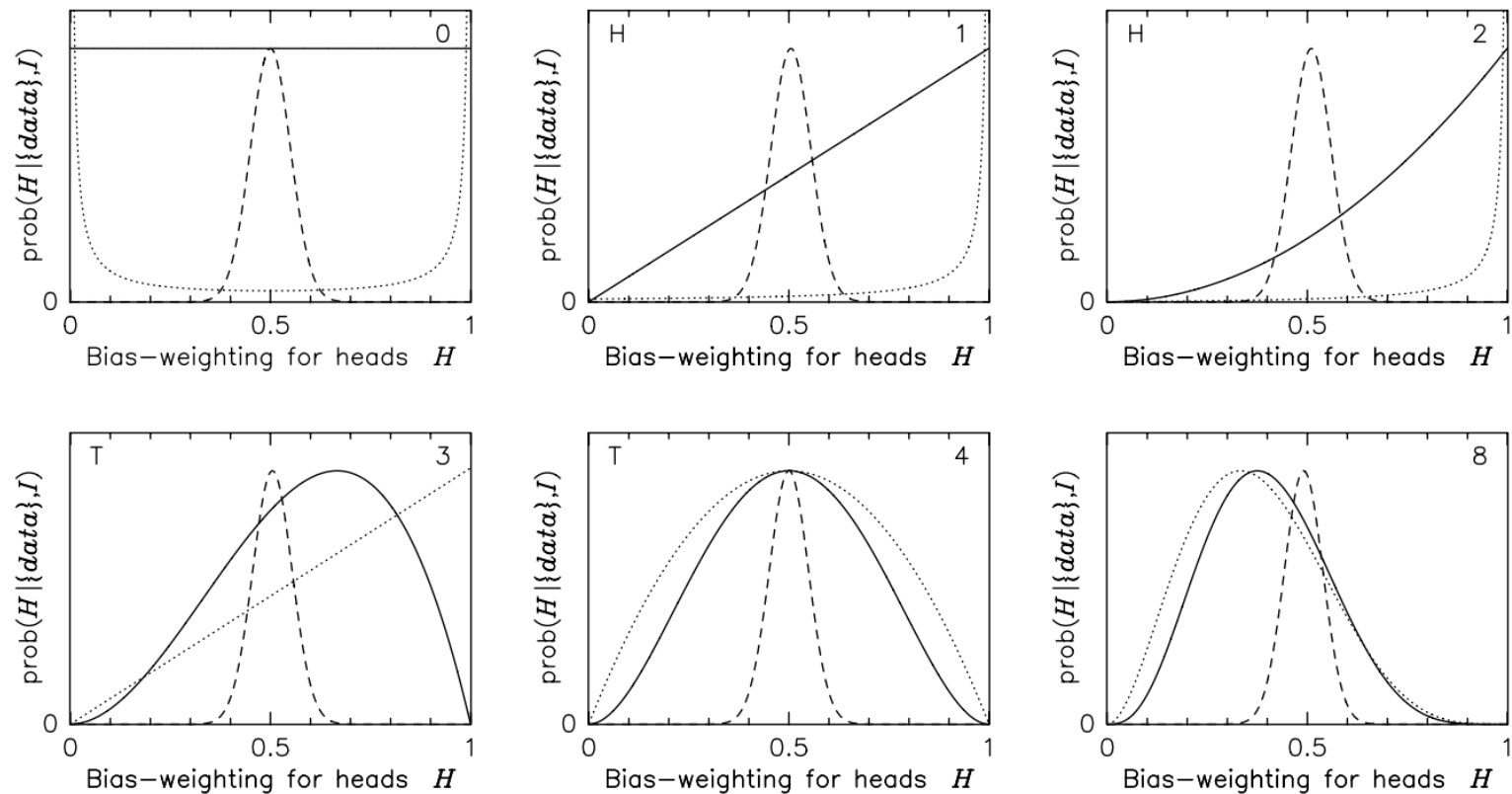
# Importance of priors

- 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

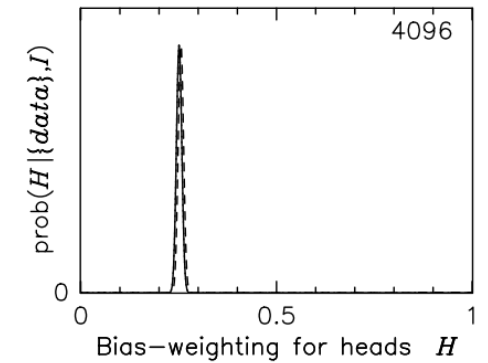
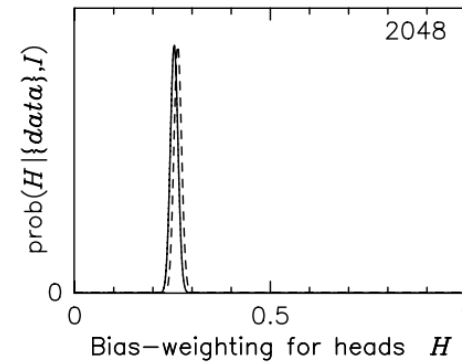
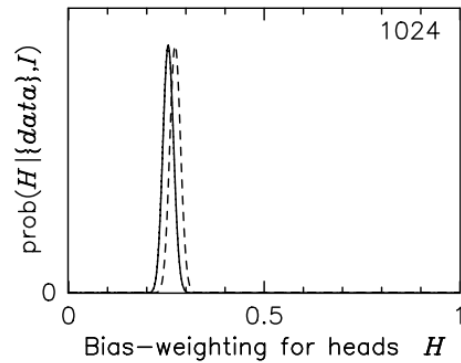
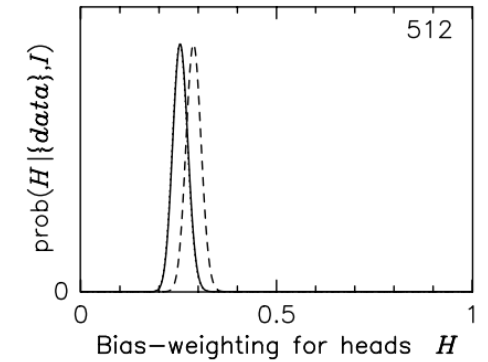
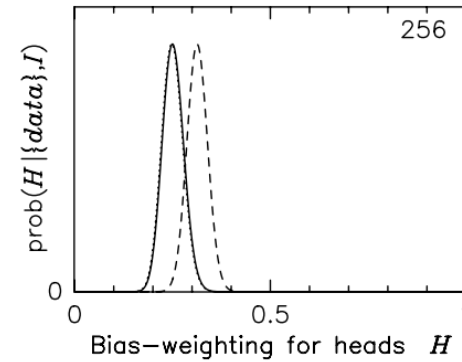
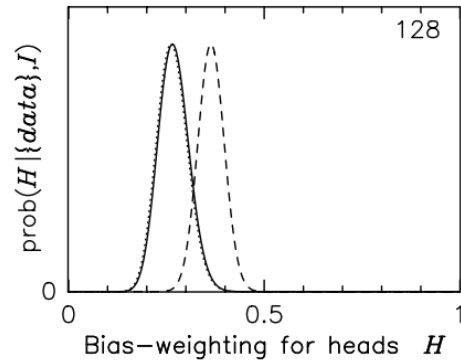
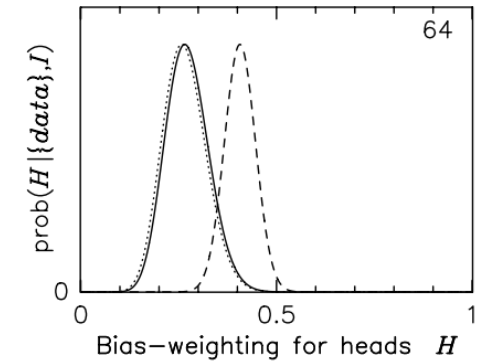
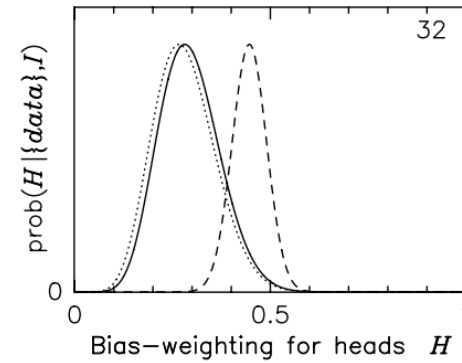
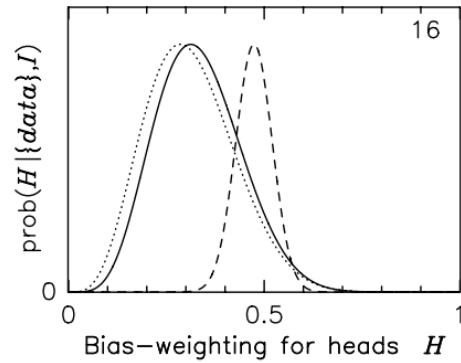


