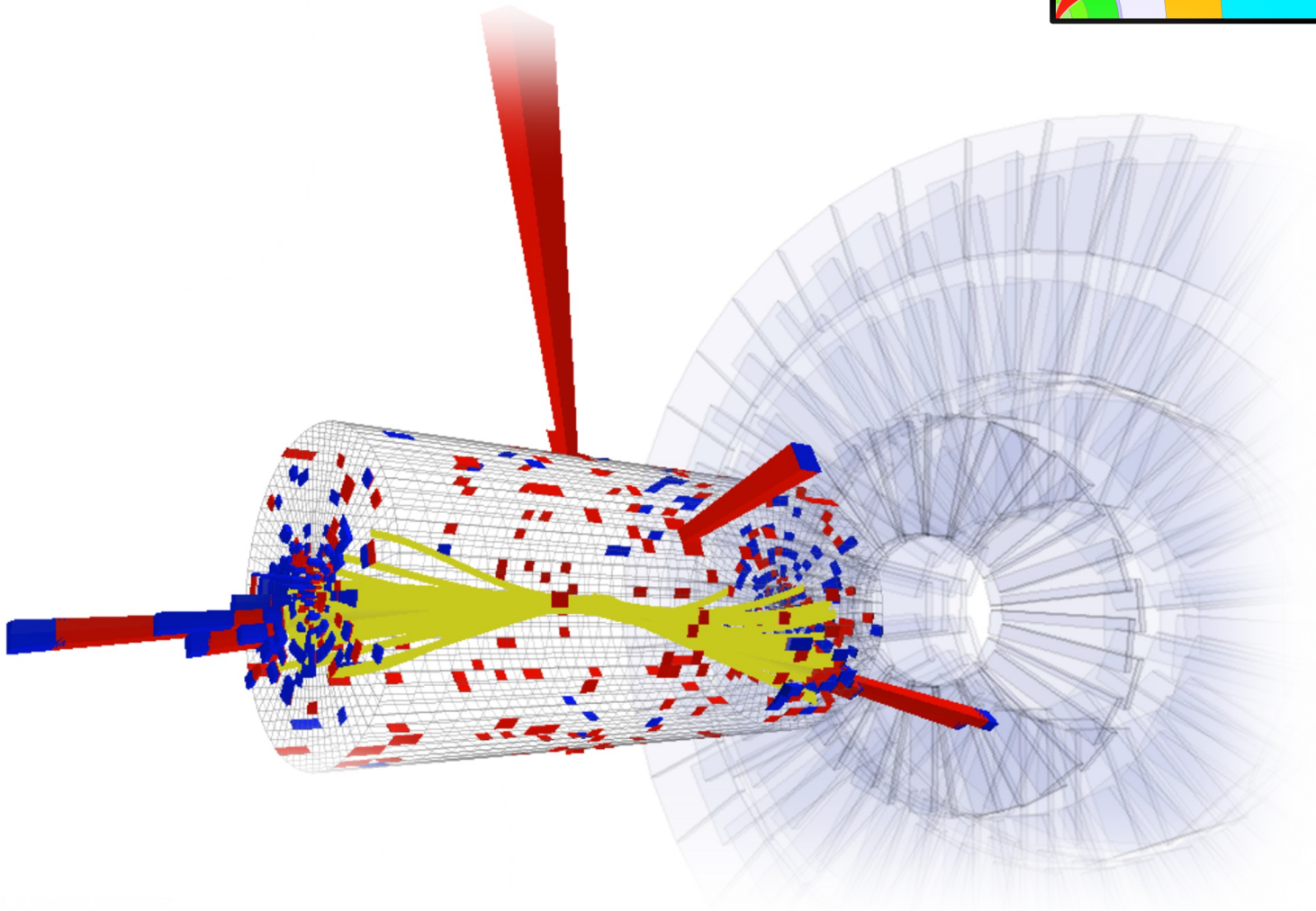
Imperial College London

Statistics for Collider Physics



Nicholas Wardle

Imperial College London

HCP Summer School

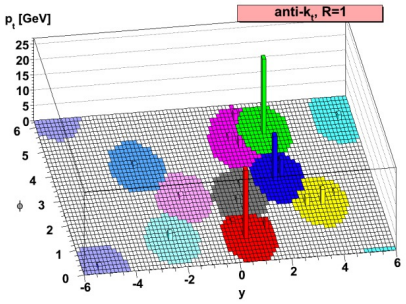
15th - 26th August 2022

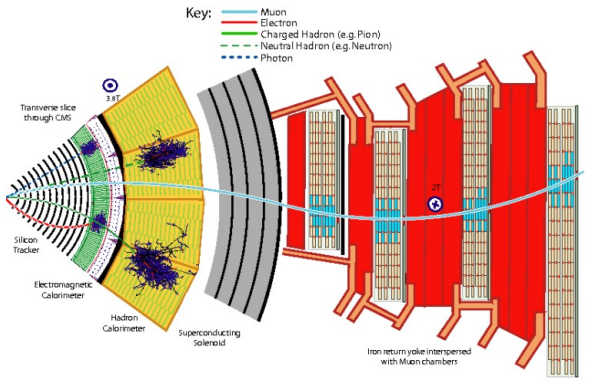
Statistics from a Particle physicist's view point

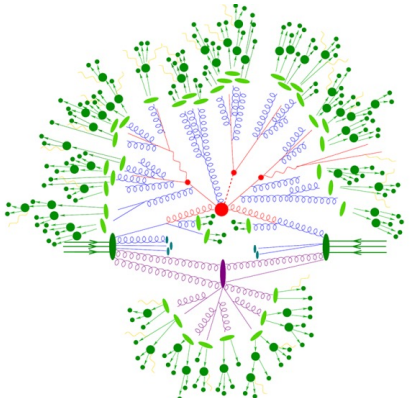
(Example from K. Cranmer)

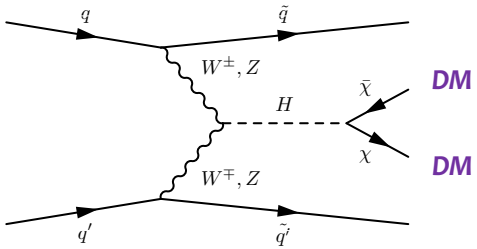
An example random number generator (the LHC)

Observables	Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
x	z_d	z_s	z_p	θ







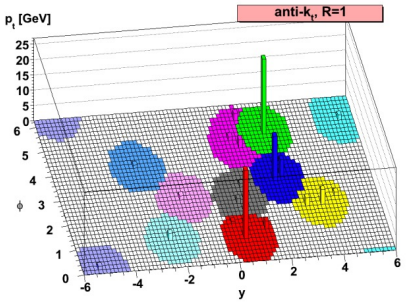


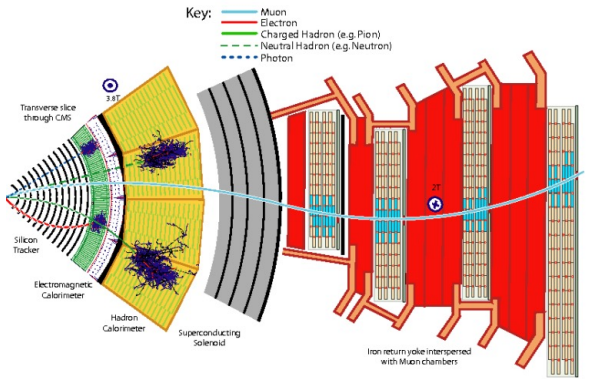
Statistics from a Particle physicist's view point

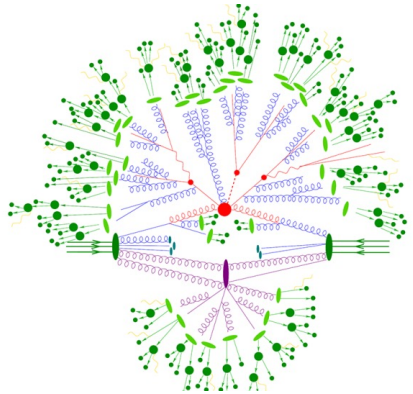
(Example from K. Cranmer)

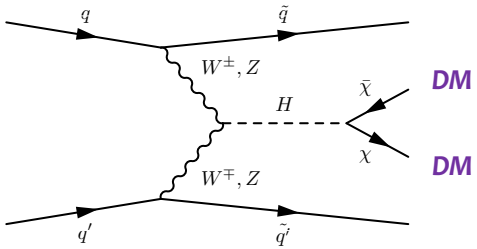
An example random number generator (the LHC)

Observables	Detector interactions	Shower splittings	Parton-level momenta	Theory parameters
x	z_d	z_s	z_p	θ









$$p(x|\theta) = \int \mathrm{d}z_d \int \mathrm{d}z_s \int \mathrm{d}z_p \; p(x|z_d) \qquad p(z_d|z_s) \qquad p(z_s|z_p) \qquad p(z_p|\theta)$$

Features — Latent Variables — Model Parameters

Outline of the course

Probability & Probability distributions

- Frequentist probability & Bayes' theorem
- Basic properties of probability
- Probability distributions
- Law of large numbers & the Central Limit Theorem

Estimators

- Likelihoods & Maximum likelihood estimators
- Properties of MLEs

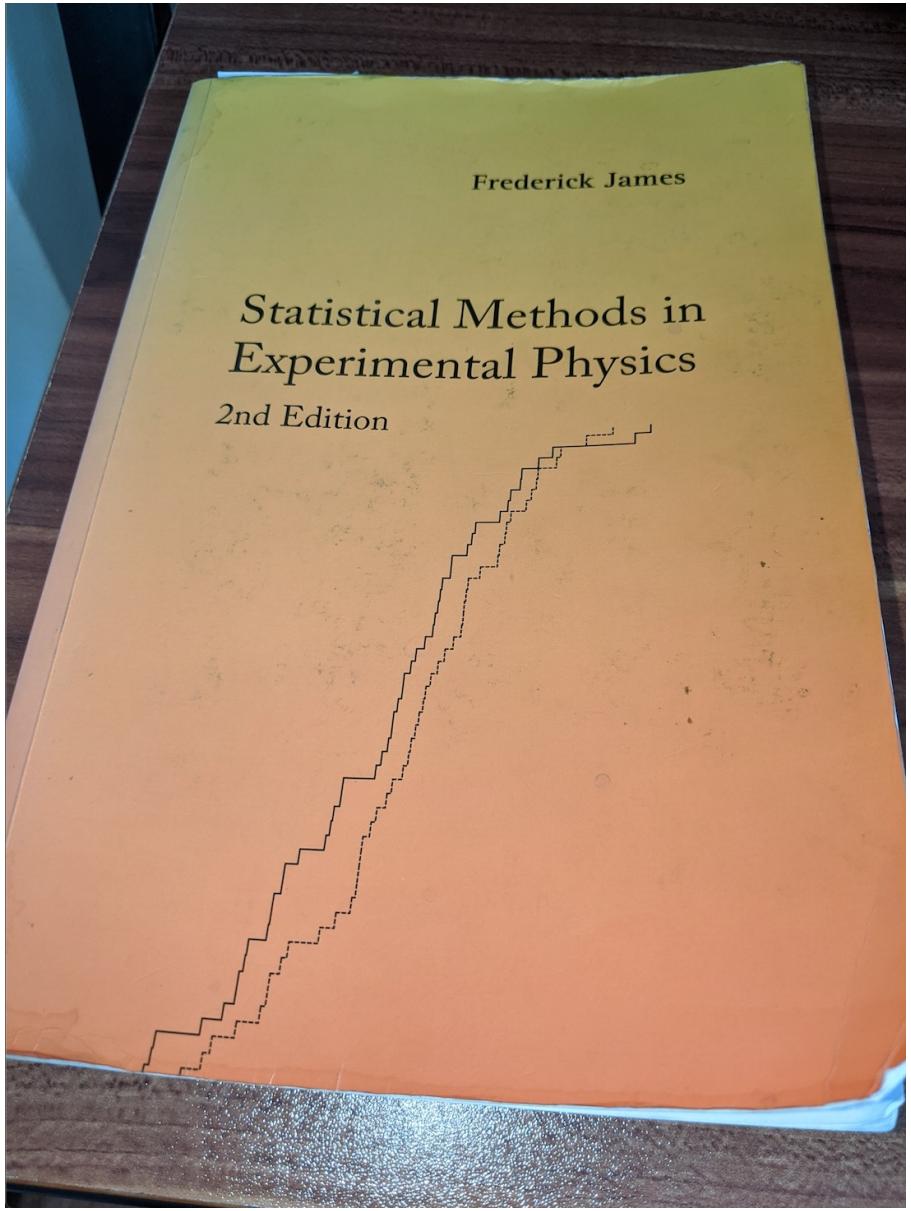
Hypothesis testing

- Type-I/II errors
- Simple and likelihood-based hypothesis tests & the Neyman-Pearson lemma
- P-values

Uncertainty & intervals

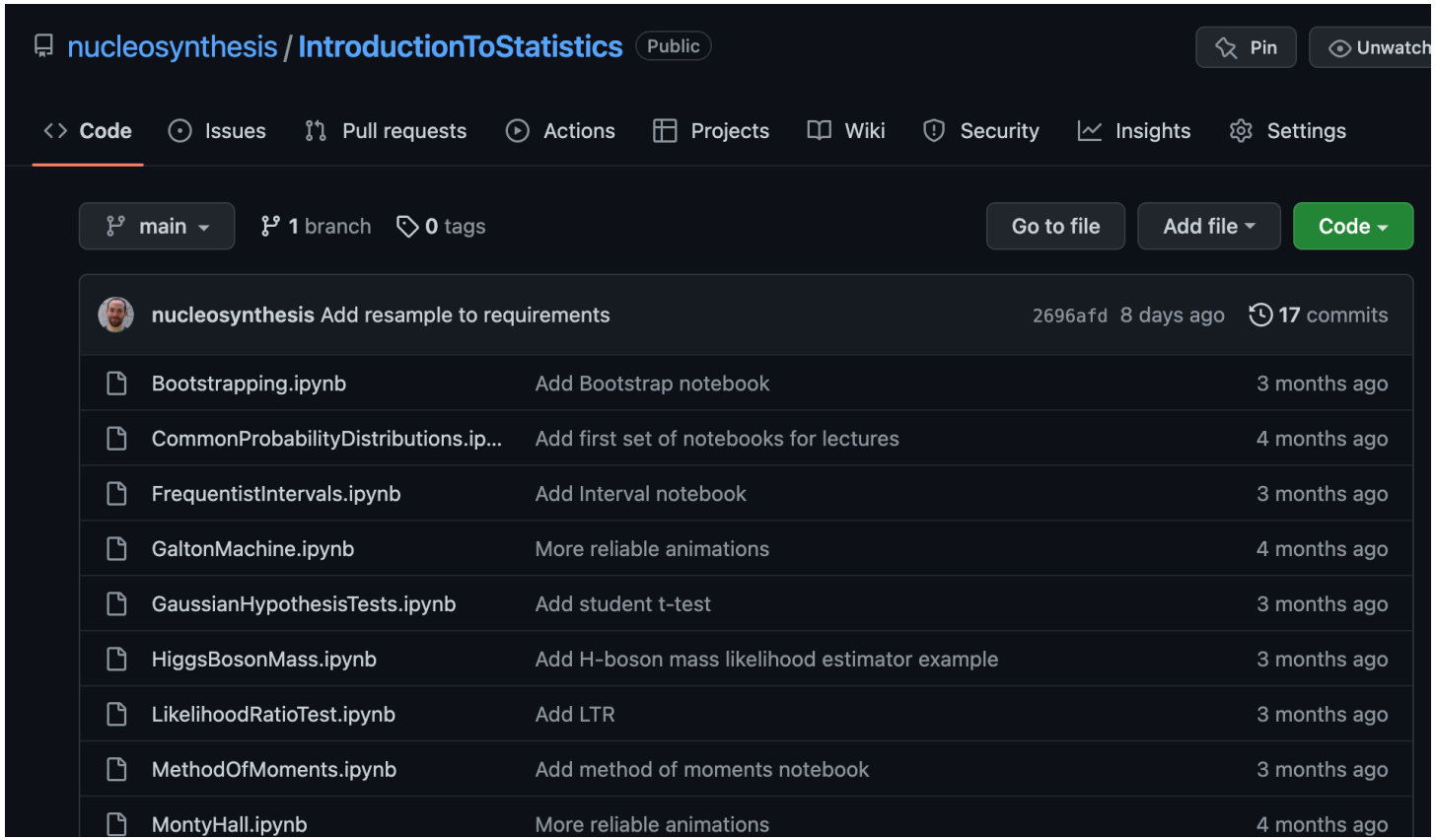
- Neyman intervals & Wilks' theorem
- Bayesian credible intervals
- Profiling and marginalizing nuisance parameters

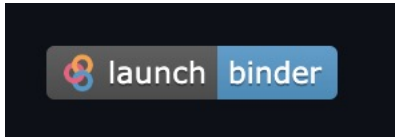
This book is great for learning statistics from a particle physics perspective



Practical examples

Throughout the lectures I plan to show a couple of practical examples / demonstrations and you can access various Jupyter notebooks in my GitHub area if you want to have a go yourselves: https://github.com/nucleosynthesis/IntroductionToStatistics

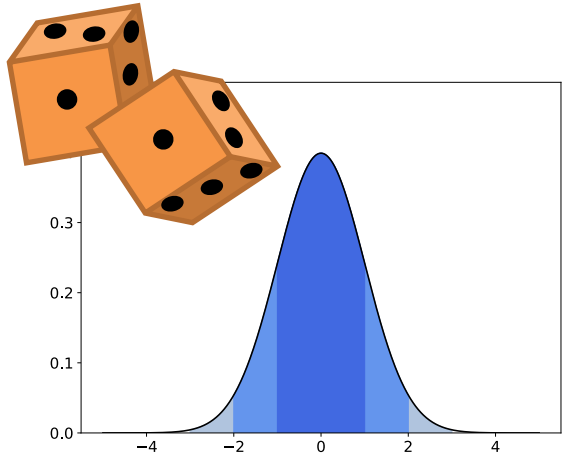




You can run the notebooks in binder if you prefer not to download them - everything uses rather basic python packages (nothing too fancy in there to install)! If you get errors with some notebooks, change "`lab`" to "`tree`" in the URL once binder has loaded

There are even a few pen-and-paper style questions there that you are more than welcome to try.

Probability & Probability Distributions



Probability

We all have an intuitive sense of what *probability* means. The mathematics developed in 1933 by Kolmogorov treats probability as something which satisfies three axioms, and therefore no specific definition is given. The Kolmogorov axioms are,

- $P(X_i) \geq 0$ for all i
- $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$ if X_i and X_j are ***exclusive***
- $\sum_\Omega P(X_i) = 1,$

Where $P(X_i)$ are **probabilities** that events/outcomes X_i occur and Ω is the set of all possible outcomes

→ E.G. when rolling an unbiased 6-sided die, $\Omega=\{1, 2, 3, 4, 5, 6\}$

The usual rules are stated:

- Sum rule : P(A or B) = P(A) + P(B)
- Product rule: P(A and B) = P(A) x P(B)

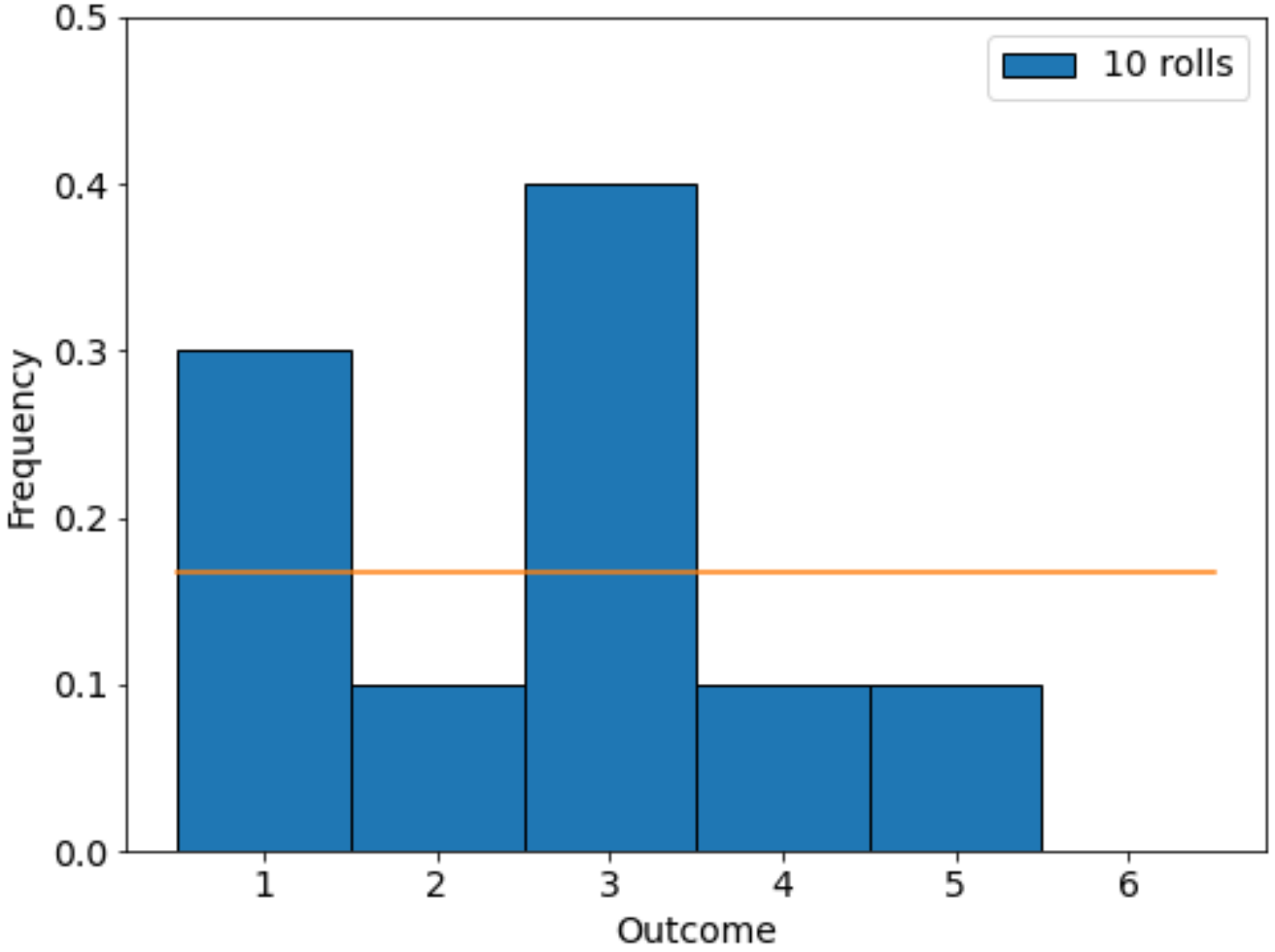
Probability interpretations - Frequentism

Frequentist (or classical) probability is the one that you are most likely familiar with.

As the name suggests, this definition of probability is related to the frequency with which an event (say the roll of a die) occurs in ***repeated trials***.

The number of times *n* that an event *X* occurs in *N* trials yields the probability as *N* gets large

$$P(X) = \lim_{N\to\infty} \frac{n}{N}.$$



Probability interpretations - Bayesian

If I asked you what you think the probability it will rain tomorrow is, answering that you'd need to repeat "tomorrow" to give me a frequency is not super helpful!

Bayesians avoid this by interpreting probability in terms of a ***degree of belief*** something will occur.

Probability interpretations - Bayesian

If I asked you what you think the probability it will rain tomorrow is, answering that you'd need to repeat "tomorrow" to give me a frequency is not super helpful!

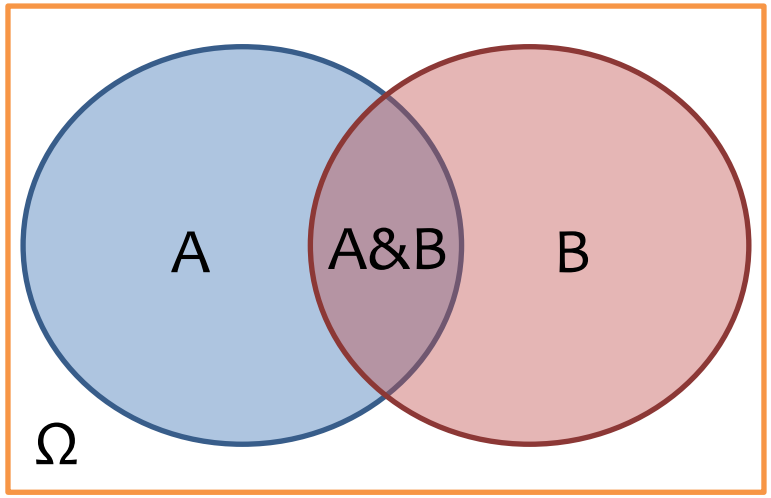
Bayesians avoid this by interpreting probability in terms of a ***degree of belief*** something will occur.

In these lectures, I'll avoid discussing Bayes' vs Frequentist interpretations but there is a useful tool that we need – **Bayes' rule!**

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$

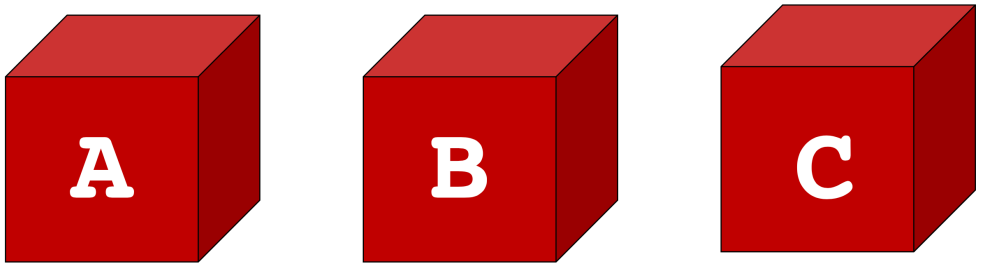
Or re-arranging

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



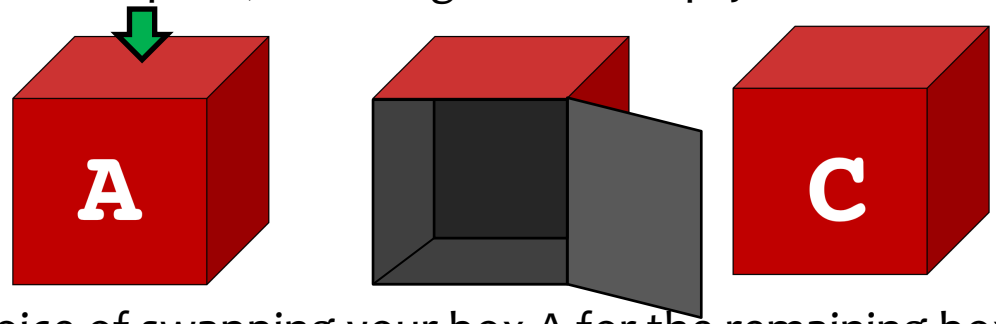
Bayes' rule in action

You are told that one of three boxes (A, B or C) contains a prize while the other two are empty.



Bayes' rule in action

You are told that one of three boxes (A, B or C) contains a prize while the other two are empty. You pick box A but before opening it, the game show host opens one of the remaining two boxes that she knows doesn't contain the prize, revealing it to be empty.



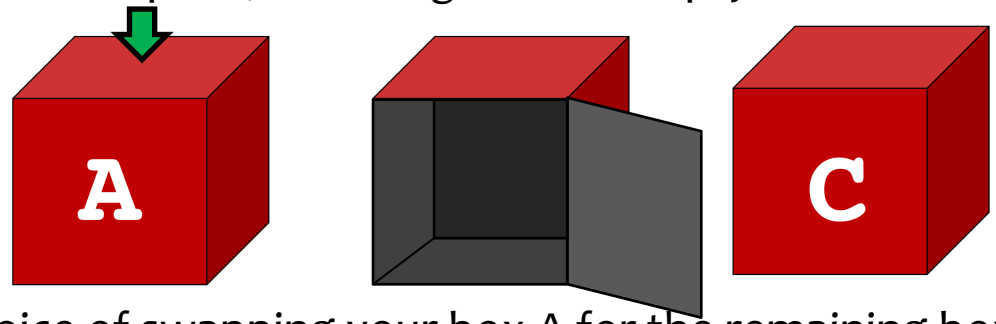
Given the choice of swapping your box A for the remaining box, do you stick with your original choice or switch?

Let P(A), P(B) and P(C) represent the prior probabilities that the prize is inside box A, B and C, respectively. Suppose the host chooses box B, then $P(A|H_B)$ is what we want

$$P(A|H_B) = \frac{P(H_B|A)P(A)}{P(H_B)}$$

Bayes' rule in action

You are told that one of three boxes (A, B or C) contains a prize while the other two are empty. You pick box A but before opening it, the game show host opens one of the remaining two boxes that she knows doesn't contain the prize, revealing it to be empty.



Given the choice of swapping your box A for the remaining box, do you stick with your original choice or switch?

Let P(A), P(B) and P(C) represent the prior probabilities that the prize is inside box A, B and C, respectively. Suppose the host chooses box B, then $P(A|H_B)$ is what we want

$$\mathbf{P(A|H_B) = 1/3}$$

$$P(A|H_B) = \frac{P(H_B|A)P(A)}{P(H_B)}$$

P(A) = 1/3 (flat prior)

$P(H_B|A)$ = P(host chooses box B if prize is in A (yours)) = 1/2

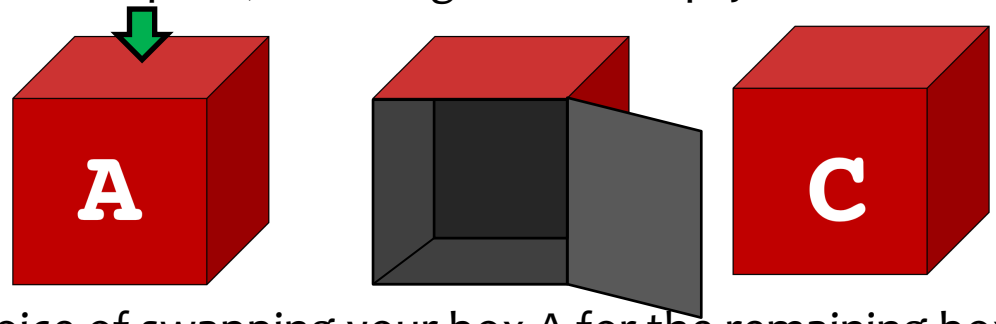
$P(H_B) = P(H_B|A)P(A) + P(H_B|B)P(B) + P(H_B|C)P(C) = 1/2$

1/2 x 1/3 0 1

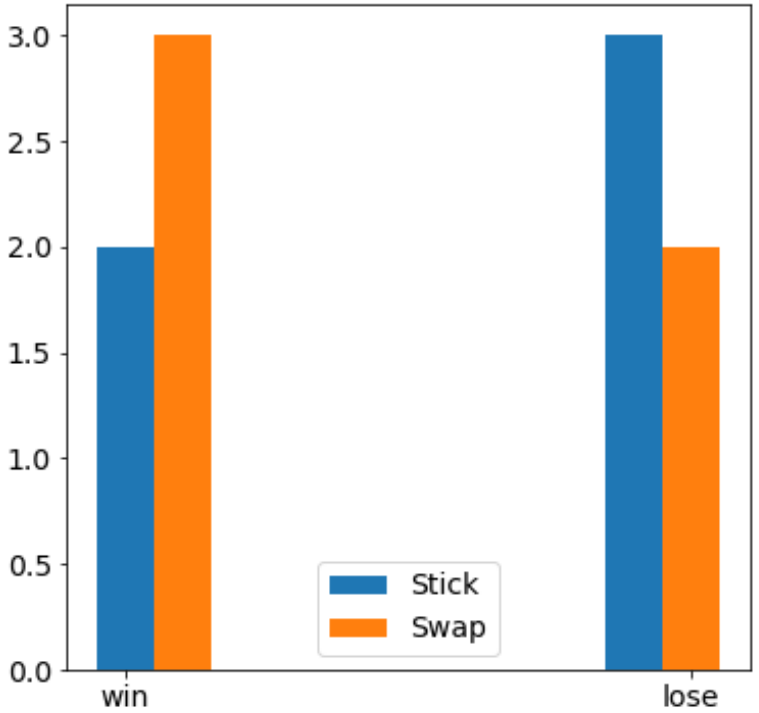
Bayes' rule in action

See MonteHall.ipynb

You are told that one of three boxes (A, B or C) contains a prize while the other two are empty. You pick box A but before opening it, the game show host opens one of the remaining two boxes that she knows doesn't contain the prize, revealing it to be empty.



Given the choice of swapping your box A for the remaining box, do you stick with your original choice or switch?



$P(A|H_B) = 1/3$

Let P(A), P(B) and P(C) represent the prior probabilities that the prize is inside box A, B and C, respectively. Suppose the host chooses box B, then $P(A|H_B)$ is what we want

$$P(A|H_B) = \frac{P(H_B|A)P(A)}{P(H_B)}$$

P(A) = 1/3 (flat prior)

$P(H_B|A)$ = P(host chooses box B if prize is in A (yours)) = 1/2

$P(H_B) = P(H_B|A)P(A) + P(H_B|B)P(B) + P(H_B|C)P(C) = 1/2$

1/2 x 1/3 0 1

Probability distributions

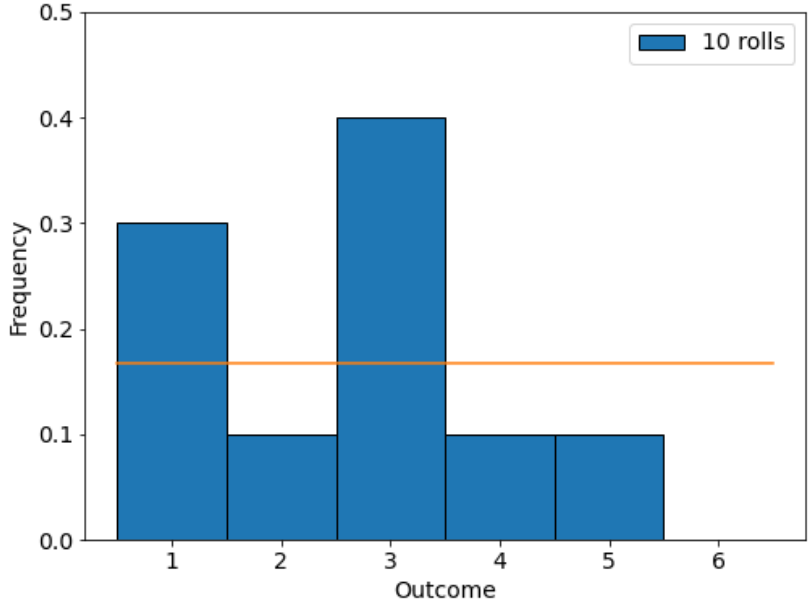
For a sequence of events X1, X2, X3..., we often use the vector notation **X**. Note however that **X** can also refer to a repeated set of experiments and each experiment will have a single observation – this is typically our "data"

The corresponding probabilities $P(X)_\Omega$ over all possible values of $X \in \Omega$ form a probability distribution.

As an example, if X is the outcome of a single die roll, then $\Omega = \{X = 1, X = 2, X = 3, X = 4, X = 5\}$ and $P(X) = 1/6$ for every possible value of X, meaning the **probability distribution is uniform**.

A single event (roll) can be then thought of a random draw from such a probability distribution, and successive rolls of the die will yield a distribution of values whose frequency converges to a uniform distribution (U).

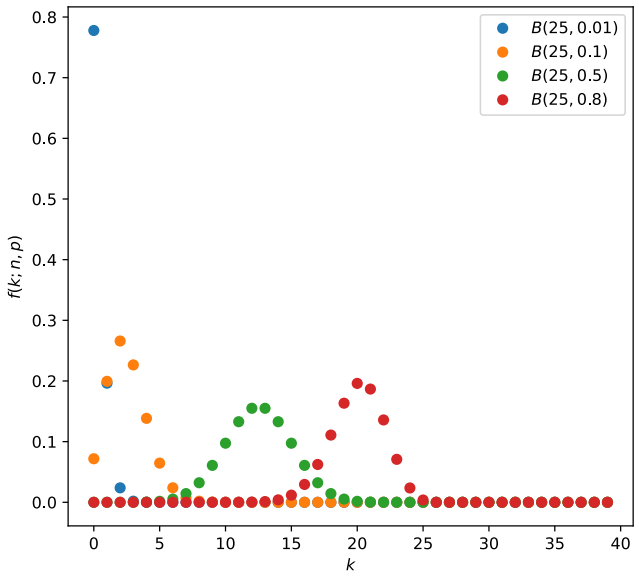
We will write this as $X \sim U(1, 6)$



Probability distributions – the binomial distribution

The binomial distribution* describes the distribution of the number of successes (k) in a sequence of n independent trials, where the probability of success in any trial is p.

More generally, any sequence of experiments, each of which results in a yes/no, 1/0 or other binary result with probabilities p and q = (1 – p) assigned to each outcome will be described by the binomial distribution,



Observed number of successes k

$$f(k; n, p) = \binom{n}{k} p^k q^{n-k}$$

Distribution f

Parameters of "model" (n, p)

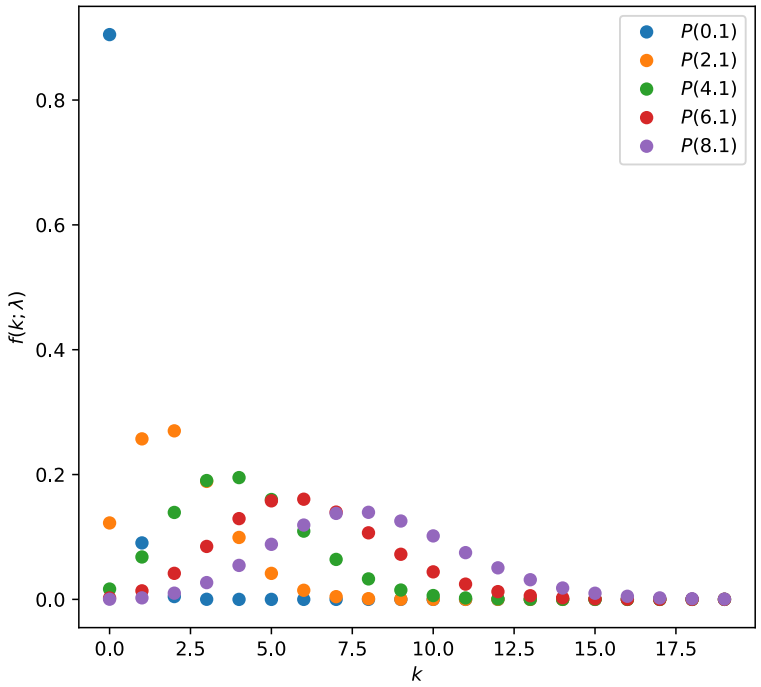
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

*If you're interested, take a look at [CommonProbabilityDistributions.ipynb]

Probability distributions – the Poisson distribution

The Poisson distribution is a limit case of the binomial distribution where n → ∞, and p → 0 such that the product **λ = np** is constant.

In particle physics, this is a **very common** distribution since it models processes that are extremely rare (p → 0) and data sets that are very large n → ∞.



$$f(k;n,p) \to f(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$$

Probability distributions – the Gaussian distribution

We can generalize probabilities (and distributions of probability) for events to probability distributions of *continuous random variables*. A continuous random variable is one whose possible values cover continuous intervals, for example X ∈ R. In this case, we need to define something called a *probability density function*

E.G : the energy E of cosmic ray particles $f(E) = \lim_{\delta E \to 0} \frac{P(E \in (E + \delta E))}{\delta E}$, $\int_{\Omega} f(E) dE = 1,$ where $E \in \Omega$.

Probability distributions – the Gaussian distribution

We can generalize probabilities (and distributions of probability) for events to probability distributions of *continuous random variables*. A continuous random variable is one whose possible values cover continuous intervals, for example X ∈ R. In this case, we need to define something called a *probability density function*

E.G : the energy E of cosmic ray particles $f(E) = \lim_{\delta E \to 0} \frac{P(E \in (E + \delta E))}{\delta E}$, $\int_{\Omega} f(E) dE = 1,$ where $E \in \Omega$.

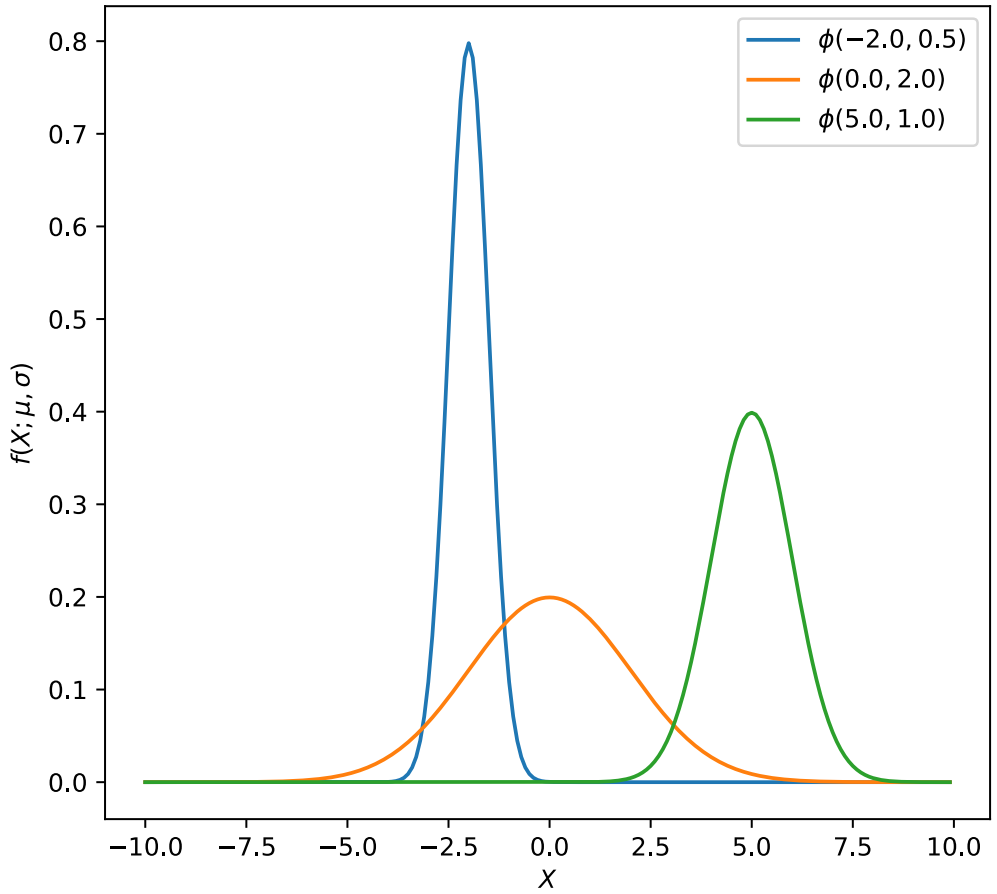
Probably (no pun intended), the Gaussian (or Normal) distribution is the most common probability distribution used in statistics. There is a good reason for this but for now, just to refresh our memories about the Gaussian distribution.

If a continuous random variable X is distributed as a Normal or Gaussian distribution, then

$$f(X; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Note: We can show that the (normalised) Poisson distribution approximates a Gaussian distribution as $\mu \to \infty$.

Probability distributions – the Gaussian distribution



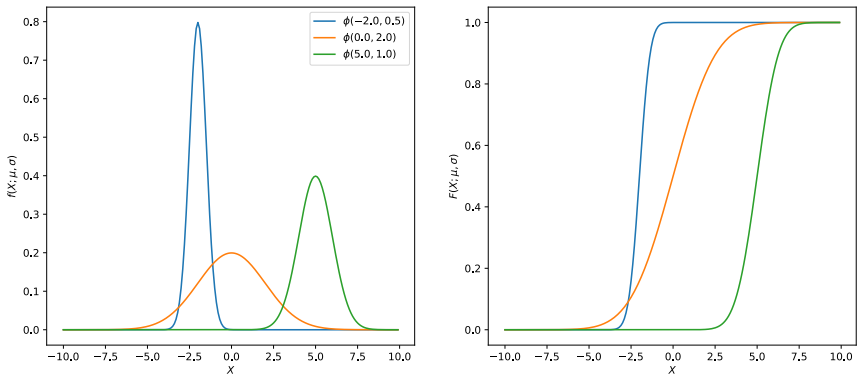
$$f(X;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Related distributions

We define distributions *related* to probability distributions, which we'll see later in the lectures,

Cumulative distribution

$$F(X) = \int_{X_a}^{X} f(X')dX'$$



Joint & conditional distributions

$$f(X, Y) = p(X|Y)h(Y) = q(Y|X)g(X)$$

Marginal distributions

$$g(X) = \int_{Y_a}^{Y_b} f(X, Y)dY$$

Probability distributions – the Multinormal distribution

The *multivariate* Gaussian probability density can be written as,

$$f(\boldsymbol{X};\boldsymbol{\mu},\mathbf{V}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{V})}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T(\mathrm{V}^{-1})(\boldsymbol{X}-\boldsymbol{\mu})}$$

Where **X** = (X_1,X_2,X_3,...,X_N) is a vector of random variables (not to be confused with a sequence of random outcomes of X), and **V** is a N × N symmetric matrix of co-variances V_{ij} = covariance(X_i,X_j).

