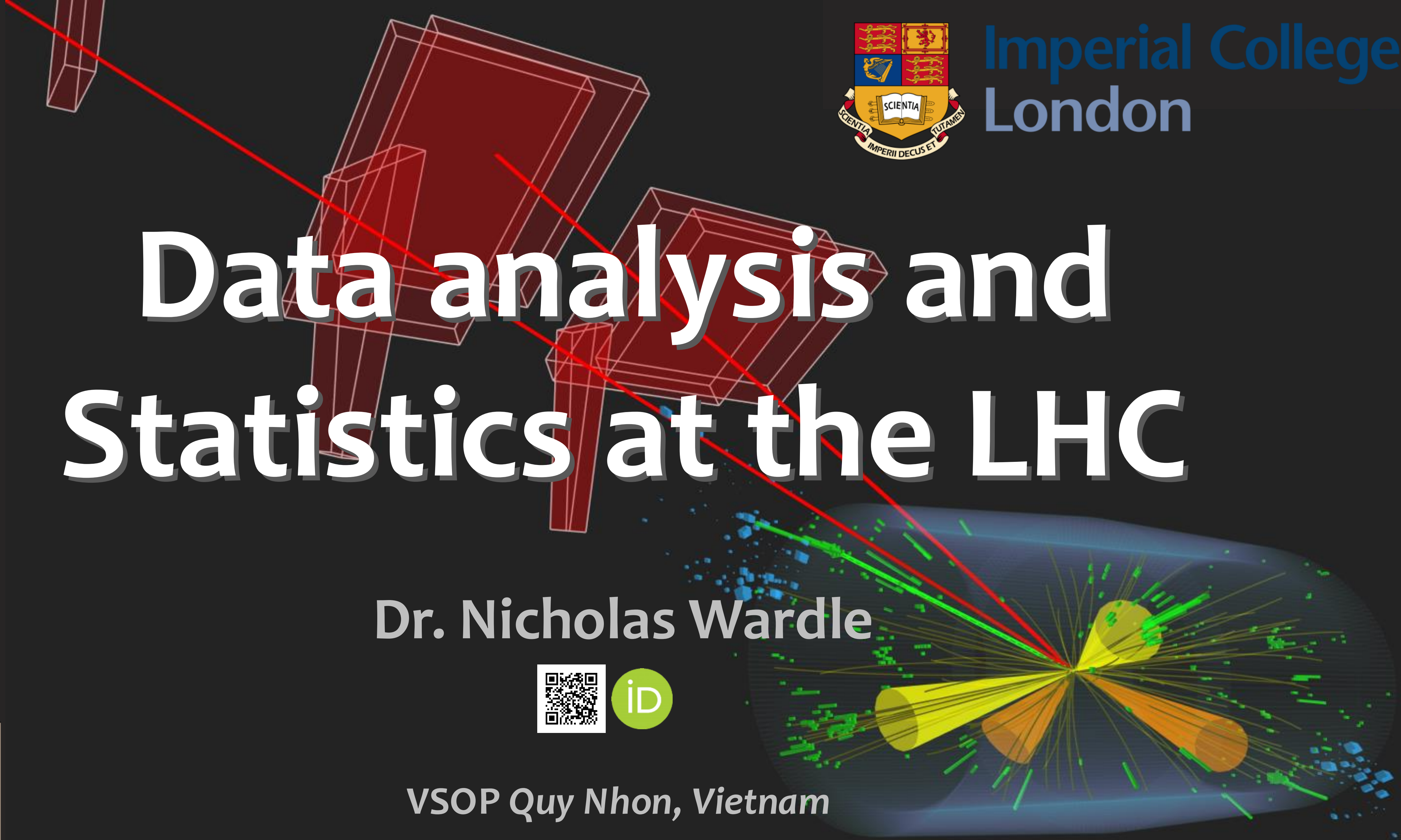


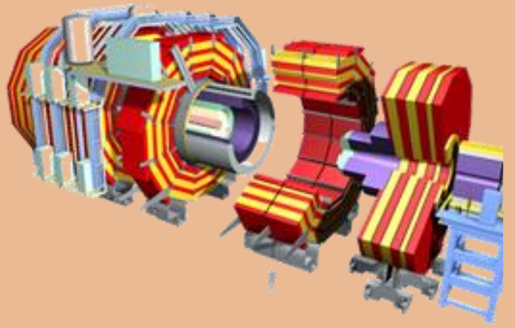
# Data analysis and Statistics at the LHC