# Effect of priors

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

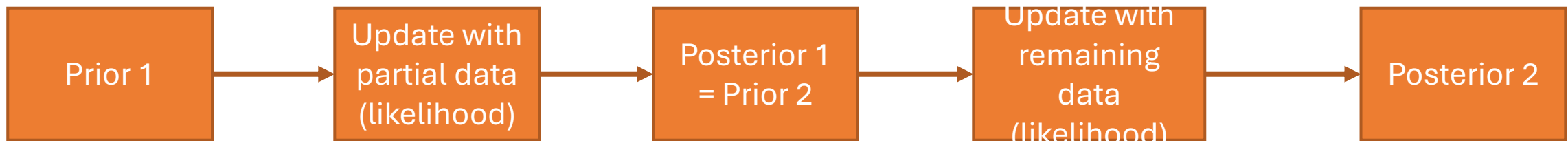
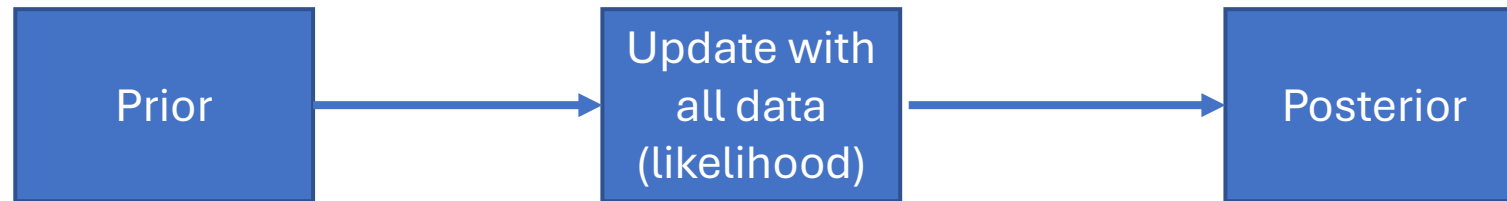


But when analyzing a larger dataset, the effect of the prior diminishes  
→ the data is more informative than the prior, and prior choices don't matter as much anymore



# 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

$$\begin{aligned} p(p|k_2, N_2) &= \frac{p(k_2|p, N_2)p(p)}{p(k_2)} = \frac{p(k_2|p, N_2)p(p)}{\int p(k_2|p, N_2)p(p) dp} \\ &= \frac{p^{k_2}(1-p)^{N_2-k_2}p^{k_1}(1-p)^{N_1-k_1}}{\int p^{k_2}(1-p)^{N_2-k_2}p^{k_1}(1-p)^{N_1-k_1}dp} = \frac{p^{k_1+k_2}(1-p)^{N_1+N_2-k_1-k_2}}{\int p^{k_1+k_2}(1-p)^{N_1+N_2-k_1-k_2}dp} = \frac{p^k(1-p)^{N-k}}{\int p^k(1-p)^{N-k}dp} \end{aligned}$$

→ 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](https://en.wikipedia.org/wiki/Conjugate_prior)

# Jeffreys Prior

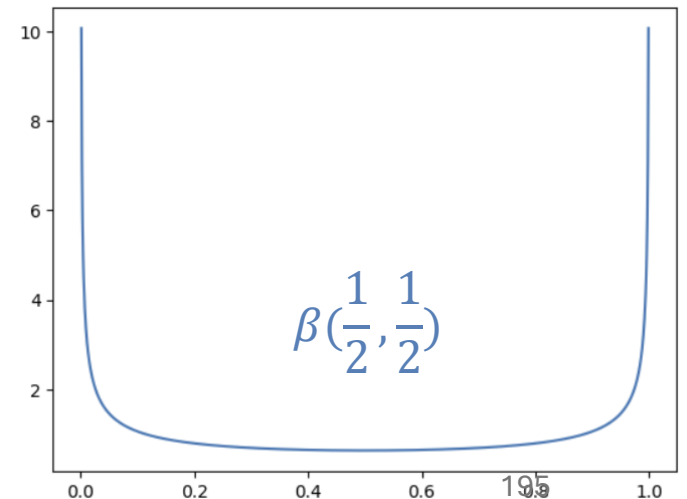
- The Jeffreys prior is intended to be a non-informative prior distribution
- It is constructed from the Fisher information  $I$  ( $\rightarrow$  see Lecture 13)
- Jeffreys prior  $p(\theta) \sim \sqrt{\det I(\theta)} = \sqrt{\mathbb{E}[(\partial\theta \log p(x|\theta))^2]}$

- Example for our Binomial:

$$\bullet I(p) = -E[\partial^2 p \log p] = \frac{Np}{p^2} + \frac{N-Np}{(1-p)^2} = \frac{N}{p(1-p)} \sim p^{-1}(1-p)^{-1}$$

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

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  $\theta \rightarrow \xi$  since

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

# 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  $\sigma$  is fixed (= known)

Let's see what we can learn about  $\mu$  from a single measurement  $x$ !

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

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

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

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

$$p(\mu|x_1, x_2, \sigma_1, \sigma_2) = \frac{1}{\sqrt{2\pi\sigma_A}} e^{-\frac{1}{2}\left(\frac{x_A - \mu}{\sigma_A}\right)^2}$$

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

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

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  $(\sum 1/\sigma_i^2)^{-1}$



# 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

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

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

# Summary of posterior

- We can construct “credible intervals” from our posterior to summarize it
  - Calculate intervals in  $\mu$  that contain a desired amount of probability, For instance the central or smallest intervals:

Probability Content (in %)	$\mu$ 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.”

# Credible vs. C.L. Regions

## Common misunderstandings [ edit ]

See also: § [Counterexamples](#)

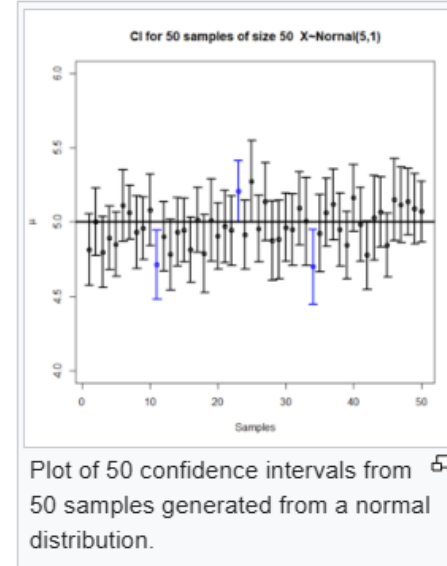
Confidence intervals and levels are frequently misunderstood, and published studies have shown that even professional scientists often misinterpret them.<sup>[12][13][14][15][16][17]</sup>

- 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).<sup>[18]</sup> 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.<sup>[19]</sup>

[Neyman](#) himself (the original proponent of confidence intervals) made this point in his original paper:<sup>[10]</sup>

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  $\alpha$ . 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  $\alpha$ ? 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.<sup>[16]</sup>