In the special case where **N = 2**, we can write

Probability distributions – the Multinormal distribution

The *multivariate* Gaussian probability density can be written as,

$$f(\boldsymbol{X};\boldsymbol{\mu},\mathbf{V}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{V})}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T(\mathrm{V}^{-1})(\boldsymbol{X}-\boldsymbol{\mu})}$$

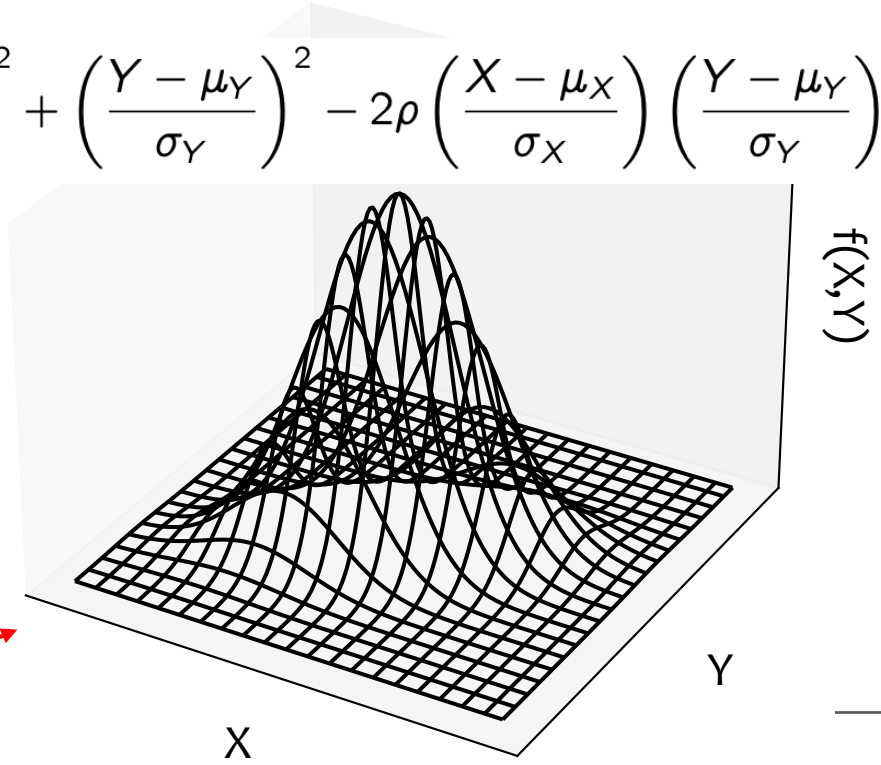
Where **X** = (X_1,X_2,X_3,...,X_N) is a vector of random variables (not to be confused with a sequence of random outcomes of X), and **V** is a N × N symmetric matrix of co-variances V_{ij} = covariance(X_i,X_j).

In the special case where **N = 2**, we can write

$$f(X,Y;\mu_X,\mu_Y,\sigma_X,\sigma_Y,\rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + \left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]\right)$$

For N=2, we can define the distribution with **5 real values**

$\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$



Correlation and independence

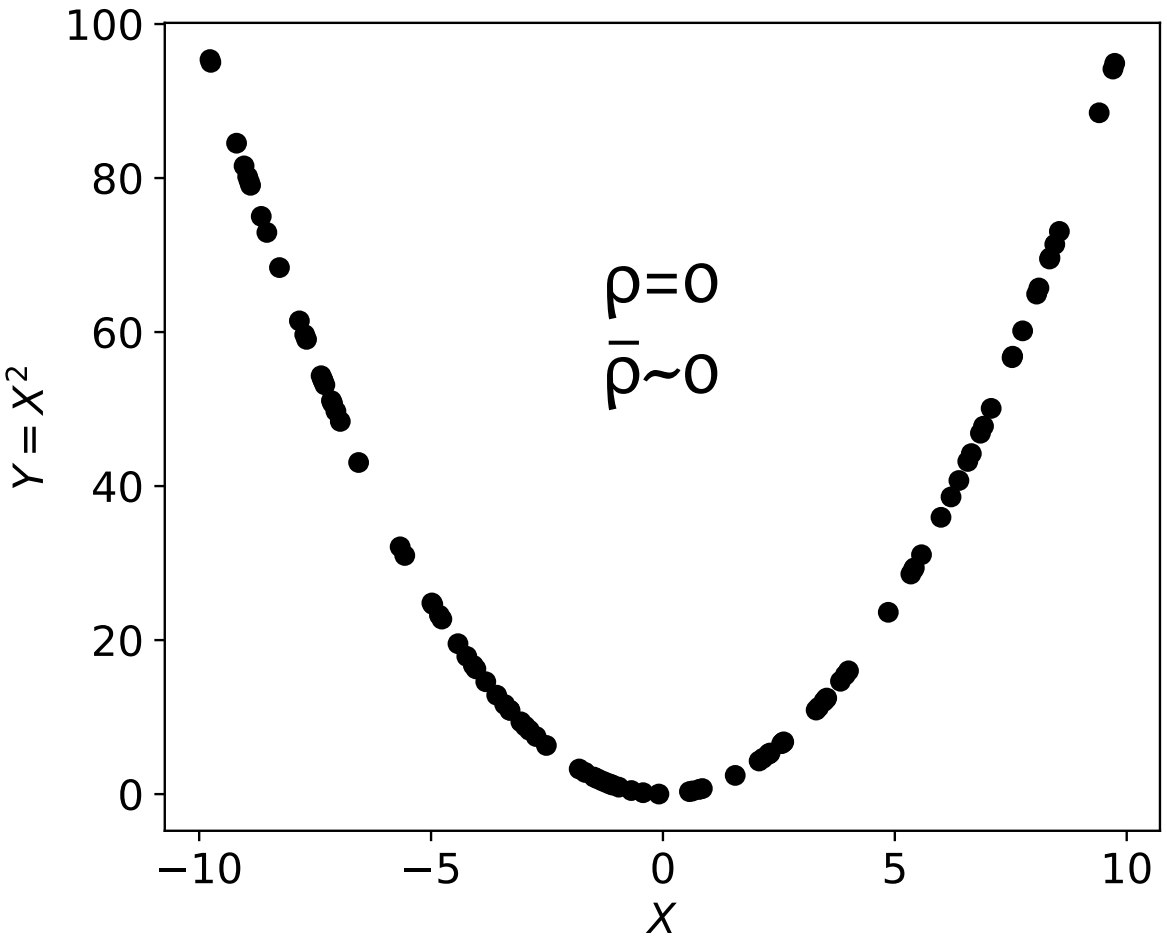
Two random variables X and Y are ***independent*** if
P(X,Y) = P(X)P(Y)

Warning! Often the covariance (or correlation) of a sample or joint distribution being zero is used to justify that two random variables are independent. **This is not sufficient!**

Independence ➔ Zero correlation coeff

Zero correlation coeff ~~➔~~ Independence

There are several better tests for independence, which we may come across later (if there's time)



Moments of probability distributions

Probability density distributions can be used to obtain useful information about a particular random variable, or functions of random variables. For example, the mean value of a random variable X (or its *expectation value* under f (X)) is given by,

$$E[X] = \int_{\Omega} X f(X) dX$$

Similarly, any function of X, g(X) has an expectation value E[g] under f (X) of $E[g] = \int_{\Omega} g(X) f(X) dX$

The expectation is a linear operator $\longrightarrow$ $E[a \cdot g(X) + b \cdot h(Y)] = a \cdot E[g(X)] + b \cdot E[h(Y)]$

Moments of probability distributions

Probability density distributions can be used to obtain useful information about a particular random variable, or functions of random variables. For example, the mean value of a random variable X (or its *expectation value* under f (X)) is given by,

$$E[X] = \int_{\Omega} X f(X) dX$$

Similarly, any function of X, g(X) has an expectation value E[g] under f (X) of $E[g] = \int_{\Omega} g(X) f(X) dX$

The expectation is a linear operator $\longrightarrow$ $E[a \cdot g(X) + b \cdot h(Y)] = a \cdot E[g(X)] + b \cdot E[h(Y)]$

In general, there are higher order moments that characterize the "*shape*" of the distribution …

$$\boxed{\begin{array}{l} \mu_n = E[X^n] \text{ is the } n^{\text{th}} \text{ algebraic moment} \\ \nu_n = E[(X - E[X])^n] \text{ is the } n^{\text{th}} \text{ central moment} \end{array}}$$

E.G: ν_2 is the **variance**

We've already come across the ***co*-variance** of two random variables

$$\text{covariance}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$$

$$\text{correlation}(X_i, X_j) = \rho(X_i, X_j) = \frac{E[(X_i - E[X_i])(X_j - E[X_j])]}{\sqrt{v_2^i v_2^j}}$$

Sample moments

Note that moments of probability distributions are **not the same** as sample moments. For a finite sample of a random variable, it is always possible to determine sample moments (unlike in the case of some probability densities)

→ they are a property of the specific data set.

For a sequence of a random variable (X_1, X_2, ..., X_N) of size N, we can define the *n*–th sample moment as, $$m_n = \frac{1}{n}\sum_i^N X_i^n.$$

You'll be familiar with the 1st such moment, which is the sample mean,

$$m_1 = \bar{X} = \frac{1}{n}\sum_i^N X_i,$$

and the second central moment, which is the sample variance,

$$\bar{V} = \frac{1}{n}\sum_i^N (X_i - \bar{X})^2.$$

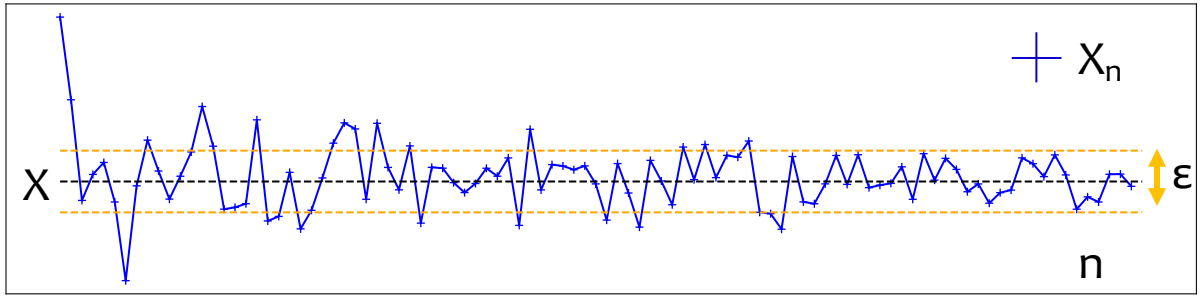
We can also define the sample covariance of two random variables X and Y

$$\text{covariance}(X, Y) = \frac{1}{n}\sum_i^N (X_i - \bar{X})(Y_i - \bar{Y})$$

Probability & Convergence

There are several definitions of convergence (which I won't go into) but two definitions are used very often in HEP …

Convergence in Probability:

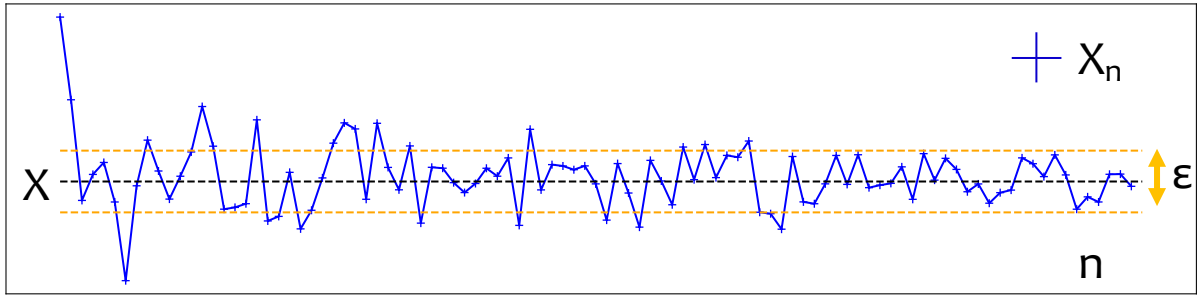


The probability that an element of the sequence X_n is far (ε) from X gets vanishingly small

Probability & Convergence

There are several definitions of convergence (which I won't go into) but two definitions are used very often in HEP …

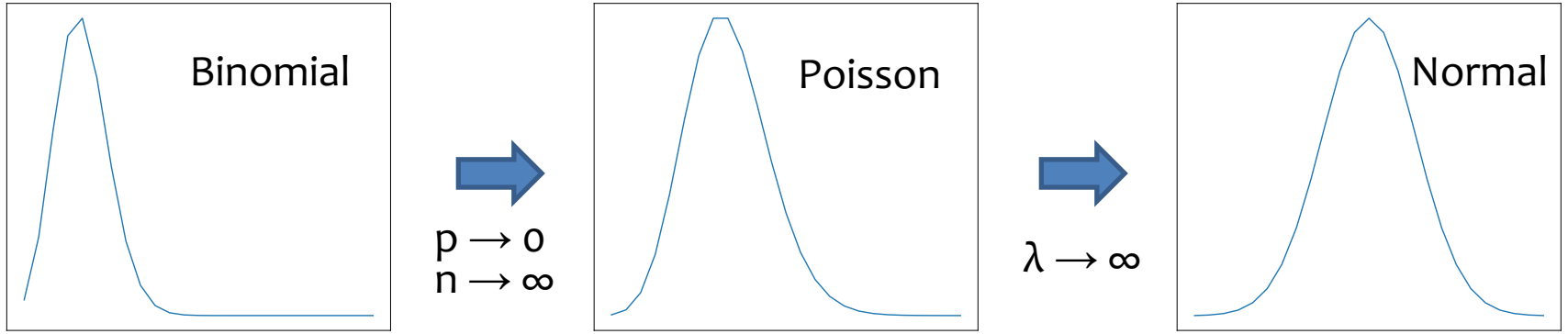
Convergence in Probability:



The probability that an element of the sequence X_n is far (ε) from X gets vanishingly small

Convergence in Distribution:

A probability distribution gets better approximated by a limiting distribution



The law of large numbers

T^hm: The law of large numbers states that the sample mean of a sequence of independent random variables whose distributions have the same mean (μ) converges in probability to μ (provided their variance is well behaved)

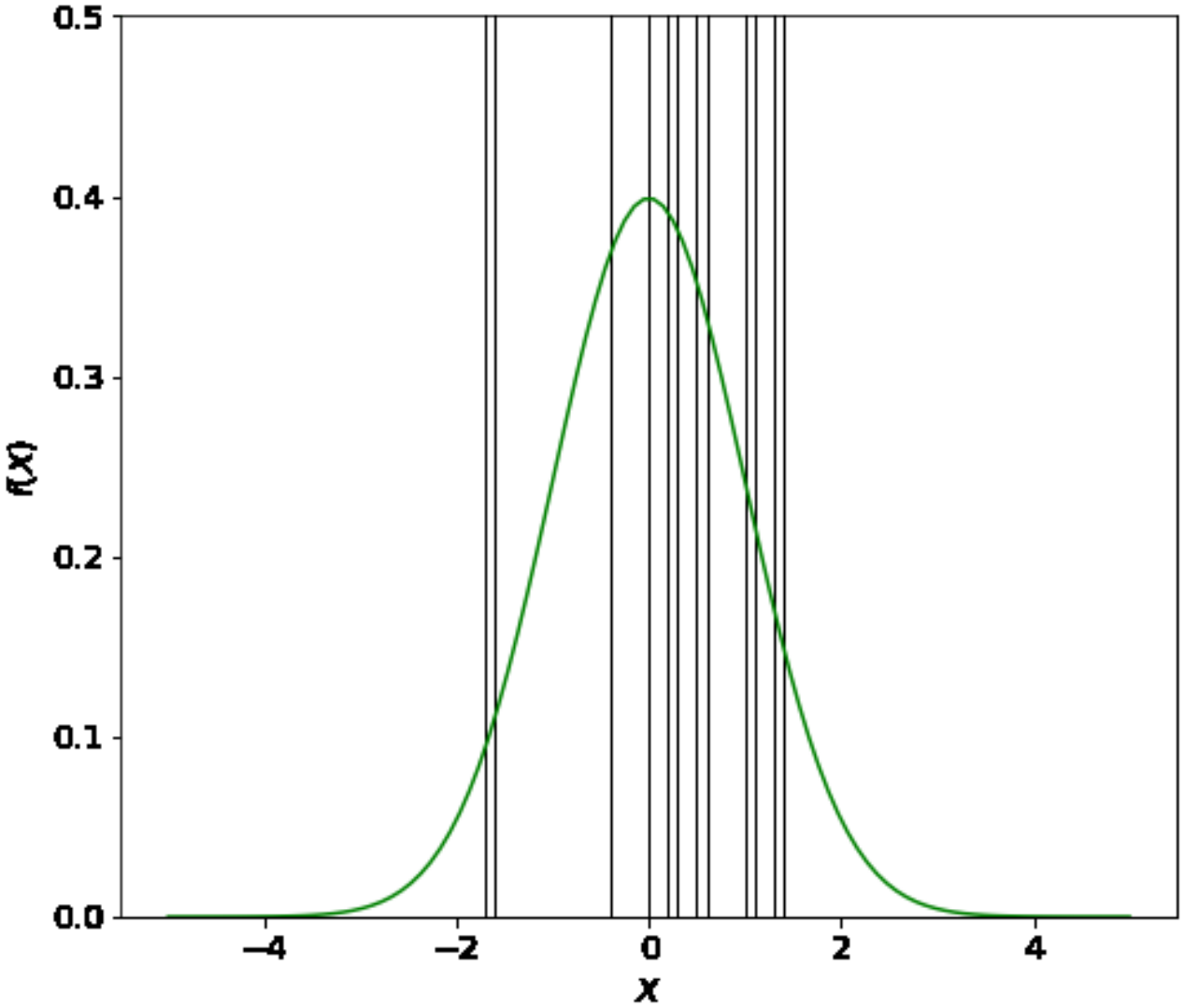
$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

This seems like its rather obvious, but note how the statement is about when a **sample mean** converges to the **mean of a distribution** (remember they are *not the same thing*)

The law of large numbers

A **histogram** is essentially an estimate of a probability density

The density in each bin is better estimated as the **number of observations in each bin** gets large!



The Central Limit Theorem

Perhaps one of the most famous results in statistics (and probably one of the most useful) is the **Central Limit Theorem.**

The central limit theorem (CLT) states how the sum of independent random variables (for example *sample means*) is distributed in the limit of large N.

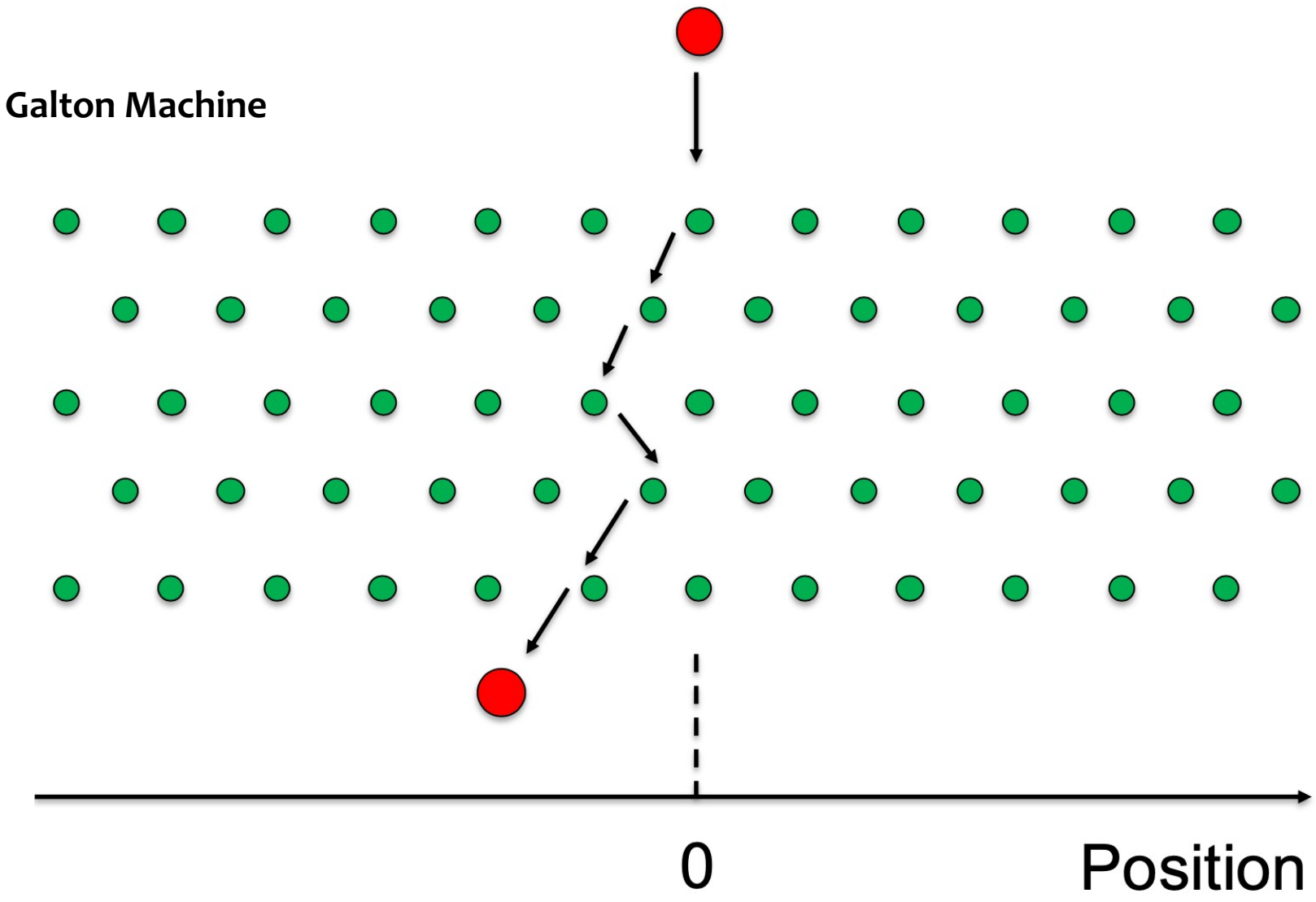
T^hm: Suppose we have a sequence of independent random variables X_i, each from a distribution with mean μ^i_1 and variance ν^i_2. Let,

$$T_N = \frac{\bar{X}N - \sum_{i=1}^{N} \mu_1^i}{\sqrt{\sum_{i=1}^{N} \nu_2^i}}.$$

Then for $T = \lim_{N\to\infty} T_N$, we have that $T \sim \varphi(T;0,1) \rightarrow T_N$ converges in distribution to a standard normal distribution

The Central Limit Theorem in action

GaltonMachine.ipynb



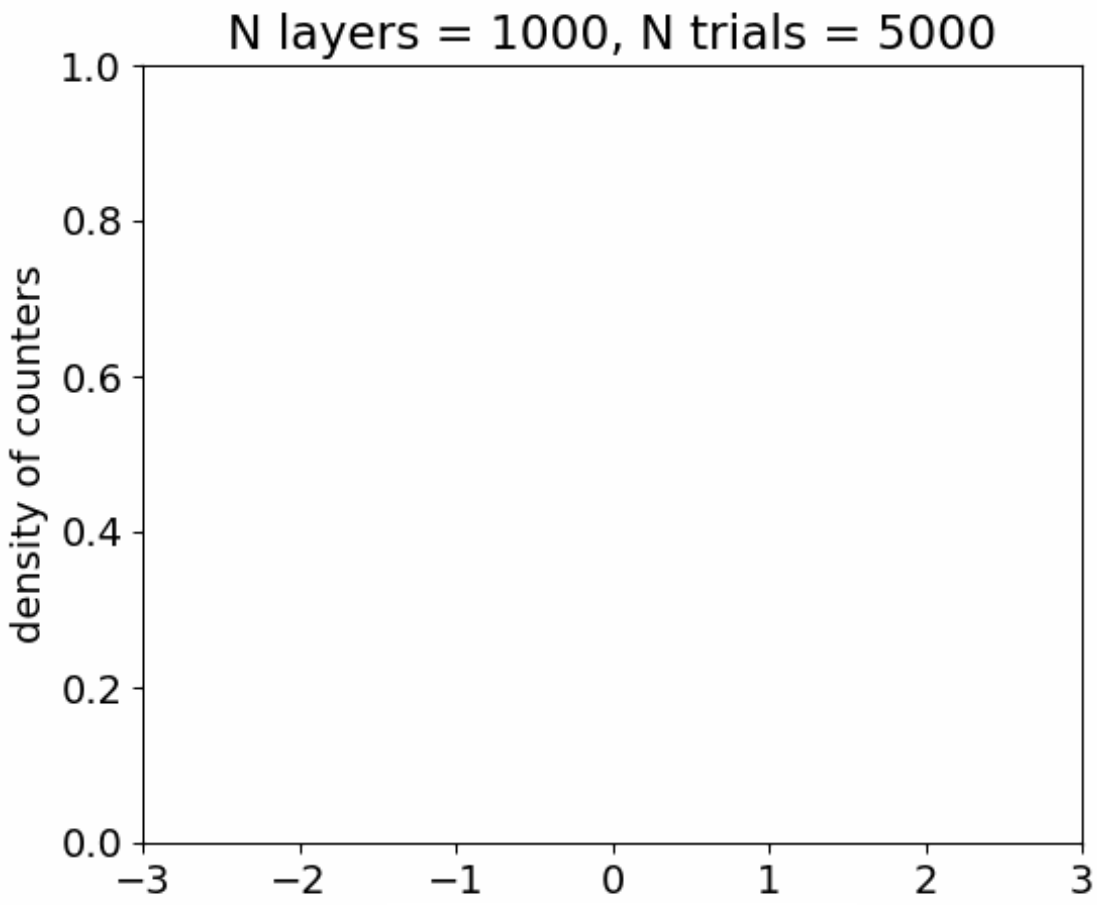
The Central Limit Theorem in action

If you think about the distribution of "position", its simply a sum over X_i where $X_i=-1$ or $X_i=+1$ with 50:50 probability. The distribution of each random variable X_i is identical - a uniform distribution in the set $\{-1,1\}$

The mean of each distribution of X_i is 0, and the variance is $(-1)^2\times1/2 +(1)^2\times1/2=1$, thus

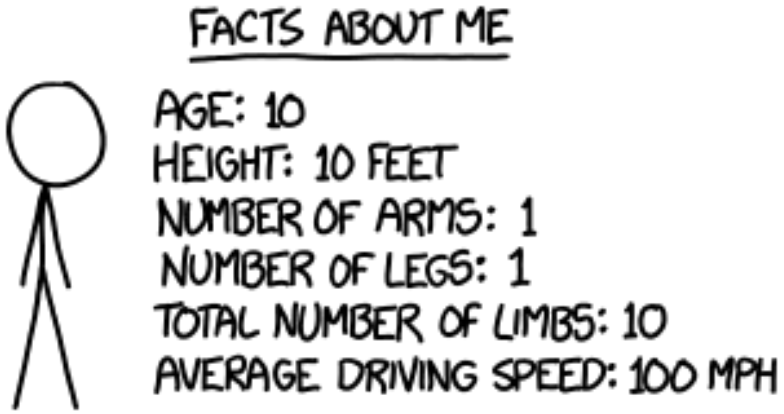
$$T_N = \frac{\bar{X}N - \sum_i \mu_1^i}{\sqrt{\sum_i v_2^i}} = \frac{\frac{N}{N}\sum_i X_i - 0}{\sqrt{N}} = P_N$$

So as $N\rightarrow+\infty$ we know from the CLT that P_N will be distributed as $\phi(0,1)$



$$P_N = \frac{1}{\sqrt{N}}\sum_i X_i.$$

Estimators



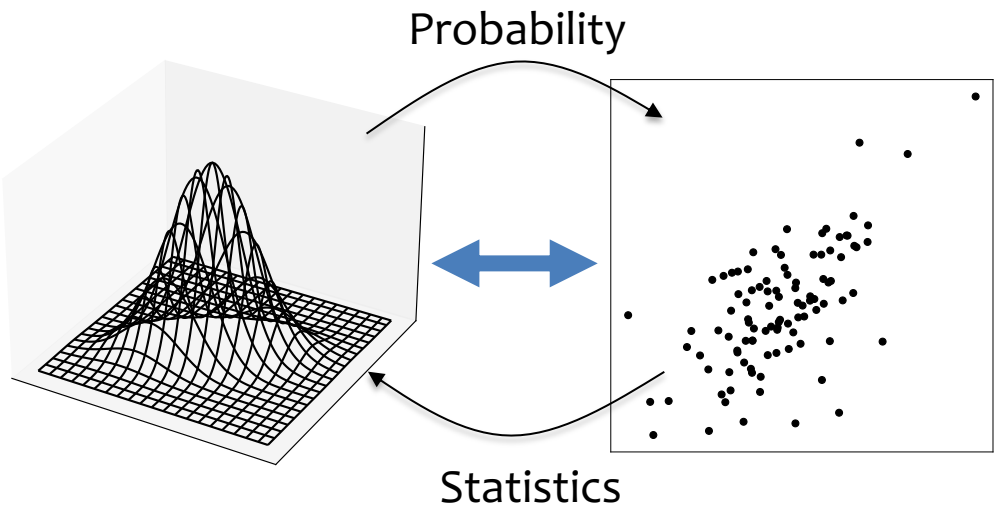
xkcd.com

Estimators

So far, we've only discussed probability densities for random variables, their properties and how these properties evolve with more and more data.

For fixed (known) values of the parameters that define those probability distribution, we can generate random variables however, statistics is really concerned with ***extracting information from the data*** – using the random observations to learn something about the model (or the parameters of the probability distributions) we're interested in!

The rest of the course will focus on this aspect and we'll start by discussing the concept (and an example) of ***estimators.***



Estimators

An estimator $\hat{\theta}$ for a quantity θ is a function of the observed data.

- θ can be a parameter of a model, a quantile or moment of a probability density function or functions of the above
- $\hat{\theta}$ is a property of a specific set of observations → its value depends on the data observed and it is itself a random variable
- The concept of a "good" estimator is not unique! We will discuss some common properties of estimators that qualify them "good" estimators but in the end it depends on which property is the most important in a given problem

There are many different estimators (too many to cover here) so I will only talk about one example estimators that is very common.

The ***maximum likelihood estimator*** is probably the most common form of parameter estimation in HEP*. We often refer to these as the "best-fit" or "fitted" parameters (which will become clear later).

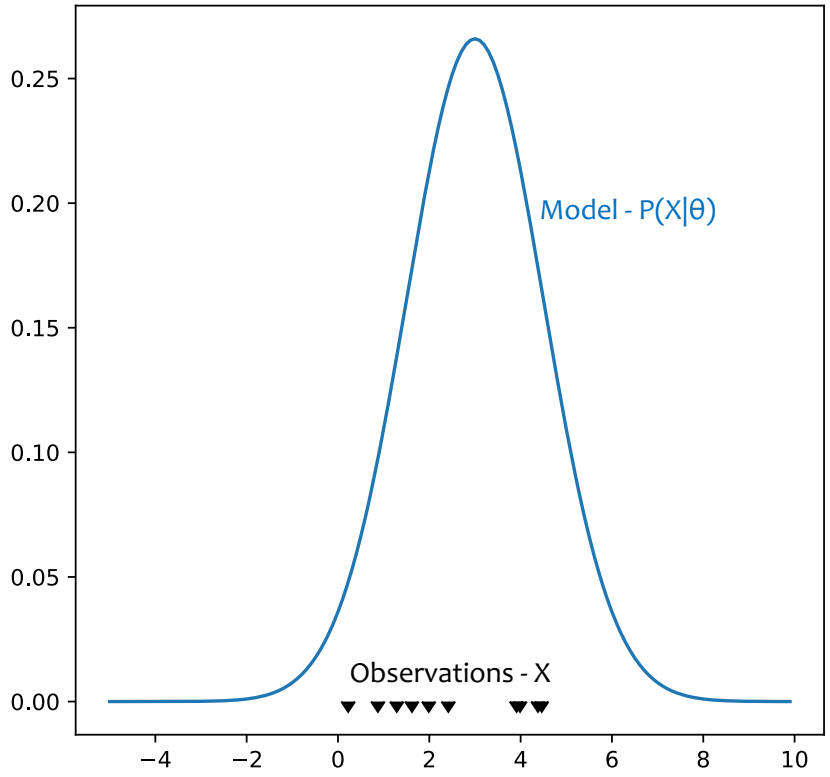
*it's not the only one, for another example see `MethodOfMoments.ipynb`

Likelihoods

The likelihood function L is defined by,

$$L(\theta) \propto P(X|\theta).$$

Where X ~ P(X|θ). Here, X is the data (or observations) we are analysing and θ are the parameter(s) we are trying to estimate.



Note that we write a proportionality rather than an equality because ***the likelihood function is not a probability density***. In practice, we often only care about ratios of likelihoods for different values of the parameter(s) and hence we almost never care about this constant of proportionality.

Likelihoods

For N ***independent observations*** $X = \{X_1, X_2, \ldots, X_N\}$, the likelihood function is,

$$L(\theta) := \prod_{i=1}^{N} f_i(X_i; \theta),$$

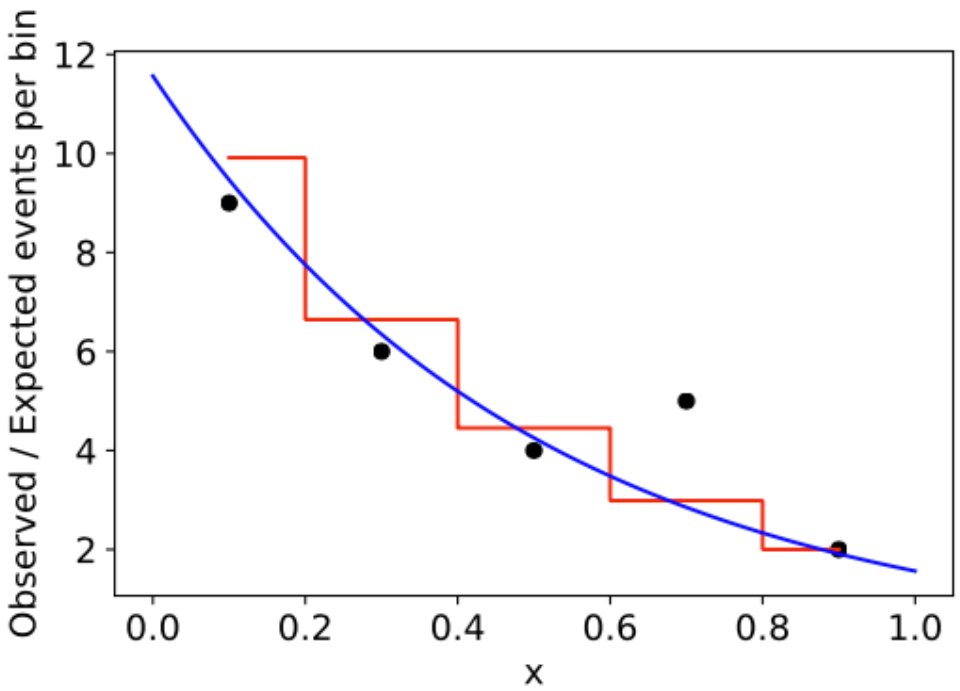
where f_i is the p.d.f for the random variable X_i.

We often make the distinction in HEP between **binned likelihoods** and **un-binned likelihoods**.

A binned likelihood is typically a product of Poisson likelihoods since the count in each bin is distributed as a Poisson random variable.

If you have some probability density function f, then the Poisson parameter λ for the bin defined as the range $[x_a, x_b]$ is defined as,

$$\lambda = \int_{x_a}^{x_b} f(x)dx$$



Maximum Likelihood Estimators

Generally we tend to take negative logs of the likelihood function and define

$$q = -\log(L)$$

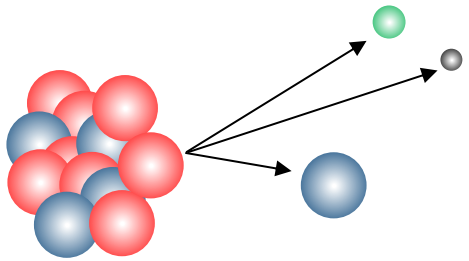
Maximum likelihood estimators (MLE) are given by the values of θ for which

$$\frac{\partial q}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0$$

Which can be solved numerically or analytically depending on the problem.

Note that **maximizing the likelihood function** (L) is equivalent to **minimizing the negative log of the likelihood function** (q)

MLE – a simple example



Example: Suppose we measure the time of decay for a large sample of unstable particles. The time of each decay will be a random variable and will be distributed **according to an exponential decay function**

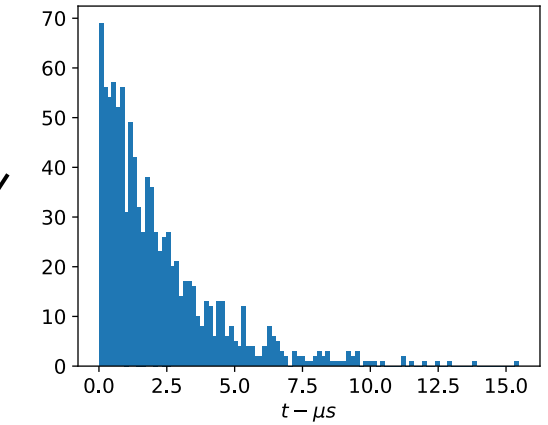
$$p(t) = \frac{1}{\tau} e^{-t/\tau}$$

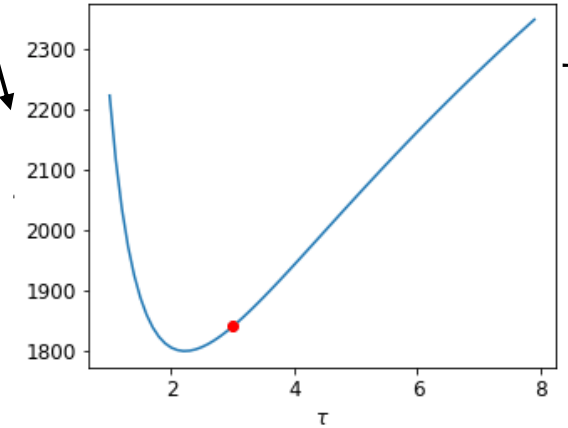
Then,

$$q = -\ln(L) = N\ln(\tau) + \sum_i^N \frac{t_i}{\tau} \quad \text{and,} \quad q' = \frac{N}{\tau} - \frac{1}{\tau^2}\sum_i^N t_i$$

Setting $q' = \frac{N}{\tau} - \frac{1}{\tau^2}\sum_i^N t_i = 0$

Yields a (hopefully) familiar expression $\hat{\tau} = \frac{1}{N}\sum_{i=1}^N t_i$





Numerical example

MLE – a special case

Suppose we want to estimate *p* parameters of a linear model and we have *n* observables,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip}, \;\; i = 1 \ldots n,$$

and each of the y_i are random variables with a normal distribution $y_i \sim \phi(\sum_{k=0}^{p} \beta_k x_{ik}, \sigma)$

Then the random vector $\boldsymbol{y}$ can be modelled using a multivariate Gaussian distribution

$$p(\boldsymbol{y}|X\boldsymbol{\beta}) = \frac{1}{\sqrt{(2\pi)^n \det(\sigma^2 I)}} e^{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^T(\sigma^2 I)^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})}$$

Where explicitly,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots x_{1p} \\ 1 & x_{21} & x_{22} & \ldots x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & x_{12} & \ldots x_{np} \end{pmatrix}$$

MLE – a special case

Suppose we want to estimate *p* parameters of a linear model and we have *n* observables,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip}, \;\; i = 1 \ldots n,$$

and each of the y_i are random variables with a normal distribution $y_i \sim \phi(\sum_{k=0}^{p} \beta_k x_{ik}, \sigma)$

Then the random vector $\boldsymbol{y}$ can be modelled using a multivariate Gaussian distribution

$$p(\boldsymbol{y}|X\boldsymbol{\beta}) = \frac{1}{\sqrt{(2\pi)^n \det(\sigma^2 I)}} e^{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^T(\sigma^2 I)^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})}$$

Remember, $L(\boldsymbol{\beta}) = p(\boldsymbol{y}|X\boldsymbol{\beta})$

And so $$q(\boldsymbol{\beta}) = -\ln L(\boldsymbol{\beta}) = \frac{n}{2}\ln(2\pi\sigma^2) + ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 \frac{1}{2\sigma^2}$$

MLE – a special case

Suppose we want to estimate *p* parameters of a linear model and we have *n* observables,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip}, \;\; i = 1 \ldots n,$$

and each of the y_i are random variables with a normal distribution $y_i \sim \phi(\sum_{k=0}^{p} \beta_k x_{ik}, \sigma)$

Then the random vector $\boldsymbol{y}$ can be modelled using a multivariate Gaussian distribution

$$p(\boldsymbol{y}|X\boldsymbol{\beta}) = \frac{1}{\sqrt{(2\pi)^n \det(\sigma^2 I)}} e^{-\frac{1}{2}(\boldsymbol{y}-X\boldsymbol{\beta})^T(\sigma^2 I)^{-1}(\boldsymbol{y}-X\boldsymbol{\beta})}$$

Remember, $L(\boldsymbol{\beta}) = p(\boldsymbol{y}|X\boldsymbol{\beta})$

And so $$q(\boldsymbol{\beta}) = -\ln L(\boldsymbol{\beta}) = \boxed{\frac{n}{2}\ln(2\pi\sigma^2)} + \boxed{||\boldsymbol{y} - X\boldsymbol{\beta}||^2}\frac{1}{2\sigma^2}$$

It should be obvious that since **this term** is a constant (since σ is a constant), then minimising q is the same thing as minimising **this**

This is ordinary least squares!

Properties of Maximum Likelihood Estimators

See `MethodOfMoments.ipynb`

Several (useful) properties of maximum likelihood estimators are:

1. Invariance: Maximum likelihood estimators are **invariant under bijective transformations** of variables and **relationships between parameters** being estimated are **preserved** in Maximum likelihood estimators

$$\alpha = g(\theta) \implies \hat{\alpha} = g(\hat{\theta})$$

Properties of Maximum Likelihood Estimators

See `MethodOfMoments.ipynb`

Several (useful) properties of maximum likelihood estimators are:

1. Invariance: Maximum likelihood estimators are **invariant under bijective transformations** of variables and **relationships between parameters** being estimated are **preserved** in Maximum likelihood estimators

$$\alpha = g(\theta) \implies \hat{\alpha} = g(\hat{\theta})$$

2. In several cases, Maximum Likelihood Estimators are ***un***biased

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Properties of Maximum Likelihood Estimators

See `MethodOfMoments.ipynb`

Several (useful) properties of maximum likelihood estimators are:

1. Invariance: Maximum likelihood estimators are **invariant under bijective transformations** of variables and **relationships between parameters** being estimated are **preserved** in Maximum likelihood estimators

$$\alpha = g(\theta) \implies \hat{\alpha} = g(\hat{\theta})$$

2. In several cases, Maximum Likelihood Estimators are ***un***biased

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

3. MLEs are **consistent estimators** (assuming uniqueness)

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0$$

as n → ∞, for arbitrary ε

Typically we have to **trade certain properties** of estimators in favor of others.

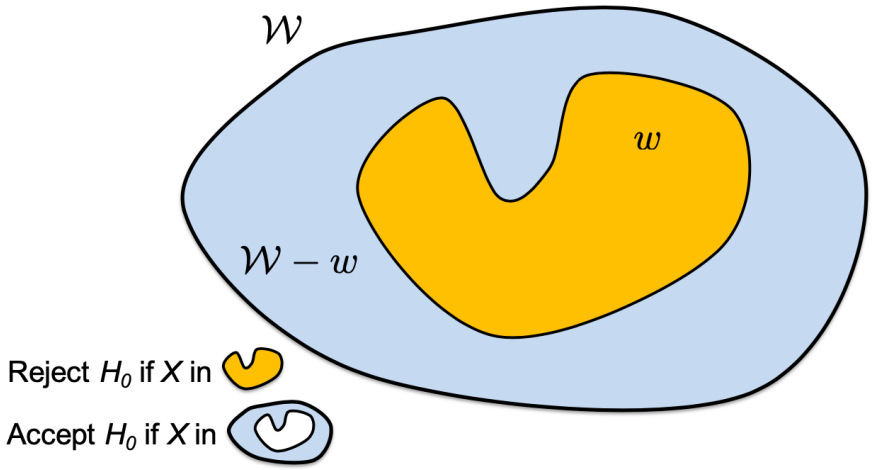
E.G 1: We might care only to minimize the "mean-squared-error"

$$MSE(\hat{\theta}) = \text{bias}(\hat{\theta})^2 + V(\hat{\theta})$$

E.G 2: We might only care about the large sample limit of the bias and variance

$$\lim_{n \to \text{inf}} \text{bias}(\hat{\theta}_n) \to 0 \quad \text{and} \quad \lim_{n \to \text{inf}} \text{V}(\hat{\theta}_n) \to 0$$

Hypothesis testing



Hypothesis testing

The concept of ***testing hypotheses*** is of course at the heart of experimental science – confronting a scientific hypothesis with data is the way we progress science after all.

Hypothesis testing is a significant part of statistics application in many areas where data is used to inform a decision making process – In fact, Hypothesis testing is part of a whole branch in statistics called ***decision theory***. We won't be able to cover decision theory in these lectures but we can get a good grounding in hypothesis testing.

Hypothesis testing

The concept of ***testing hypotheses*** is of course at the heart of experimental science – confronting a scientific hypothesis with data is the way we progress science after all.

Hypothesis testing is a significant part of statistics application in many areas where data is used to inform a decision making process – In fact, Hypothesis testing is part of a whole branch in statistics called ***decision theory***. We won't be able to cover decision theory in these lectures but we can get a good grounding in hypothesis testing.

In modelling data we often deal with the case that one or more of our hypotheses involve some unknown parameter – these are composite hypotheses.

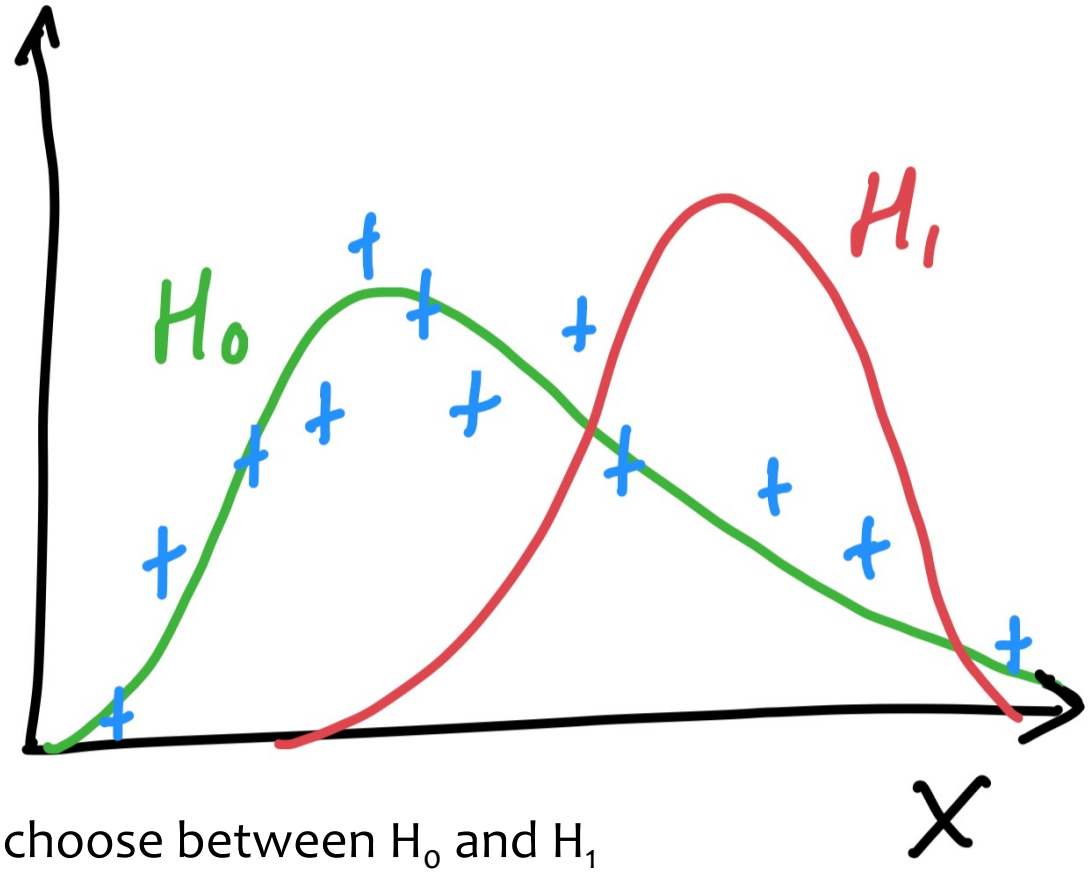
If there are no unknowns however, the hypothesis is completely specified – these are called simple hypotheses. For now, lets stick to ***simple hypotheses***.

Hypothesis testing

Suppose that we are choosing between two hypotheses to describe our experimental data labelled H0 and H1 .

We typically distinguish the two as;

H_0 := the ***null*** hypothesis
H_1 := the ***alternate*** hypothesis



X is some function of the experimental observations which is supposed to summarize the data
– this is known as a ***test statistic***.

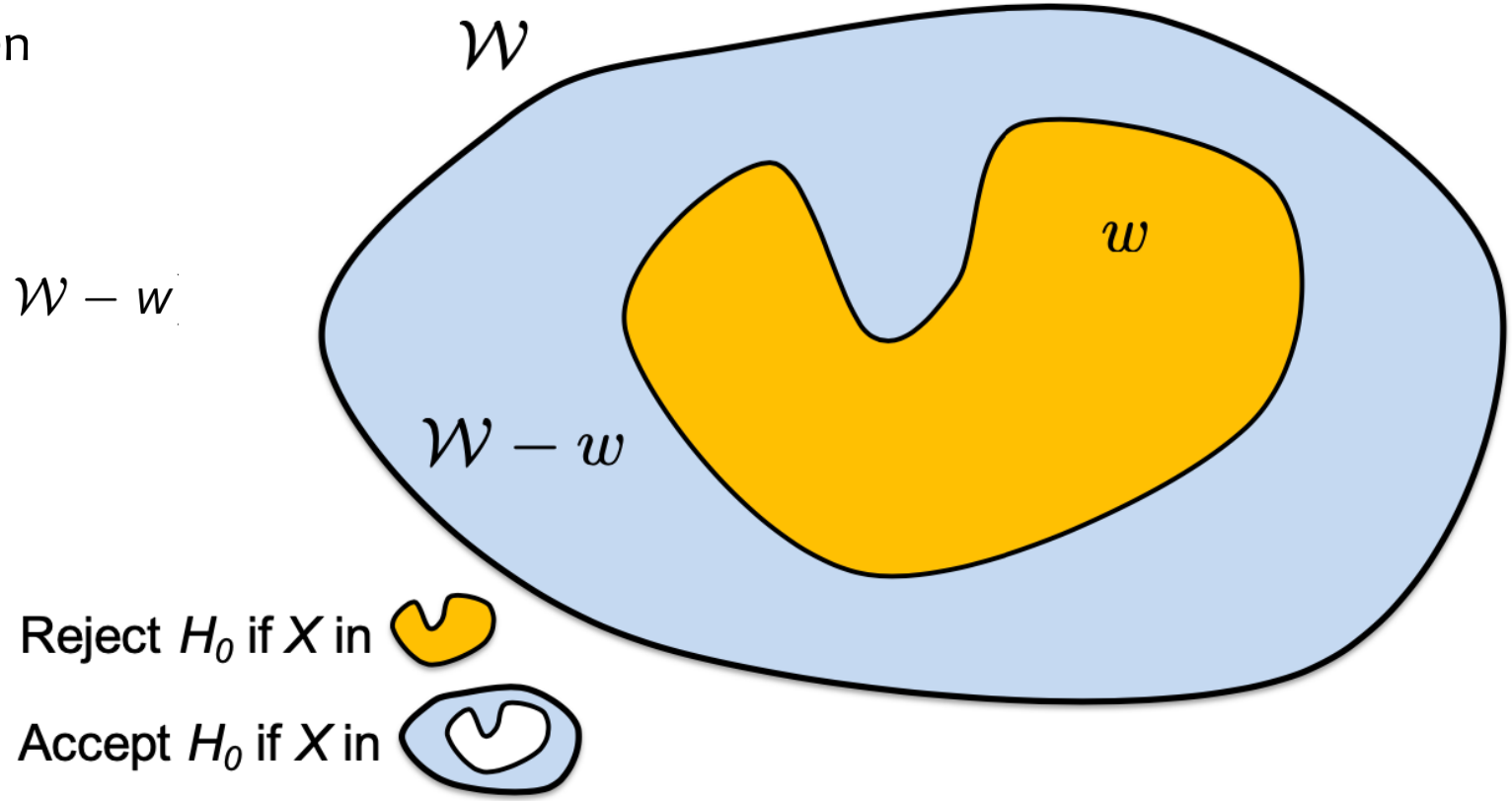
In ***frequentist statistics*** we use this test statistic to choose between H_0 and H_1 while trying to ***minimize the error*** we may make in that choice!

Type-I/Type-II errors

Suppose then that we have our chosen test statistic $X \in \mathcal{W}$

We divide this region $\mathcal{W}$ into a **critical region** w and a **region of acceptance** $\mathcal{W} - w$

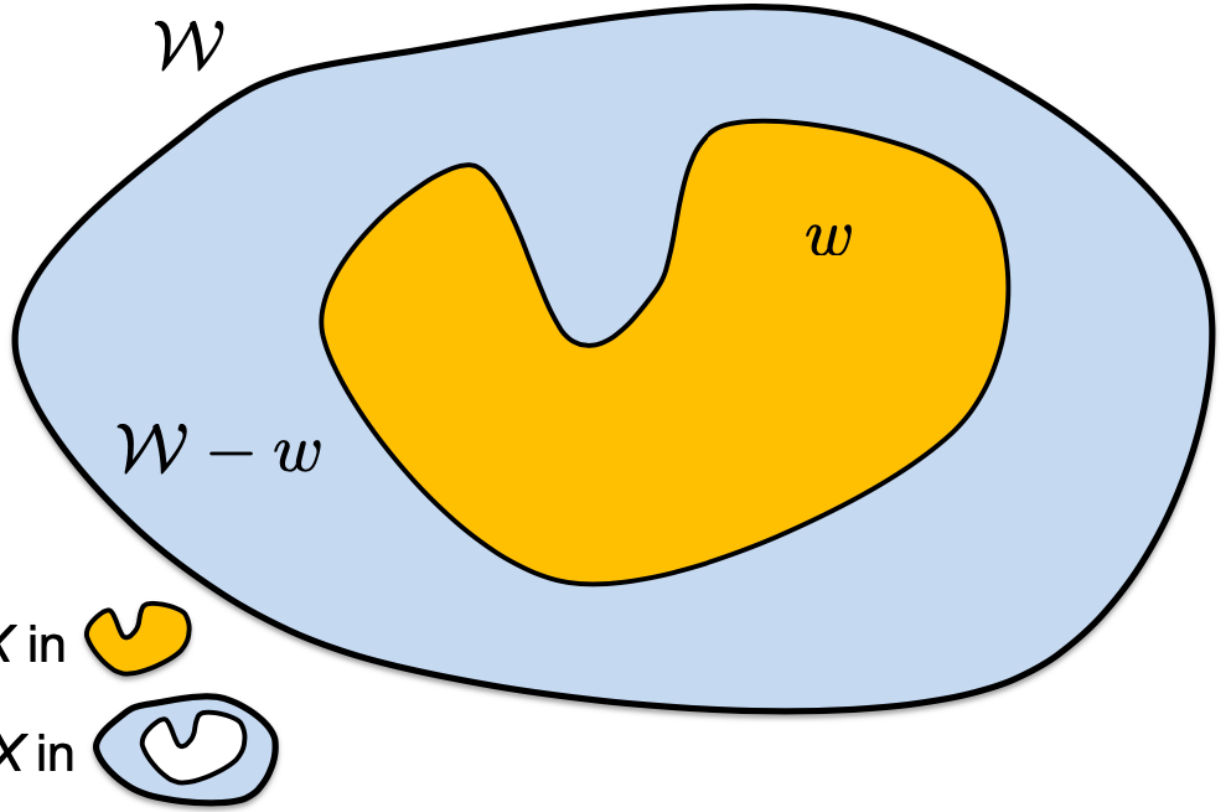
Observations of X falling into w would lead us to believe that our null hypothesis is **not true**.



Defining a test of H_0, given we've decided on our test statistic, then becomes choosing a critical region w

Type-I/Type-II errors

Type-1 Error : In practise, we often tune the critical region so as to obtain a particular probability (known as the size of the test) α that X falls into the critical region when H0 is true (we usually say "under H_0")

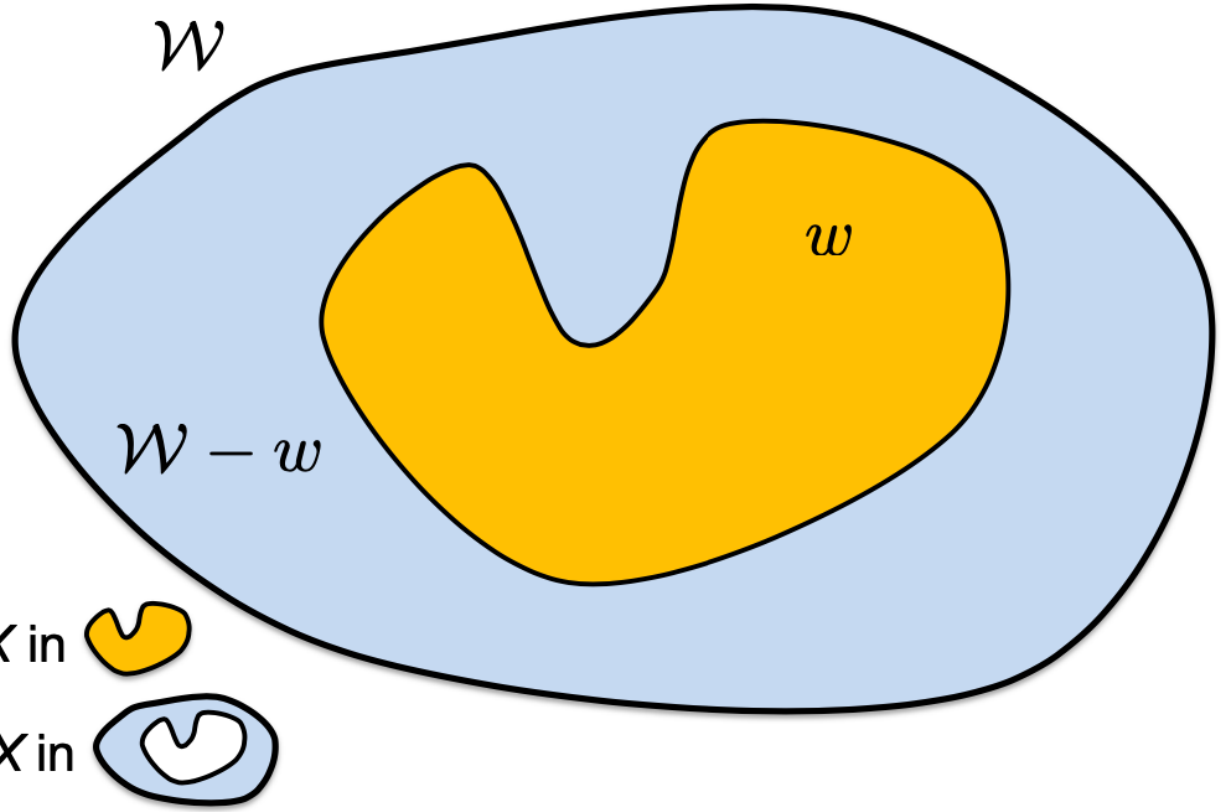


$$P(X \in w | H_0) = \alpha$$

You can see then that α is exactly **the probability to reject the null hypothesis if the null hypothesis is true**

Type-I/Type-II errors

Type-II Error: Of course, we also want to know how useful a test is at discriminating against the alternate hypothesis. This is known as the ***power of the test***, and is defined as the probability of *X* falling into the critical region if H_1 is true (under H_1),



$$P(X \in w | H_1) = 1 - \beta$$

Clearly this is related to the probability that *X* falls into the acceptance region via

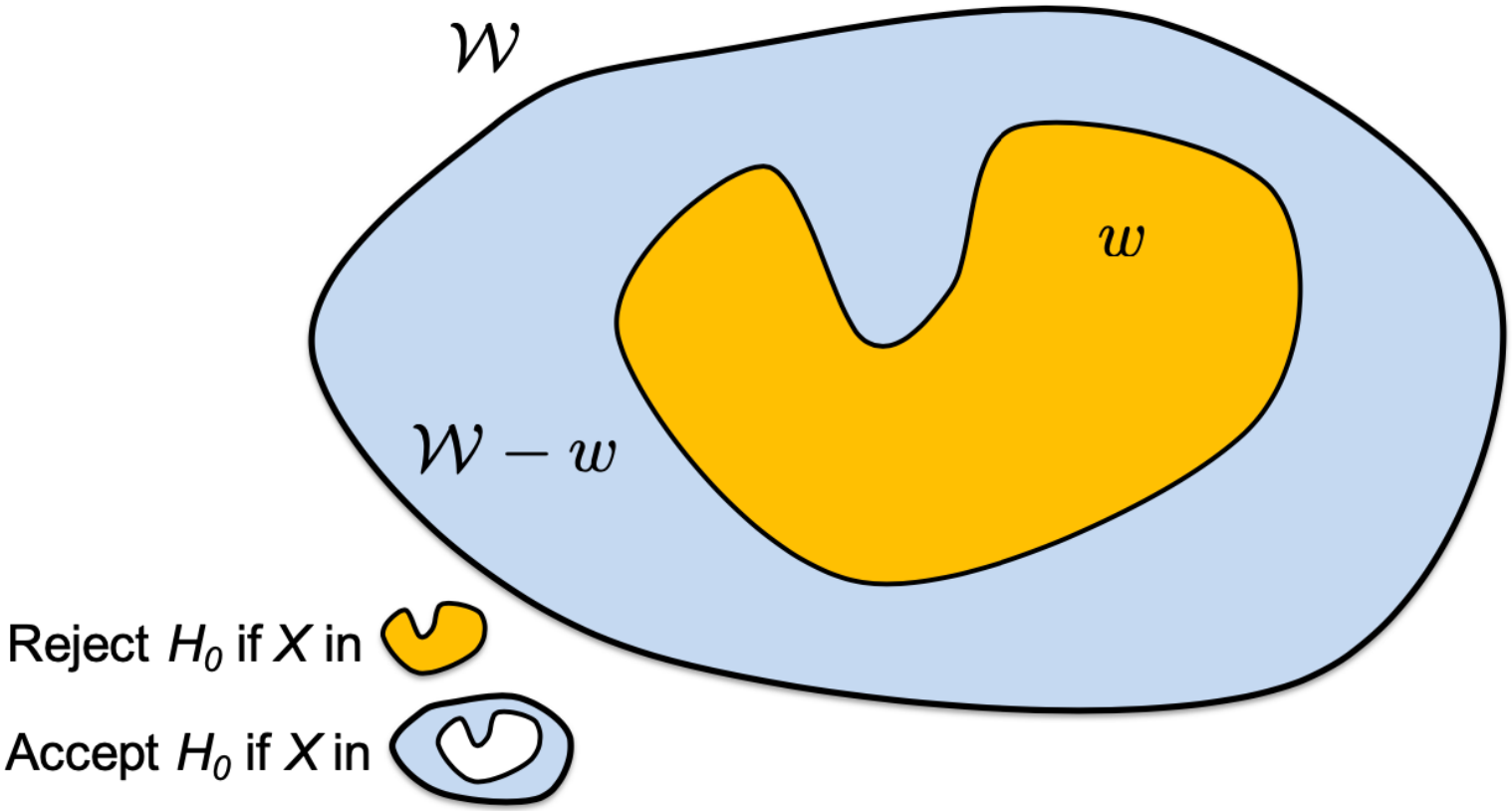
$$P(X \in \mathcal{W} - w | H_1) = 1 - P(X \in w | H_1) = \beta$$

Then **β** is the probability that we would **accept the null hypothesis when the alternative hypothesis is true**

Type-I/Type-II errors

$$P(X \in w | H_0) = \alpha \longrightarrow \text{How likely to } \textit{falsely reject } H_0$$

$$P(X \in \mathcal{W} - w | H_1) = \beta \longrightarrow \text{How likely to } \textit{falsely accept } H_0$$



Type-I error rate = α

Type-II error rate = β

Hypothesis testing

My ***personal*** way to think about frequentist hypothesis testing:

In **frequentist hypothesis testing**, we make a decision to reject a hypothesis or not and deal with the fact that **sometimes we're wrong**

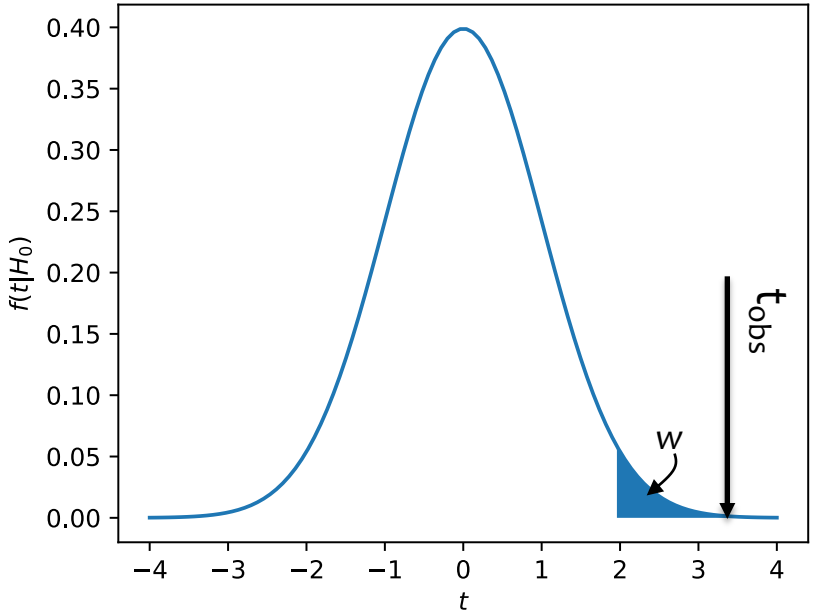
Hypothesis testing

My ***personal*** way to think about frequentist hypothesis testing:

> In **frequentist hypothesis testing**, we make a decision to reject a hypothesis or not and deal with the fact that **sometimes we're wrong**

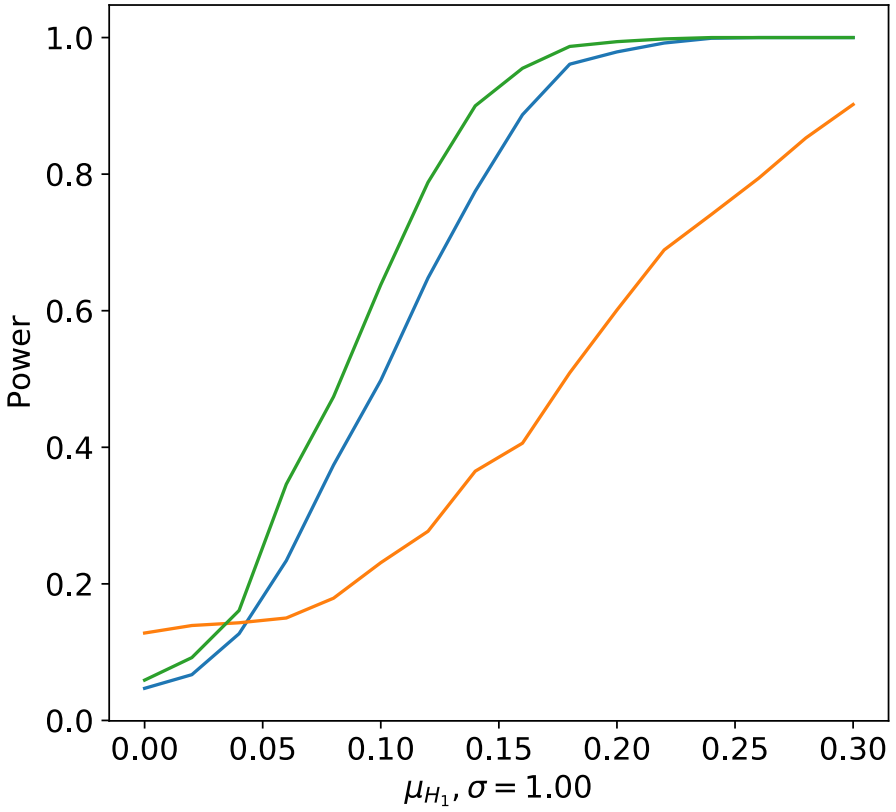
There are a huge number of hypothesis tests on the market for use in various problems, however they generally follow the same routine;

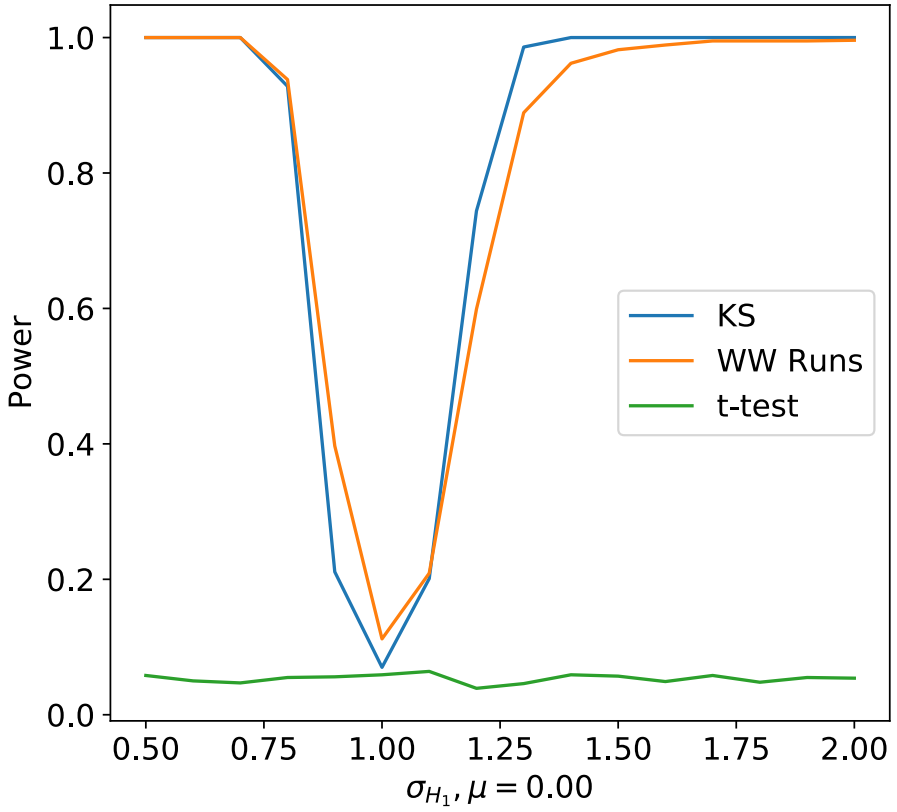
1. Define a test (summary) statistic $t \in \mathbb{R}$ that summarizes the observations and has some separation between H_0 and H_1
2. Define a critical region *w* such that
$$\int_w f(t|H_0) = \alpha$$
where α is a predefined value between 0 and 1.
3. Determine the value of *t* in the observed data,
4. Reject H_0 if $t_{\text{obs}} \in w$



Power comparison of tests

For examples a power comparison of various tests applied to the case where both H_0 and H_1 are Gaussian distributions, see [GaussianHypothesisTests.ipynb]





The Neyman-Pearson Lemma

Let's start again thinking about some observed data **X** and suppose it has a probability distribution function f(**X**;θ), where θ is used to represent our hypotheses:

$\theta = \theta_0$ represents the null hypothesis H0

$\theta = \theta_1$ represents the alternate hypothesis H1

$$f(X|H(\theta)) = f(X;\theta)$$

T^{hm}: The critical region *w* that maximises the power 1 – β for a fixed size of test α is the set of points for which $\Lambda \geq c_\alpha \in R$, where,

$$\Lambda = \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = \frac{L(\theta_1)}{L(\theta_0)}$$

and c_α satisfies $$\int_w f(\mathbf{X};\theta_0)\,d\mathbf{X} = \alpha$$

The Neyman-Pearson Lemma

Let's start again thinking about some observed data **X** and suppose it has a probability distribution function f(**X**;θ), where θ is used to represent our hypotheses:

$\theta = \theta_0$ represents the null hypothesis H0

$\theta = \theta_1$ represents the alternate hypothesis H1

$$f(X|H(\theta)) = f(X;\theta)$$

T^hm: The critical region *w* that maximises the power 1 – β for a fixed size of test α is the set of points for which $\Lambda \geq c_\alpha \in R$, where,

$$\Lambda = \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = \frac{L(\theta_1)}{L(\theta_0)}$$

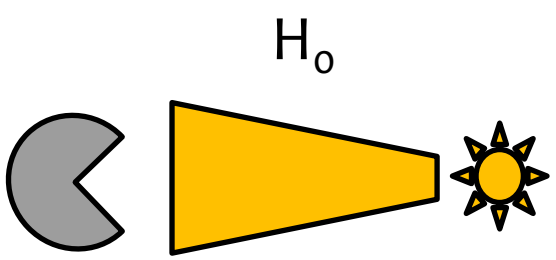
and c_α satisfies $$\int_w f(\mathbf{X};\theta_0)\,d\mathbf{X} = \alpha$$

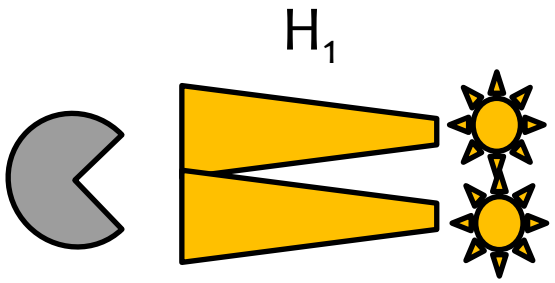
The NP-lemma says we should reject the null hypothesis for large values of the likelihood ratio Λ!

Example: Likelihood ratio test

See LikelihoodRatioTest.ipynb

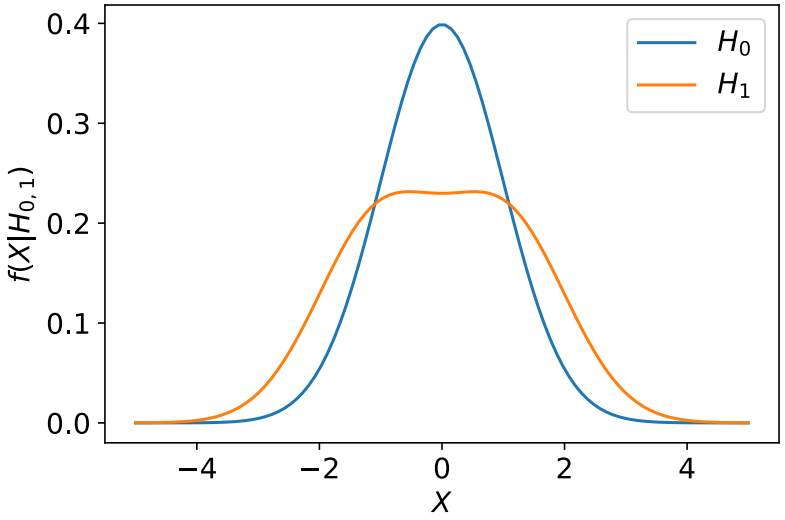
Let's look at an example where H_0 is a single Gaussian with $\mu=0$, $\sigma=1$ and H_1 is the sum of two Gaussians with $\mu_1=\mu+\delta\mu$, $\mu_2=\mu-\delta\mu$ and $\sigma_1=\sigma_2=\sigma$





The NP-lemma says we should use the likelihood ratio to separate whether a pulse originates from a single or double-source

$$\Lambda = \frac{L(H_1)}{L(H_0)} = \frac{\frac{1}{2}e^{-(X-\delta\mu)^2} + \frac{1}{2}e^{-(X+\delta\mu)^2}}{e^{-X^2}}$$



Example: Likelihood ratio test

Let's look at an example where H_0 is a single Gaussian with $\mu=0$, $\sigma=1$ and H_1 is the sum of two Gaussians with $\mu_1=\mu+\delta\mu$, $\mu_2=\mu-\delta\mu$ and $\sigma_1=\sigma_2=\sigma$

We know (from Lindsey) that machine learning models are great at separating classes (eg H_0 vs H_1)

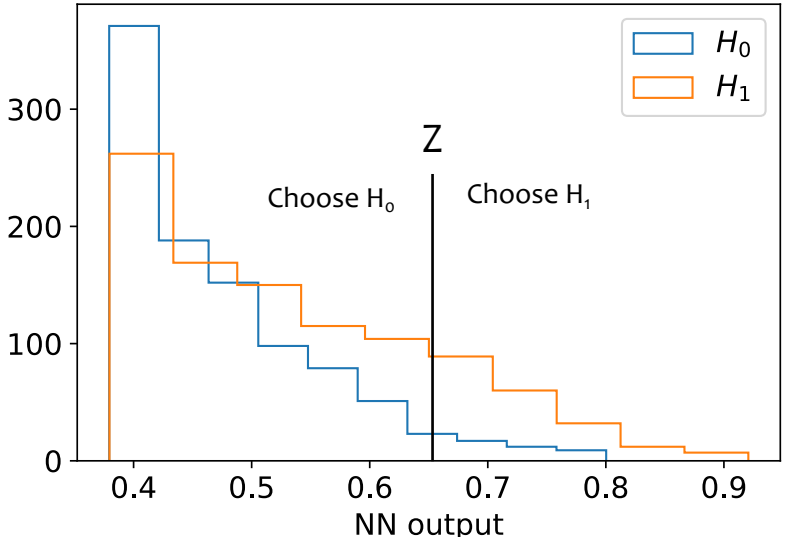
Train a NN with binary cross-entropy (BCE) loss using simulated pulses from H_0 and H_1

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \ln(p_i) + (1-y_i)\ln(1-p_i)$$

$H_0 \rightarrow y = 0$
$H_1 \rightarrow y = 1$

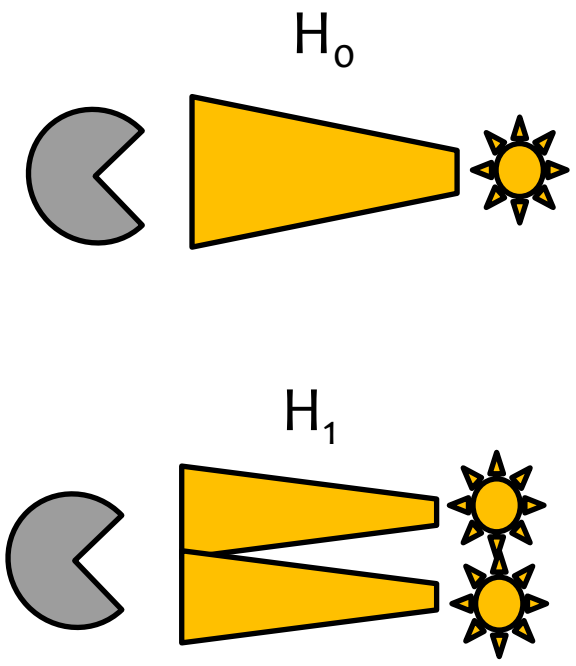
$$\alpha = \int_Z^{+\infty} f(NN|H_0)d(NN)$$

$$1-\beta = \int_Z^{+\infty} f(NN|H_1)d(NN)$$

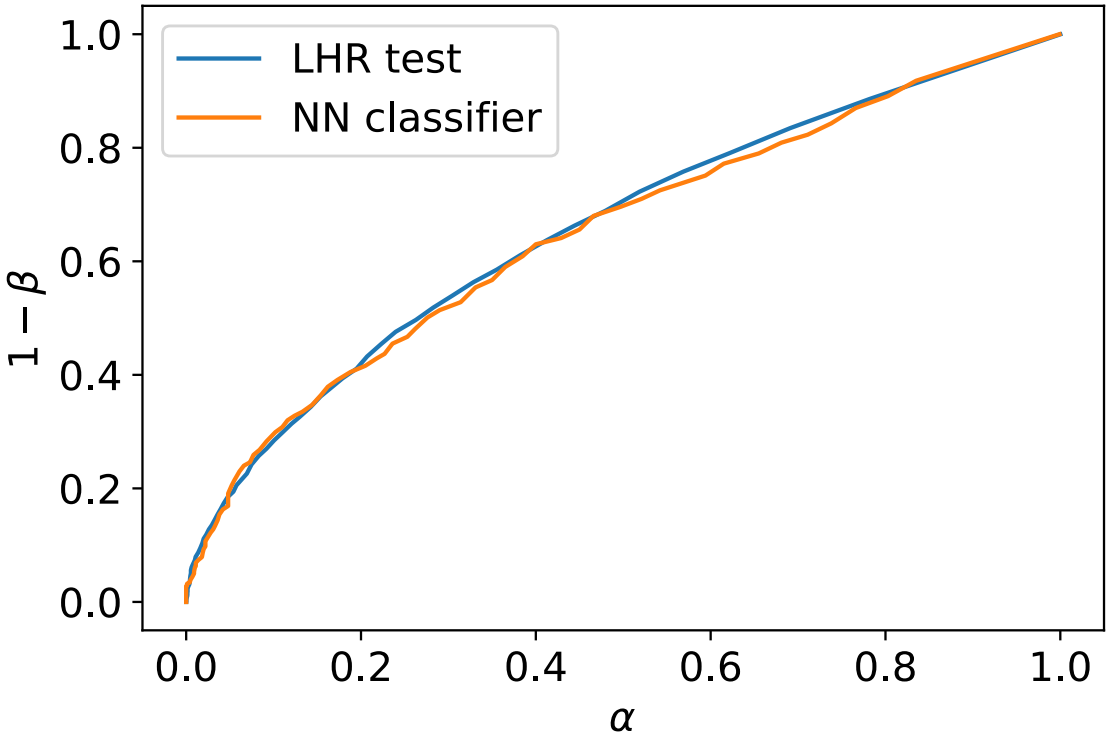


Example: Likelihood ratio test

Let's look at an example where H_0 is a single Gaussian with $\mu=0$, $\sigma=1$ and H_1 is the sum of two Gaussians with $\mu_1=\mu+\delta\mu$, $\mu_2=\mu-\delta\mu$ and $\sigma_1=\sigma_2=\sigma$



What we find is that our NN is equivalent to this Likelihood Ratio (LHR) test → since in this case the two hypotheses are *simple* , we expect this from the NP-lemma



Simple vs Composite Hypotheses

Caution! The Neyman-Pearson lemma states that the most powerful test of a fixed size is the likelihood ratio test for ***simple hypotheses*** (think parameters fixed)

More often than not, we test ***composite hypotheses*** (think parameters unknown) against one another

Composite hypotheses can be thought of as a set of hypotheses that are parameterised by one (or more) parameters, often related to parameters of a physical model we are interested in

There are some results concerning the ***average power*** against "equally distant" alternatives that suggest ***profile likelihood ratio*** tests are the most powerful asymptotically*, however, as we'll see likelihood based tests have many nice features that make them worth focusing on (even if not guaranteed to be the most powerful) for composite hypotheses too.

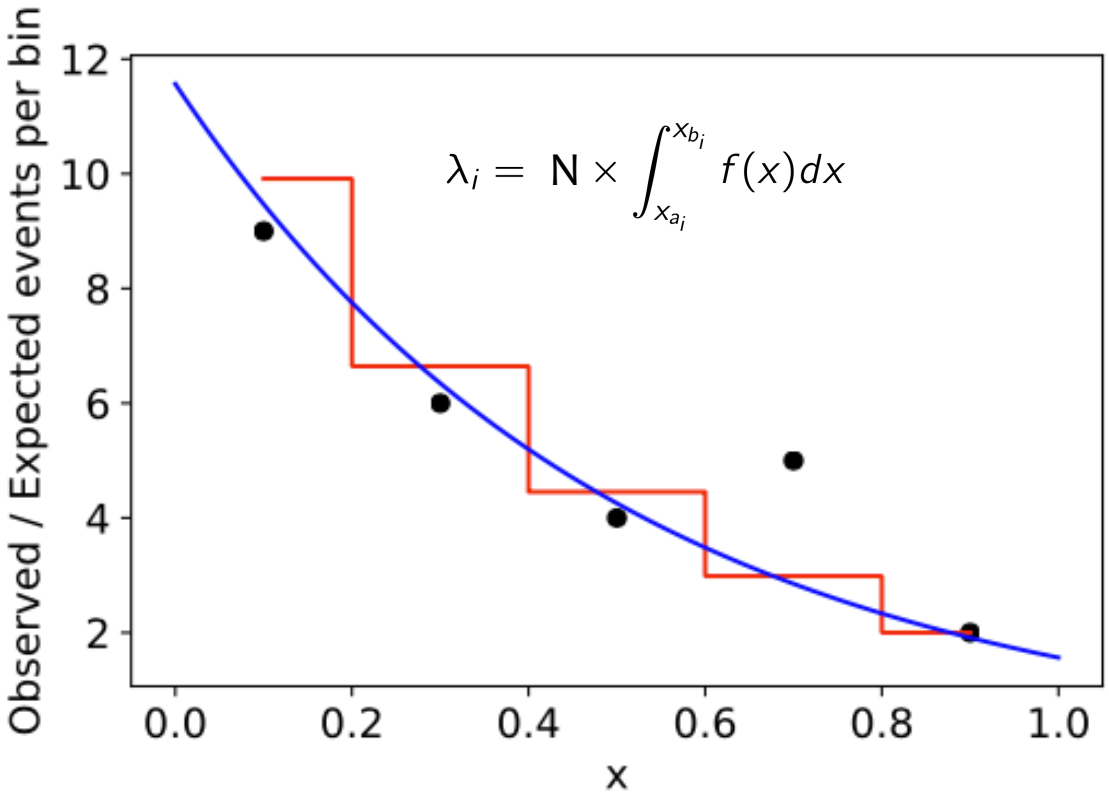
*"*Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*" A. Wald

Example – an LHR goodness of fit test

Suppose we have a histogram of data and a histogram (either derived from some analytic function or from simulation) describing what we expect to see under a given hypothesis. We can ask how well the hypothesis agrees with that data

Each bin in the histogram can be thought of as an independent Poisson random process

The observed number of events in each bin is labelled oi and the expected (Poisson mean) values are λ_i



Example – an LHR goodness of fit test

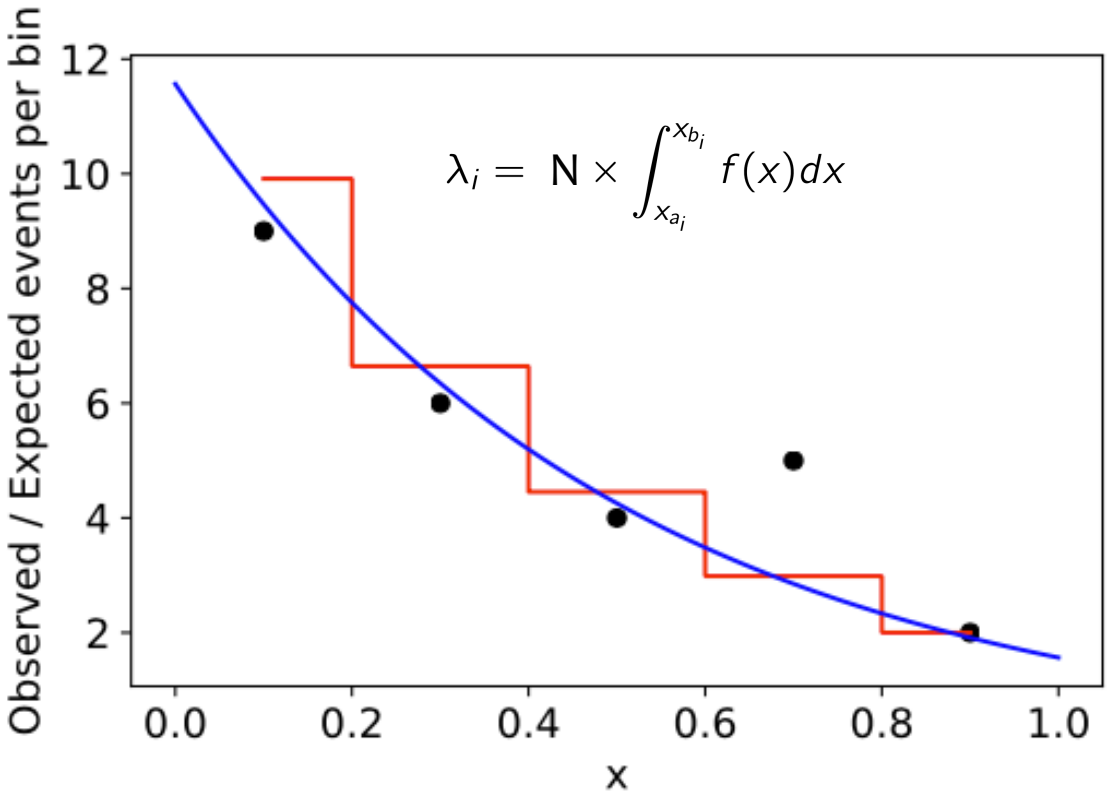
Suppose we have a histogram of data and a histogram (either derived from some analytic function or from simulation) describing what we expect to see under a given hypothesis. We can ask how well the hypothesis agrees with that data

Each bin in the histogram can be thought of as an independent Poisson random process

The observed number of events in each bin is labelled oi and the expected (Poisson mean) values are λ_i

The likelihood function is then the product over Poisson probability distributions,

$$L(H_0) = \prod_i (\lambda_i)^{o_i} e^{-\lambda_i}$$



where we've ignored the $o_i!$ terms, since in the end, this constant will drop out of the ratio of likelihoods

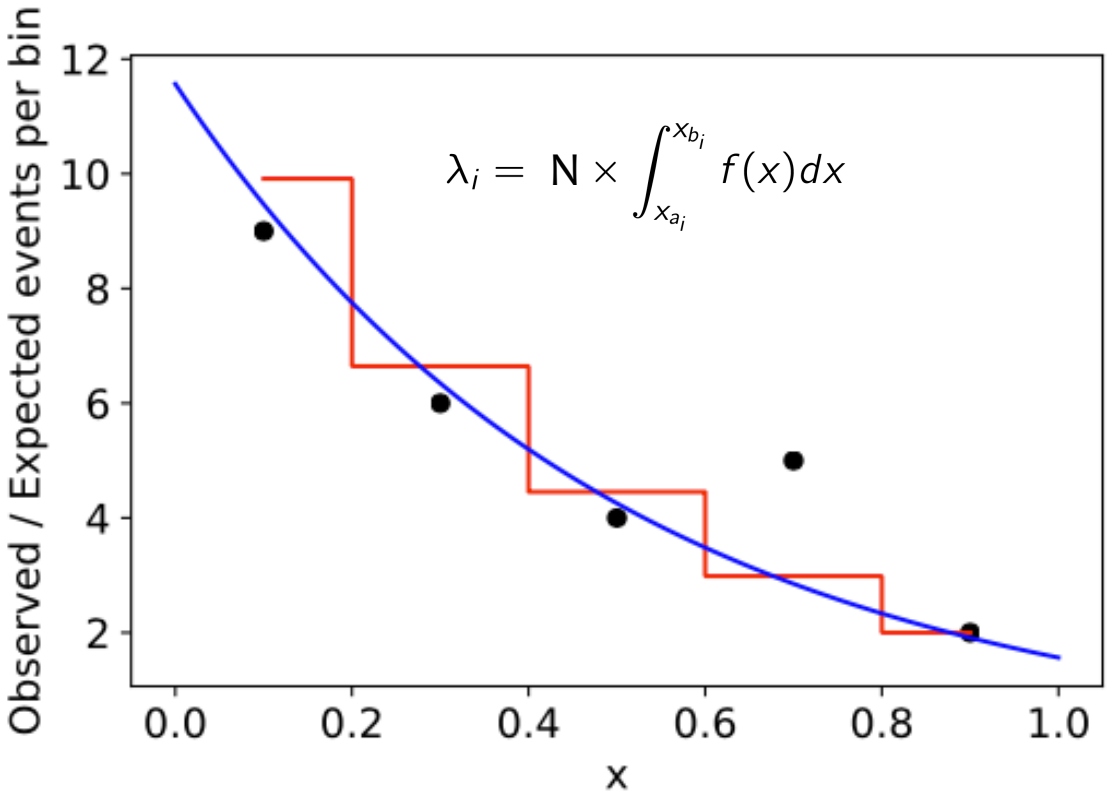
Example – an LHR goodness of fit test

We haven't specified an alternate hypothesis, but there is something we can use.