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

**Pathological Example:** Let's suppose you construct intervals for the width  $\sigma$  of a Gaussian centered at 0:

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

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

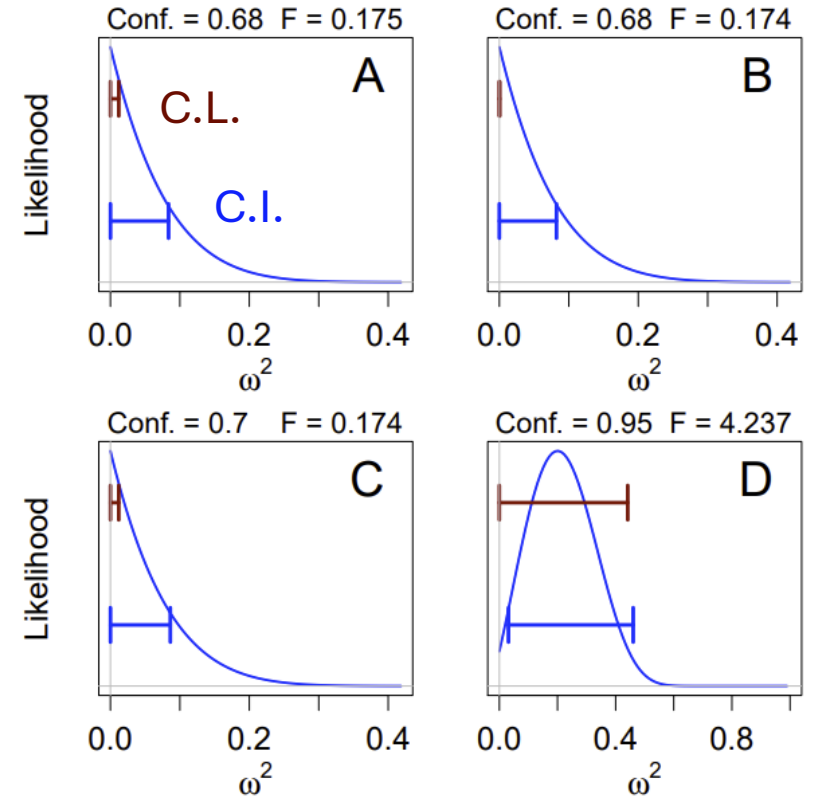
→ The true value of  $\sigma$  will be contained in exactly 50% of cases!

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

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

# 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**)



# 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 reparameterization, therefore 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

# 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  $\mu$ ?
  - The likelihood is now a function of both,  $\sigma$  and  $\mu$

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

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

## 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  $\sigma$  is fixed (= known)

Let's see what we can learn about  $\mu$  from a single measurement  $x$ !

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

# Prior

The prior now has to be a joint probability function over  $\sigma$  and  $\mu$  i.e. of the form  $p(\mu, \sigma)$

We can again choose a “flat” prior

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

Let's assume the boundaries  $\mu_{min}$ ,  $\mu_{max}$  and  $\sigma_{max}$  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  $\sigma$  and  $\mu$ !  
But we may not be interested in  $\sigma$ ...

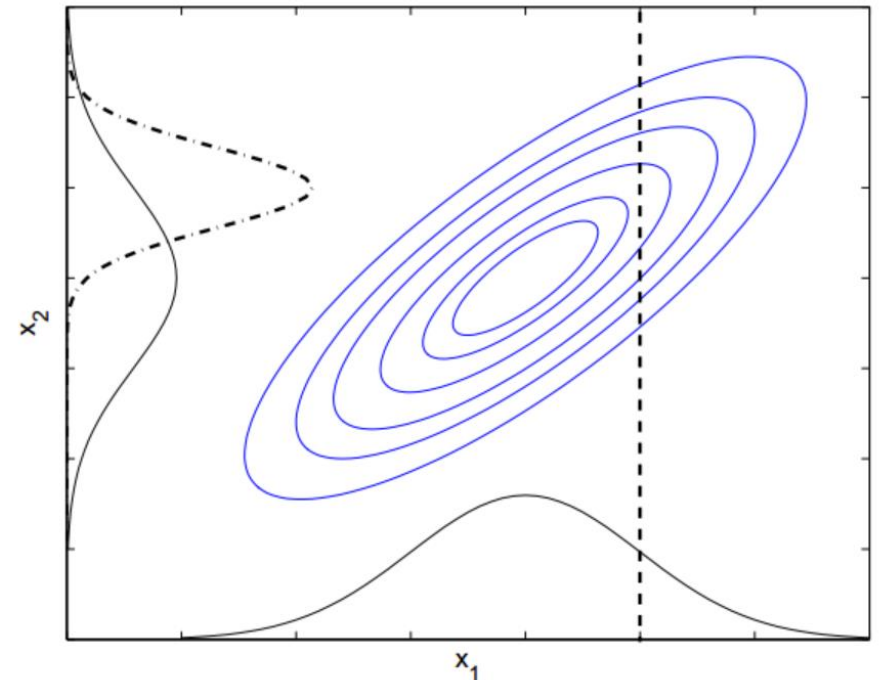
# Marginal distribution

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

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

→ 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

$$\begin{aligned} p(\mu|\{x\}) &\sim \int \prod_i \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} d\sigma \\ &= \int (\sqrt{2\pi\sigma})^{-N} e^{-\frac{1}{2\sigma^2} \sum_i (x_i-\mu)^2} d\sigma \end{aligned}$$

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

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

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

Using another substitution of  $\tau = t\sqrt{\sum_i (x_i - \mu)^2}$  makes the integral independent of  $\mu$ , 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)

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

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

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:

$$\log p(\mu) = \log p(\mu_0) + \left. \frac{d \log p}{d\mu} \right|_{\mu_0} (\mu - \mu_0) + \frac{1}{2} \left. \frac{d^2 \log p}{d\mu^2} \right|_{\mu_0} (\mu - \mu_0)^2 + \dots$$

- The first term is constant  $\rightarrow$  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:

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

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  $\mu_0$
- width  $1 / \sqrt{-\frac{d^2 \log p}{d\mu^2}}$

# 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

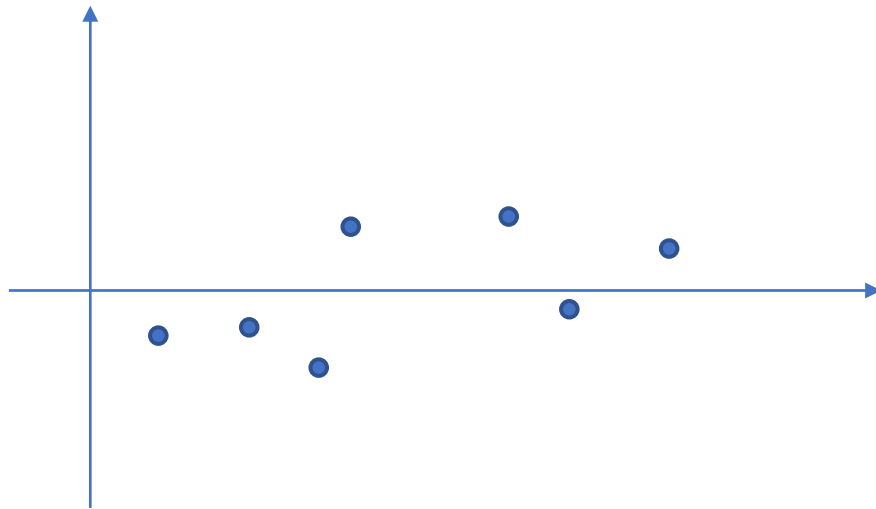
The above has the same form as the case with fixed  $\sigma$ , except that  $\sigma$  has been replaced by an estimate from the data!