The ***saturated model*** is the one for which the ***model used is exactly the observed data*** i.e $\lambda_i = o_i$

We then have

$$\Lambda = \frac{L(H_0)}{L(H_1)} = \frac{\prod_i (\lambda_i)^{o_i} e^{-\lambda_i}}{\prod_i (o_i)^{o_i} e^{-o_i}}$$



Example – an LHR goodness of fit test

We haven't specified an alternate hypothesis, but there is something we can use.

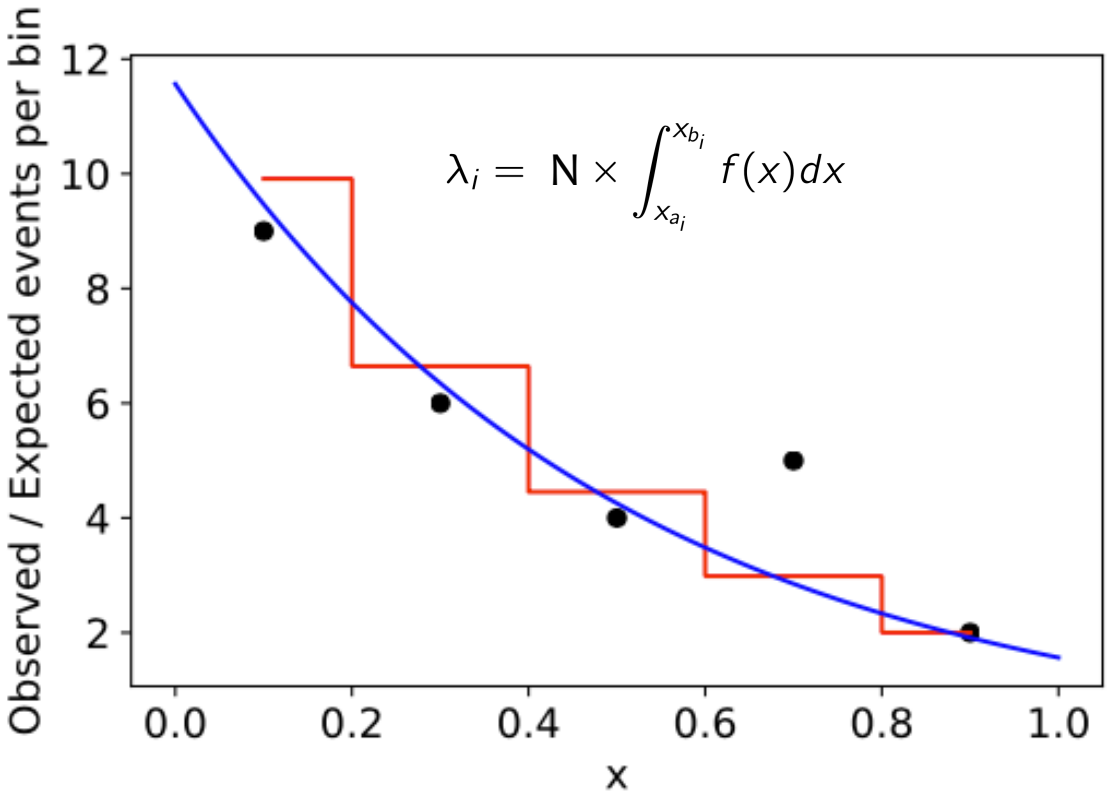
The ***saturated model [1]*** is the one for which the ***model used is exactly the observed data*** i.e $\lambda_i = o_i$

We then have

$$\Lambda = \frac{L(H_0)}{L(H_1)} = \frac{\prod_i (\lambda_i)^{o_i} e^{-\lambda_i}}{\prod_i (o_i)^{o_i} e^{-o_i}}$$

Note that L(H1) is exactly $L(\hat{\lambda}_i)$ – (MLEs)

we usually take −2× the log of the likelihood (this will become obvious why later)
so our test statistic is



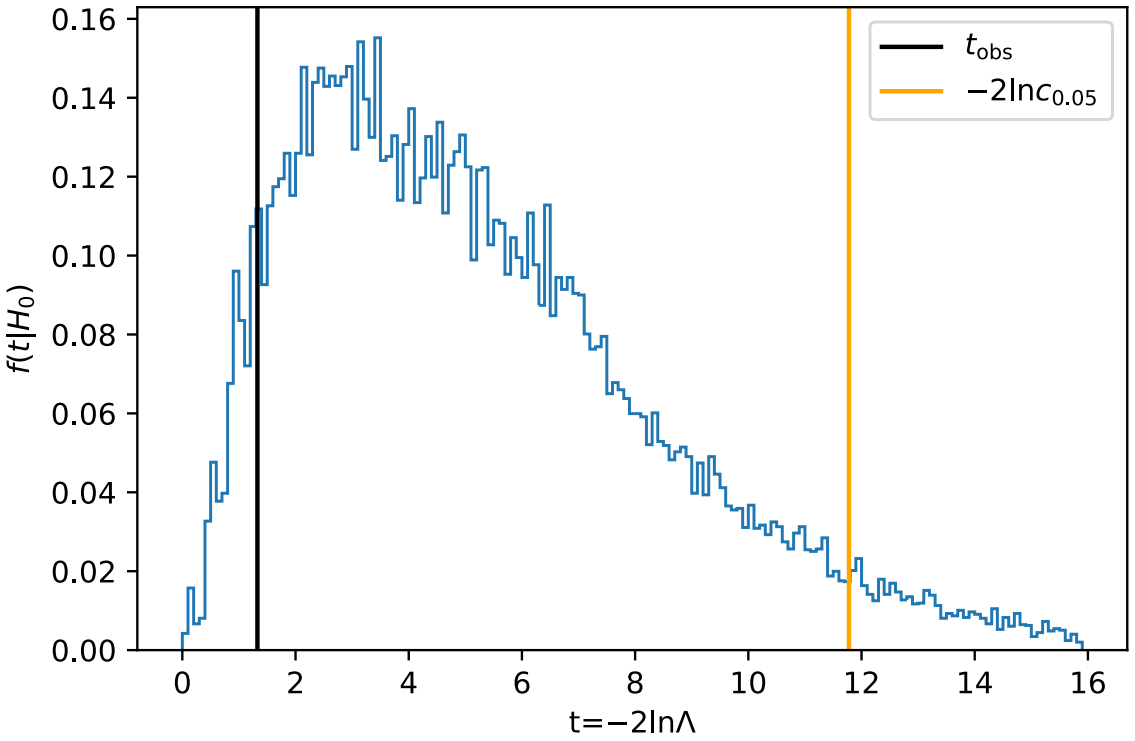
$$t = -2\ln\Lambda = -2\ln\frac{L(H_0)}{L(H_1)} = -2\sum_i \left(o_i \ln(\lambda_i) - \lambda_i - o_i \ln(o_i) + o_i\right)$$

[1] https://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf

Example – an LHR goodness of fit test

The distribution of the test statistic under the null hypothesis can be determined with Monte Carlo simulation

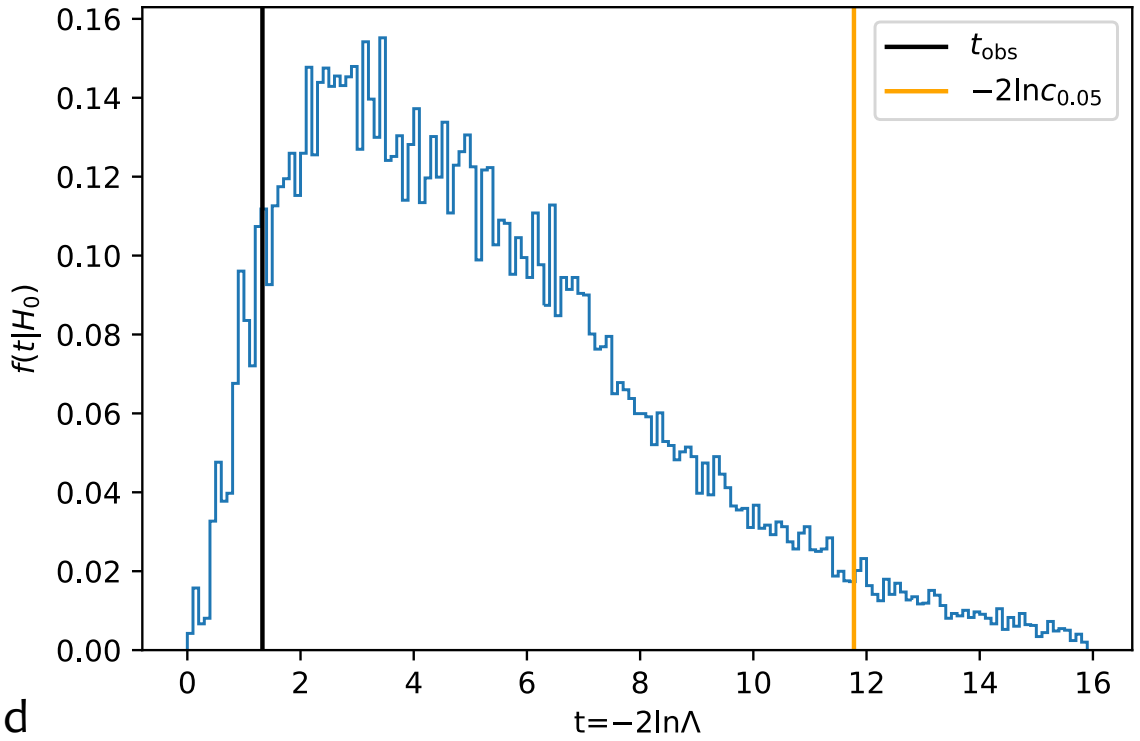
In this case, the observed value is clearly not in the critical region for α=0.05 and so we can conclude that the model is a good description of our data and we'd be wrong with an error rate of 5%



Example – an LHR goodness of fit test

The distribution of the test statistic under the null hypothesis can be determined with Monte Carlo simulation

In this case, the observed value is clearly not in the critical region for α=0.05 and so we can conclude that the model is a good description of our data and we'd be wrong with an error rate of 5%



Note: If each of the bins were Gaussian distributed (i.e if the number of events in each bin were large) with means $\mu=\lambda_i$ and standard deviations σ_i , then

$$L(H_0) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{o_i-\lambda_i}{\sigma_i}\right)^2}$$

$$L(H_1) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{o_i-o_i}{\sigma_i}\right)^2}$$

$$-2\ln\Lambda = -2\ln\left(\frac{L(H_0)}{L(H_1)}\right) = \sum_i \left(\frac{o_i-\lambda_i}{\sigma_i}\right)^2$$

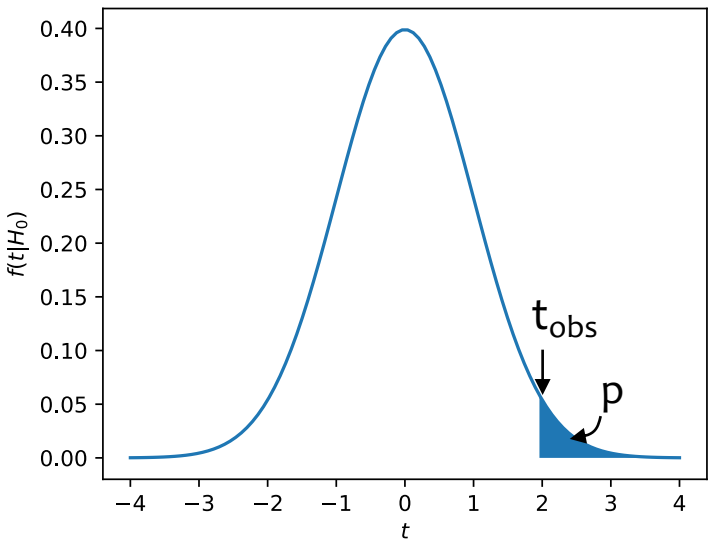
Which is the (hopefully) familiar formula for the Pearson χ^2 test!

P-values (an aside)

P-values often get a bad reputation in the statistics literature due to their ***misuse***

A p-value can be defined with respect to any hypothesis but more often than not it is defined with respect to the null hypothesis;

$$p = \int_{t_{\text{obs}}}^{+\infty} f(t|H_0)\,dt$$



The p-value is

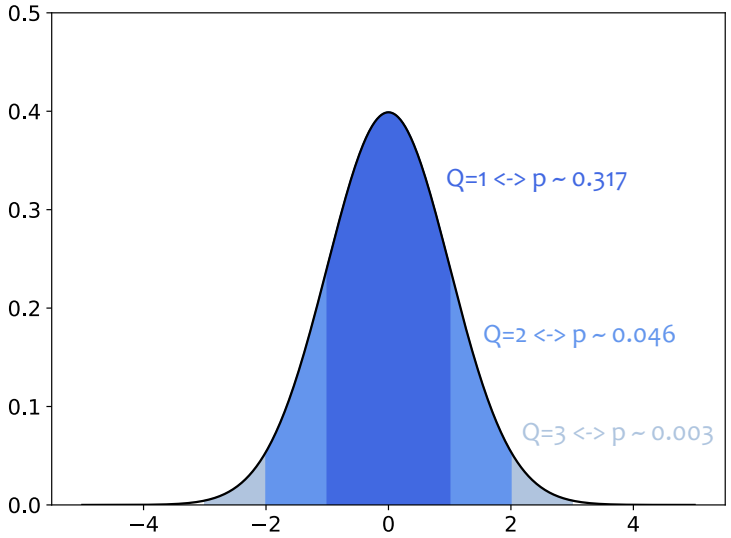
- A random variable that depends on the observed data (it's a **post-observation quantity**)
- Distributed uniformly between 0 and 1 under the null hypothesis

P-values (an aside)

P-values often get a bad reputation in the statistics literature due to their ***misuse***

A p-value can be defined with respect to any hypothesis but more often than not it is defined with respect to the null hypothesis;

$$p = \int_{t_{\text{obs}}}^{+\infty} f(t|H_0)dt$$



The p-value is

- A random variable that depends on the observed data (it's a **post-observation quantity**)
- Distributed uniformly between 0 and 1 under the null hypothesis

Typically, p-values are converted into "significance"

This can be obtained by finding Q for which

$$p = 1 - \int_{-Q}^{+Q} \phi(X; \mu = 0, \sigma = 1)dX$$

Then Q is the "number of sigma" or ***significance*** of an observation

This is **just a conversion** and does not assume anything about how the test-statistic is distributed (no Gaussian assumption)

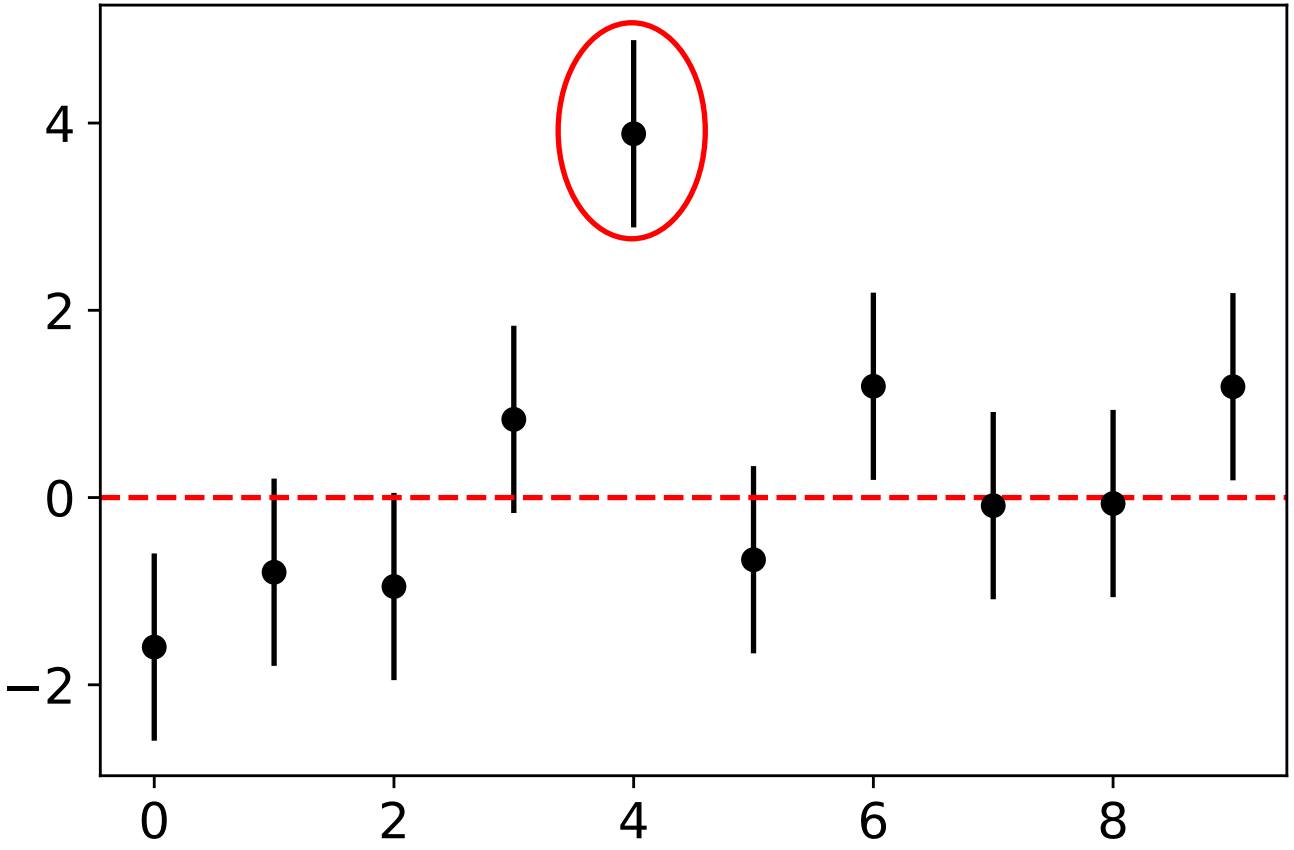
P-values (an aside)

Be careful to correct for numerous trials when reporting p-values/significance

Suppose we had 10 gaussian random variables (measurements/observations…) and we saw the following …

We might get enthusiastic about this particular data point being far from the expectation (zero)

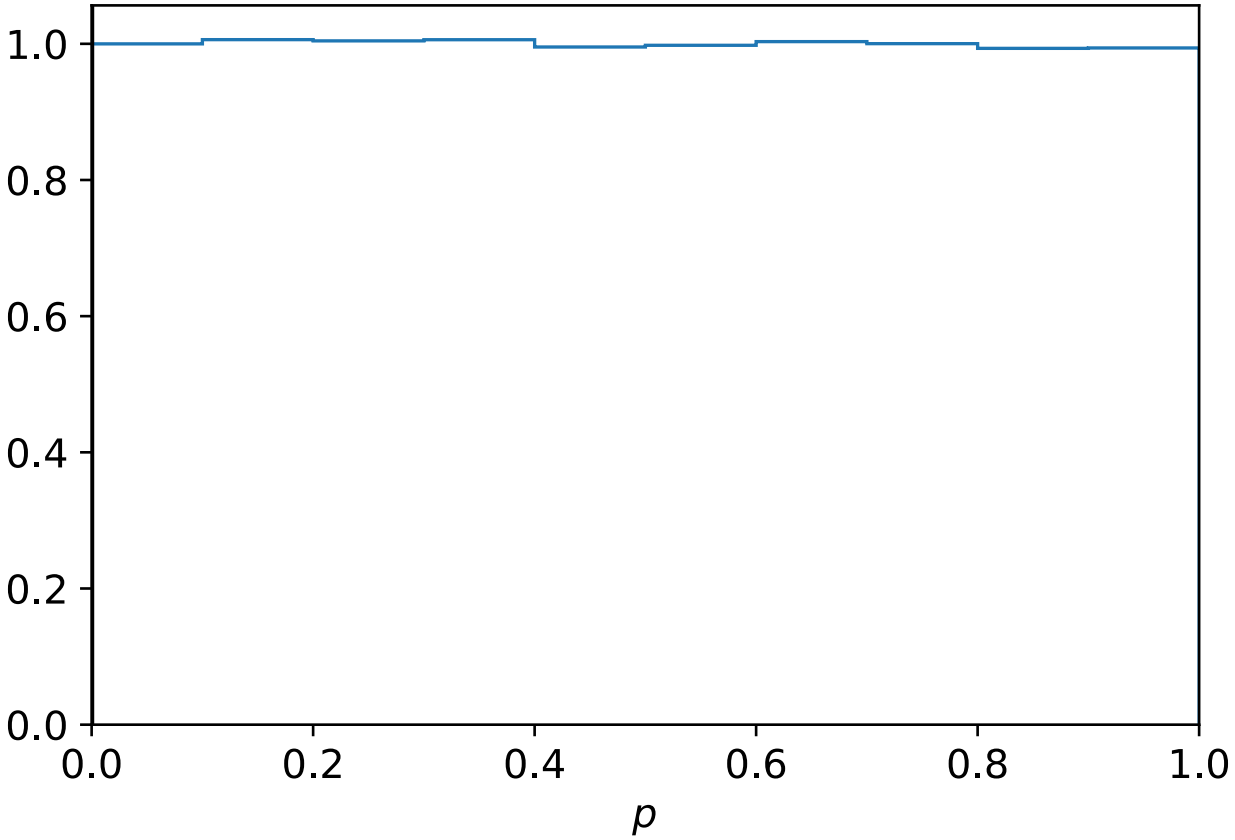
Using a LHR test would yield a p-value of 0.0001!



P-values (an aside)

It's easy to check how often we would get a p-value as small as this for that measurement (the distribution is flat under the null hypothesis)

The probability to see a p-value at least as small as the one we observed for ***that measurement*** under the null hypothesis is 1 in 10,000!



P-values (an aside)

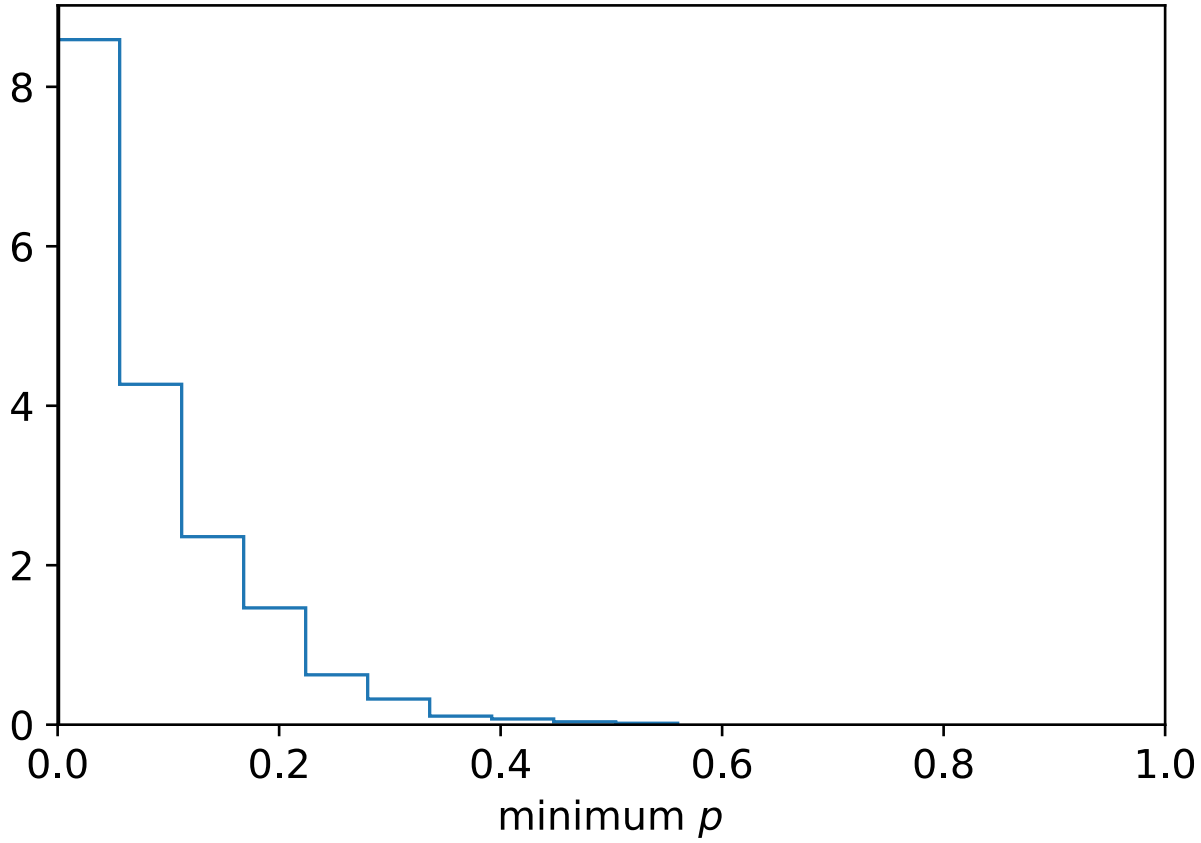
What about the other measurements? What is the probability to see a p-value at least as small as that in ***any of the measurements***?

The probability to see a p-value at least as small as the one we observed for ***any measurement*** under the null hypothesis is 1 in 1000.

For this example, it's easy to show that

$$P(\text{minimum } p < p_{\text{obs}}) = \frac{P(p < p_{\text{obs}})}{n \text{ measurements}}$$

This is known as a "***trials factor***" (or Bonferroni correction)



P-values (an aside)

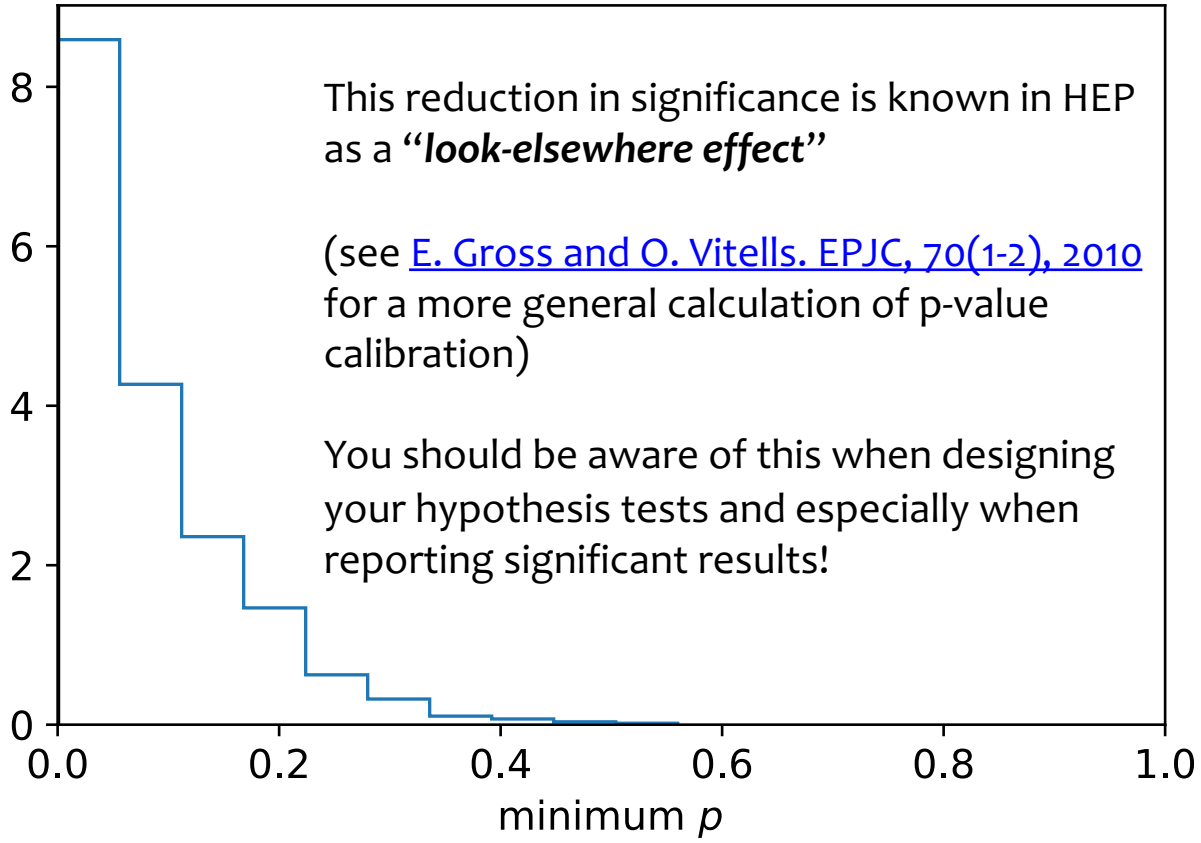
What about the other measurements? What is the probability to see a p-value at least as small as that in ***any of the measurements***?

The probability to see a p-value at least as small as the one we observed for ***any measurement*** under the null hypothesis is 1 in 1000.

For this example, it's easy to show that

$$P(\text{minimum } p < p_{\text{obs}}) = \frac{P(p < p_{\text{obs}})}{n \text{ measurements}}$$

This is known as a "***trials factor***" (or Bonferroni correction)

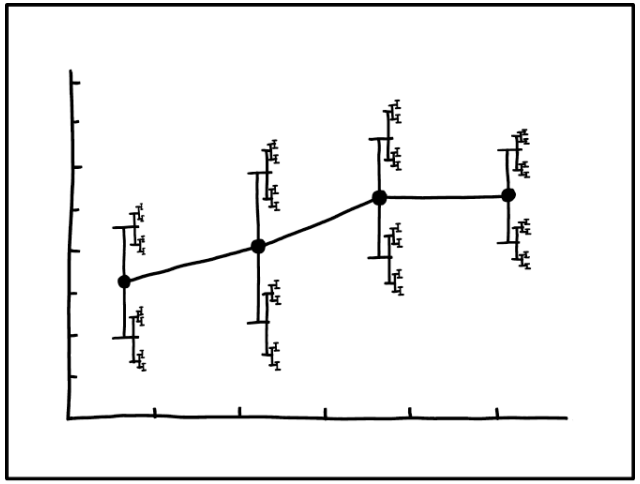


This reduction in significance is known in HEP as a "***look-elsewhere effect***"

(see E. Gross and O. Vitells. EPJC, 70(1-2), 2010 for a more general calculation of p-value calibration)

You should be aware of this when designing your hypothesis tests and especially when reporting significant results!

Uncertainty (confidence & credible intervals)



xkcd.com

I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

Frequentist Intervals

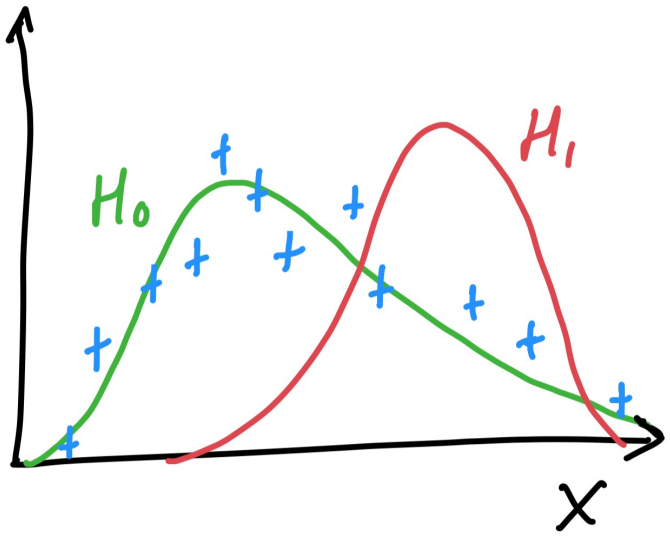
I've so far avoided talking about uncertainty in these lectures (instead we've talked about variance) because uncertainty in (frequentist) statistics is really all about ***hypothesis testing***

Instead of comparing simple hypotheses, we want to test composite (parametric) hypothesis to determine frequentist ***confidence intervals***

Simple hypothesis testing

H_0 vs H_1

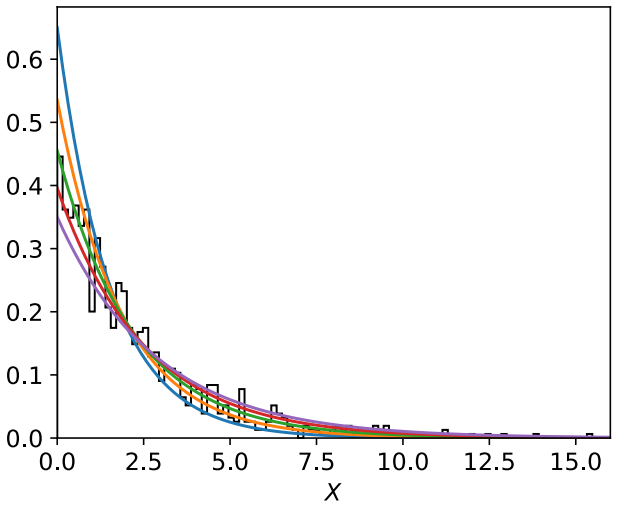
Type-I /II error control



Composite hypothesis testing

$H(\theta) \rightarrow$ parametric hypotheses

Control ***coverage***



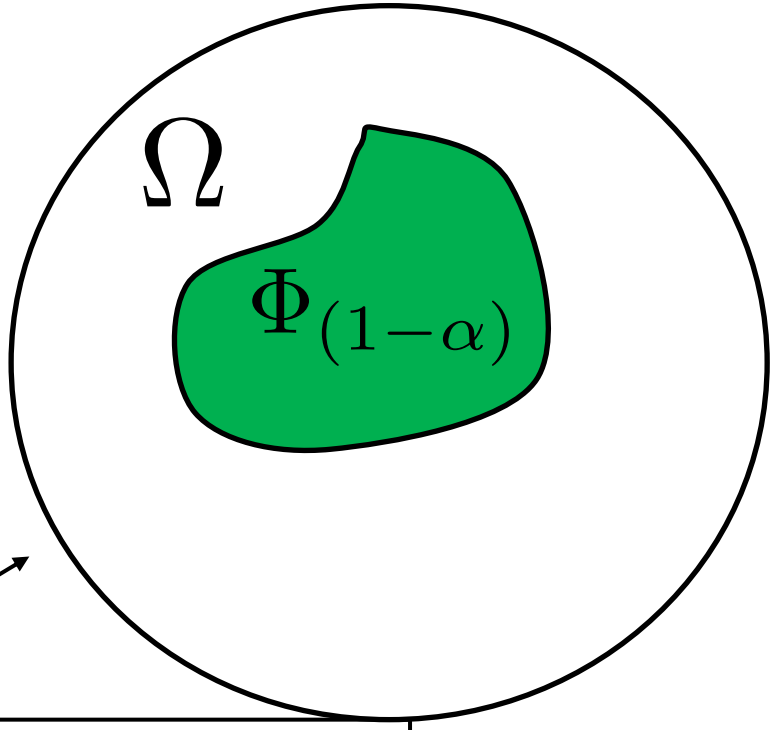
Confidence Intervals

A confidence interval (or region) is a set of parameter values $\theta \in \Phi_{(1-\alpha)}$ at a specified confidence level 100x(1-α)

Confidence intervals are determined ***after the observations → They are random***

Specifically, the confidence region at a confidence level 100 x (1-α) is a region which is constructed such that the true values of the true parameters θ_0 is contained in the region with a probability (1-α)

E.G: a **68% confidence interval** in θ will contain the true value θ_0 in **68% of outcomes in data**

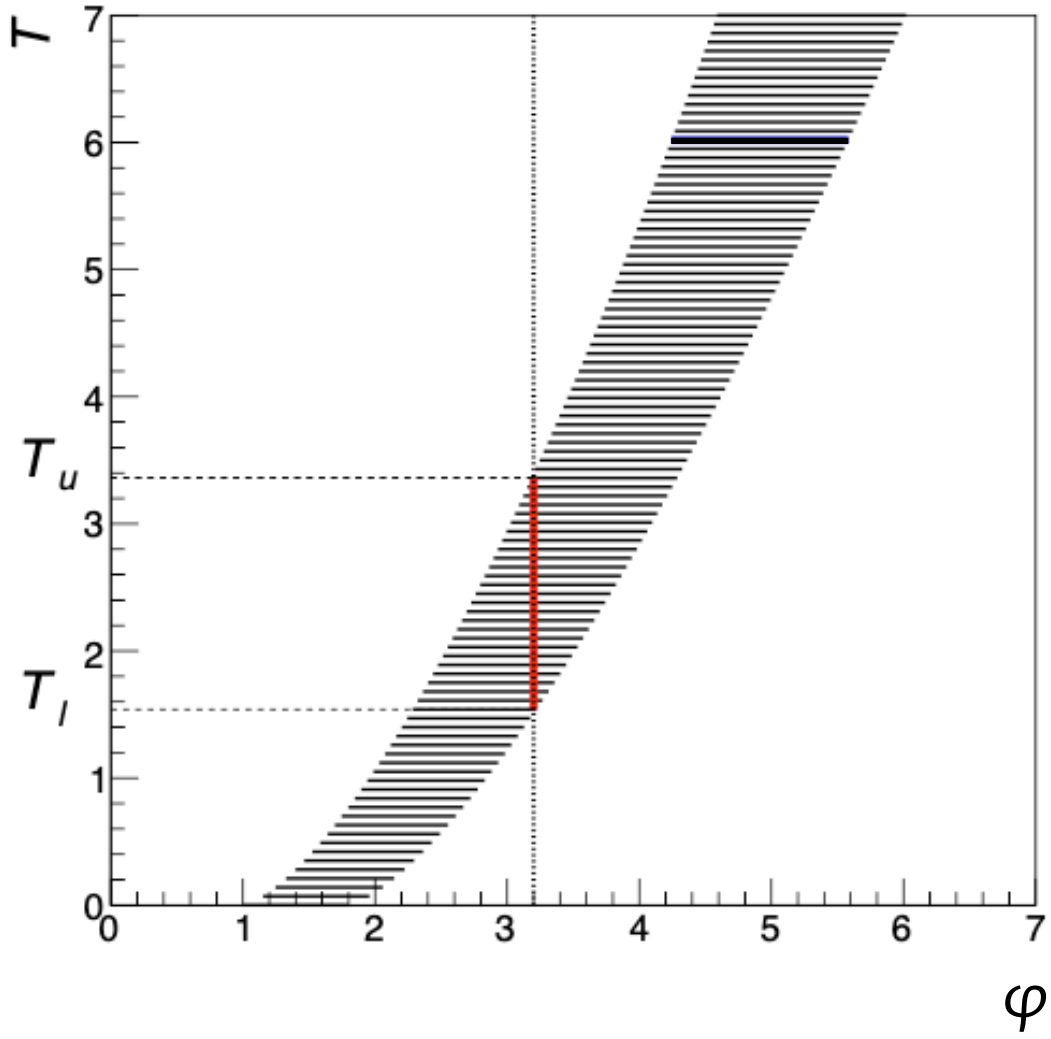


Alternatively, we can think of rejecting all H(θ) with $\theta \notin \Phi_{1-\alpha}$ at an error rate of at most α

Neyman Construction

The ***Neyman construction*** of frequentist intervals is an elegant way to achieve intervals with exactly that property

Suppose we have a single parameter T that parameterizes a (composite) hypothesis H(T) and a test statistic φ

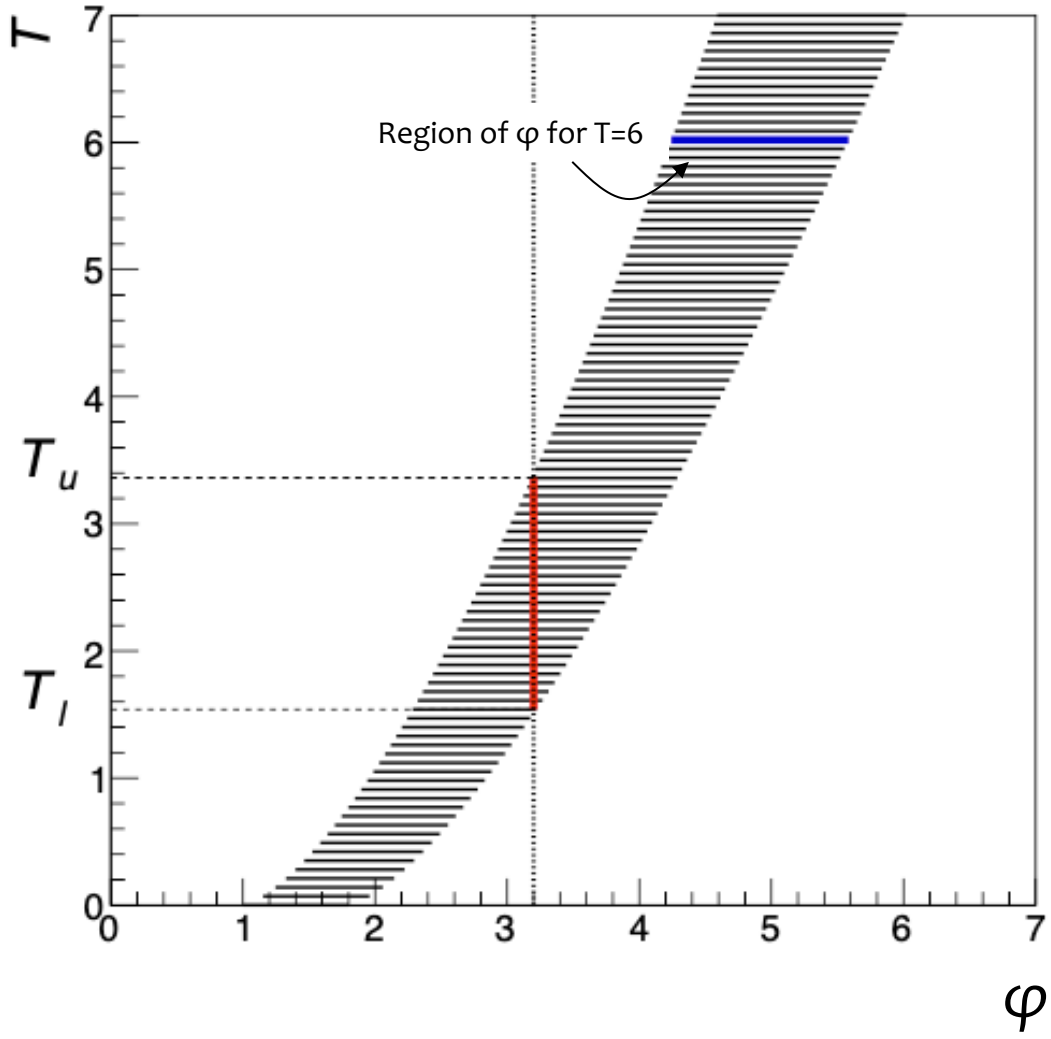


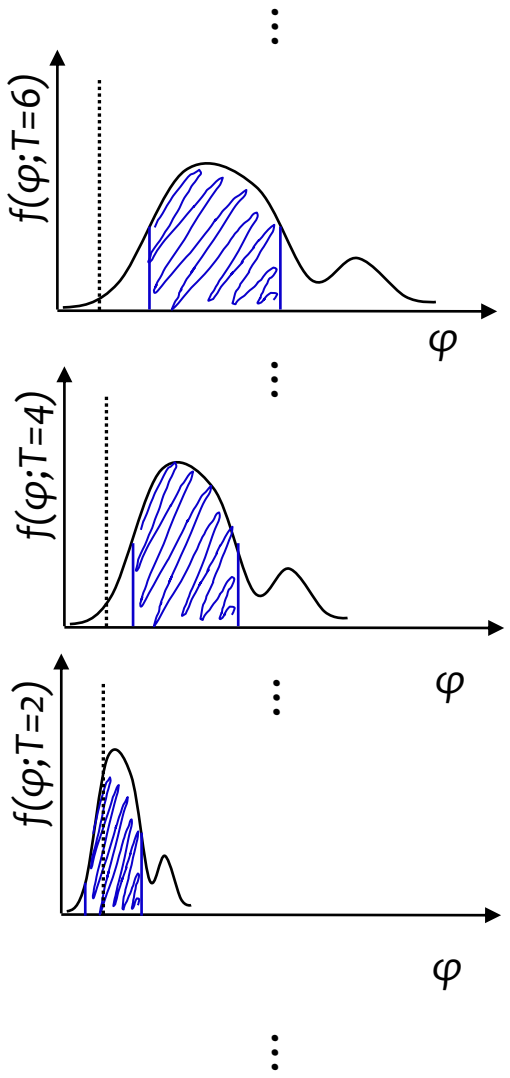
Neyman Construction

The ***Neyman construction*** of frequentist intervals is an elegant way to achieve intervals with exactly that property

Suppose we have a single parameter T that parameterizes a (composite) hypothesis H(T) and a test statistic φ

For each T, we derive the distribution of φ and choose a region of φ that contains 100 x (1-α) of the distribution





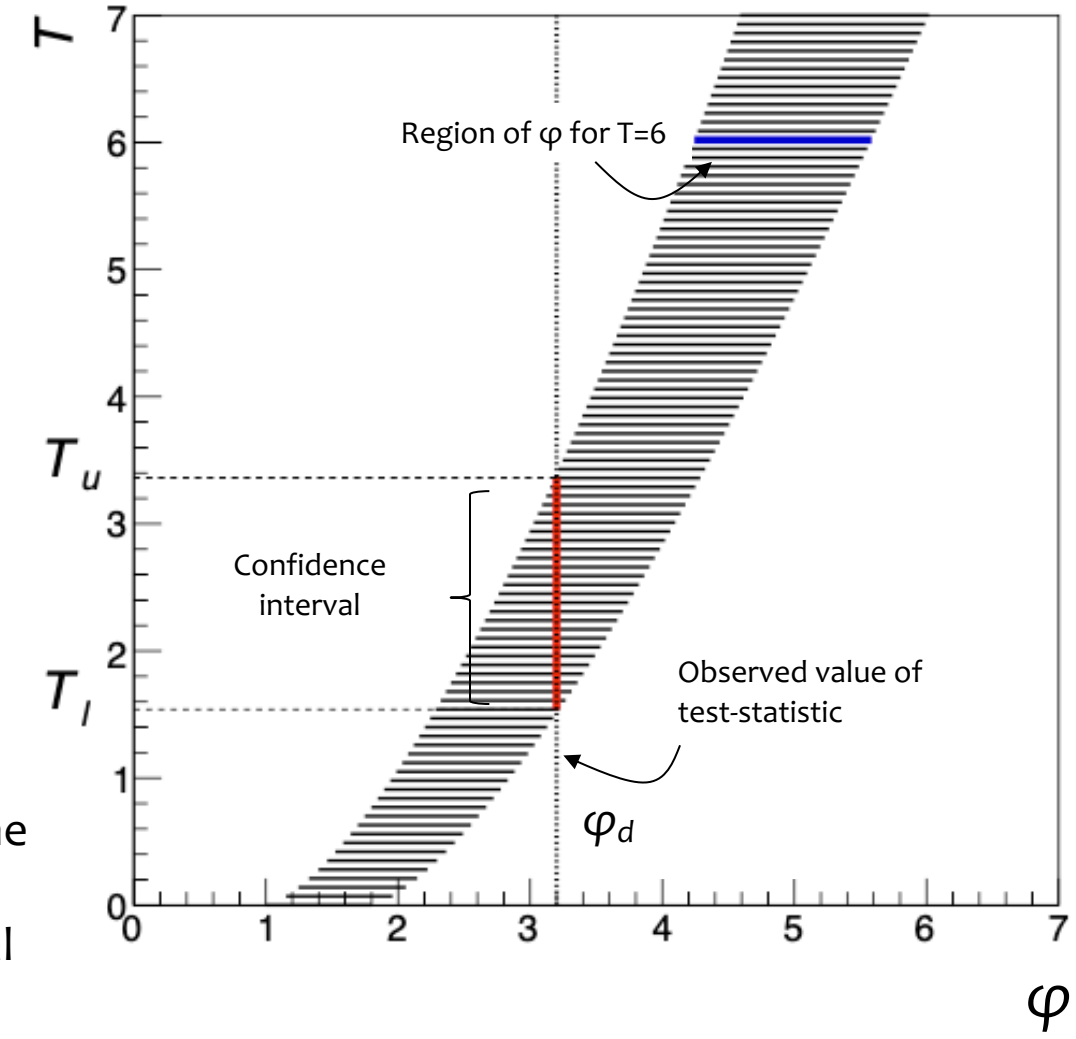
Neyman Construction

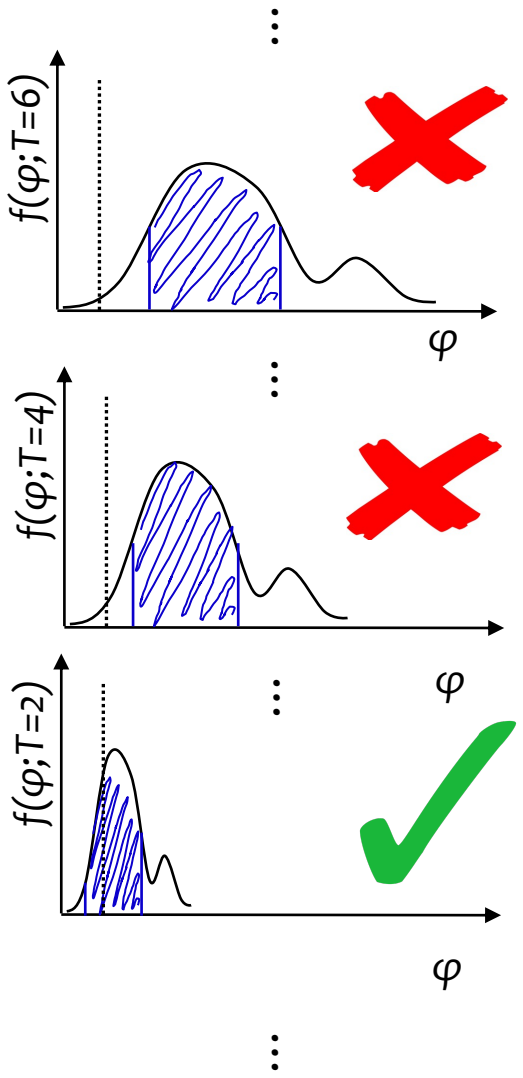
The ***Neyman construction*** of frequentist intervals is an elegant way to achieve intervals with exactly that property

Suppose we have a single parameter T that parameterizes a (composite) hypothesis H(T) and a test statistic φ

For each T, we derive the distribution of φ and choose a region of φ that contains 100 x (1-α) of the distribution

Any point for which the observed value of the test statistic φ_d is contained in that region is included in the 100 x (1-α) confidence interval $[T_l, T_u]$





Profile Likelihood Ratio ordering

For simple hypotheses, we know that the likelihood ratio test statistic yields the ***most powerful*** test

Similarly for composite hypotheses, it can be shown that the ***profile likelihood ratio*** will yield the test with the ***largest average power***

$$\zeta_\theta = -2\ln\frac{L(\theta)}{L(\hat{\theta})} = q(\theta) - q(\hat{\theta})$$

Recall $\hat{\theta}$ is the MLE of θ

We should include the outcomes for which ζ_θ is large or (equivalently) we use ζ_θ as the test statistic and choose the largest values when constructing the (1-α) region

→ Find the value of $\zeta_\theta^{(1-\alpha)}$ that satisfies $\int_0^{\zeta_\theta^{(1-\alpha)}} f(\zeta_\theta|H(\theta))d\zeta_\theta = 1-\alpha$

Note that in this case, the test-statistic ζ_θ will depend on the value of θ and so the observed value of the test statistic will be a curve → the **profile likelihood ratio curve**

Example – Poisson mean

Consider a very simple example of a Poisson random variable *n*

$$H(\lambda) \rightarrow f(n;\lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$$

$$\zeta_\lambda = q(\lambda) - q(\hat{\lambda}) = -2\left(\ln(\lambda) - \lambda\right) + 2\left(\ln(n) + n\right)$$

The MLE for λ is given by

$$\frac{dq}{d\lambda}|_{\lambda=\hat{\lambda}} = 0 \implies \hat{\lambda} = n$$

Suppose we **observe n=4** and want to construct a 68% frequentist confidence interval in λ

Example – Poisson mean

Consider a very simple example of a Poisson random variable *n*

$$H(\lambda) \rightarrow f(n;\lambda) = \frac{\lambda^n}{n!}e^{-\lambda}$$

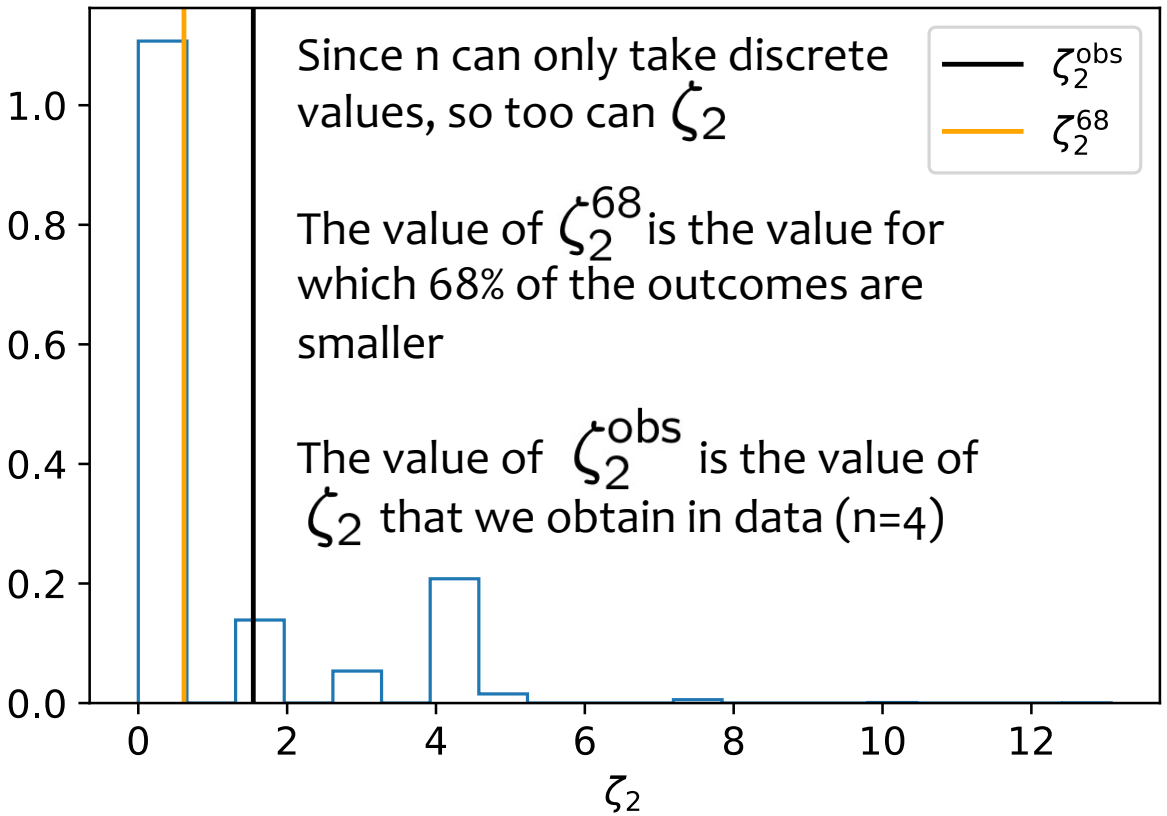
$$\zeta_\lambda = q(\lambda) - q(\hat{\lambda}) = -2\left(\ln(\lambda) - \lambda\right) + 2\left(\ln(n) + n\right)$$

The MLE for λ is given by

$$\frac{dq}{d\lambda}|_{\lambda=\hat{\lambda}} = 0 \implies \hat{\lambda} = n$$

Suppose we **observe n=4** and want to construct a 68% frequentist confidence interval in λ

e.g λ=2



Since n can only take discrete values, so too can ζ_2

The value of ζ_2^{68} is the value for which 68% of the outcomes are smaller

The value of ζ_2^{obs} is the value of ζ_2 that we obtain in data (n=4)

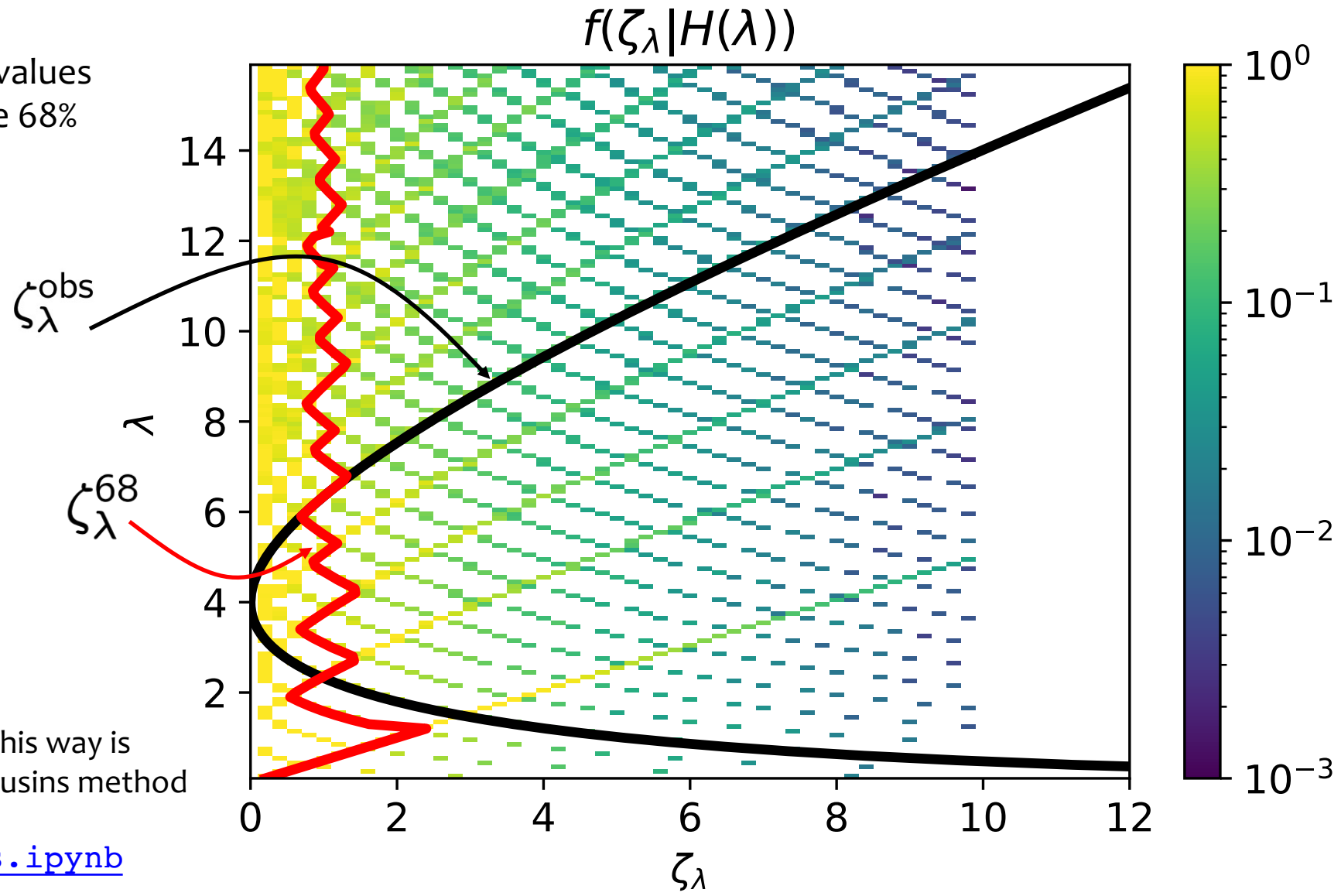
We can generate the distribution of ζ_λ by generating random values of n under a hypothesis H(λ) and calculating the test statistic ζ_λ

Example – Poisson mean

Repeating this for other values of λ allows us to build the 68% confidence region for λ

We include any value of λ for which

$$\zeta_\lambda^{\text{obs}} \leq \zeta_\lambda^{68}$$



Using the likelihood ratio in this way is what we call the Feldman-Cousins method

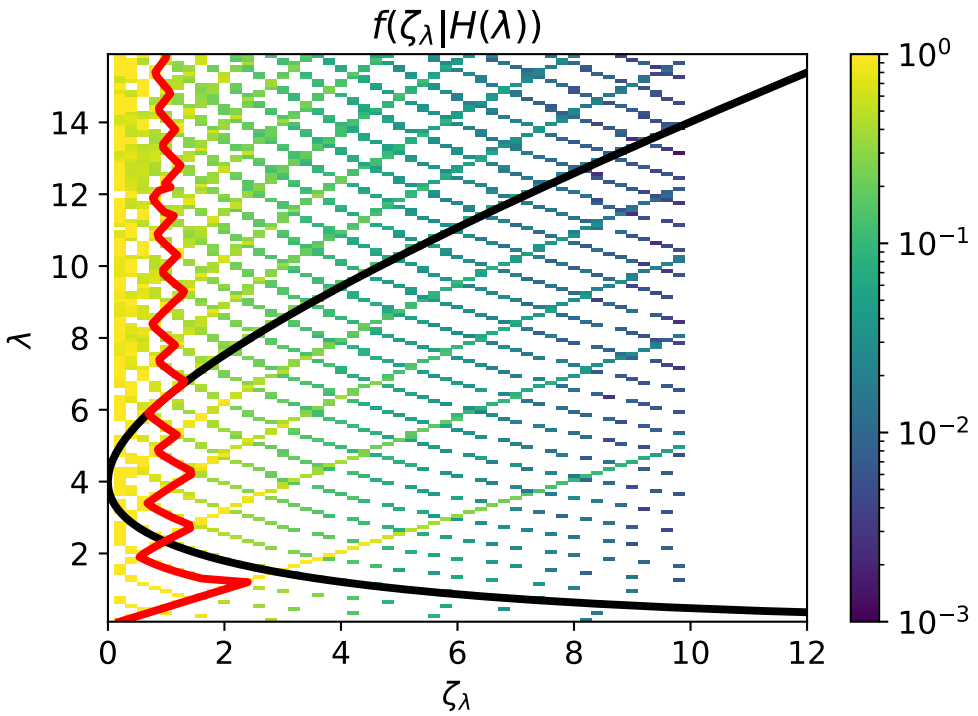
FrequentistIntervals.ipynb

Wilks' theorem

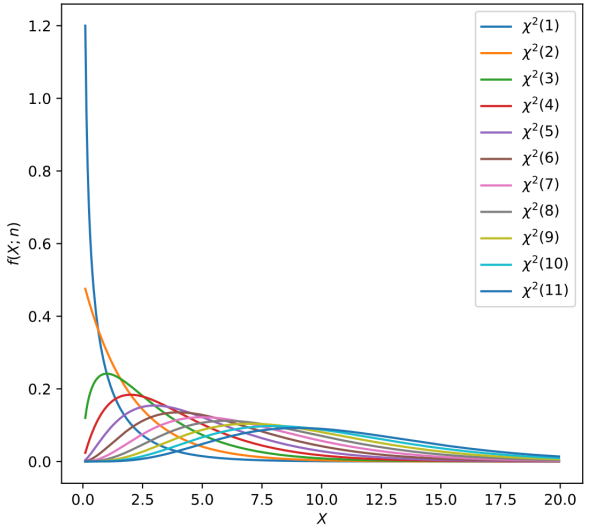
You might notice that as λ grows large, the value of ζ_λ^{68} starts to settle down to a constant value

This is due to Wilks' theorem:

T^hm: The probability distribution of ζ_θ under the hypothesis H(θ) approaches (converges in distribution) to a chi-square (χ^2) distribution with number of degrees of freedom equal to the length of the vector θ as the data sample size tends to infinity



$$f(\zeta_{\theta_1,\ldots,\theta_n}|H(\theta_1,\ldots,\theta_n)) \rightarrow \chi^2(\zeta_{\theta_1,\ldots,\theta_n};n)$$

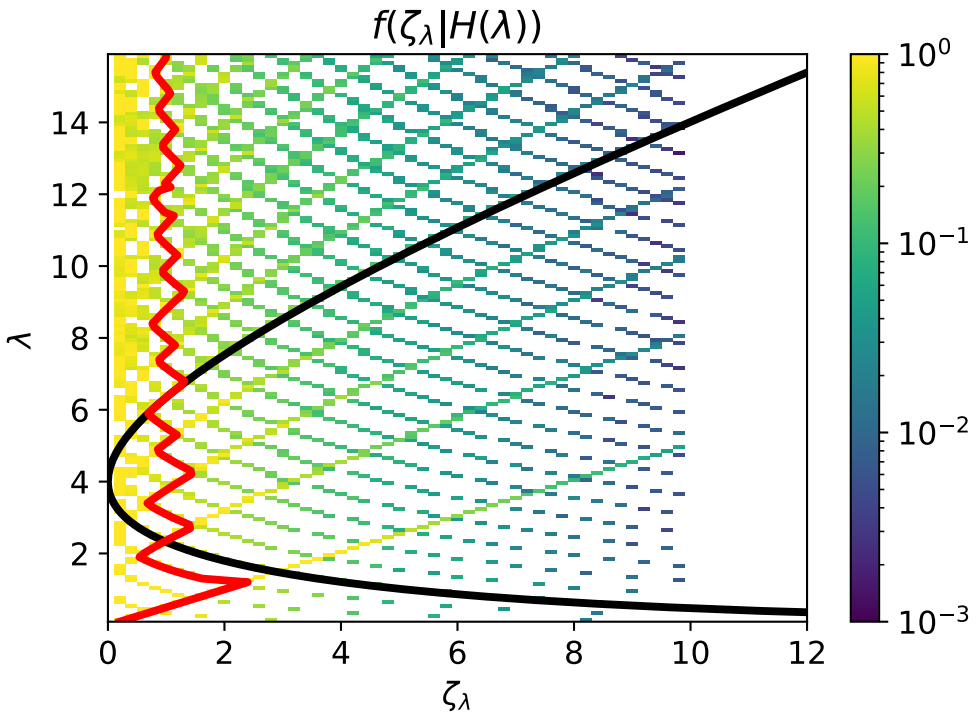


Wilks' theorem

You might notice that as λ grows large, the value of ζ_λ^{68} starts to settle down to a constant value

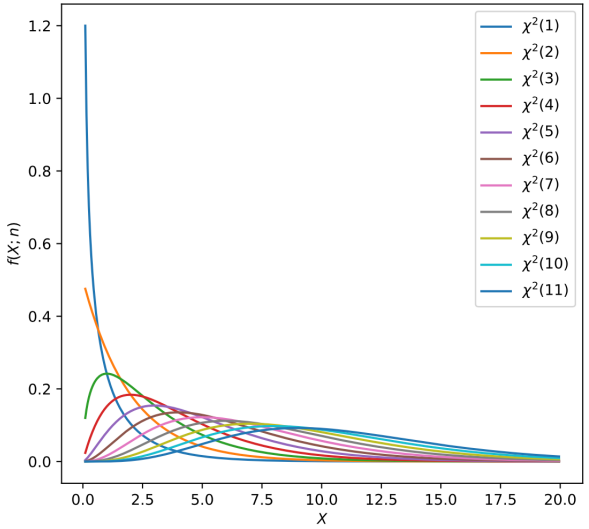
This is due to Wilks' theorem:

T^hm: The probability distribution of ζ_θ under the hypothesis H(θ) approaches (converges in distribution) to a chi-square (χ^2) distribution with number of degrees of freedom equal to the length of the vector θ as the data sample size tends to infinity



$$f(\zeta_{\theta_1,\ldots,\theta_n}|H(\theta_1,\ldots,\theta_n)) \rightarrow \chi^2(\zeta_{\theta_1,\ldots,\theta_n};n)$$

Then we can just read off the value of ζ_θ^{68}



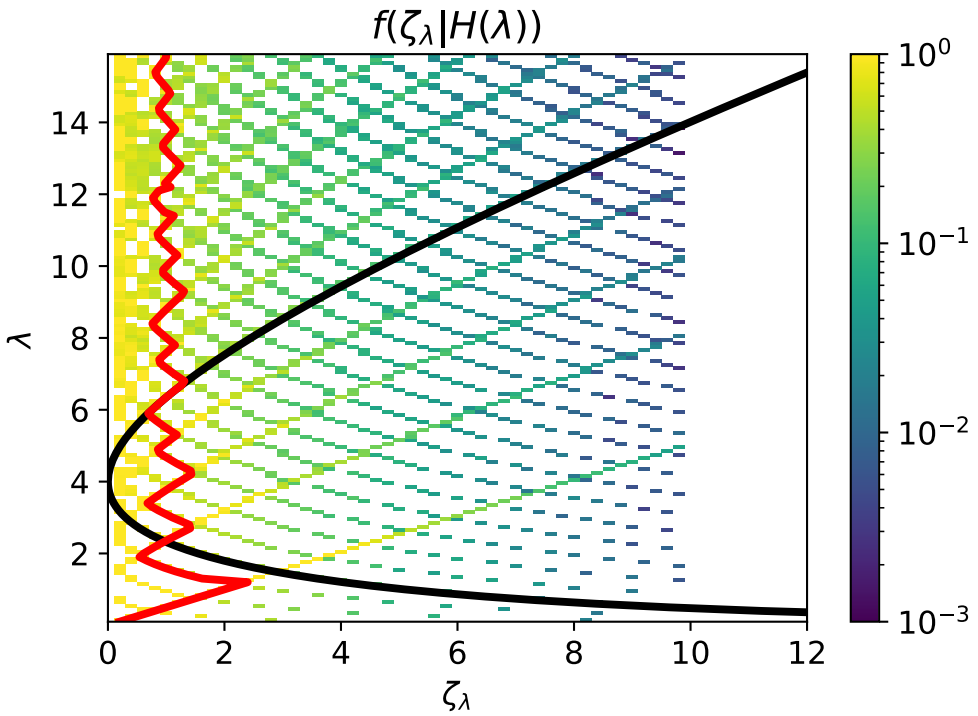
n	ζ_θ^{68}
1	~1
2	~2.3
3	~3.5
4	~4.7
5	~5.9

Wilks' theorem

You might notice that as λ grows large, the value of ζ_λ^{68} starts to settle down to a constant value

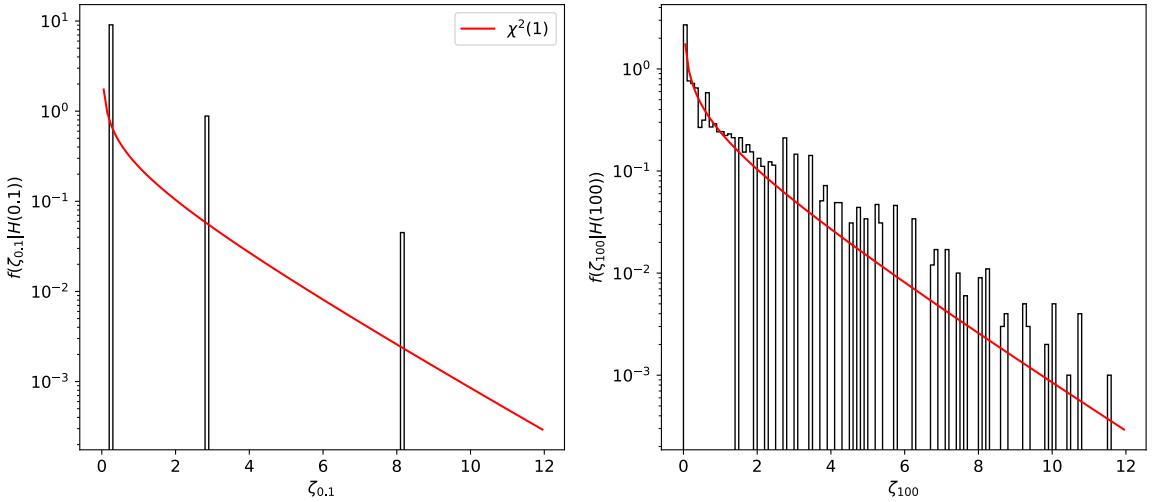
This is due to Wilks' theorem:

T^hm: The probability distribution of ζ_θ under the hypothesis H(θ) approaches (converges in distribution) to a chi-square (χ^2) distribution with number of degrees of freedom equal to the length of the vector θ as the data sample size tends to infinity



In the simple Poisson mean case, n=1 so we can compare the distribution of ζ_λ to a $\chi^2(1)$ and $\zeta_\lambda^{68} = 1$

Clearly the agreement is much better for large values of λ where the number of events will be large, than for small values.



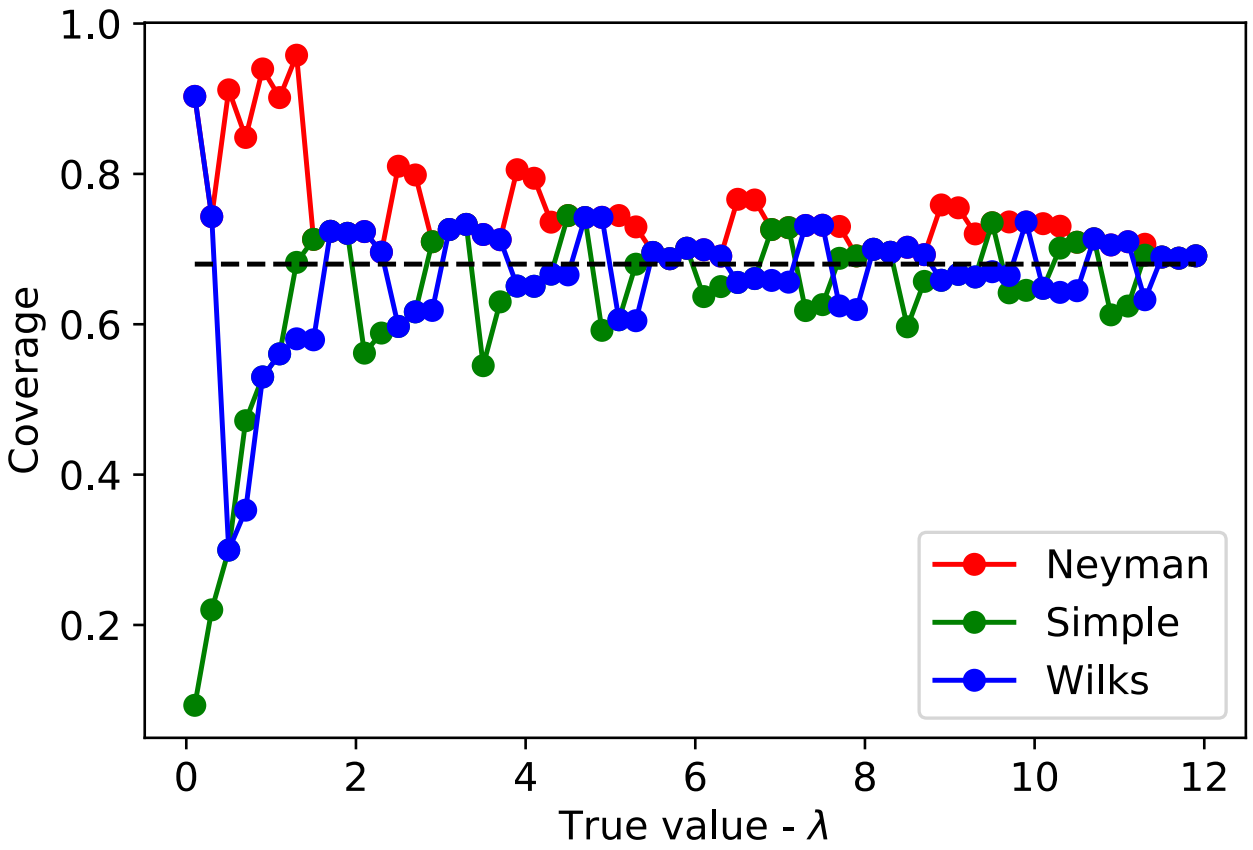
Coverage

We can calculate the probability that the true value of λ is contained in the confidence interval with the value of the pre-defined confidence level → this is called the "***coverage***" of our interval setting procedure

The Neyman construction overcovers where the discrete nature of the Poisson makes it hard to define ζ_λ^{68}

The 68% confidence intervals using Wilks' theorem converge as λ grows

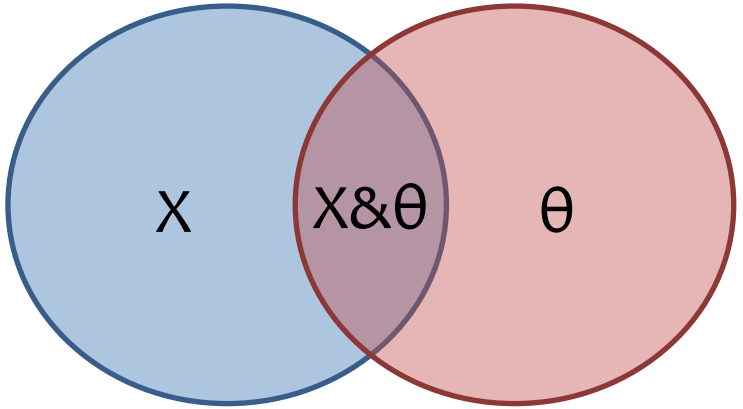
A simple estimate of a 68% confidence interval using the formula
$n_{\text{obs}} - \sqrt{n_{\text{obs}}} < \lambda < n_{\text{obs}} + \sqrt{n_{\text{obs}}}$
performs poorly at small λ



Bayesian Credible Intervals

In Bayesian statistics, we can use Bayes' rule to determine ***credible intervals***

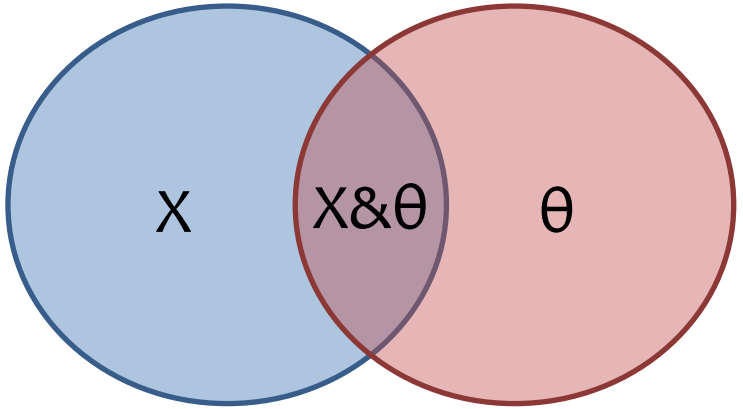
$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$



Bayesian Credible Intervals

In Bayesian statistics, we can use Bayes' rule to determine ***credible intervals***

$$p(\theta|X) = \frac{\boxed{p(X|\theta)}p(\theta)}{p(X)}$$

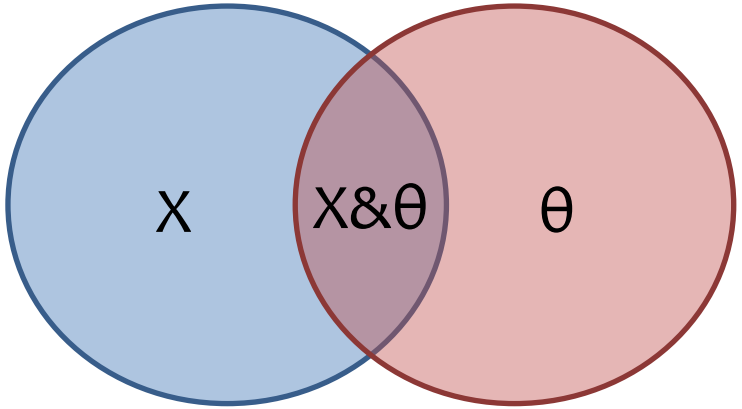


We've seen **this term** already – this is just the likelihood of the data (X) given the parameter value θ

Bayesian Credible Intervals

In Bayesian statistics, we can use Bayes' rule to determine ***credible intervals***

$$\boxed{p(\theta|X)} = \frac{\boxed{p(X|\theta)}\boxed{p(\theta)}}{p(X)}$$



We've seen **this term** already – this is just the likelihood of the data (X) given the parameter value θ

This term is called the prior distribution of θ and **this term** is the posterior distribution of θ

Bayesian Credible Intervals

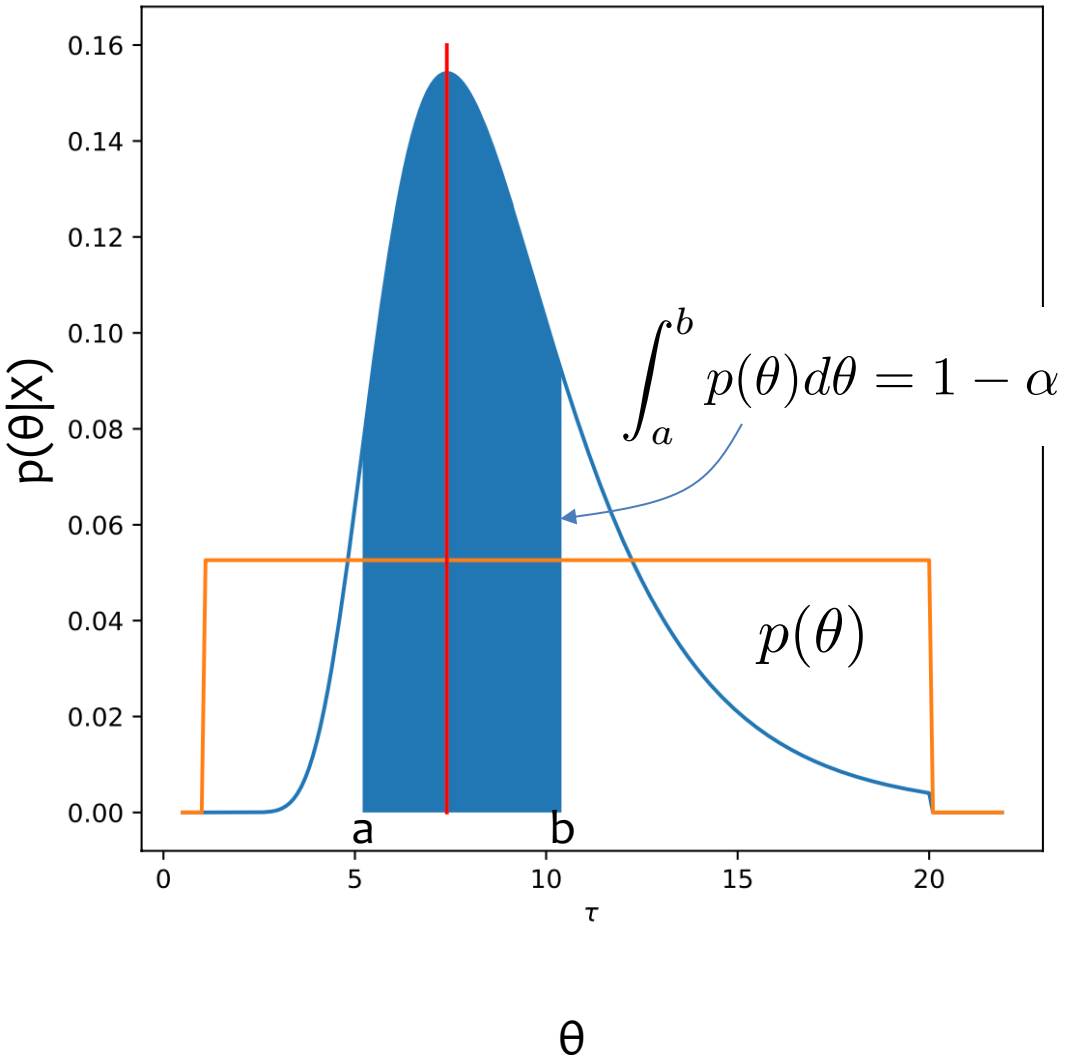
In Bayesian statistics, we can use Bayes' rule to determine ***credible intervals***

$$\boxed{p(\theta|X)} = \frac{\boxed{p(X|\theta)}\boxed{p(\theta)}}{p(X)}$$

We've seen **this term** already – this is just the likelihood of the data (X) given the parameter value θ

This term is called the prior distribution of θ and **this term** is the posterior distribution of θ

p(X) is a normalisation term $p(X) = \int p(X|\theta)p(\theta)d\theta$



A **100x(1-α)% credible interval** (a, b) in θ is the region of the posterior distribution that contains **1-α of the distribution. NOTE:** the result may depend on what we assume for our prior knowledge on θ

Systematic uncertainties

Systematic uncertainties come in roughly 3 groups*

- **The Good**: Your own calibrations → basically statistical
- **The Bad**: Using other peoples results, poorly modelled data or analysis technique, model assumptions, …
- **The Ugly**: Different theoretical estimates, theory with limited number of terms, …

2021 Phystat-Systematics workshop: https://indico.cern.ch/event/1051224/

**Pekka Sinervo – PHYSTAT 2003*

Systematic uncertainties - example

For the first type, both frequentist and Bayesian methods incorporate systematic uncertainties as ***nuisance parameters***

Let's take our simple counting experiment but this time our Poisson mean will be a function of two parameters + a constant background (B)

$$\lambda(\mu, \eta) = \mu\sigma(pp \to \mathrm{X})A\epsilon l_0 \cdot (1 + \kappa)^{\eta} + B$$

- **μ** – a signal strength to control the cross-section of some "signal process"
- **η** – our parameterised knowledge of luminosity (from a luminosity calibration measurement)

Systematic uncertainties - example

For the first type, both frequentist and Bayesian methods incorporate systematic uncertainties as ***nuisance parameters***

Let's take our simple counting experiment but this time our Poisson mean will be a function of two parameters + a constant background (B)

$$\lambda(\mu, \eta) = \mu\sigma(pp \to \mathrm{X})A\epsilon l_0 \cdot (1+\kappa)^{\eta} + B$$

- **μ** – a signal strength to control the cross-section of some "signal process"
- **η** – our parameterised knowledge of luminosity (from a luminosity calibration measurement)

Frequentist

$$q(\mu, \eta) = -2\ln L(\mu, \eta) = \eta^2 + 2\lambda(\mu, \eta) - 2n\ln\lambda(\mu, \eta)$$

Likelihood includes random variable η distributed as

$$\eta \sim \varphi(0,1) \rightarrow \hat{\eta}=0$$

Systematic uncertainties - example

For the first type, both frequentist and Bayesian methods incorporate systematic uncertainties as ***nuisance parameters***

Let's take our simple counting experiment but this time our Poisson mean will be a function of two parameters + a constant background (B)

$$\lambda(\mu, \eta) = \mu\sigma(pp \to \mathrm{X})A\epsilon l_0 \cdot (1+\kappa)^{\eta} + B$$

- **μ** – a signal strength to control the cross-section of some "signal process"
- **η** – our parameterised knowledge of luminosity (from a luminosity calibration measurement)

Frequentist

$$q(\mu, \eta) = -2\ln L(\mu, \eta) = \eta^2 + 2\lambda(\mu, \eta) - 2n\ln\lambda(\mu, \eta)$$

Likelihood includes random variable η distributed as

$\eta \sim \varphi(0,1) \rightarrow \hat{\eta}=0$

Bayesian

$$p(\mu, \eta|n) = \frac{p(n|\mu, \eta)p(\mu, \eta)}{p(n)} = \frac{p(n|\mu, \eta)p(\mu)p(\eta)}{p(n)}$$

Nuisance parameter has prior term

$p(\eta) = \varphi(0,1)$

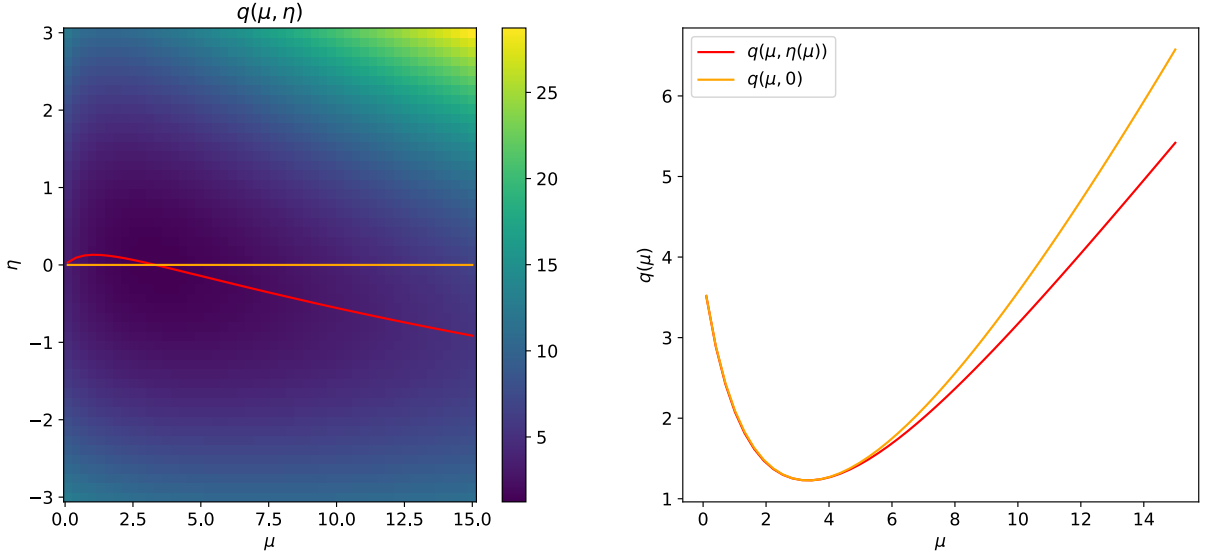
Systematic uncertainties - example

σ_{TH} = 0.01, A = 0.5, ε = 0.9, I_0 = 100, **k = 0.3**, B = 0.5, n = 2

We don't want to set confidence/credible intervals on η. We have two approaches to dealing with nuisance parameters

Frequentist: profile the nuisance parameters

$$q(\mu, \eta) \rightarrow q(\mu, \eta(\mu)) = q(\mu, \hat{\eta}_\mu)$$



ProfilingVsMarginalisation.ipynb

Systematic uncertainties - example

$\sigma_{TH} = 0.01$, $A = 0.5$, $\varepsilon = 0.9$, $I_0 = 100$, $\mathbf{k = 0.3}$, $B = 0.5$, $n = 2$

We don't want to set confidence/credible intervals on η. We have two approaches to dealing with nuisance parameters