# 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: ...

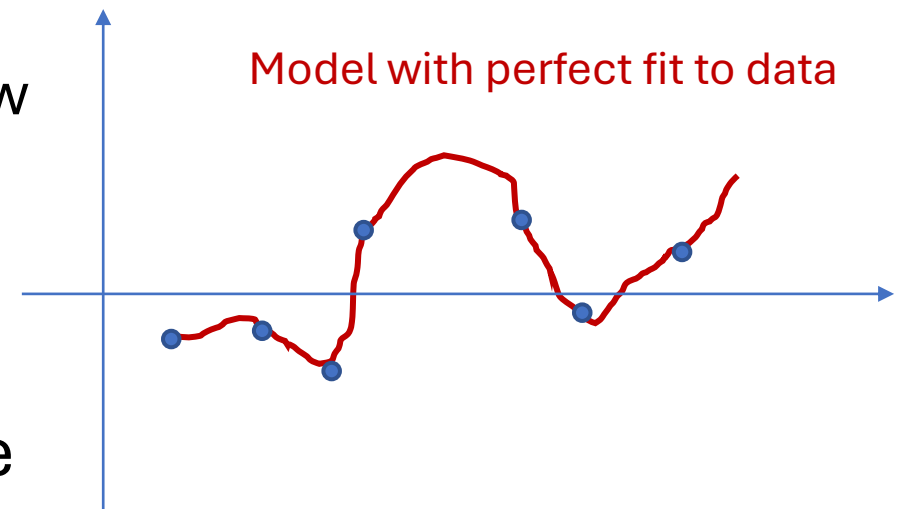
?

# 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 than 1)

# Using Bayes

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

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

(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  $\lambda$ ?
- 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  $\lambda$  under model B



# Example

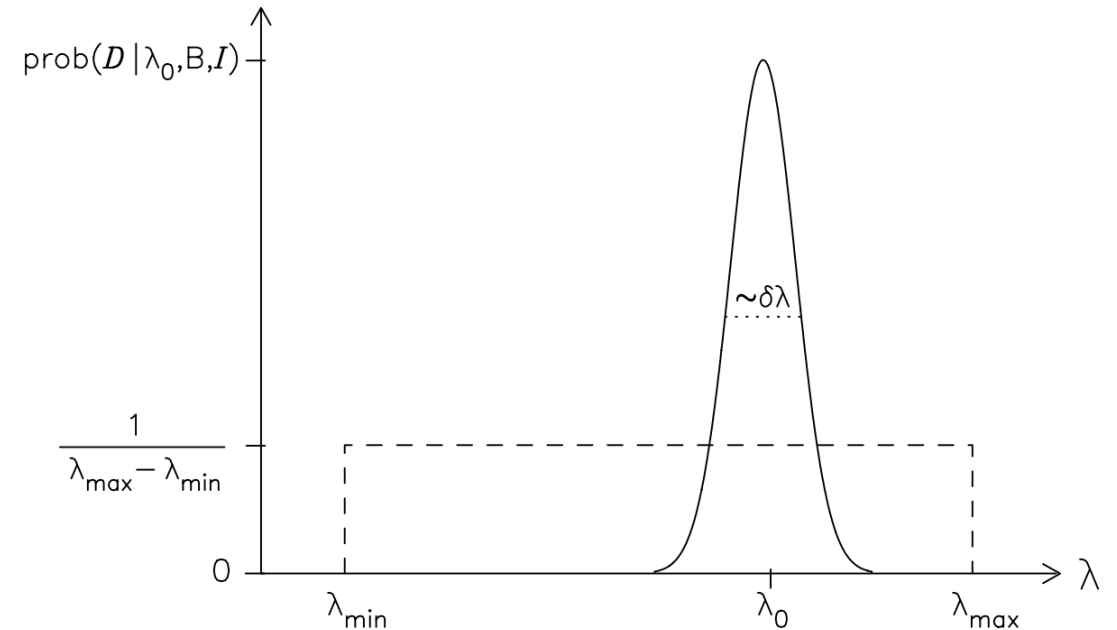
- Let's use a uniform prior for  $\lambda$ :

$$\bullet p(\lambda|B) = \frac{1}{\lambda_{max} - \lambda_{min}}$$

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

- Let's also assume that the likelihood is normal around the MLE  $\lambda_0$  (i.e. parabolic in LLH) with a width (uncertainty) of  $\delta\lambda$

$$\bullet p(D|\lambda, B) = p(D|\lambda_0, B) \times e^{-\frac{(\lambda - \lambda_0)^2}{2\delta\lambda^2}}$$



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

# 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

Posterior

Likelihood

Prior

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

Evidence (Marginal Likelihood)

# Solving the integral...

- 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**

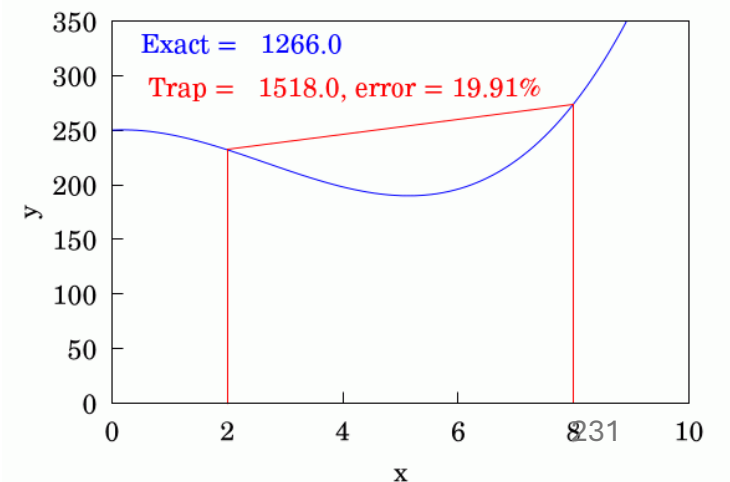
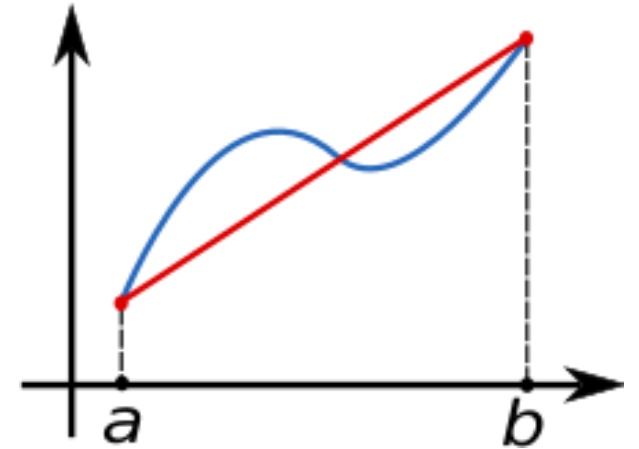
# Numerical Methods

# Trapezoidal rule

- The trapezoid method makes an approximation of the function  $f(x)$  to be integrated as a trapezoid

$$I = \int_a^b f(x) dx \approx (b - a) \frac{1}{2} (f(a) + f(b))$$

- Can approximate integral values arbitrarily well by chaining together trapezoids over  $\Delta x$  and going to smaller grids
- Works well in 1d, breaks down in higher dimensions (curse of dimensionality)

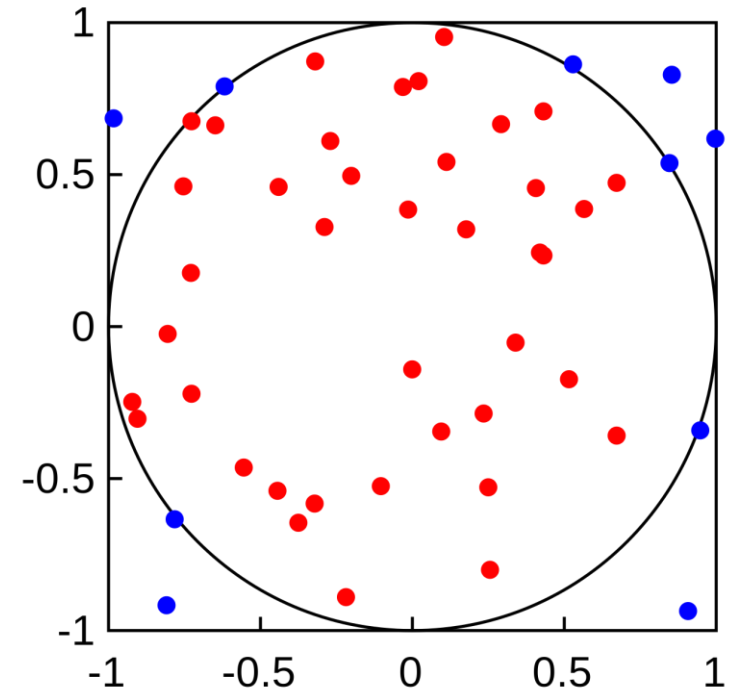


# Hit-or-miss Integration

- Integration method based on random numbers, i.e. a Monte Carlo (MC) integration technique
- Encapsulating the function  $f(x)$  under a uniform function  $g(x)$  (= a box) for which we know the integral value

$$I = \int_a^b f(x)dx = r \int_a^b g(x)dx, \text{ where } r \approx \frac{H}{N}$$

→ The error on  $I$  scales with  $\frac{1}{\sqrt{n}}$



H = number of hits in N trials



# Geometry in higher dimensions

Integration with „boxes“ – i.e. hypercubes, can become problematic in higher dimensions

- Volume of a unit hypercube (side lengths = 1) is always  $V_{box} = 1$
- Volume of a sphere in  $n$  dimensions is:

$$V_{sphere}(r) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2} + 1)} r^{\frac{n}{2}}$$

→ For  $2d$ , our efficiency of hit-or-miss is  $\frac{\pi}{4} \approx 78\%$

→ For  $5d$ , the efficiency is  $\approx 16\%$

→ For  $10d$ , our efficiency drops already to  $\approx 0.2\%$  (!!)

High-dimensional hypercubes have almost all Volume in the corners

# Sample mean MC

An integral can be rephrased as an expectation value over a distribution with pdf  $g$

$$I = \int_a^b f(x) dx = \int_a^b \frac{f(x)}{g(x)} g(x) dx = \mathbb{E}_g \left[ \frac{f(x)}{g(x)} \right]$$

So long as  $g(x) > 0$  when  $f(x) \neq 0$

We can simply use a flat distribution (a box) again with large enough boundaries.

→ This let's us estimate the expectation value as the sample mean!

$$\hat{I} = \frac{1}{n} \sum_i \frac{f(x_i)}{g(x_i)}$$

Where the samples  $x_i$  are generated according to  $g(x)$

The error of this method also scales as  $\frac{1}{\sqrt{n}}$

# Importance sampling

We can improve the precision of the sample mean MC when choosing a function  $g(x)$  that has more probability mass where it is “more important”

→  $g(x)$  does not need to be uniform

→ This is called importance sampling

Minimum variance is achieved when the function  $g$  follows the shape of  $f$  very closely, i.e.  $g(x) \approx \frac{f(x)}{\int f(x)dx} = \frac{f(x)}{I}$  (here assuming  $f(x) > 0 \forall x$ )

In the limit, this requires knowledge of  $I$ , which is what we’re interested in in the first place

# Harmonic Mean

- Let's define a normalized function:

$$\tilde{f}(x) = \frac{f(x)}{\int_a^b f(x)dx} = \frac{f(x)}{I}$$

- Calculating the expectation value

$$\mathbb{E}_{\tilde{f}} \left[ \frac{1}{f(x)} \right] = \int_a^b \frac{1}{f(x)} \tilde{f}(x) dx = \frac{V}{I} = \frac{b-a}{I}$$

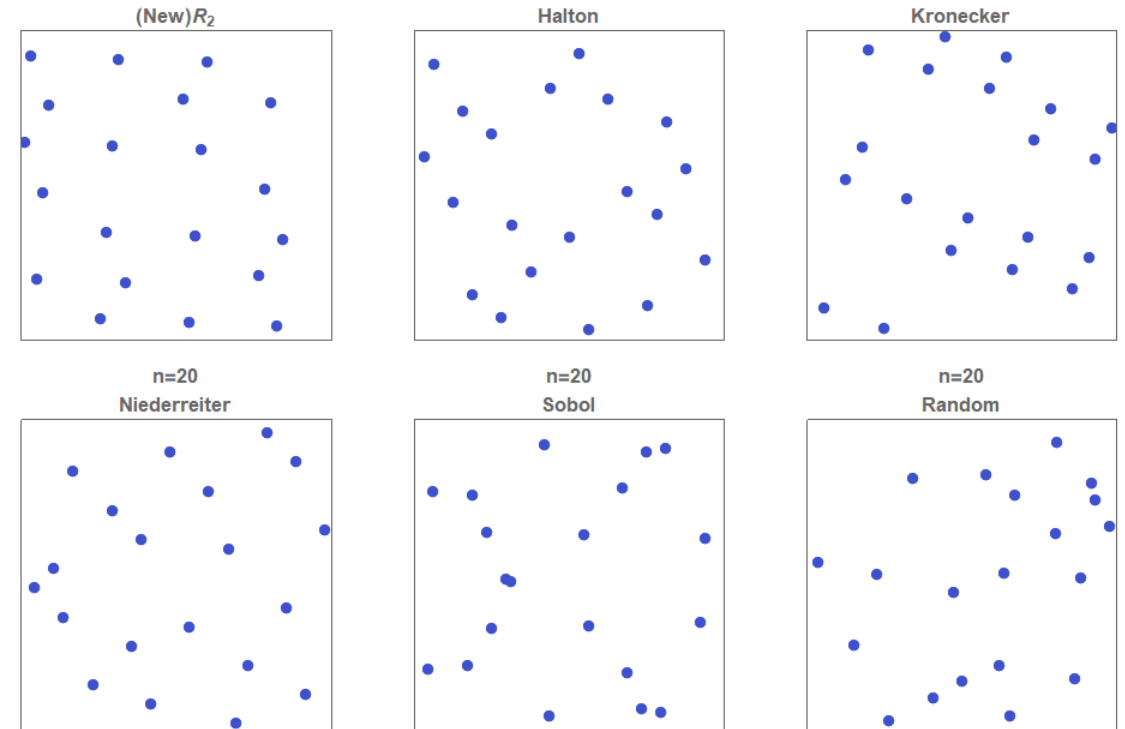
→ Our integral can be estimated as  $\hat{I} = \frac{b-a}{\frac{1}{n} \sum_i \frac{1}{f(x_i)}}$  where  $x_i$  are distributed according to  $\tilde{f}$

(this is a useful technique if we have access to samples of  $\tilde{f}$ , e.g. the posterior)

This method has also been called 'The worst Monte Carlo algorithm ever' (Radford Neal), because it's easy for  $\frac{1}{f}$  to diverge. (See our paper for more <https://arxiv.org/abs/1808.08051>)

# Quasi Random Sampling

- MC sampling has some limitations:
  - Parts of the phase space may be empty (no samples)
  - Other parts may have several samples close together (clumping)
- In rel. low numbers of dimensions, this can be improved by using quasi random numbers instead of pseudo random numbers
  - A.k.a. “Low discrepancy sequences”
- Further reading:  
<https://extremelearning.com.au/unreasonable-effectiveness-of-quasirandom-sequences/>



# Summary of today

- Bayes in more than one parameter
    - Marginalization of nuisance parameters: incorporates our ignorance
    - Conditional on nuisance parameters: assumes perfect knowledge
  - Approximation of posterior by Gaussian
    - Can give us the MAP plus error bars
  - Bayes for Model selection:
    - Compare posterior odds or Bayes factor
    - Additional degrees of freedom and large priors are penalized
  - Numerical methods for integration
    - Becomes a difficult problem in higher dimensions
    - Saw a few basic algorithms
- Next week we'll discuss Markov chain Monte Carlo (MCMC)

# MCMC Sampling

# Monte Carlo Methods

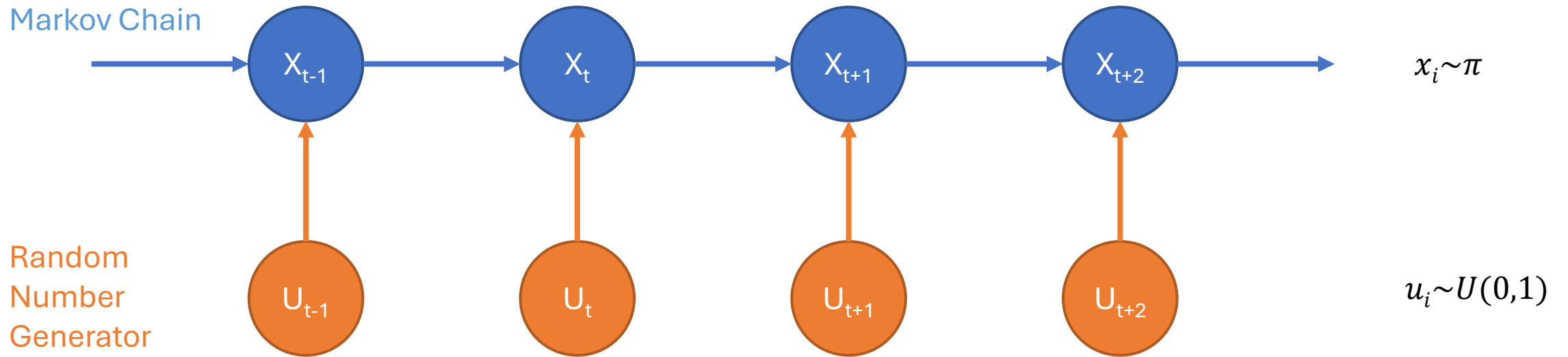
- We had already talked about MC methods for generating random numbers
  - Uniform random numbers between 0 and 1
  - Random numbers distributed according to other densities via transformation
  - Accept-Reject Sampling
  - ...
- We have used those also for computing integrals
  - Last lecture: sample mean MC, Harmonic mean, ...