Frequentist: profile the nuisance parameters

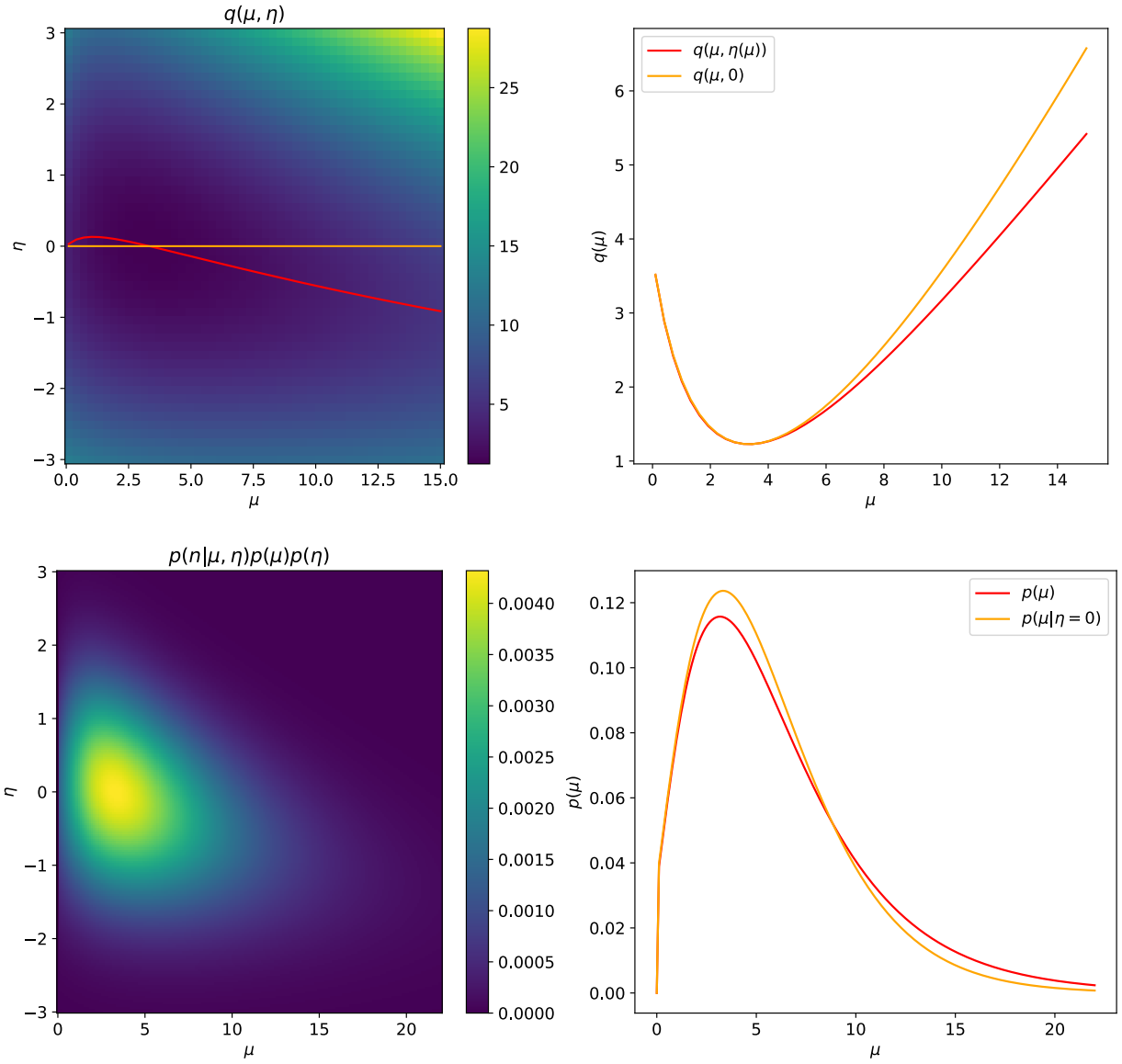
$$q(\mu, \eta) \rightarrow q(\mu, \eta(\mu)) = q(\mu, \hat{\eta}_\mu)$$

Bayesian: marginalize the probability distribution

$$p(\mu|n) = \int \frac{p(n|\mu, \eta)p(\mu)p(\eta)}{p(n)} d\eta$$

In both cases, the effect of the systematic uncertainty will be to widen the confidence/credible intervals we set
→ add additional uncertainty to our measurements

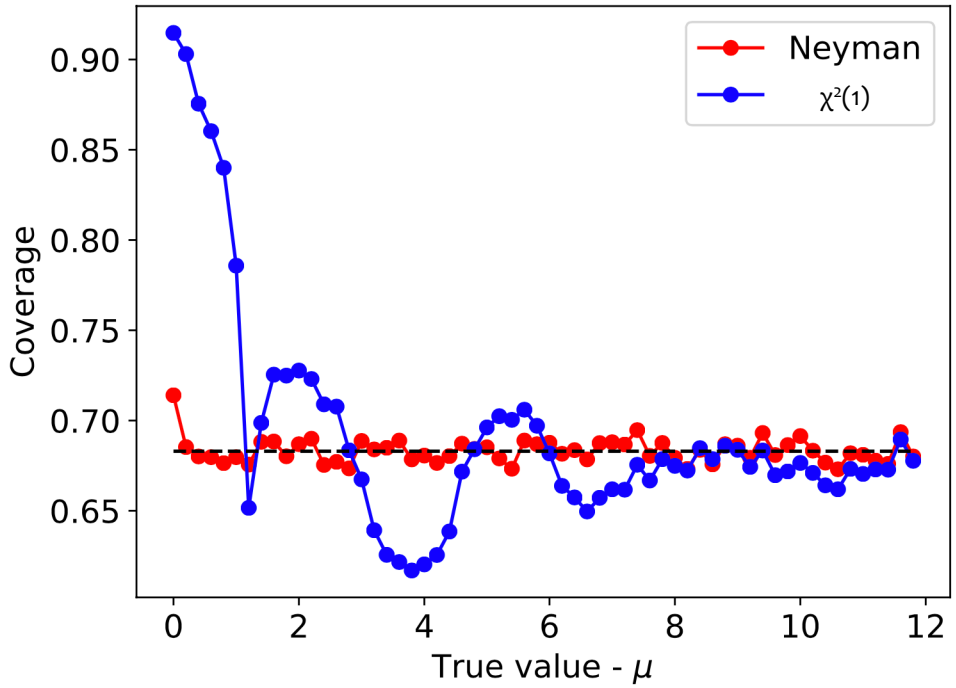
ProfilingVsMarginalisation.ipynb

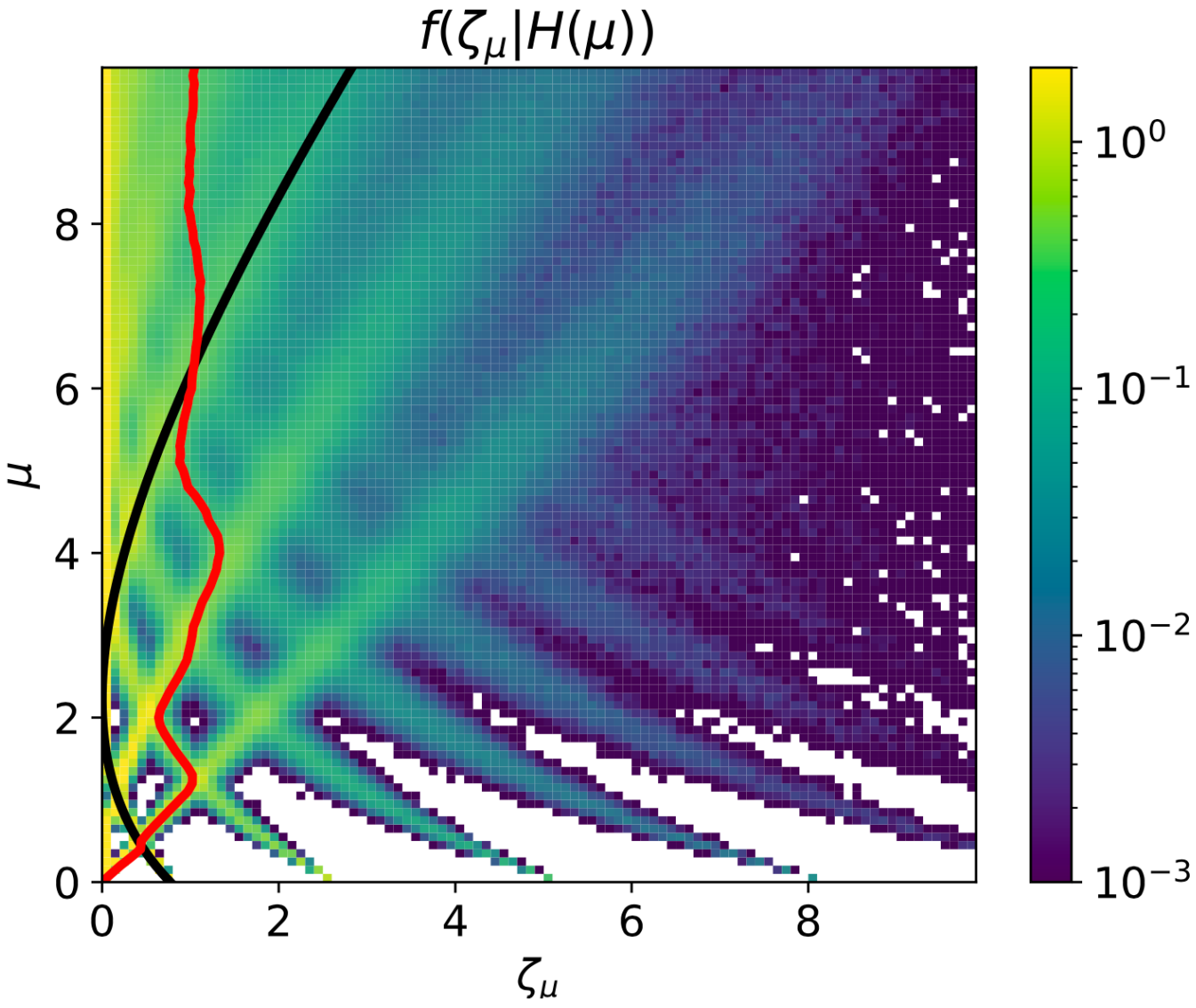


Systematic uncertainties - example

Measurements at the LHC can involve hundreds-thousands of nuisance parameters!

→ $\chi^2(n)$ still used in presence of nuisance parameters.
→ n = number of parameters of interest (doesn't include nuisance parameters)
→ Good idea to check coverage at profiled nuisance parameter values & nearby values





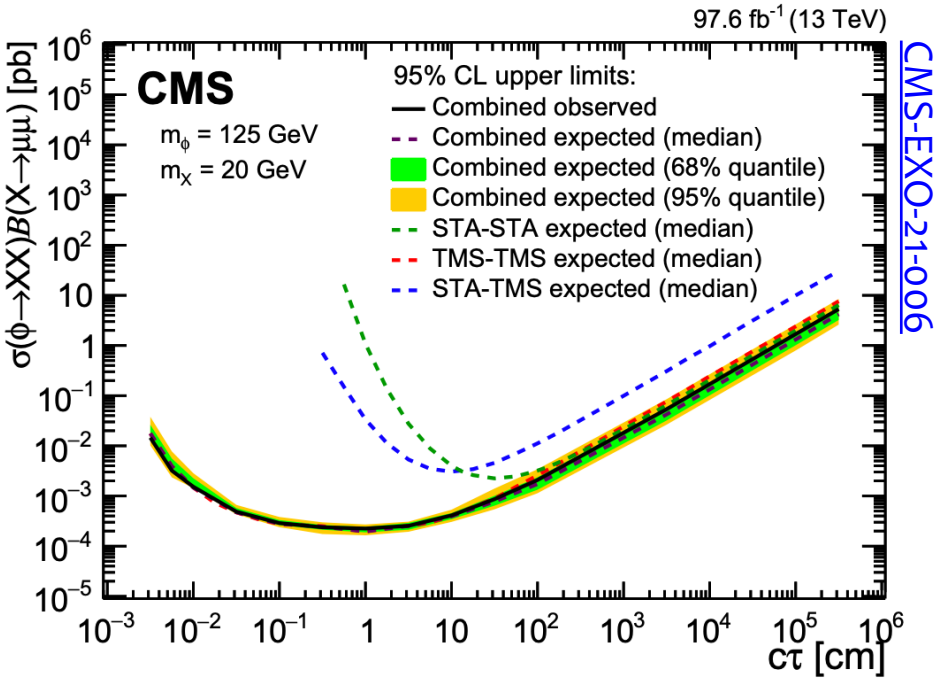
Upper limits at the LHC & CL_s

Most of the time (perhaps unfortunately) we set upper limits on cross-sections for BSM processes in searches for new physics

We use the modified test-statistic

$$t_\mu = \begin{cases} q(\mu, \hat{\eta}_\mu) - q(\hat{0}, \hat{\eta}_0) & \hat{\mu} < 0 \\ q(\mu, \hat{\eta}_\mu) - q(\hat{\mu}, \hat{\eta}) & \hat{\mu} \in (0, \mu] \\ 0 & \hat{\mu} > \mu, \end{cases}$$

To set upper limits (μ=0 always included). This test-statistic also has a well-known distribution in the limit of large data sets [1]



CMS-EXO-21-006

μ_{up} x $(\sigma.B)_{TH}$ = upper limit on (σ.B)

[1] Eur.Phys.J.C71:1554,2011

Upper limits at the LHC & CL_s

Most of the time (perhaps unfortunately) we set upper limits on cross-sections for BSM processes in searches for new physics

We use the modified test-statistic

$$t_\mu = \begin{cases} q(\mu, \hat{\eta}_\mu) - q(\hat{0}, \hat{\eta}_0) & \hat{\mu} < 0 \\ q(\mu, \hat{\eta}_\mu) - q(\hat{\mu}, \hat{\eta}) & \hat{\mu} \in (0, \mu] \\ 0 & \hat{\mu} > \mu, \end{cases}$$

To set upper limits (μ=0 always included). This test-statistic also has a well-known distribution in the limit of large data sets [1]

At the LHC, we typically use what is known as the CLs criterion[2] to avoid excluding the BSM signal hypothesis when the observed data don't support the background only (SM) either

We define $p_\mu = \int_{t_\mu^{\rm obs}}^{+\infty} f(t_\mu|H(\mu))dt_\mu$

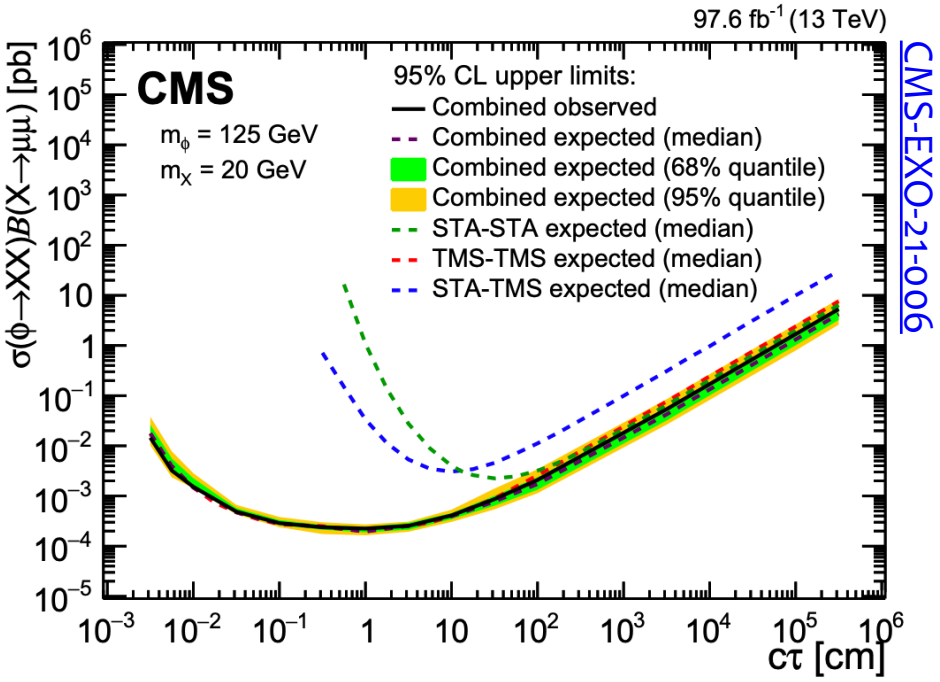
and

$$p_b = \int_0^{t_\mu^{\rm obs}} f(t_\mu|H(0))dt_\mu$$

$$CL_S = \frac{p_\mu}{1-p_b}$$

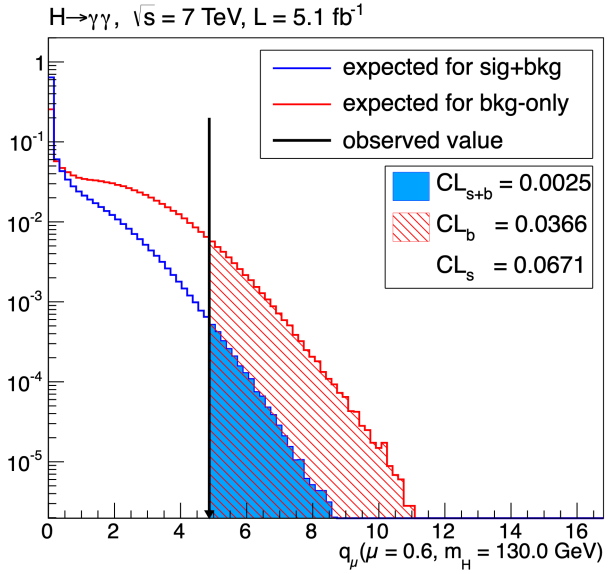
$\mu^{95\%}_{up}$ = smallest value of μ for which CLs < 0.05

[1] Eur.Phys.J.C71:1554,2011 [2] J.Phys.G 28 (2002) 2693-2704

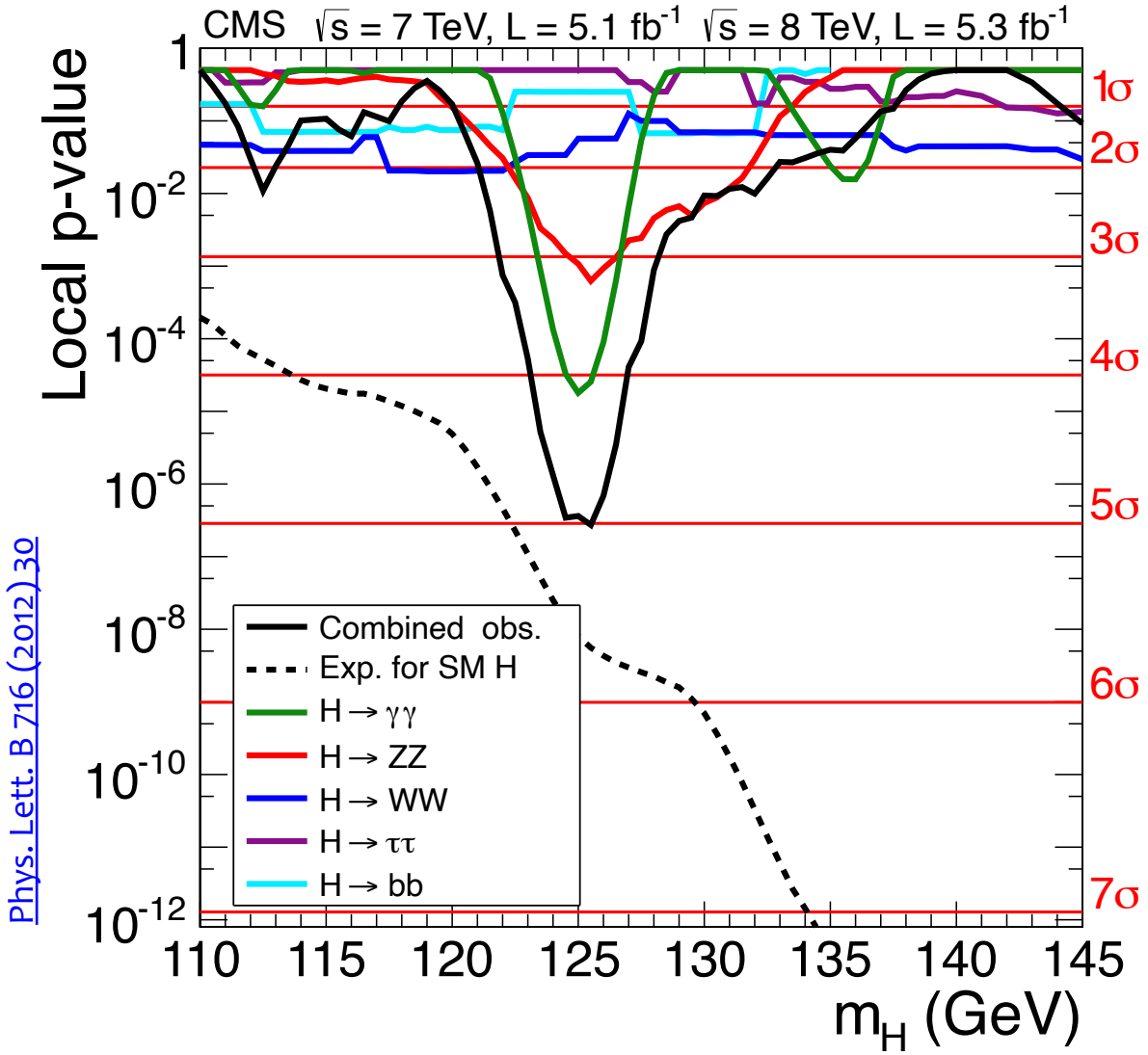


CMS-EXO-21-006

μ_{up} x $(\sigma.B)_{TH}$ = upper limit on (σ.B)



P-values (again) at the LHC



Phys. Lett. B 716 (2012) 30

We define the test-statistic for observation (as used in the Higgs discovery)

$$t_0 = \begin{cases} q(0, \hat{\eta}_0) - q(\hat{\mu}, \hat{\eta}) & \hat{\mu} > 0 \\ 0 & \hat{\mu} \leq 0, \end{cases}$$

Which similarly has a known distribution* in the limit of large data sample size, so that the p-value and corresponding significance can be calculated

$$p = \int_Q^{+\infty} \left(\frac{1}{2}\delta(0) + \frac{1}{2}\chi^2(X;1) \right) dX$$

$$\mathrm{Z} = \sqrt{Q}$$

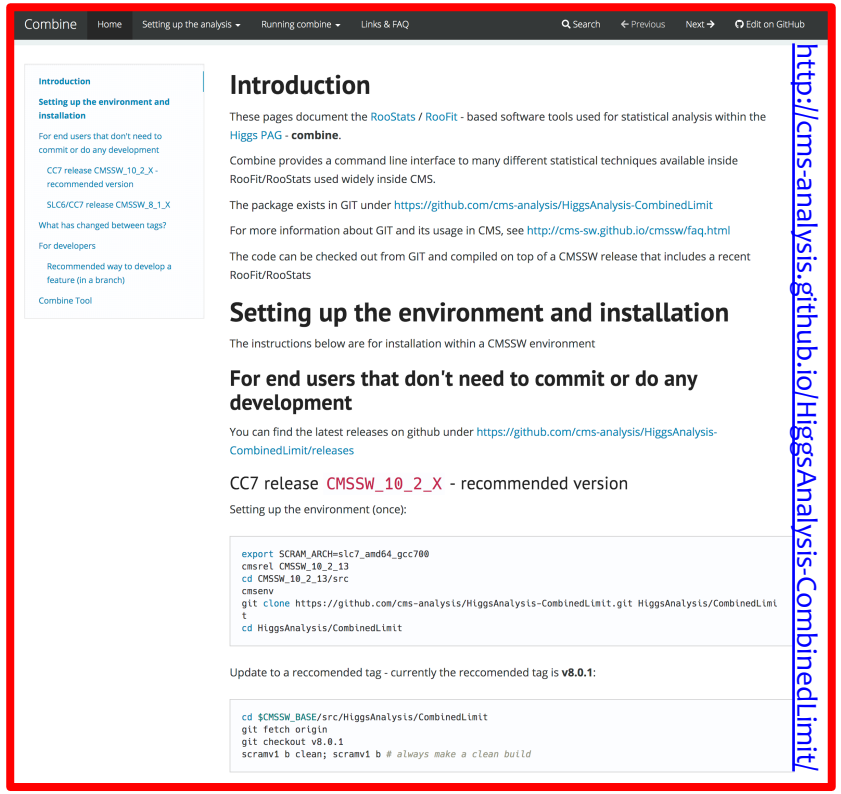
Again, there is a 1-1 correspondence between the p-value and the significance

* Eur.Phys.J.C71:1554,2011

LHC toolkits

Several excellent high-level tools to make **model building and running various statistical analyses** for LHC data simple!

No need for you to re-invent the wheel but I hope these lectures have helped with **understanding what they are doing under the hood!**

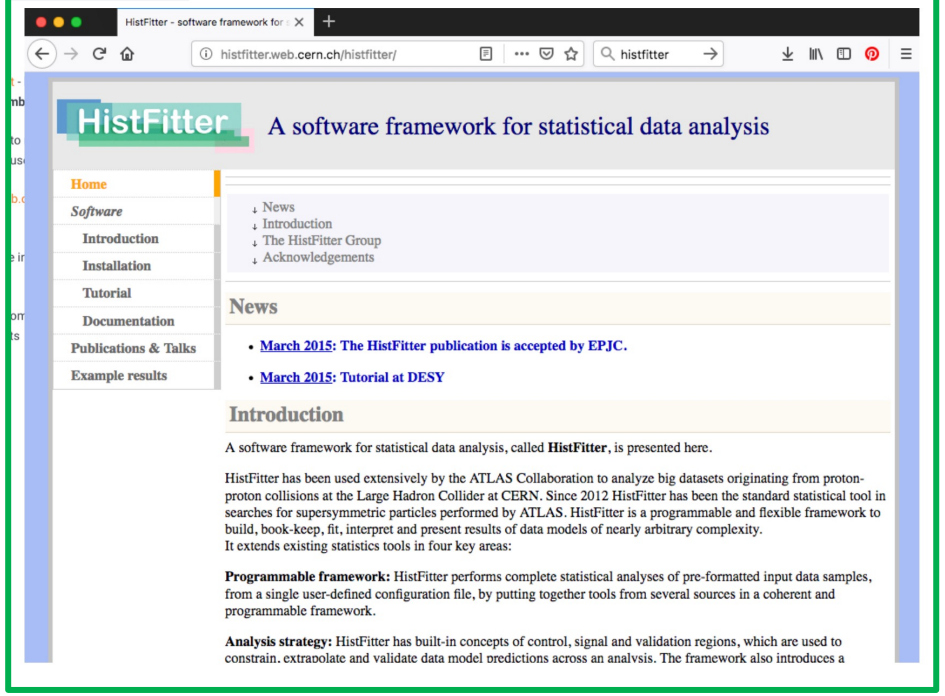


http://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/

https://github.com/scikit-hep/pyhf

https://github.com/zfit/zfit

http://histfitter.web.cern.ch/histfitter/



Summary

We've made it to the end of this (short) introduction to statistics!

We covered a (slightly biased) range of topics related to statistical inference (estimators, hypothesis testing and measurements) suitable for using Collider data analyses

At the least, I hope you have a better understanding of where some of these techniques come from

Probability & Probability distributions
- Frequentist probability & Bayes' theorem
- Basic properties of probability
- Probability distributions
- Law of large numbers & the Central Limit Theorem

Estimators
- Likelihoods & Maximum likelihood estimators
- Properties of MLEs

Hypothesis testing
- Type-I/II errors
- Simple and likelihood-based hypothesis tests & the Neyman-Pearson lemma
- P-values

Uncertainty & intervals
- Neyman intervals & Wilks' theorem
- Bayesian credible intervals
- Profiling and marginalizing nuisance parameters

Summary

Further reading:

- L. Lyons, N. Wardle, "***Statistical issues in searches for new phenomena in High Energy Physics***", Journal of Physics G: Nuclear and Particle Physics, Volume 45, Number 3.
- G. Cowan, "**Statistics" (section 39) in "Review of particle physics**", Chin. Phys. C 40, 100001 (2016).
- O. Behnke, K. Kroninger, G. Schott, T. Schorner-Sadenius, "***Data Analysis in High Energy Physics: A***
- ***Practical Guide to Statistical Methods***", ISBN: 978-3-527-41058-3 (2013).
- K. Cramner, "***Practical Statistics for the LHC***", Proceedings, 2011 European School of High-Energy Physics, (2011).
- G. Cowan, K. Cranmer, E. Gross, O. Vitells, "***Asymptotic formulae for likelihood-based tests of new physics***", EPJC, 71 (2011)
- G. J. Feldman, R. D. Cousins, "***A Unified approach to the classical statistical analysis of small signals***", Phys. Rev. D57 (1998).
- G. Cowan, "***Statistical Data Analysis***", ISBN: 978-0-198-50155-8 (1998).
- L. Lista, "***Statistical Methods for Data Analysis in Particle Physics***", ISBN 978-3-319-20176-4, (2015).
- F. James, "***Statistical Methods in Experimental Physics",*** ISBN: 978-9-812-70527-3 (2006).
- A. Stuart, K. Ord, S. Arnold, "***Kendall's Advanced theory of Statistics***", Vol 2A: Classical inference and the linear model, ISBN: 978-0-470-68924-0 (2010).
- Some practical examples in https://github.com/nucleosynthesis/IntroductionToStatistics

Backup

Compositions of probability distributions

Simple algebraic operations on random variables can yield non-intuitive distributions. Composition of random variables is not as simple as composition of numbers.

Example : at school you probably learned that if you measure two quantities X, Y with "errors*" σ_X and σ_Y, then the error on the sum X+Y is given by,

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

This comes from the fact that the sum of two Gaussian distributed random variables X and Y will be distributed as a Gaussian with a variance given by the sum of the variances of X and Y.

*we didn't even start to think about what this means yet …

Compositions of probability distributions

Suppose X ~ f(X) and Y ~ g(Y) are two independent random variables and U = X +Y. What is the probability distribution p(U) of U?

First, we change variables … $U = X + Y, \; V = X.$

The probability distribution h(U, V) is defined by, $h(U, V)dUdV = f(X)g(Y)dXdY,$

Compositions of probability distributions

Suppose X ~ f(X) and Y ~ g(Y) are two independent random variables and U = X +Y. What is the probability distribution p(U) of U?

First, we change variables … $U = X + Y, \; V = X.$

The probability distribution h(U, V) is defined by, $h(U, V)dUdV = f(X)g(Y)dXdY,$

Probability (not necessarily the density however) must be invariant under a change of variables.

For a change of variables, the infinitesimal volume dXdY changes as

$$dXdY = \left| \det \begin{bmatrix} \frac{\partial U}{\partial X} & \frac{\partial U}{\partial Y} \\ \frac{\partial V}{\partial X} & \frac{\partial V}{\partial Y} \end{bmatrix} \right|^{-1} dUdV = dUdV$$

Compositions of probability distributions

$$h(U,V)dUdV = f(V)g(U-V)dUdV \implies h(U,V) = f(V)g(U-V)$$

To obtain the marginal density p(U), we integrate over V

$$p(U) = \int_{-\infty}^{+\infty} h(U,V)dV = \int_{-\infty}^{+\infty} f(U)g(U-V)dV.$$

Compositions of probability distributions

$$h(U,V)dUdV = f(V)g(U-V)dUdV \implies h(U,V) = f(V)g(U-V)$$

To obtain the marginal density p(U), we integrate over V

$$p(U) = \int_{-\infty}^{+\infty} h(U,V)dV = \int_{-\infty}^{+\infty} f(V)g(U-V)dV.$$

If we plug in $X \sim \phi(X;\mu_X,\sigma_X)$ h $Y \sim \phi(Y;\mu_Y,\sigma_Y)$

$$p(U) = \frac{1}{\sqrt{2\pi}\sigma_U}\exp\left[-\frac{(U-(\mu_X+\mu_Y)^2)}{2\sigma_U^2}\right]$$

Compositions of probability distributions

Suppose now we want to take the ratio of two random variables U = X/Y. Again, at school you learn…

$$\frac{\sigma_U}{U} = \sqrt{\left(\frac{\sigma_X}{X}\right)^2 + \left(\frac{\sigma_Y}{Y}\right)^2}$$

Compositions of probability distributions

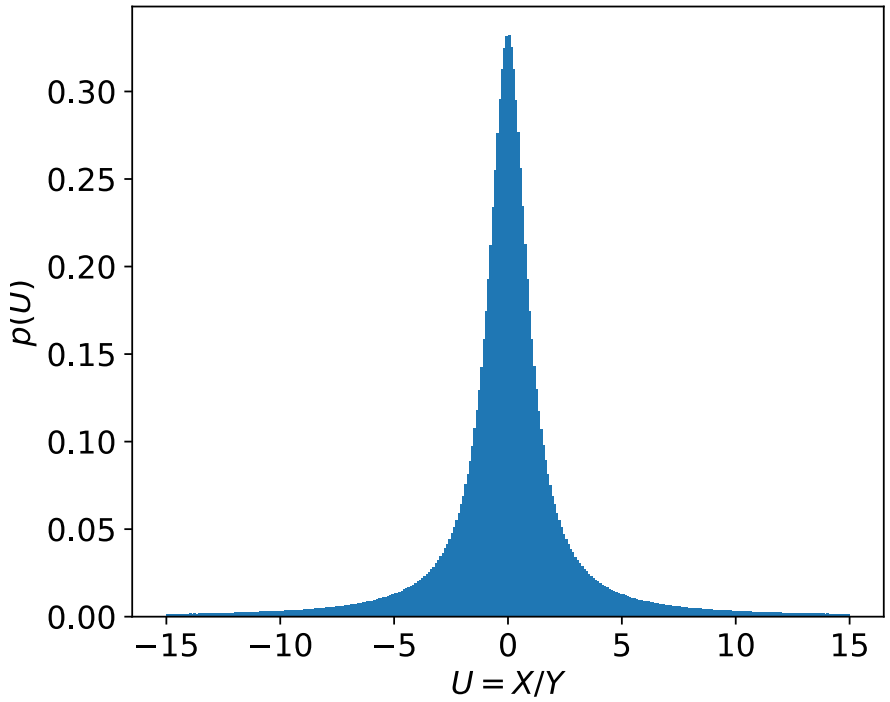
Suppose now we want to take the ratio of two random variables U = X/Y. Again, at school you learn…

$$\cancel{\frac{\sigma_U}{U} = \sqrt{\left(\frac{\sigma_X}{X}\right)^2 + \left(\frac{\sigma_Y}{Y}\right)^2}}$$

If $X \sim \phi(X; 0, 1) \quad Y \sim \phi(Y; 0, 1)$
then,

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

This is a *Cauchy distribution* which has an ill-defined mean and infinite variance!



Convergence in Distribution

Convergence in distribution (in English):

Consider a sequence of random variables $\{X_1, X_2, ..., X_n\}$ ($X_i \varepsilon R$) with cumulative distribution functions $\{F_1(x), F_2(x), ..., F_n(x)\}$. The sequence X_n converges in distribution as $n \rightarrow \infty$ to X, with cumulative distribution **F** if for every point $x \varepsilon R$ where F(x) is continuous,

$$\lim_{n\rightarrow\infty} F_n(x) = F(x)$$

Where F(x) is the CDF of X.

Convergence in Distribution

Convergence in distribution (in English):

A set of random variables $X_1, X_2, \ldots, X_n \sim f_n(x)$ converging to a random variable $X \sim f(x)$ where $\lim_{n\to\infty} F_n(x) = F(x)$ means that the distributions of X_n will get closer to the distribution of X.

Example: Convergence in Distribution

Let $X_n \sim \varphi(X; 0, 1/n)$ for each n, where φ denotes a Gaussian distribution. What is the limiting behaviour as $n \to \infty$?

Example: Convergence in Distribution

Let $X_n \sim \varphi(X; 0, 1/n)$ for each n, where φ denotes a Gaussian distribution. What is the limiting behaviour as $n \to \infty$?

The cumulative distribution function at each n for any point X is given by

$$F_n(X) = \int_{-\infty}^{X} \frac{n}{\sqrt{2\pi}} e^{-\frac{t^2 n^2}{2}} \, dt$$

Example: Convergence in Distribution

Let $X_n \sim \varphi(X; 0, 1/n)$ for each n, where φ denotes a Gaussian distribution. What is the limiting behaviour as $n \to \infty$?

The cumulative distribution function at each n for any point X is given by

$$F_n(X) = \int_{-\infty}^{X} \frac{n}{\sqrt{2\pi}} e^{-\frac{t^2 n^2}{2}} dt$$

Which can be shown to be*

$$F_n(X) = \begin{cases} \frac{1}{2} + \left[\frac{1}{2}\mathrm{erf}\left(\frac{\mathrm{xn}}{\sqrt{2}}\right)\right]_0^X & X > 0 \\ 1 - \frac{1}{2} + \left[\frac{1}{2}\mathrm{erf}\left(\frac{\mathrm{xn}}{\sqrt{2}}\right)\right]_0^X & X < 0, \end{cases}$$

* Feel free to do so as an exercise

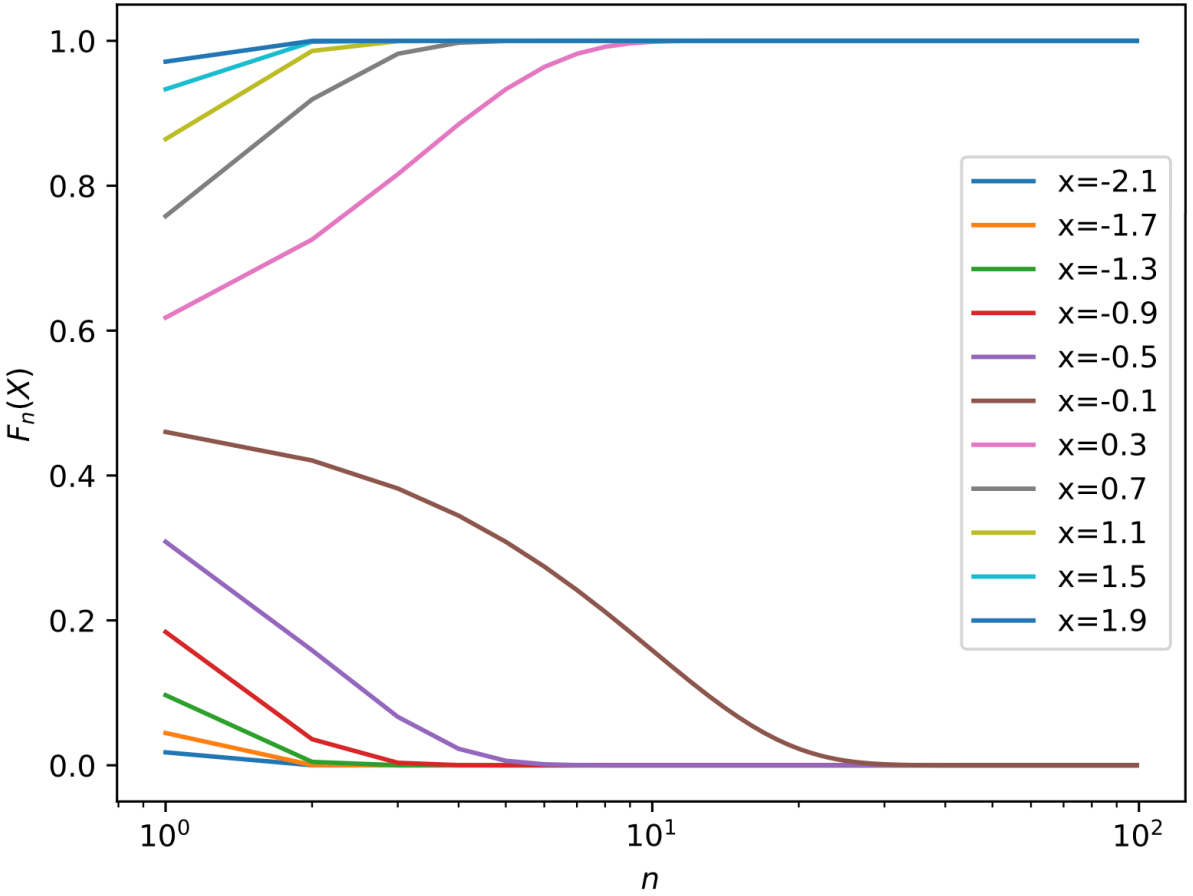
Example: Convergence in Distribution

Let $X_n \sim \varphi(X; 0, 1/n)$ for each n, where φ denotes a Gaussian distribution. What is the limiting behaviour as $n \to \infty$?

Its easy to see that for large n, we end up with

$$F_n(X) \to F(X) = \begin{cases} 1 & X > 0 \\ 0 & X < 0, \end{cases}$$

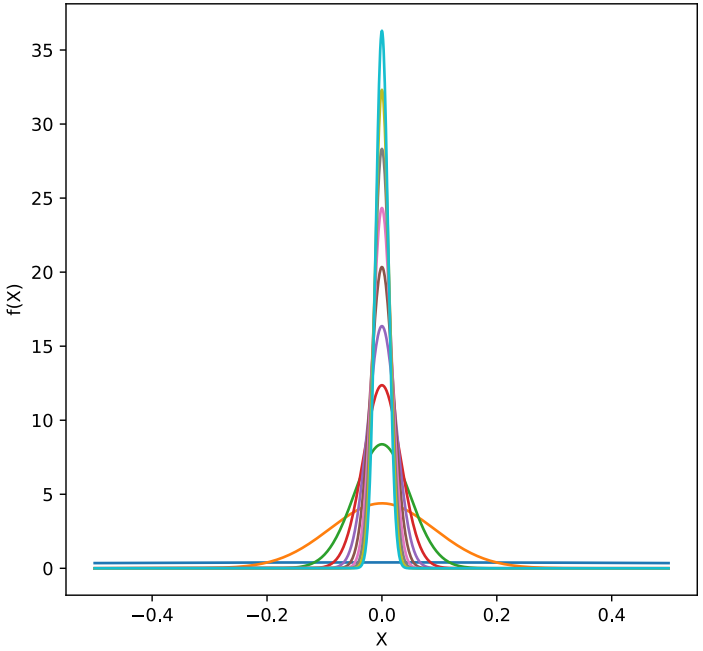
Which is the definition of the dirac-delta function!

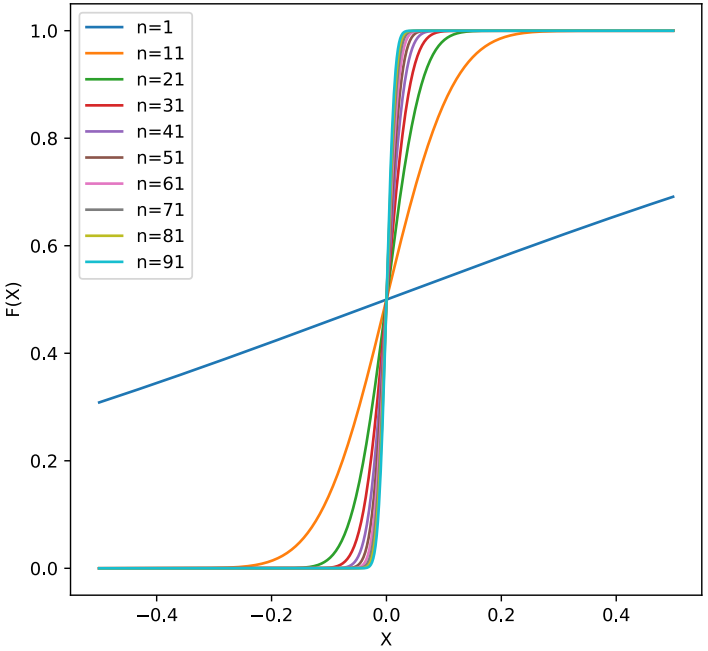


Example: Convergence in Distribution

Let $X_n \sim \varphi(X; 0, 1/n)$ for each n, where φ denotes a Gaussian distribution. What is the limiting behaviour as $n \to \infty$?

This could even have been obvious by just plotting the distributions…





Convergence in Probability

Convergence in probability:

We say that the sequence of random variables $\{X_1, X_2, \ldots, X_n\}$, converges in probability to X if for any $\varepsilon > 0$ and $\delta > 0$, there exists a value of N such that,

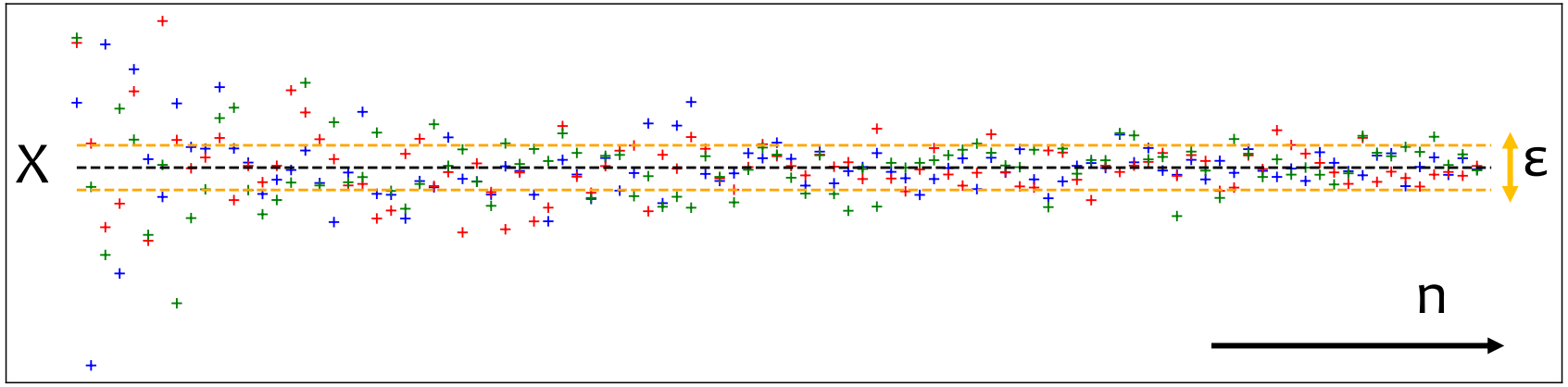
$$P(|X_n - X| > \epsilon) < \delta,$$

for all $n \geq N$.

Convergence in Probability

Convergence in probability (in English):

A sequence of random variables $\{X_1, X_2, \ldots, X_n\}$, converges in probability to X if eventually the probability that X_n is far from X gets vanishingly small.



This definition of convergence is stronger than convergence in distribution since nothing about the distance between X and X_n is implied in the other definition. Convergence in probability implies convergence in distribution.

The law of large numbers

The most important application of convergence is in the **law of large numbers**, which concerns the convergence of the sample mean.

Let $\{X_1,\ldots X_N\}$ be a sequence of independent random variables, each having the same mean μ_1 and variances ν_2^i . We determine the sample mean,

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Then if $\lim_{N\to\infty}\left[\frac{1}{N^2}\sum_{i=1}^{N}\nu_2^i\right] = 0$ then $\bar{X}$ converges in probability to μ_1

The Central Limit Theorem

Perhaps one of the most famous results in statistics (and probably one of the most useful) is the **Central Limit Theorem.** The central limit theorem (CLT) states how the sum of independent random variables (for example sample means) is distributed in the limit of large N.