# Today

- Use Markov Chain MC to generate samples  $x_i$  that are distributed according to a target distribution  $\pi$ 
  - This could for instance be our posterior distribution
  - But is not limited to the case of Bayes
- Pros:
  - This new method will work well in high numbers of dimensions
  - Will allow to sample from arbitrary distributions
- Cons:
  - The generated samples will no longer be *i. i. d.* but can be correlated
    - The effective sample size can be smaller than  $n$

# Markov Chain



# Markov Process

The defining property of a **Markov Process** is, that the probability distribution of  $x$  depends only on the current state, i.e.:

$$p(a < x < b | x_1, x_2, \dots, x_n) = p(a < x < b | x_n)$$

→ This means any previous behavior does not matter

**Example:** A random walk, where each step is going forward a fixed length left or right from the current position is a Markov Process

For a discrete state space, this is called a **Markov Chain**

# Transition Kernel & Stationarity

The **transition kernel** (transition probability function) is the key to a Markov chain

→ It tells us where to go from a state to the next one, i.e. it defines:

$$p(x_t|x_s), \quad t > s$$

**Definition:** If the joint probability distribution of

$\{x_{t_1}, x_{t_2}, x_{t_3}, \dots, x_{t_n}\}$  and  $\{x_{t_1+h}, x_{t_2+h}, \dots, x_{t_n+h}\}$

are the same for any  $t$  and offset  $h$  we have a **stationary** process

# Transition matrix

- For a Markov chain (discrete steps), we can express the probability to go from a state  $i$  to to state  $j$  in one step as the matrix  $P_{ij}$ 
  - $P_{ij}$  is called the “**1-step transition probability matrix**”
  - This matrix + some initial conditions fully define a Markov Chain

- **Chapman-Kolmogorov Relation:**

If our Markov chain is in the state of a stationary process, then:

$$P^n_{ij} = \sum_k P^r_{ik} P^s_{kj}, \text{ where } n = r + s$$

( $P^n$  is the n-step transition probability)

→ Transition probability can be decomposed into intermediate steps

# Irreducible process

If all possible states that we can be in *communicate* with each other, they are all in the same equivalence class

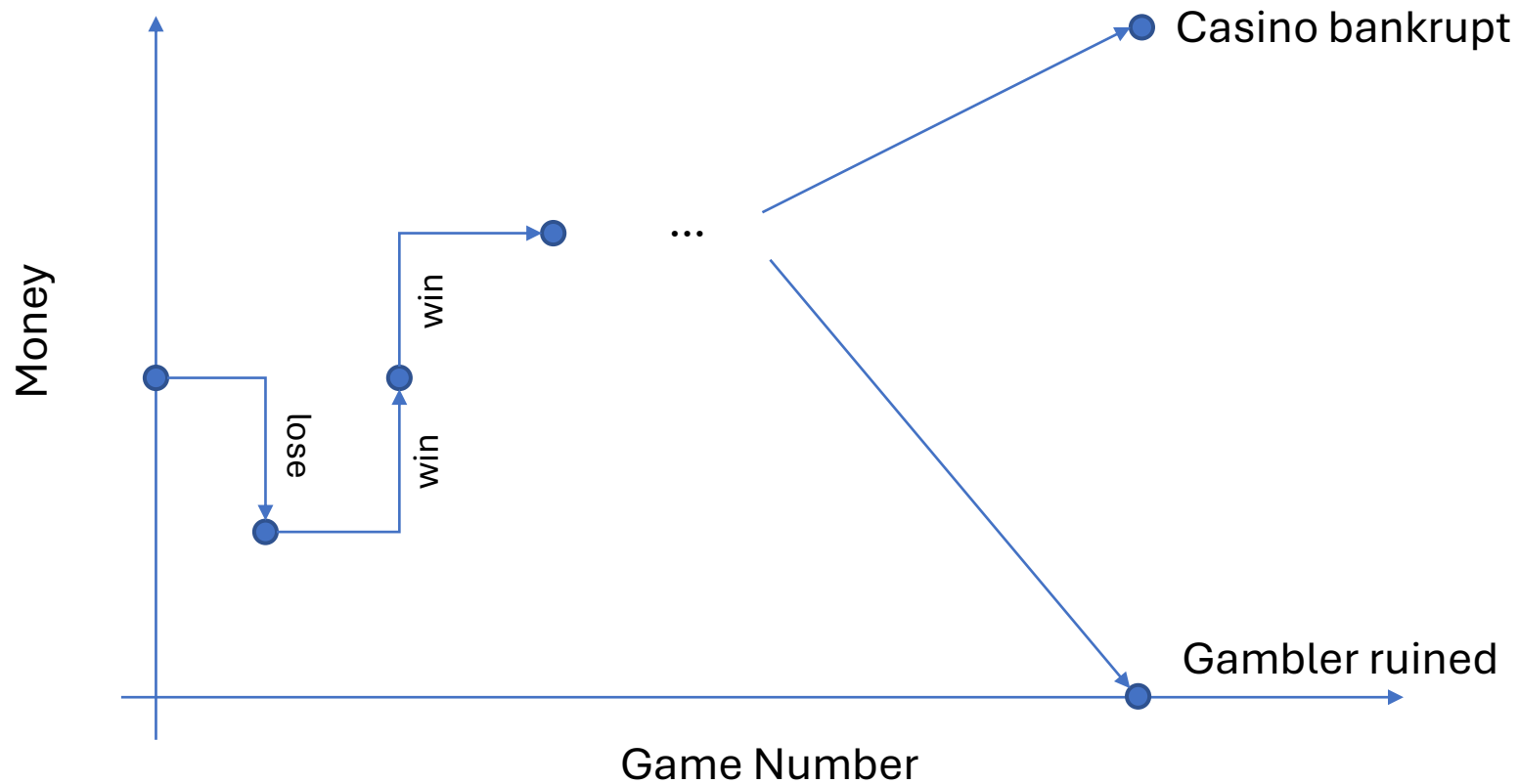
- This means any state  $i$  can be reached from any other state  $j$  in finite time
- (And vice versa  $j \rightarrow i$ )

**Definition:** An **irreducible** Markov Chain contains only one equivalence class

→ If this would not be the case, our chain would get “stuck” somewhere and cannot reach other parts, so it cannot sample from our target distribution correctly

# Example: Random Walks in 1d

- Gambler's Ruin



**Rules:**

- Start with some Money
- Play games with e.g. 50/50 chance of winning/losing
- Play until you have no money left (Gambler's ruin) or you bankrupted the casino

# Gambler's Ruin

- The 1-step transition matrix will look like:

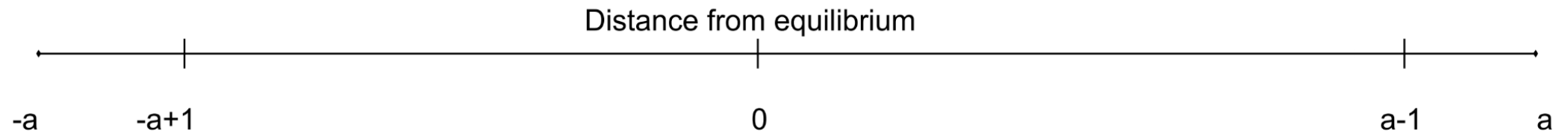
$$P_{ij} = \begin{pmatrix} 1 & 0 & 0 & \dots & & & \\ q & 0 & p & 0 & \dots & & \\ 0 & q & 0 & p & 0 & \dots & \\ \cdot & & & & & & \\ \cdot & & & & & & \\ \cdot & \dots & & & & & 1 \end{pmatrix}$$

- The two end points are absorbing:
  - You cannot play anymore if you (or the casino) run out of money!
  - The endpoints can be reach from anywhere, but the system remains in such a state once it is reached
  - The system contains three classes and therefor is not irreducible!



# Ehrenfest Model

Consider a different 1d random walk, where there is an equilibrium position and a restoring force proportional to the distance to the equilibrium



- 1-step transition Matrix become:

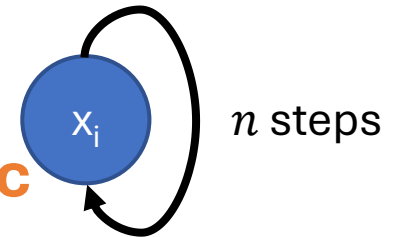
- The endpoints are “reflecting”
- This system is irreducible

$$P_{ij} = \begin{cases} \frac{a-i}{2a} & j = i+1 \\ \frac{a+i}{2a} & j = i-1 \\ 0 & \text{else} \end{cases} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \frac{1}{2a} & 0 & \frac{2a-1}{2a} & 0 & \\ 0 & \frac{2}{2a} & 0 & \frac{2a-2}{2a} & \\ \vdots & \frac{2a}{2a} & & 0 & \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \begin{matrix} i = -a \\ i = -a+1 \\ i = -a+2 \\ \\ i = a \end{matrix}$$

# Aperiodic process

- The period  $d$  of a state is the greatest common divisor for  $n \geq 1$  for which  $P^n_{ii} > 0$   
This means the probability to come back to state  $i$  after taking  $n$  steps is non zero

**Definition:** If the period  $d = 1$ , our chain is said to be **aperiodic**



## Example:

- Random walk with probability  $p$  to go left and  $q = 1 - p$  to go right:  
period  $d = 2 \rightarrow$  **not aperiodic**
- Random walk with a probability  $> 0$  to remain  $(1 - p - q)$  **is aperiodic**

# Recurrence

**Definition:** A state is said to be **recurrent** if

$$\sum_{n=1}^{\infty} P^n_{ii} = \infty$$

This means that we are guaranteed to be back to state  $i$  in finite time.

**Example:**

- Random walk in 1d and 2d is recurrent
- Random walk in 3d is *not* recurrent  
(see e.g. [https://en.wikipedia.org/wiki/Random\\_walk](https://en.wikipedia.org/wiki/Random_walk) for details)

# Basic Limit Theorem

If we only consider **irreducible**, **recurrent**, and **aperiodic** Markov chains  
→ Then the basic limit theorem holds:

$$\lim_{n \rightarrow \infty} P^n_{ji} = P_{ii}$$

This means that after a large number of steps, it does not matter where we started from!

→ Independent of the initial conditions, we will eventually reach the stationary distribution

Such a Markov Chain is also called “**ergodic**”

# Ergodicity

For an ergodic process with stationary probability distribution  $\pi$ :

$$\lim_{n \rightarrow \infty} P_{jj}^n = \pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}$$

→ Our goal is to find a Markov chain with stationary distribution  $\pi =$  our desired pdf of interest (e.g. a posterior probability function)

# Convergence

How do we know if we reached the stationary distribution?  
(i.e. how do we know if  $n$  is large enough?)

**Detailed balance** is a sufficient condition:

$$\pi_i P_{ij} = \pi_j P_{ji}$$

Since:  $\sum_{i=0}^{\infty} \pi_i P_{ij} = \sum_{i=0}^{\infty} \pi_j P_{ji} = \pi_j \sum_{i=0}^{\infty} P_{ji} = \pi_j$

# Actual MCMC Algorithms

# Metropolis Algorithm

Original MCMC algorithm (1953):

1. Suppose we are in state  $x$ , generate a new state  $y$  according to a symmetric function  $g(y|x) = g(x|y)$  (= proposal distribution)
2. Calculate  $r = \frac{\pi(y)}{\pi(x)}$
3. Draw a uniform random number  $u \sim U(0,1)$   
If  $u < r$  accept new state  $y$ . Otherwise remain in state  $x$

(So a step with  $\pi(y) > \pi(x)$  is always accepted)



# Example: Normal from Uniform

## Metropolis MC:

```
# target density:  
p = stats.norm()  
  
# proposal function:  
width = 0.1  
g = stats.uniform(-width, 2*width)
```

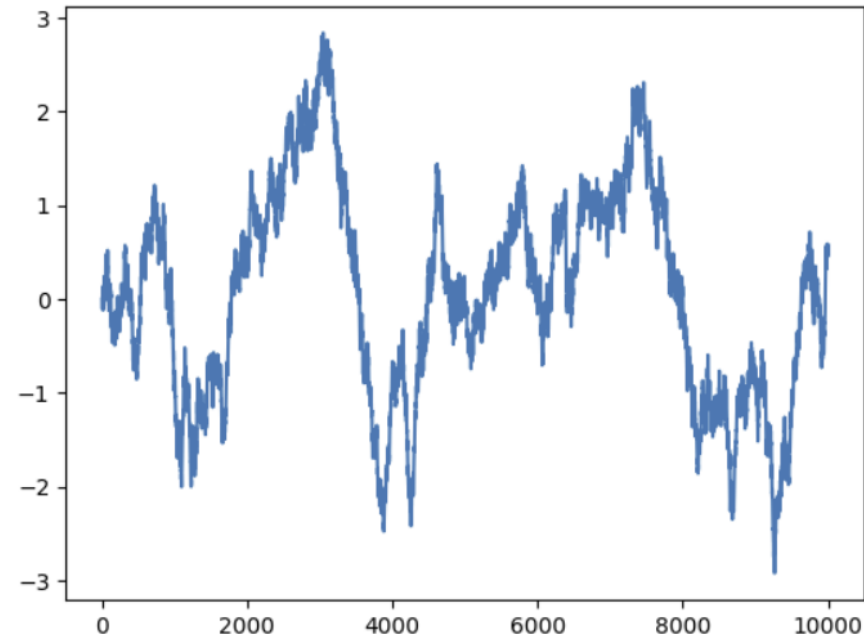
```
samples = [0, ]
```

```
for i in range(10_000):  
    x = samples[-1]  
    y = x + g.rvs()  
    r = p.pdf(y)/p.pdf(x)  
    u = stats.uniform().rvs()  
    if u < r:  
        samples.append(y)  
    else:  
        samples.append(x)
```

## Trace of the Chain:

```
plt.plot(samples)
```

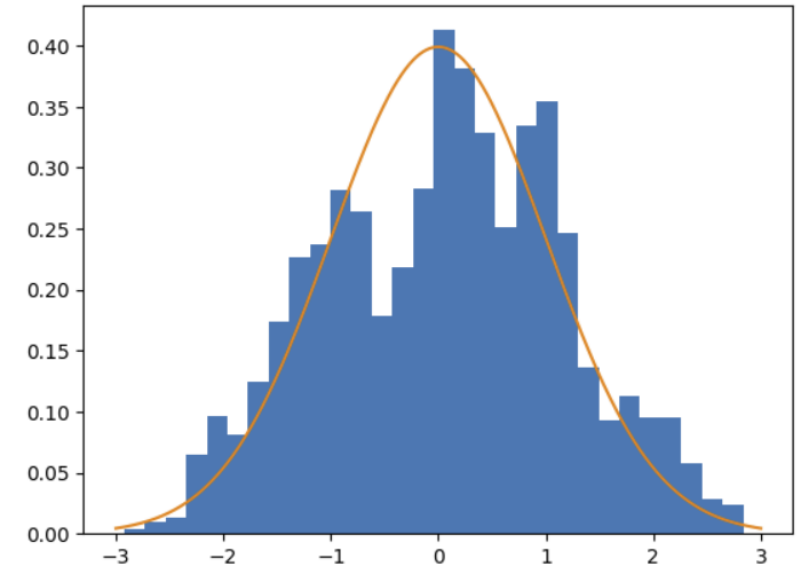
```
[<matplotlib.lines.Line2D at 0x7f73e0264160>]
```