Suppose we have a sequence of independent random variables X_i, each from a distribution with mean μ^i_1 and variance ν^i_2. Now let,

$$T_N = \frac{\bar{X}N - \sum_{i=1}^{N} \mu_1^i}{\sqrt{\sum_{i=1}^{N} \nu_2^i}}.$$

Then for $T = \lim_{N\to\infty} T_N$, we have that $T \sim \varphi(T;0,1) \rightarrow T_N$ converges in distribution to a standard normal distribution

Proof of simplified CLT

Example: Let $\{X_1, ..., X_N\}$ be independent and identically distributed random variables with $E(X) = 0$ and $V(X) = \nu \in \mathbb{R}$. Then,

$$T_N = \frac{X_1 + ... + X_N}{\sqrt{\nu N}} = \sum_{i=1}^{N} \frac{Y_i}{\sqrt{N}}.$$

where $Y_i = X_i/\sqrt{\nu}$. We can say that the characteristic function of T_N is given by,

$$\varphi_{T_N} = \varphi_{Y_1}\left(\frac{t}{\sqrt{N}}\right) \cdot \varphi_{Y_2}\left(\frac{t}{\sqrt{N}}\right) \cdot ... \cdot \varphi_{Y_N}\left(\frac{t}{\sqrt{N}}\right) = \left(\phi_{Y_1}\left(\frac{t}{\sqrt{N}}\right)\right)^N$$

from Eqn. 79 and the linearity property of the expectation. Recall that the characteristic function is given by the expectation so,

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{N}}\right) = E\left[e^{i\frac{t}{\sqrt{N}}Y_1}\right] = E\left[\sum_{r=0}^{\infty} \frac{\left(i\frac{t}{\sqrt{N}}Y_1\right)^r}{r!}\right] = \sum_{r=0}^{\infty} \frac{\left(i\frac{t}{\sqrt{N}}\right)^r}{r!} E\left[Y_1^r\right]$$

where we have Taylor expanded the exponential. Now, we know that the mean of Y_1 is 0 (since also $\mu = 0$) and the variance of Y_1 will be the same as the variance of X_1 divided by itself, hence $E\left[Y^2\right] = V(Y_1) = 1$. So we have,

$$\varphi_{Y_1}\left(\frac{t}{\sqrt{N}}\right) = 1 - \frac{t^2}{2N} + \mathcal{O}\left(\frac{t^3}{N^{\frac{3}{2}}}\right).$$

Substituting this back into Eqn. 81, we have that as $N \to \infty$, $\varphi_{T_N} \to e^{-\frac{1}{2}t^2}$, which is the characteristic function of a standard normal distribution. From the Paul Levy theorem, this means that the limit of T_N, T, is distributed as a standard normal distribution – $T \sim \phi(T; 0, 1)$.

Invariance of MLEs

Two (useful) properties of maximum likelihood estimators are immediately clear from the definition:

1. Maximum likelihood estimators are invariant under bijective transformations of variables

2. Relationships between parameters being estimated are preserved in Maximum likelihood estimators

These are rather straight forward to show…

Invariance of MLEs

1. Maximum likelihood estimators are invariant under bijective transformations of variables

Recall that under a change of variables X → Y = g(X)

$$f_Y(Y)dY = f_X(X)dX \implies f_Y(Y) = \frac{f_X(X)}{|g'(X)|}$$

Invariance of MLEs

1. Maximum likelihood estimators are invariant under bijective transformations of variables

Recall that under a change of variables $X \rightarrow Y = g(X)$

$$f_Y(Y)dY = f_X(X)dX \implies f_Y(Y) = \frac{f_X(X)}{\boxed{|g'(X)|}}$$

Since the likelihood is equal to the probability density for a random variable, the maximum of it (or minimum of its log) will not change when multiplying (adding) by a **constant**!

Invariance of MLEs

2. Relationships between parameters being estimated are preserved in Maximum likelihood estimators

Suppose $\hat{\theta}$ is the maximum likelihood estimate for θ and $\alpha = g(\theta)$ is some function of θ, then the maximum likelihood estimator of α will be,

$$\hat{\alpha} = g(\hat{\theta}).$$

This is obvious since

$$0 = \frac{\partial L}{\partial \theta}|_{\theta=\hat{\theta}} = \frac{\partial L}{\partial \alpha}\frac{dg}{d\theta}|_{\alpha=\hat{\alpha}} = \frac{\partial L}{\partial \alpha}g'(\alpha)|_{\alpha=\hat{\alpha}} \implies \frac{\partial L}{\partial \alpha}|_{\alpha=\hat{\alpha}} = 0$$

Invariance of MLEs

Note that this is not universally true for estimators.

For example, suppose I was using a method of moments and estimating the first moment of a probability distribution (μ) using the sample mean with some random variable X.

The square of the mean is given by $g(\mu) = \mu^2$

while generally $E[X^2] \neq (E[X])^2$

so in this case, the relationship between parameters is not preserved in the estimators.

What makes a "good" estimator?

It's clear that the two estimators we've discussed so far will have advantages and disadvantages, which applications they are suited for, what the ultimate goal for the estimate is and even in terms of their (numerical) implementation.

Statisticians however usually have the following 3 criteria in mind when deciding whether that an estimator is "***good***"

- **Unbiased** → The expectation of the estimator agrees with the true value
- **Consistent** → The value of the estimator converges to the true value as more data are included
- **Efficient** → The variance of the estimator is as small as theoretically possible

Bias of an estimator

The bias of an estimator is defined by the difference between the expectation of the estimator and the true parameter value:

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

We say that an estimator is unbiased if the bias is zero.

Bias of an estimator

The bias of an estimator is defined by the difference between the expectation of the estimator and the true parameter value:

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

We say that an estimator is unbiased if the bias is zero.

In the method of moments for the mean and standard deviation (eg when estimating the parameters of a Gaussian distribution), we find that the estimator for the mean is ***unbiased:***

$$E[\hat{\mu}] = E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu = \mu$$

Linearity of E[.]

Bias of an estimator

The estimator for σ however ***is biased.*** We can see this by first considering the expectation of the estimator $\hat{\sigma}^2 = \bar{V}$

$$E\left[\hat{\sigma}^2\right] = E\left[\bar{V}\right] \quad = \quad E\left[\frac{1}{n}\sum_i (X_i - \bar{X})^2\right]$$

By the def^n of the sample variance

Bias of an estimator

The estimator for σ however ***is biased.*** We can see this by first considering the expectation of the estimator $\hat{\sigma}^2 = \bar{V}$

$$E\left[\hat{\sigma}^2\right] = E\left[\bar{V}\right] = E\left[\frac{1}{n}\sum_i (X_i - \bar{X})^2\right]$$

By the def^n of the sample variance

$$= E\left[\frac{1}{n}\sum_i \left((X_i - \mu) - (\bar{X} - \mu)\right)^2\right]$$

$$= E\left[\frac{1}{n}\sum_i \left((X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu)\right)\right]$$

Bias of an estimator

$$
\begin{aligned}
E\left[\hat{\sigma}^2\right] = E\left[\bar{V}\right] &= E\left[\frac{1}{n}\sum_i \left((X_i-\mu)^2 + (\bar{X}-\mu)^2 - 2(X_i-\mu)(\bar{X}-\mu)\right)\right] \\
&= E\left[\frac{1}{n}\sum_i (X_i-\mu)^2 + (\bar{X}-\mu)^2 - \frac{2}{n}\sum_i (X_i-\mu)(\bar{X}-\mu)\right]
\end{aligned}
$$

$\frac{1}{n}\sum_{i=1}^{n} 1 = 1$

Bias of an estimator

$$
\begin{aligned}
E\left[\hat{\sigma}^2\right] = E\left[\bar{V}\right] &= E\left[\frac{1}{n}\sum_i \left((X_i-\mu)^2 + (\bar{X}-\mu)^2 - 2(X_i-\mu)(\bar{X}-\mu)\right)\right] \\
&= E\left[\frac{1}{n}\sum_i (X_i-\mu)^2 + (\bar{X}-\mu)^2 - \frac{2}{n}\sum_i (X_i-\mu)(\bar{X}-\mu)\right] \\
&= E\left[\frac{1}{n}\sum_i (X_i-\mu)^2 + (\bar{X}-\mu)^2 - 2(\bar{X}-\mu)^2\right] \\
&= E\left[\frac{1}{n}\sum_i (X_i-\mu)^2 - (\bar{X}-\mu)^2\right]
\end{aligned}
$$

$\frac{1}{n}\sum_{i=1}^{n} 1 = 1$

$\bar{X}-\mu = \frac{1}{n}\sum_i (X_i-\mu)$

Bias of an estimator

$$E\left[\hat{\sigma}^2\right] = E\left[\bar{V}\right] = E\left[\frac{1}{n}\sum_i (X_i - \mu)^2 - (\bar{X} - \mu)^2\right]$$

Linearity of E[]

$$= E\left[\frac{1}{n}\sum_i (X_i - \mu)^2\right] - E\left[(\bar{X} - \mu)^2\right]$$

Bias of an estimator

$$
\begin{aligned}
E\left[\hat{\sigma}^2\right] = E\left[\bar{V}\right] &= E\left[\frac{1}{n}\sum_i (X_i-\mu)^2 - (\bar{X}-\mu)^2\right] \\
&= E\left[\frac{1}{n}\sum_i (X_i-\mu)^2\right] - E\left[(\bar{X}-\mu)^2\right] \\
&= \frac{1}{n}\sum_i E\left[(X_i-\mu)^2\right] - V(\bar{X}) \\
&= \frac{n}{n}V(X) - V(\bar{X}) \\
&= \sigma^2 - \frac{1}{n}\sigma^2 < \sigma^2.
\end{aligned}
$$

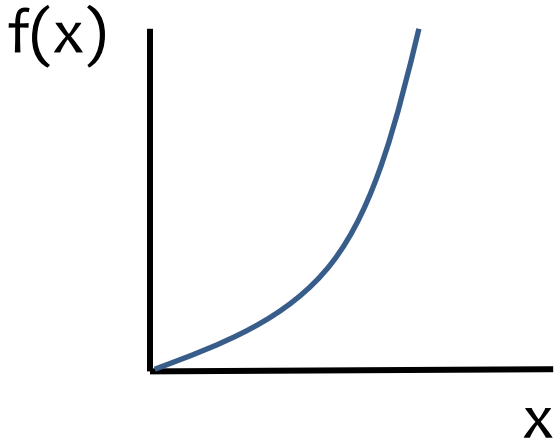
Linearity of E[]

$V(\frac{1}{n}X) = \frac{1}{n^2}V(X)$ & $V(\sum_i X_i) = \sum_i V(X_i) = n\sigma^2$

For independent variables X_i

Bias of an estimator

This shows that the estimator for the variance is biased*! To obtain the result for the estimator of σ, we can use a result known as Jensen's inequality, that states for any convex function f (x), $E[f(x)] \geq f(E[x])$



$$(E[\hat{\sigma}])^2 \leq E\left[\hat{\sigma}^2\right] < \sigma^2 \implies E[\hat{\sigma}] < \sigma$$

Jensen's inequality

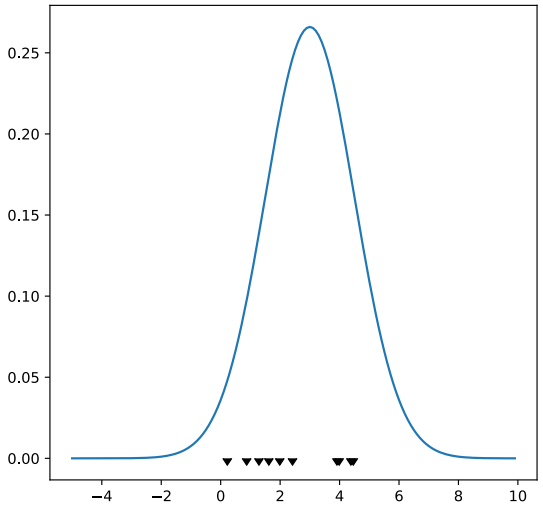
so the bias of the estimator $\hat{\sigma}$ is clearly not zero proving that the estimator for the σ parameter is a biased estimator!

*Instead the estimator $s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ is an unbiased estimator of σ^2

Bias of an estimator

What about the Maximum Likelihood Estimator?

Let's take a look at the same example as we did for the method of moments, i.e the case where X ~ φ(μ,σ)



$$L(\mu,\sigma) = \prod_i \phi(X_i;\mu,\sigma) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$q(\mu,\sigma) = -\ln L(\mu,\sigma) = \sum_i \ln(\sqrt{2\pi}\sigma) + \sum_i \frac{1}{2}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

Bias of an estimator

The MLEs are defined by $\frac{\partial q(\mu,\sigma)}{\partial \mu}|_{\mu=\hat{\mu}} = 0$ and $\frac{\partial q(\mu,\sigma)}{\partial \sigma}|_{\sigma=\hat{\sigma}} = 0$

The MLEs are defined by $\boxed{\frac{\partial q(\mu,\sigma)}{\partial \mu}|_{\mu=\hat{\mu}} = 0}$ and $\frac{\partial q(\mu,\sigma)}{\partial \sigma}|_{\sigma=\hat{\sigma}} = 0$

$$
\begin{aligned}
0 = \frac{\partial q(\mu,\sigma)}{\partial \mu} &= \frac{1}{2}\sum_i \frac{-2}{\sigma}\left(\frac{X_i-\mu}{\sigma}\right) \\
0 &= \frac{1}{\sigma^2}\sum_i X_i - \frac{1}{\sigma^2}\sum_i \mu \\
0 &= \frac{1}{\sigma^2}\sum_i X_i - \frac{n}{\sigma^2}\mu
\end{aligned}
$$

Bias of an estimator

The MLEs are defined by $\boxed{\frac{\partial q(\mu,\sigma)}{\partial \mu}|_{\mu=\hat{\mu}} = 0}$ and $\frac{\partial q(\mu,\sigma)}{\partial \sigma}|_{\sigma=\hat{\sigma}} = 0$

$$0 = \frac{\partial q(\mu,\sigma)}{\partial \mu} = \frac{1}{2}\sum_i \frac{-2}{\sigma}\left(\frac{X_i-\mu}{\sigma}\right)$$

$$0 = \frac{1}{\sigma^2}\sum_i X_i - \frac{1}{\sigma^2}\sum_i \mu$$

$$0 = \frac{1}{\sigma^2}\sum_i X_i - \frac{n}{\sigma^2}\mu \quad \Rightarrow \quad \boxed{\hat{\mu} = \frac{1}{n}\sum_i X_i}$$

This is just the ***sample mean***
→ the MLE is the same as the method of moments estimator!

The MLEs are defined by $\frac{\partial q(\mu,\sigma)}{\partial \mu}|_{\mu=\hat{\mu}} = 0$ and $\boxed{\frac{\partial q(\mu,\sigma)}{\partial \sigma}|_{\sigma=\hat{\sigma}} = 0}$

$$0 = \frac{\partial q(\mu,\sigma)}{\partial \sigma} = \sum_i \frac{\sqrt{2\pi}}{\sigma\sqrt{2\pi}} - \sum_i (X_i - \mu)^2 \frac{1}{\sigma^3}$$

$$0 = \frac{n}{\sigma} - \frac{1}{\sigma^3} \sum_i (X_i - \mu)^2$$

Bias of an estimator

The MLEs are defined by $\frac{\partial q(\mu,\sigma)}{\partial \mu}|_{\mu=\hat{\mu}} = 0$ and $\boxed{\frac{\partial q(\mu,\sigma)}{\partial \sigma}|_{\sigma=\hat{\sigma}} = 0}$

$$0 = \frac{\partial q(\mu,\sigma)}{\partial \sigma} = \sum_i \frac{\sqrt{2\pi}}{\sigma\sqrt{2\pi}} - \sum_i (X_i - \mu)^2 \frac{1}{\sigma^3}$$

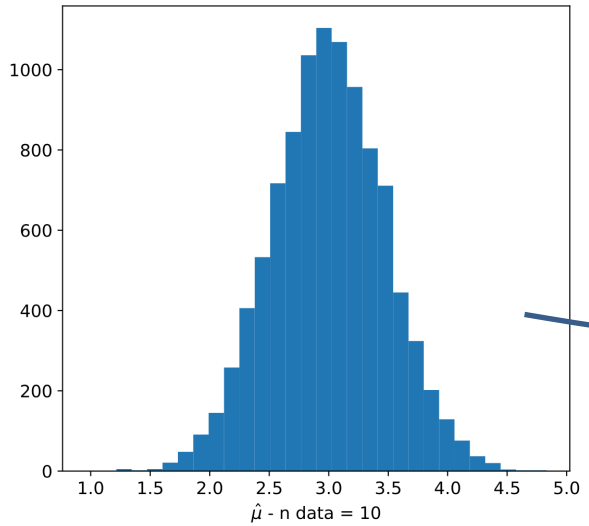
$$0 = \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_i (X_i - \mu)^2 \quad \Rightarrow \quad \boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \mu)^2}$$

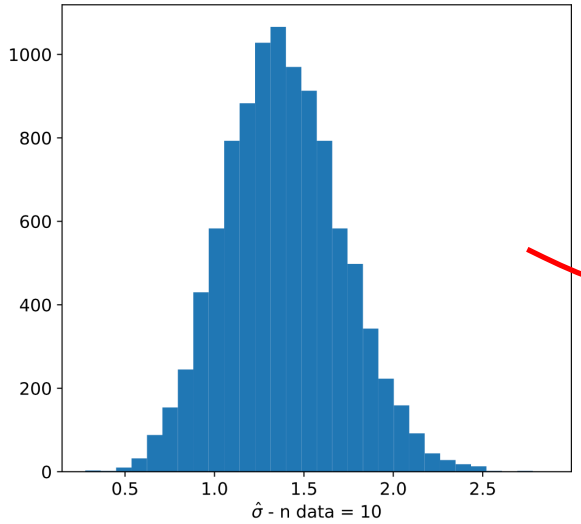
In this case, we can see that ***MLEs can also be biased estimators***

This is just the ***sample variance***
→ the MLE is the same as the method of mo
estimator (for the variance σ^2)!

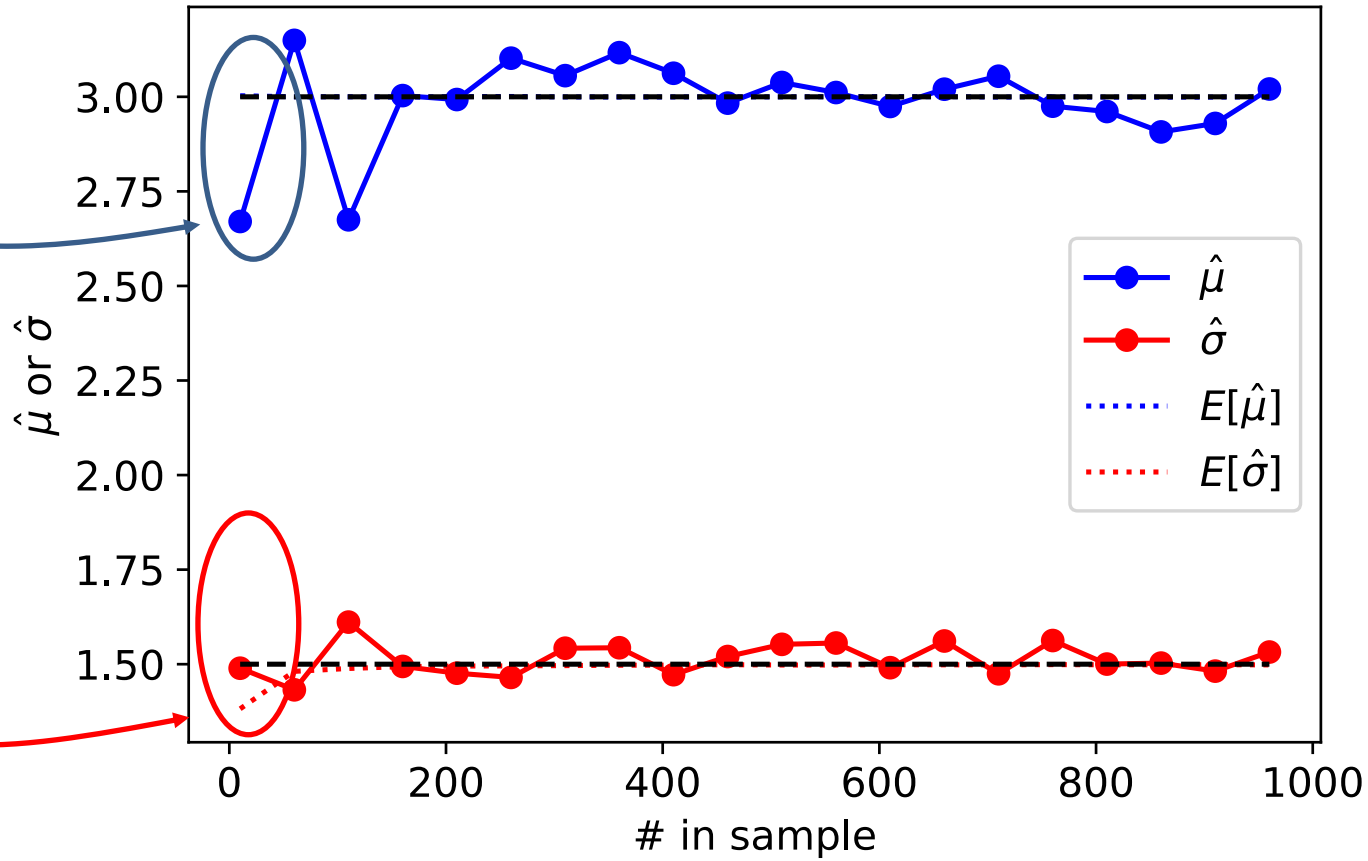
Bias in the Gaussian parameter estimators

See MethodOfMoments.ipynb





In our example for X ~ φ(1.5,3.0):



The estimator for **μ** is unbiased, but the estimator for **σ** ***is biased*** at least for smaller sample sizes.

Consistent estimators

A ***consistent estimator*** is one in which the value of the **estimator converges to the true value** as more data are included ($n \to \infty$)

Formally: for a sequence of estimators $\hat{\theta}_n$, obtained from a sample of size n, the estimator is consistent if the sequence converges in probability to the true value of θ as $n \to \infty$.

$$P(|\theta_n - \theta| > \epsilon) \to 0$$ for any $\varepsilon > 0$

Consistent estimators

A ***consistent estimator*** is one in which the value of the **estimator converges to the true value** as more data are included ($n \to \infty$)

Formally: for a sequence of estimators $\hat{\theta}_n$, obtained from a sample of size n, the estimator is consistent if the sequence converges in probability to the true value of θ as $n \to \infty$.

$$P(|\theta_n - \theta| > \epsilon) \to 0$$ for any $\varepsilon > 0$

We can use the law of large numbers to help us check if an estimator is consistent:

Remember this law says that if $\lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\nu_2^i\right) = 0$ then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu_1 = \mu$

Consistent estimators

Example: the method of moments estimator for the μ parameter is a consistent estimator. For a given sample size n, the variance of the sample mean is given by,

$$V(\bar{X}_n) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n}\sigma^2.$$

Consistent estimators

Example: the method of moments estimator for the μ parameter is a consistent estimator. For a given sample size n, the variance of the sample mean is given by,

$$V(\bar{X}_n) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n}\sigma^2.$$

Taking the sum,

$$\frac{1}{n^2}\sum_{i=1}^{n} \nu_2^i = \frac{1}{n}^2 \sum_{i=1}^{n} \frac{1}{n}\sigma^2 = \frac{1}{n^2}\sigma^2 \rightarrow 0$$

This means that by the law of large numbers we must have that $\hat{\mu}_n = \bar{X}_n \rightarrow \mu$

In fact, the method of moments estimator is consistent if it results from a (unique) solution to

$$g(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

Consistent estimators

Example: Maximum likelihood estimators are consistent estimators. This can be seen by first considering,

$$l_N(\theta) = -\frac{1}{N} \ln L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \ln f(X_i; \theta)$$

Probability density function

Consistent estimators

Example: Maximum likelihood estimators are consistent estimators. This can be seen by first considering,

$$l_N(\theta) = -\frac{1}{N}\ln L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\ln f(X_i;\theta)$$

Probability density function

By the law of large numbers, for all θ, $l_N(\theta) \to l(\theta)$ as N $\to \infty$, where

$$l(\theta) = E\left[-\ln(f(X;\theta))\right]_{\theta=\theta_0} = \int -\ln(f(X;\theta))f(X;\theta_0)dX$$

True parameter value

We need to check that θ_0 is the value of θ that minimises $l(\theta)$

$$
\begin{aligned}
I(\theta) - I(\theta_0) &= \int -\ln(f(X;\theta))f(X;\theta_0)dX - \int -\ln(f(X;\theta_0))f(X;\theta_0)dX \\
&= \int -\ln\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right) f(X;\theta_0)dX \\
&= E\left[-\ln\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right]_{\theta=\theta_0}
\end{aligned}
$$

Consistent estimators

$$
\begin{aligned}
I(\theta) - I(\theta_0) &= \int -\ln(f(X;\theta))f(X;\theta_0)dX - \int -\ln(f(X;\theta_0))f(X;\theta_0)dX \\
&= \int -\ln\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right) f(X;\theta_0)dX \\
&= E\left[-\ln\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right]_{\theta=\theta_0} \geq -\ln\left(E\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right]_{\theta=\theta_0}\right)
\end{aligned}
$$

Jensen's inequality –ln(x) is convex

Consistent estimators

Now,

$$\ln\left(E\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right]_{\theta=\theta_0}\right) = \ln\left(\int \frac{f(X;\theta)}{f(X;\theta_0)} f(X;\theta_0)dX\right) = \ln\left(\int f(X;\theta)dX\right) = \ln(1) = 0,$$

Which means that $I(\theta) - I(\theta_0) \geq 0$ the value that minimizes l(θ)

Consistent estimators

Now,

$$\ln\left(E\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right]_{\theta=\theta_0}\right) = \ln\left(\int \frac{f(X;\theta)}{f(X;\theta_0)} f(X;\theta_0)dX\right) = \ln\left(\int f(X;\theta)dX\right) = \ln(1) = 0,$$

Which means that $I(\theta) - I(\theta_0) \geq 0$ the value that minimizes $l(\theta)$

Now we can say that for each N, we have a value $\hat{\theta}_N$ that minimizes $l_N(\theta)$ and we know that $l_N(\theta) \rightarrow l(\theta)$.

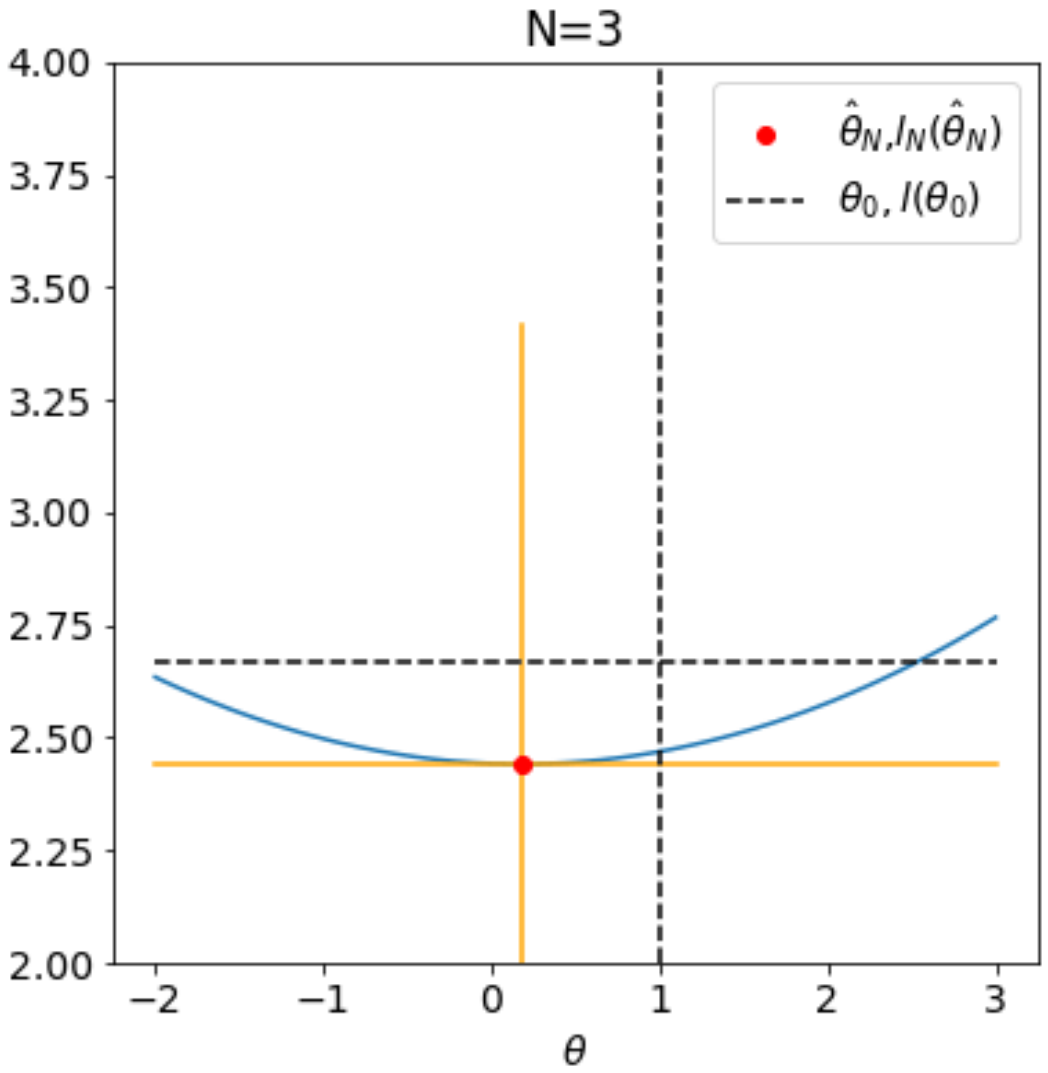
It would seem sensible that since the sequence of likelihood functions $l_N(\theta)$ have minima that get arbitrarily close to the minimum of $l(\theta)$, if ***only one*** value of θ minimises these functions, these values should also get arbitrarily close to θ_o

Consistent estimators

Proof by demonstration:

You can (hopefully) see that as the minima of the curves approach their limit, the estimators approach θ_0

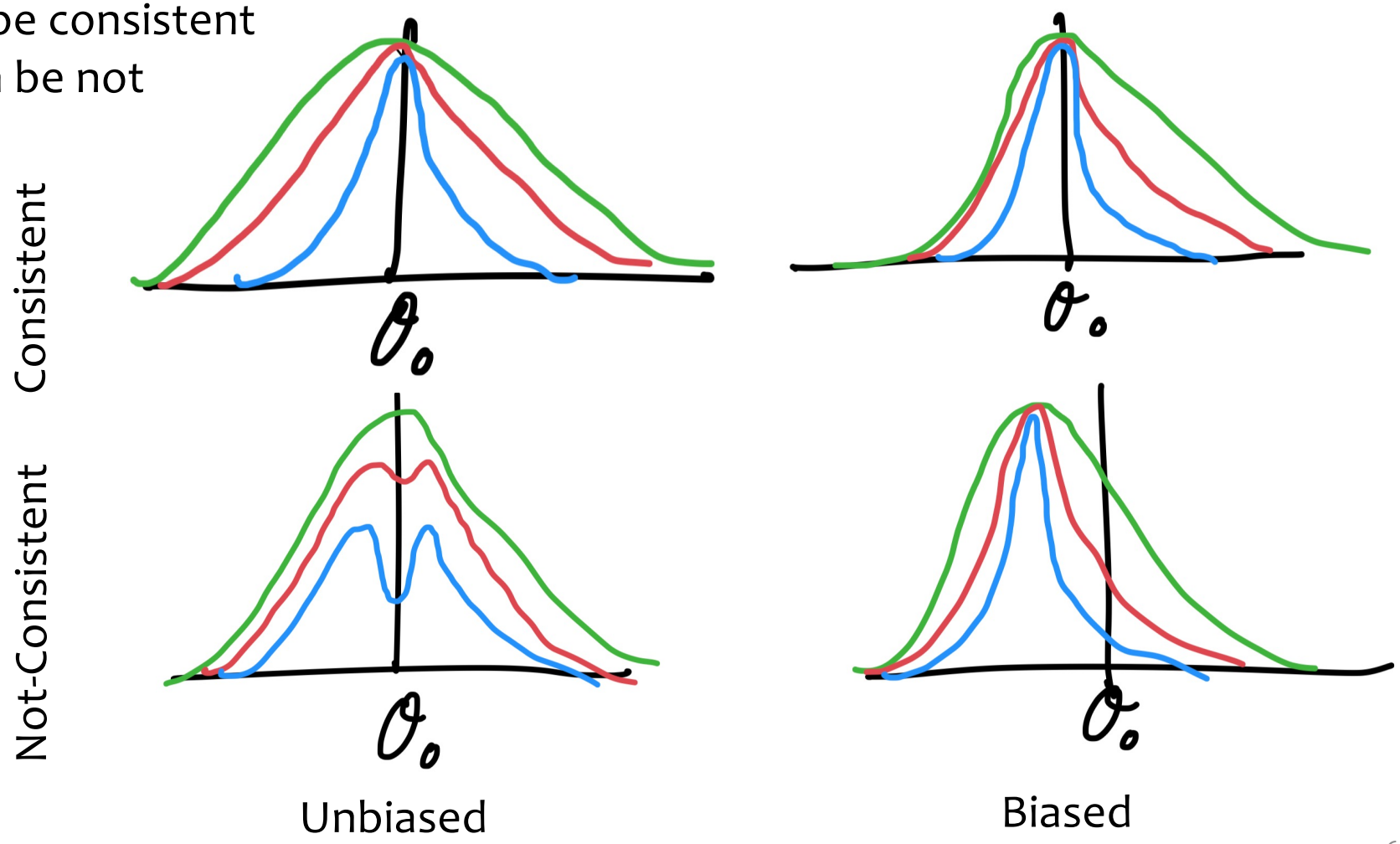
MLEs are consistent estimators (provided they are unique)



Bias and consistency

You should remember that the ***bias*** and ***consistency*** of estimators are not interchangeable.

Unbiased estimators can be consistent and biased estimators can be not consistent!



Efficiency

An ***efficient*** estimator is one who's variance is as small as theoretically possible

The variance of an estimator gives us an idea of how accurate it is at estimating some parameter (or moment/quantile etc) of a model → we want this to be small!

Efficiency

An ***efficient*** estimator is one who's variance is as small as theoretically possible

The variance of an estimator gives us an idea of how accurate it is at estimating some parameter (or moment/quantile etc) of a model → we want this to be small!

A result of Cramer and Rao states that the variance of an unbiased estimator of a parameter θ, V (θ) is bounded below by

$$V(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2] \geq \frac{1}{I(\theta)}$$

Where I(θ) is the ***Fisher Information:***

$$I(\theta) = E\left[\left(\frac{d(\ln f(X;\theta))}{d\theta}\right)^2\right]_{f(X;\theta)} = E\left[-\frac{d^2(\ln f(X;\theta))}{d\theta^2}\right]_{f(X;\theta)}$$

Efficiency

An ***efficient*** estimator is one who's variance is as small as theoretically possible

The variance of an estimator gives us an idea of how accurate it is at estimating some parameter (or moment/quantile etc) of a model → we want this to be small!

A result of Cramer and Rao states that the variance of an unbiased estimator of a parameter θ, V (θ) is bounded below by

$$V(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2] \geq \frac{1}{I(\theta)}$$

Where I(θ) is the ***Fisher Information:***

$$I(\theta) = E\left[\left(\frac{d(\ln f(X;\theta))}{d\theta}\right)^2\right]_{f(X;\theta)} = E\left[-\frac{d^2(\ln f(X;\theta))}{d\theta^2}\right]_{f(X;\theta)}$$

$$\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2} = \frac{\frac{\partial^2}{\partial\theta^2} f(X;\theta)}{f(X;\theta)} - \left(\frac{\frac{\partial}{\partial\theta} f(X;\theta)}{f(X;\theta)}\right)^2 = \boxed{\frac{\frac{\partial^2}{\partial\theta^2} f(X;\theta)}{f(X;\theta)}} - \left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2$$

E[] =0

Efficiency

An efficient (unbiased) estimator is one for which $V(\hat{\theta}) = \frac{1}{I(\theta)}$

For **MLEs** the following is true:

- If the MLE is unbiased in the limit $N \to \infty$, then that MLE becomes efficient in the limit of large data (we will see later this is the case when we get to Wilks' theorem)
- Of a given set of estimators, the MLE is typically the most efficient
- For a specified model, **any** unbiased efficient estimator **is the MLE**

Note that typically the method of moments yields inefficient estimators since often there is more information in the probability density that isn't captured by a fixed amount of moments.

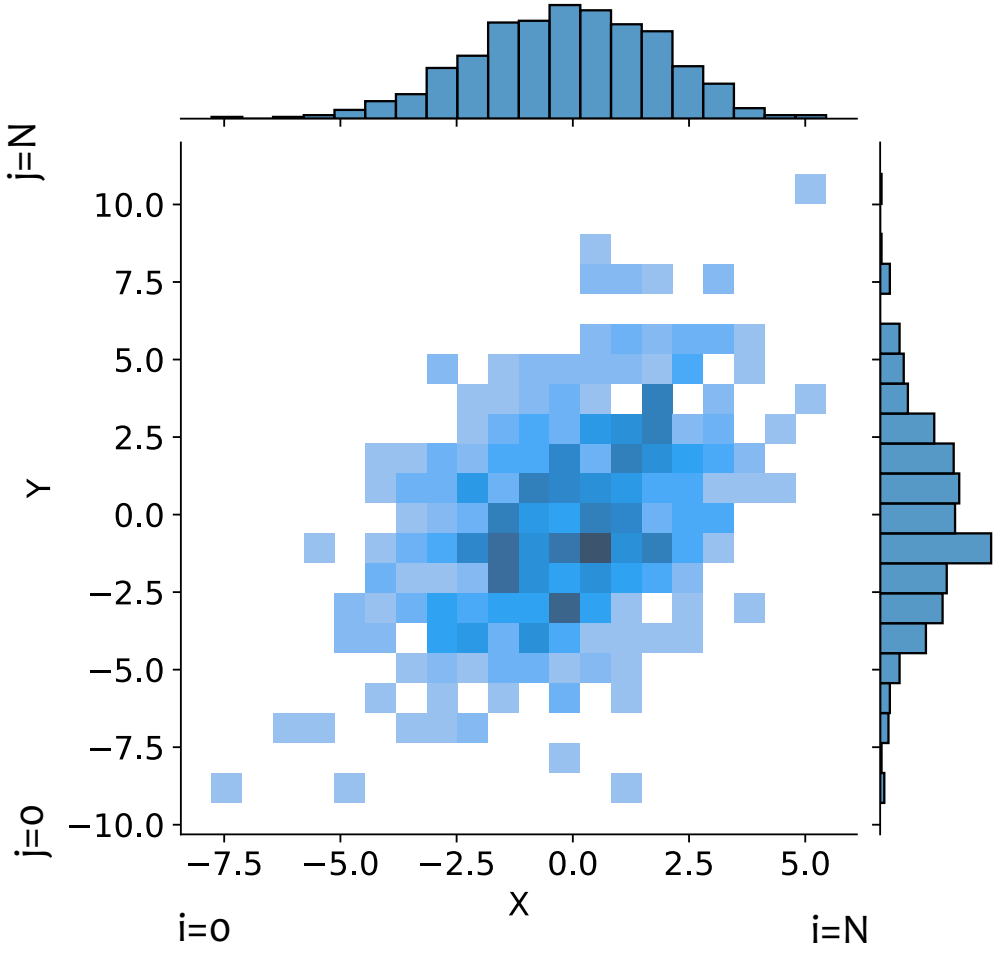
Recap

Estimator	Unbiased?	Consistent?	Efficient?
MLE	Not generally (yes for large N)	Always (assume uniqueness)	Yes as $N \to \infty$, usually most efficient
Method of Moments	Not generally (often corrections exist)	Very often	Not often

Table based on a mix of theorems and personal experience with uses ☺

The USP test

There are a number of ***tests of independence*** that are much better than checking for a zero linear correlation



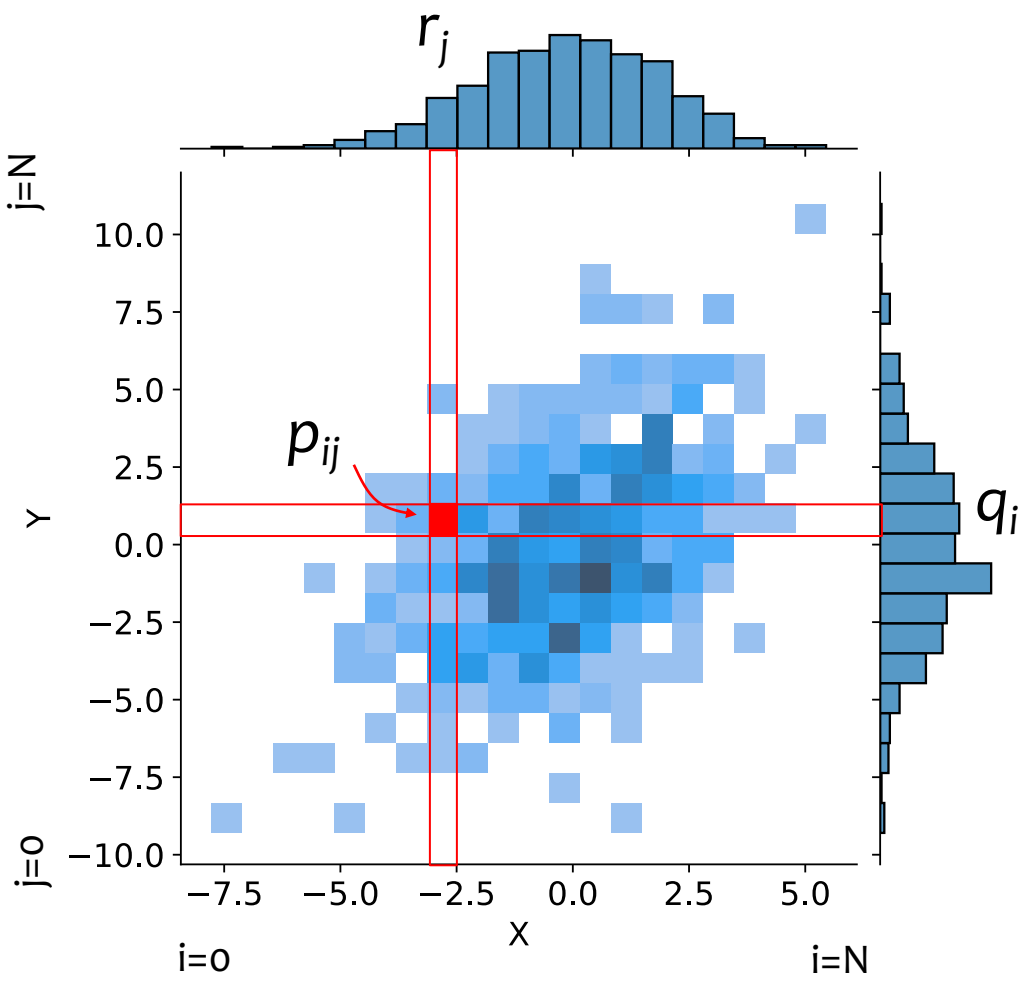
The USP test

There are a number of ***tests of independence*** that are much better than checking for a zero linear correlation

Example: The U-statistic permutation test

The USP is based on the ***divergence of binned distributions*** (sometimes referred to as tables)

$$D = \sum_{i=1}^{N} \sum_{j=1}^{M} (p_{ij} - q_i r_j)^2$$



The USP test

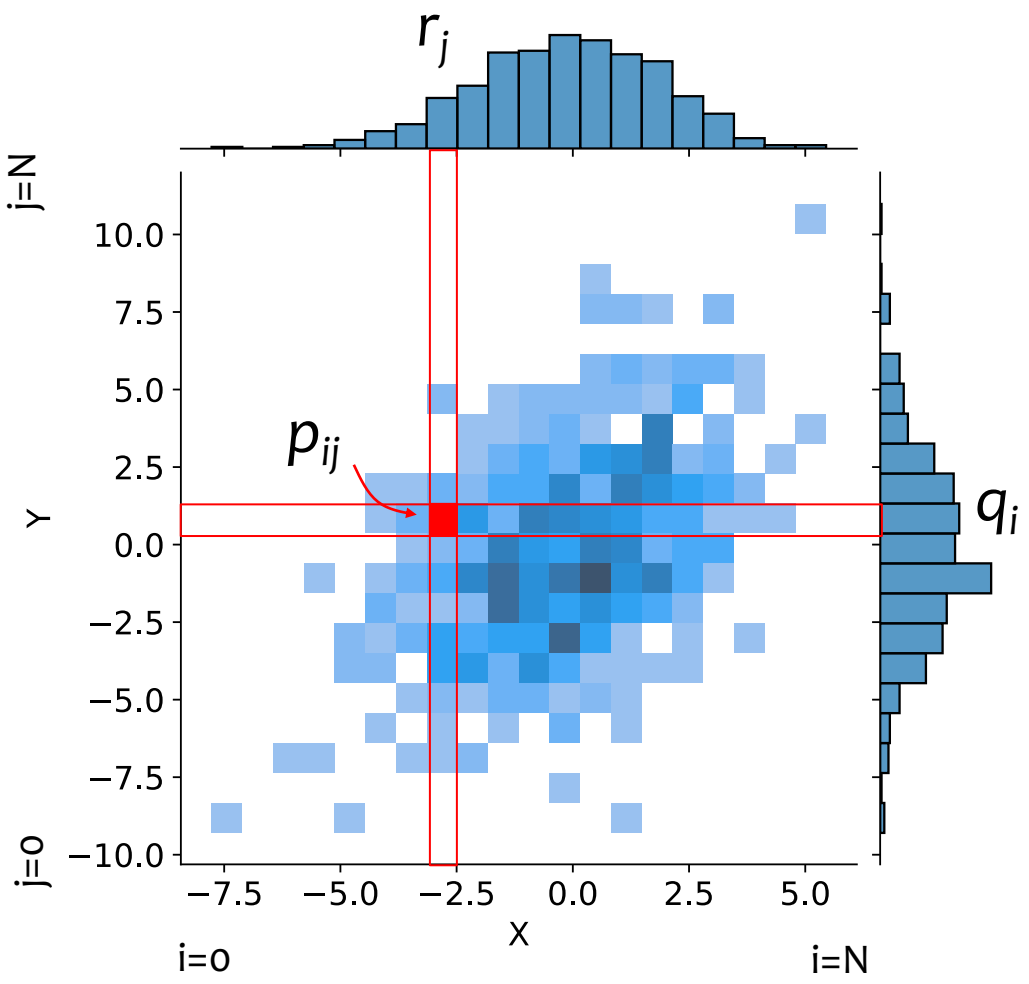
There are a number of ***tests of independence*** that are much better than checking for a zero linear correlation