## Density of generated Samples:

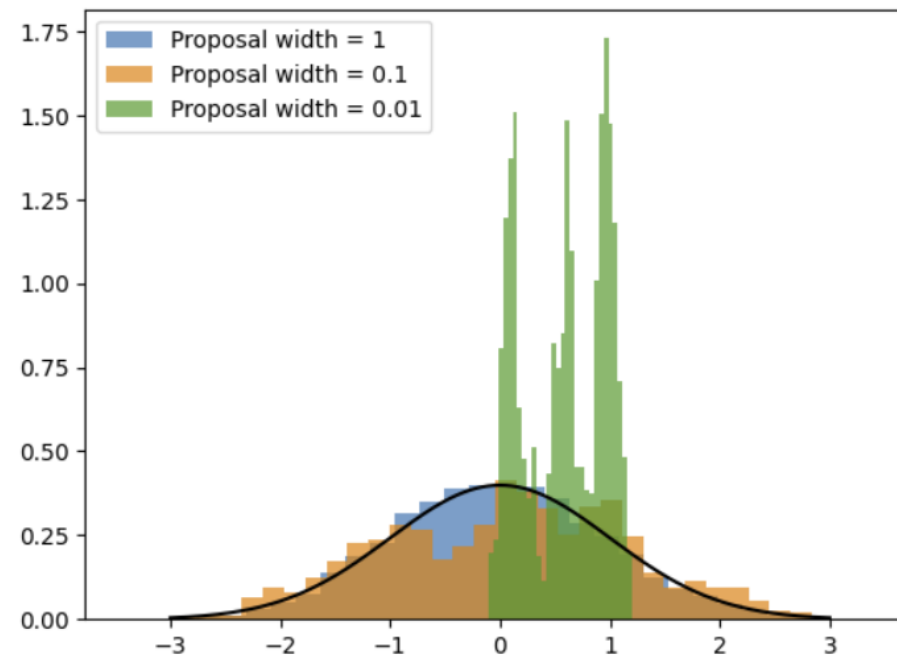
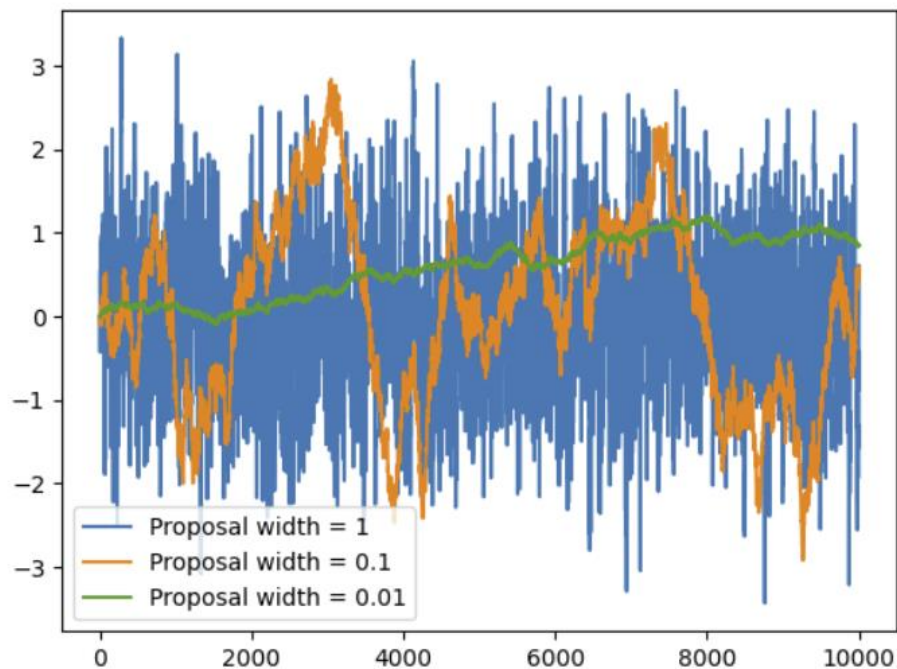
```
plt.hist(samples, bins=30, density=True)  
l = np.linspace(-3, 3, 1000)  
plt.plot(l, p.pdf(l))
```

```
[<matplotlib.lines.Line2D at 0x7f73e01e9bb0>]
```



# Importance of proposal Function

- The speed with which we reach the stationary distribution depends on our choice of proposal distribution
  - The acceptance fraction should not be too low



# Metropolis-Hastings Algorithm

Generalization of Metropolis Algorithm for non-symmetrical proposal functions  $g$ :

1. Suppose we are in state  $x$ , generate a new state  $y$  according to proposal distribution  $g(y|x)$
2. Calculate  $r = \min \left\{ \frac{\pi(y)g(x|y)}{\pi(x)g(y|x)}, 1 \right\}$
3. Draw a uniform random number  $u \sim U(0,1)$   
If  $u < r$  accept new state  $y$ . Otherwise remain in state  $x$

# MCMC for Bayes

In Bayesian Inference, we want to know properties of the posterior probability:

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

Using Metropolis-(Hastings), we only need the **ratio** for the transition probability from state  $\theta_{old} \rightarrow \theta_{new}$ :

$$\frac{p(\theta_{new} | \mathbf{x}, M)}{p(\theta_{old} | \mathbf{x}, M)} = \frac{p(\mathbf{x} | \theta_{new}, M) p(\theta_{new} | M)}{p(\mathbf{x} | \theta_{old}, M) p(\theta_{old} | M)}$$

The evidence cancels out!!! (it is independent of  $\theta$ )

# Practical considerations

- We start our Markov Chain from some arbitrary starting (seed) point
- It takes a while until:
  - A) the proposal function is tuned
  - B) we reach equilibrium (the chain is sampling from the stationary distribution), i.e. convergence

→ First “bunch” of samples in a chain are discarded (burn-in)

How many is a non-trivial question

- Some standard diagnostics include auto-correlation lengths
- Or start multiple chains and compare their trace

# Other Algorithms

- We have discussed the most vanilla MCMC algorithms
  - Metropolis
  - Metropolis-Hastings
- There exist a multitude of other algorithms
  - More basic: Gibbs sampler, slice sampler, ...
  - Using gradient information: Hamiltonian MC (only works for smooth, differentiable, unimodal target densities, but up to very high dimensions!)
  - Nested Sampling: different approach, similar to Lebesgue integration (Also gives Evidence value in addition to samples!)

# A few libraries to consider

- General Libraries:
  - PyMC3: <https://github.com/pymc-devs/pymc>
  - BAT: <https://github.com/bat/bat> (Python via <https://github.com/bat/batty>)
- Affine sampler:
  - Emcee: <https://emcee.readthedocs.io/>
- Hamiltonian MC:
  - STAN: <https://mc-stan.org/>
- Nested sampling:
  - Multinest: <https://github.com/rjw57/MultiNest>
  - Ultranest: <https://johannesbuchner.github.io/UltraNest/index.html>
  - Dynesty: <https://dynesty.readthedocs.io/en/stable/>
  - Polychord: <https://github.com/PolyChord/PolyChordLite>
- And many, many more...

# Summary of MCMC

- MCMC sampling offers a way to generate samples according to a target density
  - Needs to fulfil some criteria: irreducible, aperiodic & recurrent (+ ergodic) to be guaranteed to converge to the stationary distribution
  - Detailed balance sufficient criteria
- Example Algorithm Metropolis-(Hastings):
  - Needs only a choice of (symmetric) proposal Distribution
  - And the ratio of the target density probability between current and proposed state
- Other (more advanced) algorithms exist: HMC, Nested Sampling, ...