Example: The U-statistic permutation test

The USP is based on the ***divergence of binned distributions*** (sometimes referred to as tables)

$$D = \sum_{i=1}^{N} \sum_{j=1}^{M} (p_{ij} - q_i r_j)^2$$

H_0 := X and Y independent ➔ D = 0

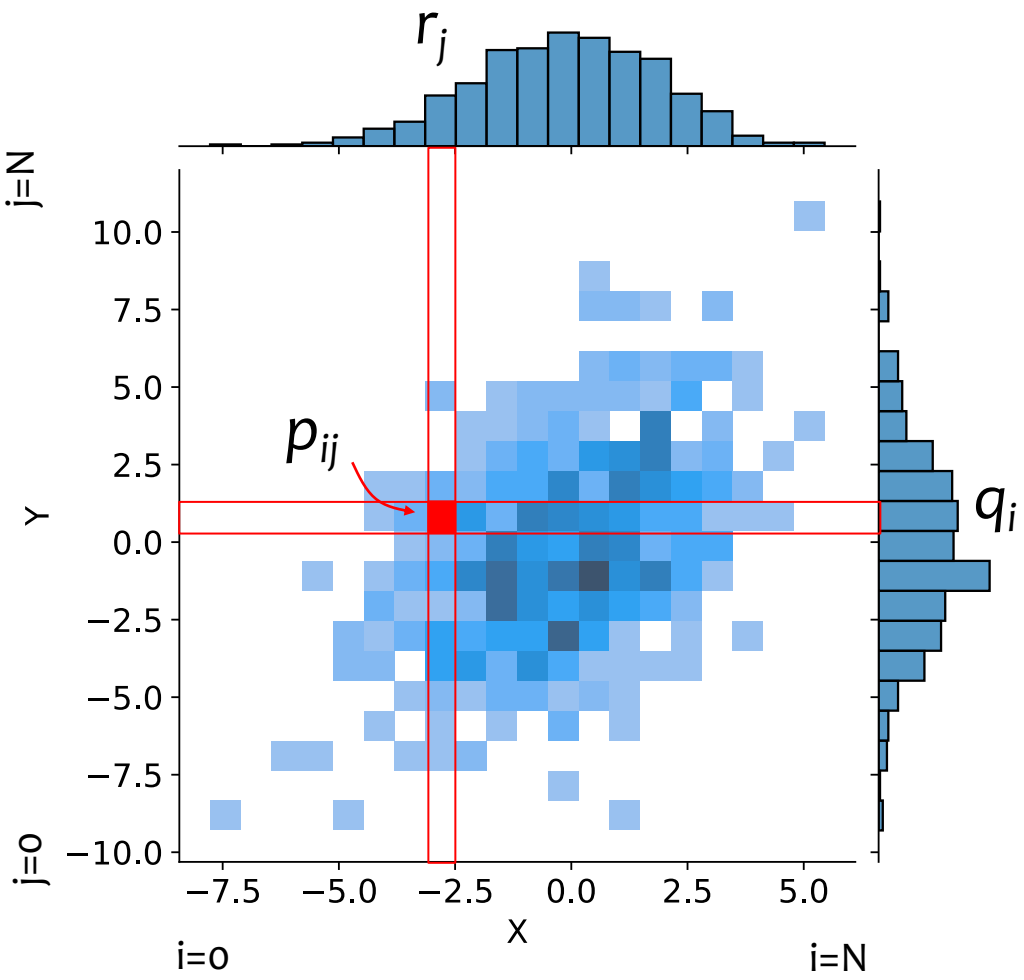


The USP test

With only the observed data, we have good (unbiased) estimators for the densities p_{ij} and $q_i r_j$

$$\hat{p_{ij}} = \frac{o_{ij}}{n} \qquad \hat{q_i r_j} = \frac{o_i o_j}{n}$$

Where n is the total number of events and O_{ij} and $O_{i/j}$ are the observed number of events in the ij-th bin of the 2D histogram, and the i-th or j-th bin of each 1D histogram



see [R. Soc. A 477: 20210549]

The USP test

With only the observed data, we have good (unbiased) estimators for the densities p_{ij} and $q_i r_j$

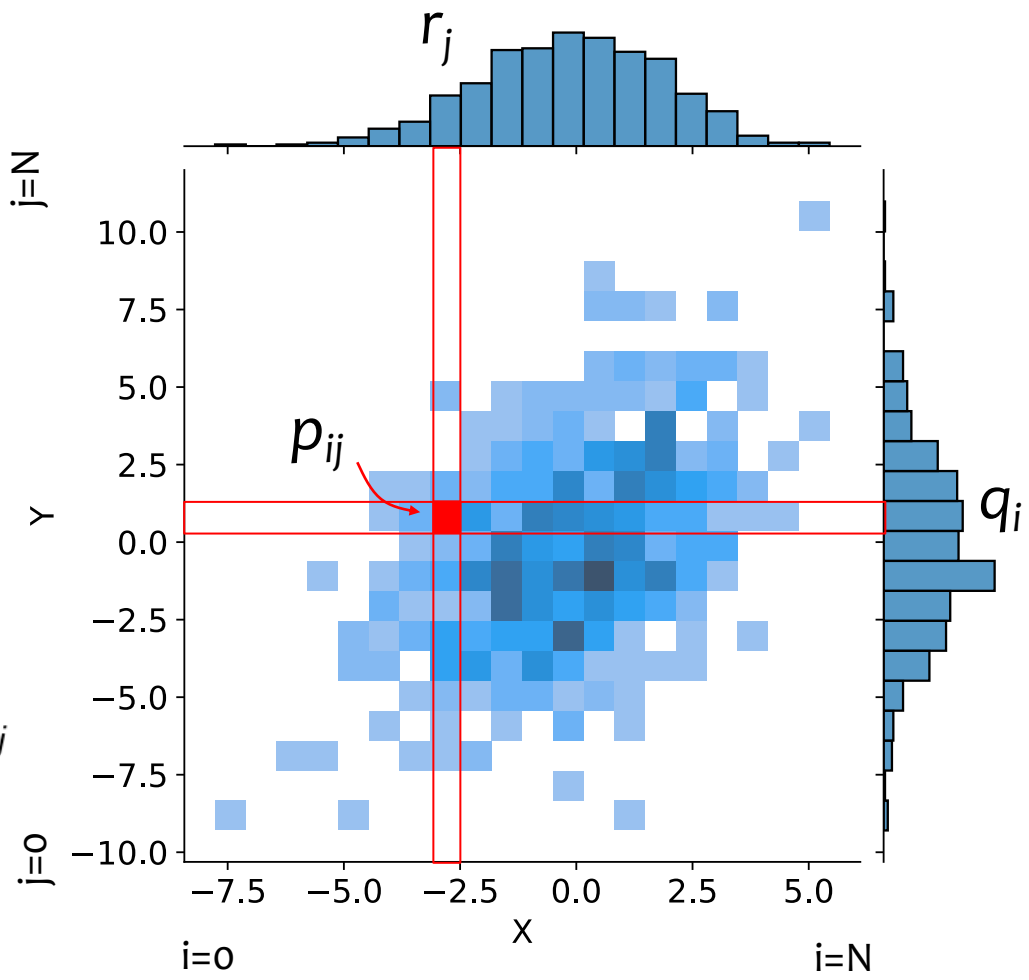
$$\hat{p_{ij}} = \frac{o_{ij}}{n} \qquad \hat{q_i r_j} = \frac{o_i o_j}{n}$$

Where n is the total number of events and O_{ij} and $O_{i/j}$ are the observed number of events in the ij-th bin of the 2D histogram, and the i-th or j-th bin of each 1D histogram

We define the test statistic U

$$U = \frac{1}{n(n-3)} \sum_{i=1}^{N} \sum_{j=1}^{M} (o_{ij} - o_i o_j)^2 - \frac{4}{n(n-2)(n-3)} \sum_{i=1}^{N} \sum_{j=1}^{M} o_{ij} o_i o_j$$

see R. Soc. A 477: 20210549



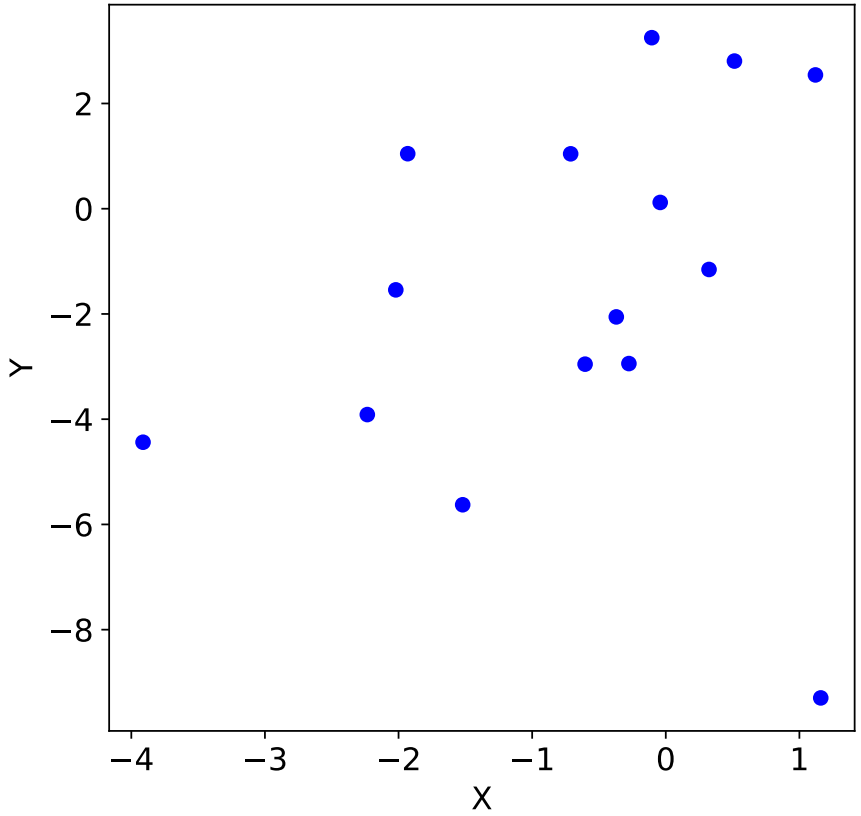
The USP test

To estimate the distribution of *U*, we can use a ***permutation bootstrap*** method!

We can construct bootstrap samples that are independent in X and Y but preserve the marginal distributions q(X) and r(Y)

The original data is $\{(X_1,Y_1), (X_2,Y_2),\ldots,(X_n,Y_n)\}$

- Generate a permutation σ of $\{1,\ldots,n\}$ and create a new dataset $\{(X_1,Y_{\sigma(1)}), (X_2,Y_{\sigma(2)}),\ldots,(X_n,Y_{\sigma(n)})\}$



R. Soc. A 477: 20210549

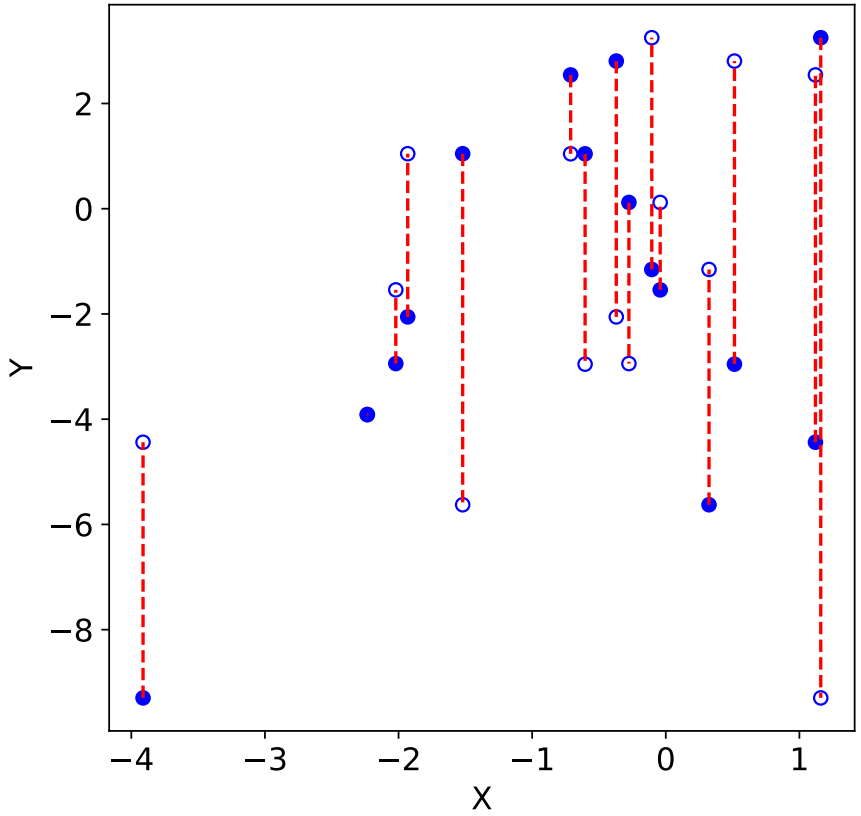
The USP test

To estimate the distribution of *U*, we can use a ***permutation bootstrap*** method!

We can construct bootstrap samples that are independent in X and Y but preserve the marginal distributions q(X) and r(Y)

The original data is $\{(X_1,Y_1), (X_2,Y_2),\dots,(X_n,Y_n)\}$

- Generate a permutation σ of $\{1,\dots,n\}$ and create a new dataset $\{(X_1,Y_{\sigma(1)}), (X_2,Y_{\sigma(2)}),\dots,(X_n,Y_{\sigma(n)})\}$
- Bin the data in the original histogram and calculate the test statistic *U* for this new dataset
- Repeat a large number of times (there are *n!* available samples)



R. Soc. A 477: 20210549

The USP test

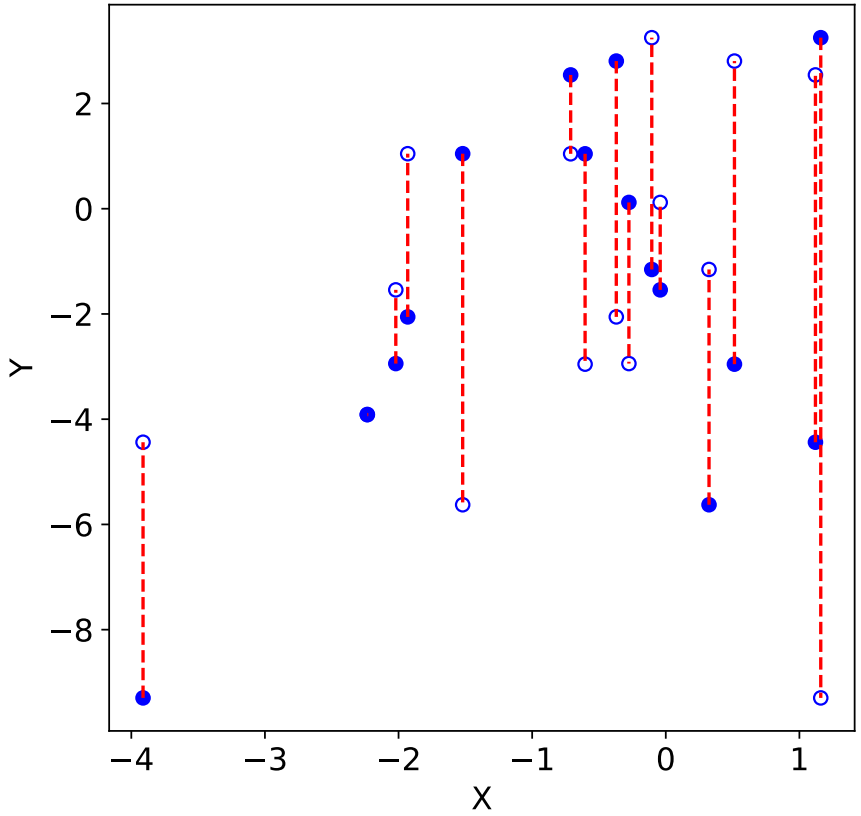
To estimate the distribution of *U*, we can use a ***permutation bootstrap*** method!

We can construct bootstrap samples that are independent in X and Y but preserve the marginal distributions q(X) and r(Y)

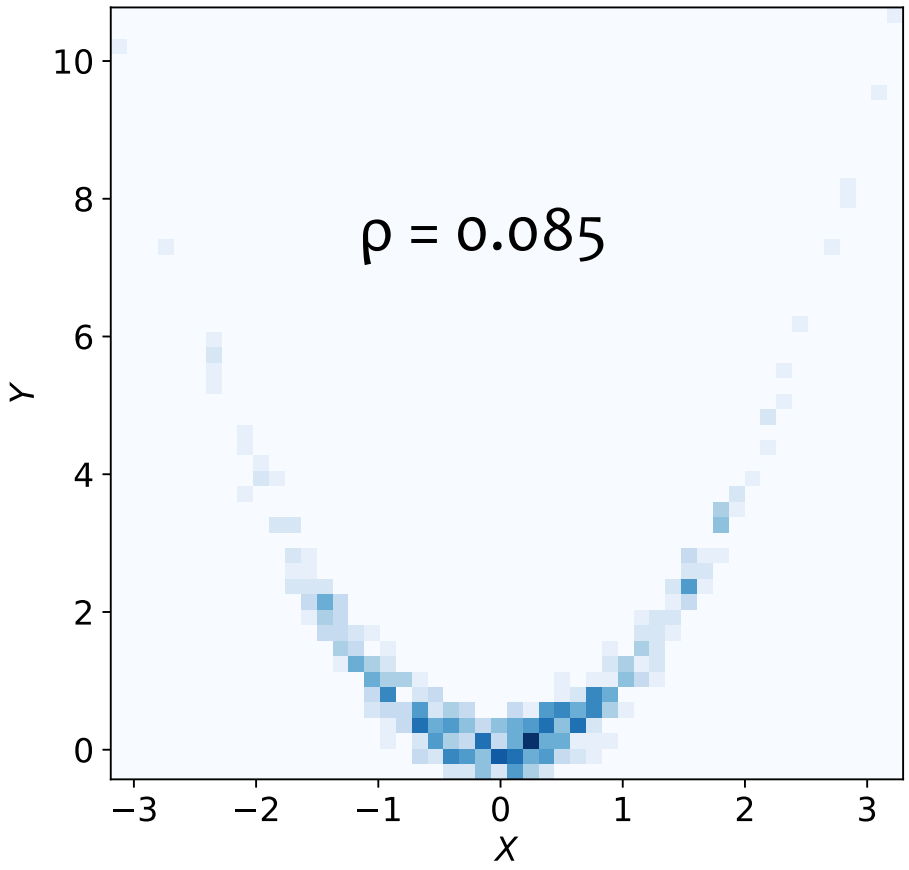
The original data is $\{(X_1,Y_1), (X_2,Y_2),\ldots,(X_n,Y_n)\}$

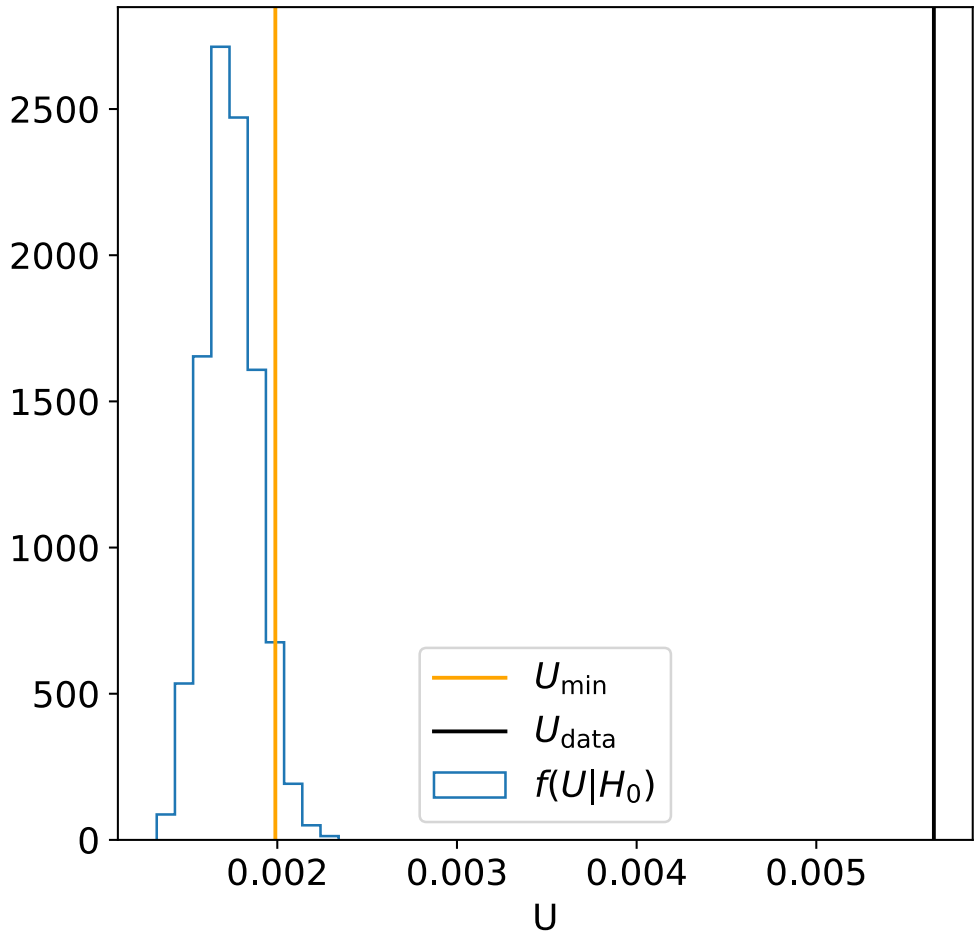
- Generate a permutation σ of $\{1,\ldots,n\}$ and create a new dataset $\{(X_1,Y_{\sigma(1)}), (X_2,Y_{\sigma(2)}),\ldots,(X_n,Y_{\sigma(n)})\}$
- Bin the data in the original histogram and calculate the test statistic *U* for this new dataset
- Repeat a large number of times (there are *n!* available samples)

See [UStatisticPermutationTest.ipynb]



The USP test


ρ = 0.085
Y
X
10
8
6
4
2
0
−3
−2
−1
0
1
2
3


U_{min}
U_{data}
$f(U|H_0)$
U
2500
2000
1500
1000
500
0
0.002
0.003
0.004
0.005

The most powerful test

For a specific size of test α we choose w such that,

$$\int_w f(\mathbf{X};\theta_0)\,d\mathbf{X} = \alpha$$

and we want to find the region w which maximises 1 – β

$$\begin{aligned} 1-\beta &= \int_w f(\mathbf{X};\theta_1)\,d\mathbf{X} \\ &= \int_w \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} f(\mathbf{X};\theta_0)\,d\mathbf{X} \\ &= E\left[\frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)}\middle|\theta=\theta_0\right]_w \end{aligned}$$

Expectation value in a restricted space *w*

The most powerful test

We write the quantity

$$\Lambda = \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = \frac{L(\theta_1)}{L(\theta_0)}$$

This quantity is the **ratio of the likelihood function,** evaluated under the two hypotheses

The most powerful test

We write the quantity

$$\Lambda = \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = \frac{L(\theta_1)}{L(\theta_0)}$$

This quantity is the **ratio of the likelihood function**, evaluated under the two hypotheses

The expectation will be maximal when w is chosen to contain the largest values of Λ

→ The best critical region is the set of points for which $\Lambda \geq c_\alpha \in R$, where c_α satisfies

$$\int_w f(\mathbf{X};\theta_0)\,d\mathbf{X} = \alpha$$

→ If $\Lambda > c_\alpha$, we would choose H_1, while $\Lambda \leq c_\alpha$ leads us to choose H_0!

The most powerful test

We write the quantity

$$\Lambda = \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = \frac{L(\theta_1)}{L(\theta_0)}$$

This quantity is the **ratio of the likelihood function**, evaluated under the two hypotheses

The expectation will be maximal when w is chosen to contain the largest values of Λ

→ The best critical region is the set of points for which $\Lambda \geq c_\alpha \in R$, where c_α satisfies

$$\int_w f(\mathbf{X};\theta_0)\,d\mathbf{X} = \alpha$$

→ If $\Lambda > c_\alpha$, we would choose H_1, while $\Lambda \leq c_\alpha$ leads us to choose H_0!

This is the Neyman-Pearson Lemma for simple hypotheses!

P-values

The p-value **is**

- A random variable that depends on the observed data (it's a post-observation quantity)
- Distributed uniformly between 0 and 1 under the null hypothesis

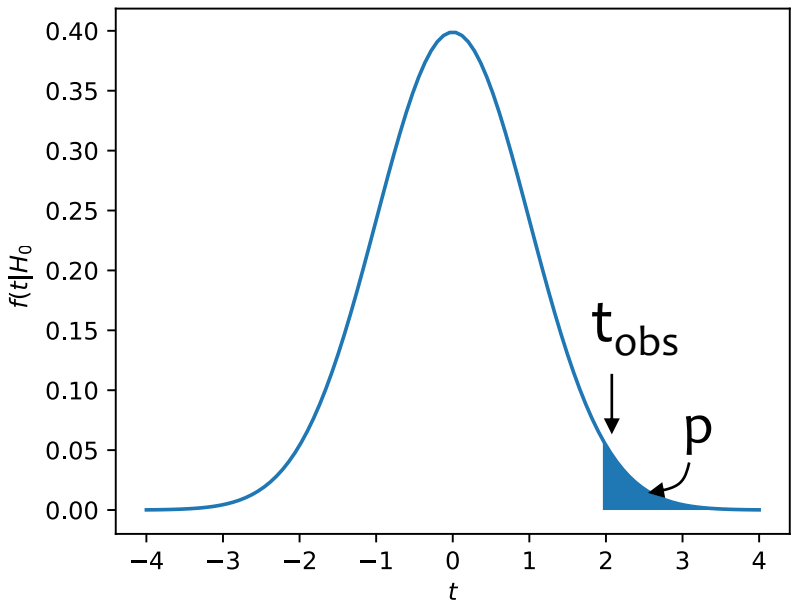
$$p = P(t > t_{\text{obs}}|H_0) = 1 - P(t < t_{\text{obs}}|H_0) = 1 - F(t)$$

$$1 - F(t) = P(t > t_{\text{obs}}|H_0) = P(F(t) > F(t_{\text{obs}})|H_0) = 1 - P(F(t) < F(t_{\text{obs}})|H_0)$$

Since F(.) is monotonic and increasing

Hence $F(t) = P(F(t) < F(t_{\text{obs}})|H_0)$

Which is true for any t_{obs}



Wilks' theorem

Wilks' theorem provides a very powerful result that allows us to calculate the distribution of ζ_θ under the hypothesis H(θ) in the limit of a large sample size

Here I'll go over the simplest case where we have a single parameter of interest θ

Wilks' theorem

Wilks' theorem provides a very powerful result that allows us to calculate the distribution of ζ_θ under the hypothesis H(θ) in the limit of a large sample size

Here I'll go over the simplest case where we have a single parameter of interest θ

As usual, for N observations of a random variable X ~ f(X;θ), we define,

$$\zeta_{N,\theta} = q_N(\theta) - q_N(\hat{\theta}) = -2\sum_{i=1}^{N} \ln f(X_i;\theta) + 2\sum_{i=1}^{N} \ln f(X_i;\hat{\theta}).$$

Wilks' theorem

Providing that the derivatives of *f* exist, we can Taylor expand the first derivative of q_N to approximate the value at $\hat{\theta}$

$$0 = \left.\frac{dq_N}{d\theta}\right|_{\theta=\hat{\theta}} = q'_N(\hat{\theta}) = q'_N(\theta) + q''_N(\theta)(\hat{\theta} - \theta)$$

Wilks' theorem

Providing that the derivatives of *f* exist, we can Taylor expand the first derivative of q_N to approximate the value at $\hat{\theta}$

$$0 = \left.\frac{dq_N}{d\theta}\right|_{\theta=\hat{\theta}} = q'_N(\hat{\theta}) = q'_N(\theta) + q''_N(\theta)(\hat{\theta} - \theta)$$

Wilks' theorem

Providing that the derivatives of *f* exist, we can Taylor expand the first derivative of q_N to approximate the value at $\hat{\theta}$

$$0 = \left.\frac{dq_N}{d\theta}\right|_{\theta=\hat{\theta}} = q'_N(\hat{\theta}) = q'_N(\theta) + q''_N(\theta)(\hat{\theta} - \theta)$$

which we can re-write as

$$(\hat{\theta} - \theta)\sqrt{q''_N(\theta)} = -\frac{q'_N(\theta)}{\sqrt{q''_N(\theta)}}$$

Which seems like an odd thing to do but there is a good reason

Wilks' theorem

$$(\hat{\theta} - \theta)\sqrt{q_N''(\theta)} = \boxed{-\frac{q_N'(\theta)}{\sqrt{q_N''(\theta)}}}$$

Focusing on this term

$$-q_N'(\theta) = \frac{d}{d\theta}\sum_{i=1}^{N} 2\ln f(X_i;\theta) = \sum_{i=1}^{N} 2\frac{d}{d\theta}\left(\ln f(X_i;\theta)\right)$$

Wilks' theorem

$$(\hat{\theta} - \theta)\sqrt{q_N''(\theta)} = \boxed{-\frac{q_N'(\theta)}{\sqrt{q_N''(\theta)}}}$$

Focusing on this term

$$-q_N'(\theta) = \frac{d}{d\theta}\sum_{i=1}^{N} 2\ln f(X_i;\theta) = \sum_{i=1}^{N} 2\frac{d}{d\theta}(\ln f(X_i;\theta))$$

Now we define a new random variable $u = 2\frac{d}{d\theta}\ln f(X|\theta)$

then $-q_N'(\theta) = N\bar{u}$

Wilks' theorem

$$(\hat{\theta} - \theta)\sqrt{q''_N(\theta)} = \boxed{-\frac{q'_N(\theta)}{\sqrt{q''_N(\theta)}}}$$

Focusing on this term

$$-q'_N(\theta) = \frac{d}{d\theta}\sum_{i=1}^{N} 2\ln f(X_i;\theta) = \sum_{i=1}^{N} 2\frac{d}{d\theta}(\ln f(X_i;\theta))$$

Now we define a new random variable $u = 2\frac{d}{d\theta}\ln f(X|\theta)$

then $-q'_N(\theta) = N\boxed{\bar{u}}$

The sample mean of *u*

Wilks' theorem

The expectation value of *u* can be found by considering that for any value of θ

$$1 = \int f(X;\theta)dX$$

Taking derivatives

$$0 = \frac{d}{d\theta}\int f(X;\theta)dX = \int \frac{d}{d\theta} f(X;\theta)dX = \int \frac{d}{d\theta}\left(\ln f(X;\theta)\right)\cdot f(X;\theta)dX$$

Wilks' theorem

The expectation value of *u* can be found by considering that for any value of θ

$$1 = \int f(X;\theta)dX$$

Taking derivatives

$$0 = \frac{d}{d\theta}\int f(X;\theta)dX = \int \frac{d}{d\theta} f(X;\theta)dX = \int \frac{d}{d\theta}\left(\ln f(X;\theta)\right)\cdot f(X;\theta)dX$$

$$= E\left[\frac{d}{d\theta}\ln f(X;\theta)\right]_\theta = \frac{1}{2}E\left[u\right]_\theta$$

$$u = 2\frac{d}{d\theta}\ln f(X|\theta)$$

The expectation of *u* under f(X;θ) is 0

Wilks' theorem

Now, we can define a quantity T_N,

$$T_{N,\theta} = \frac{N\bar{u} - NE[u]_\theta}{\sqrt{N \cdot V(u)_\theta}}$$

Wilks' theorem

Now, we can define a quantity T_N,

$$T_{N,\theta} = \frac{N\bar{u} - NE[u]_\theta}{\sqrt{N \cdot \boxed{V(u)_\theta}}}$$

Let's take a look at $V(u)_\theta$. We have by definition that,

$$V(u)_\theta \;=\; E\left[(u - E[u]_\theta)^2\right]_\theta = E\left[(u)^2\right]_\theta$$

Wilks' theorem

Now, we can define a quantity T_N,

$$T_{N,\theta} = \frac{N\bar{u} - NE[u]_\theta}{\sqrt{N \cdot \boxed{V(u)_\theta}}}$$

Let's take a look at $V(u)_\theta$. We have by definition that,

$$V(u)_\theta = E\left[(u - E[u]_\theta)^2\right]_\theta = E\left[(u)^2\right]_\theta$$

$E[u]_\theta = 0$

$$= 4\int \left(\frac{d}{d\theta} \ln f(X;\theta)\right)^2 f(X;\theta)dX$$

Wilks' theorem

Again, we can take

$$1 = \int f(X;\theta)dX$$

Differentiating twice, we have

$$\begin{aligned} 0 &= \int \frac{d}{d\theta}\left(\frac{d}{d\theta}\ln f(x;\theta)\cdot f(X;\theta)\right)dX \\ &= \int \frac{d^2}{d\theta^2}\ln f(X;\theta)\cdot f(X;\theta)dX + \int \frac{d}{d\theta}\ln f(X;\theta)\cdot \frac{d}{d\theta}f(X;\theta)dX \end{aligned}$$

Again, we can take

$$1 = \int f(X;\theta)dX$$

Differentiating twice, we have

$$\begin{aligned} 0 &= \int \frac{d}{d\theta}\left(\frac{d}{d\theta}\ln f(x;\theta)\cdot f(X;\theta)\right)dX \\ &= \int \frac{d^2}{d\theta^2}\ln f(X;\theta)\cdot f(X;\theta)dX + \int \frac{d}{d\theta}\ln f(X;\theta)\cdot \frac{d}{d\theta}f(X;\theta)dX \\ &= \int \frac{d^2}{d\theta^2}\ln f(X;\theta)\cdot f(X;\theta)dX + \int \left(\frac{d}{d\theta}\ln f(X;\theta)\right)^2 f(X;\theta)dX \end{aligned}$$

$$\frac{d}{d\theta}f(\theta) = \frac{d}{d\theta}(\ln f(\theta))\, f(\theta)$$

Wilks' theorem

$$
\begin{aligned}
0 &= \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX + \int \left(\frac{d}{d\theta} \ln f(X;\theta) \right)^2 f(X;\theta) dX \\
&= \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX + \frac{1}{4} V(u)_\theta
\end{aligned}
$$

Wilks' theorem

$$
\begin{aligned}
0 &= \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX + \int \left(\frac{d}{d\theta} \ln f(X;\theta) \right)^2 f(X;\theta) dX \\
&= \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX + \frac{1}{4} V(u)_\theta
\end{aligned}
$$

Rearranging gives

$$
\frac{1}{4} V(u)_\theta = - \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX = -E \left[\frac{d^2}{d\theta^2} \ln f(X;\theta) \right]_\theta
$$

Wilks' theorem

$$0 = \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX + \int \left(\frac{d}{d\theta} \ln f(X;\theta) \right)^2 f(X;\theta) dX$$

$$= \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX + \frac{1}{4} V(u)_\theta$$

Rearranging gives

$$\frac{1}{4} V(u)_\theta = - \int \frac{d^2}{d\theta^2} \ln f(X;\theta) \cdot f(X;\theta) dX = -E \left[\frac{d^2}{d\theta^2} \ln f(X;\theta) \right]_\theta$$

From the def^n of q(θ)

$$q''_N(\theta) = -2 \sum_{i=1}^{N} \frac{d^2}{d\theta^2} \ln f(X_i;\theta)$$

Wilks' theorem

Now we have $-\dfrac{q''(\theta)}{N}$ is a sample mean of some quantity $\dfrac{d^2}{d\theta^2}\ln f(X;\theta)$

The law of large numbers tells us that this should converge to the expectation of that quantity so,

$$\frac{-q''_N(\theta)}{N} \to 2E\left[\frac{d^2}{d\theta^2}\ln f(X;\theta)\right]_\theta \qquad \text{as } N \to \infty$$

Wilks' theorem

Now we have $-\dfrac{q''(\theta)}{N}$ is a sample mean of some quantity $\dfrac{d^2}{d\theta^2}\ln f(X;\theta)$

The law of large numbers tells us that this should converge to the expectation of that quantity so,

$$\frac{-q''_N(\theta)}{N} \to 2E\left[\frac{d^2}{d\theta^2}\ln f(X;\theta)\right]_\theta \qquad \text{as } N \to \infty$$

But this is just $-\dfrac{1}{2}V(u)_\theta$ previous slide!

so then, $$\frac{q''_N(\theta)}{N} \to \frac{1}{2}V(u)_\theta$$

Wilks' theorem

Back to our T_N variable

$$\frac{1}{\sqrt{2}}(\hat{\theta}-\theta)\sqrt{q_N''(\theta)} = -\frac{q_N'(\theta)}{\sqrt{2q_N''(\theta)}}$$

Wilks' theorem

Back to our T_N variable

$$\frac{1}{\sqrt{2}}(\hat{\theta}-\theta)\sqrt{q''_N(\theta)} = -\frac{q'_N(\theta)}{\sqrt{2q''_N(\theta)}} \rightarrow \frac{q'_N(\theta)}{\sqrt{N\cdot V(u)_\theta}} = \frac{N\bar{u} - NE[u]_\theta}{\sqrt{N\cdot V(u)_\theta}}$$

$$= T_{N,\theta}$$

By the central limit theorem* then we have

$$T_{N,\theta} \rightarrow T_\theta \sim \phi(T_\theta, 0, 1)$$

* We've really used a result due to Slutsky that says this limit holds even though the numerator converges in distribution and the denominator converges in probability

Now, back to the definition of $\zeta_{N,\theta}$

Taylor expanding again, we have

$$q_N(\theta) = q(\hat{\theta}) + \cancelto{0}{q'_N(\hat{\theta})}(\theta - \hat{\theta}) + \frac{1}{2} q''_N(\hat{\theta})(\theta - \hat{\theta})^2$$

So we have

$$\zeta_{N,\theta} = \frac{1}{2} q''_N(\hat{\theta})(\theta - \hat{\theta})^2 = \left(\frac{1}{\sqrt{2}} (\hat{\theta} - \theta) \sqrt{q''_N(\hat{\theta})} \right)^2$$

Now, back to the definition of $\zeta_{N,\theta}$

Taylor expanding again, we have

$$q_N(\theta) = q(\hat{\theta}) + \cancelto{0}{q'_N(\hat{\theta})}(\theta - \hat{\theta}) + \frac{1}{2} q''_N(\hat{\theta})(\theta - \hat{\theta})^2$$

So we have

$$\zeta_{N,\theta} = \frac{1}{2} q''_N(\hat{\theta})(\theta - \hat{\theta})^2 = \left(\boxed{\frac{1}{\sqrt{2}}(\hat{\theta} - \theta)\sqrt{q''_N(\hat{\theta})}} \right)^2$$

This is very similar to $\frac{1}{\sqrt{2}}(\hat{\theta} - \theta)\sqrt{q''_N(\theta)} = T_{N,\theta}$

Wilks' theorem

But remember the MLE is a ***consistent estimator*** so we know that $\hat{\theta} \to \theta_0$
so evaluating at $\theta = \theta_0$

$$\zeta_{N,\theta_0} = (T_{N,\theta_0})^2 \quad \text{as} \quad N \to \infty$$

But we know that T_{N,θ_0} is distributed as a unit normal $\varphi(T; 0, 1)$ under the hypothesis $H(\theta_0)$ in the limit of large N

Therefore, when $\theta = \theta_0$ (the true value), ζ_{θ_0} is distributed as a chi-square distribution with 1 degree of freedom!

$$\chi^2(X; 1) = \frac{1}{\sqrt{2\pi X}} e^{-\frac{X}{2}}$$

Hessian approximation

If we had a random variable that was distributed as a Gaussian with $\theta \sim \varphi(\theta, \sigma)$, we would find that that the twice the negative log-likelihood would be

$$-2\left(\ln(\phi(\theta,\sigma)) - \ln(\phi(\hat{\theta},\sigma))\right) = -2\left(-\frac{1}{2}\left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{\hat{\theta}-\hat{\theta}}{\sigma}\right)^2\right) = \left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2$$

Hessian approximation

If we had a random variable that was distributed as a Gaussian with θ ~ φ(θ, σ), we would find that that the twice the negative log-likelihood would be

$$-2\left(\ln(\phi(\theta,\sigma)) - \ln(\phi(\hat{\theta},\sigma))\right) = -2\left(-\frac{1}{2}\left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{\hat{\theta}-\hat{\theta}}{\sigma}\right)^2\right) = \left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2$$

Comparing with the previous, more general, result

$$(q_N(\theta) - q(\hat{\theta})) = \frac{1}{2}q_N''(\hat{\theta})(\theta-\hat{\theta})^2$$

We must have

$$\frac{1}{\sigma^2} = \frac{1}{2}q_N''(\hat{\theta}).$$

Hessian approximation

If we had a random variable that was distributed as a Gaussian with θ ~ φ(θ, σ), we would find that that the twice the negative log-likelihood would be

$$-2\left(\ln(\phi(\theta,\sigma)) - \ln(\phi(\hat{\theta},\sigma))\right) = -2\left(-\frac{1}{2}\left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2 + \frac{1}{2}\left(\frac{\hat{\theta}-\hat{\theta}}{\sigma}\right)^2\right) = \left(\frac{\theta-\hat{\theta}}{\sigma}\right)^2$$

Comparing with the previous, more general, result

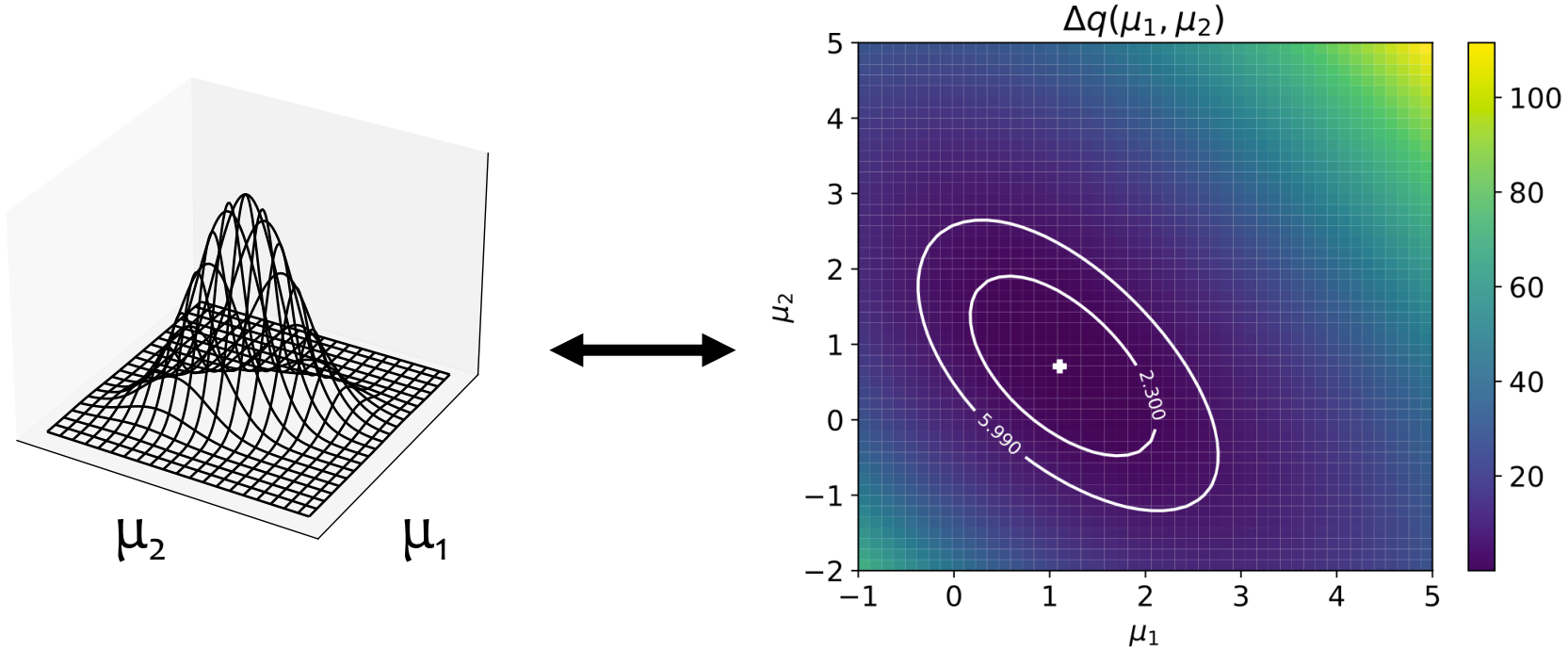
$$(q_N(\theta) - q(\hat{\theta})) = \frac{1}{2}q_N''(\hat{\theta})(\theta-\hat{\theta})^2$$

We must have

$$\boxed{\frac{1}{\sigma^2} = \frac{1}{2}q_N''(\hat{\theta})}$$

This may be familiar → the variance can be obtained by the 2nd derivative of the log-likelihood

Hessian approximation

More generally for a multi-normal distribution

$$\nu_{1,1}^{ij} = \left(\frac{1}{2} \left[\frac{\partial^2 q}{\partial \theta_k \partial \theta_l} \right] \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^{-1}_{ij}$$