Dr. Nicholas Wardle

VSOP *Quy Nhon, Vietnam*

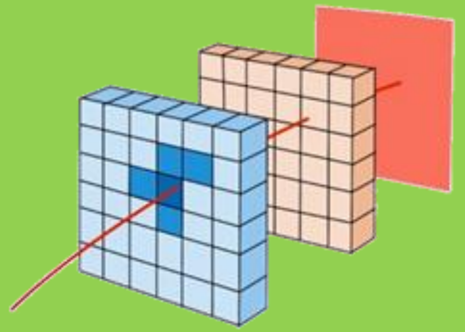
15-26 July 2024

## Data

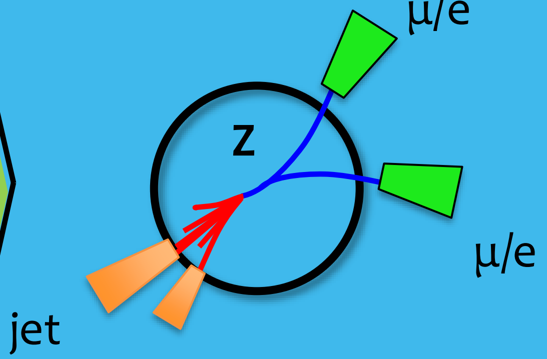


## Simulation

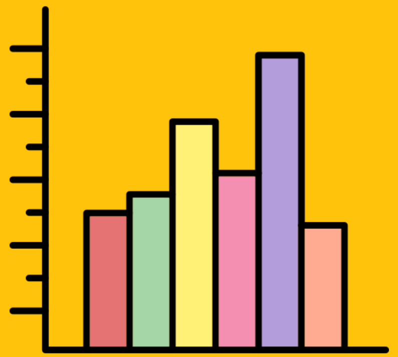
## Reconstruction & Particle ID



## Calibrations



## Event Selections & Distributions

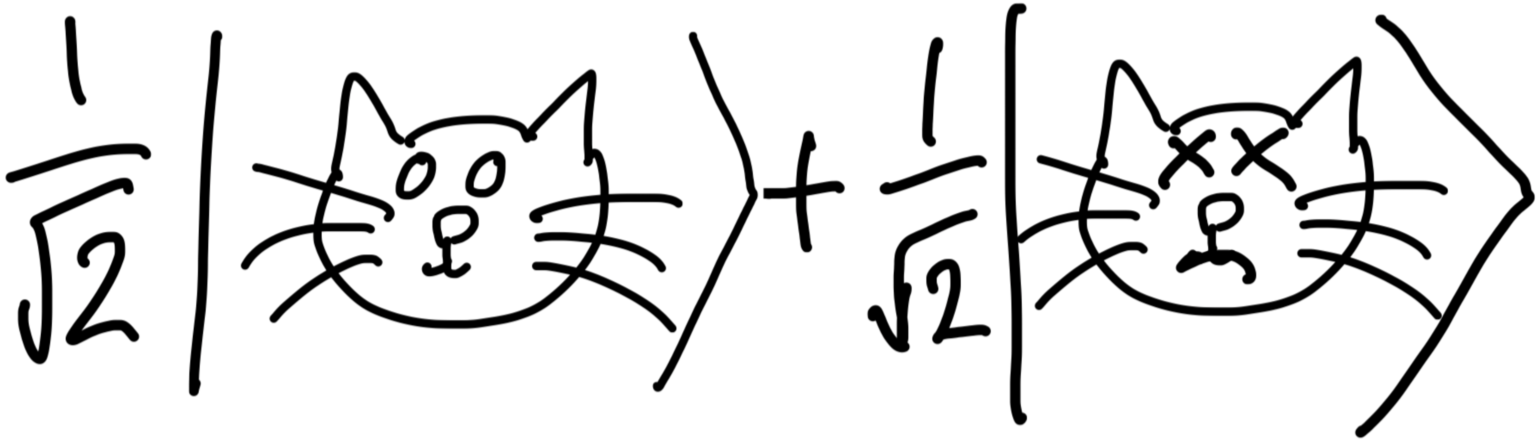


## RESULTS

# Statistical Analysis at the LHC

In particle physics,

1. Physical laws (and the observations we make) are ***probabilistic in nature*** due to quantum mechanics
2. The way we perform experiments (e.g collider experiments) means that ***events are statistically independent***

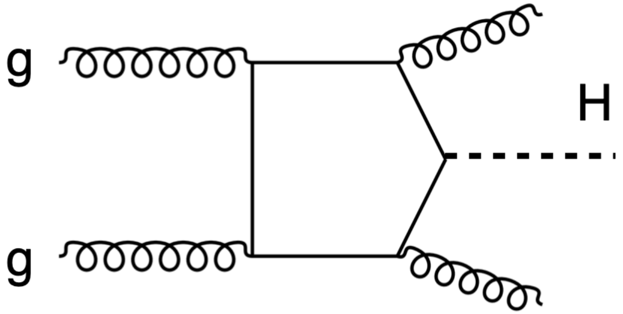
$\Psi =$ 

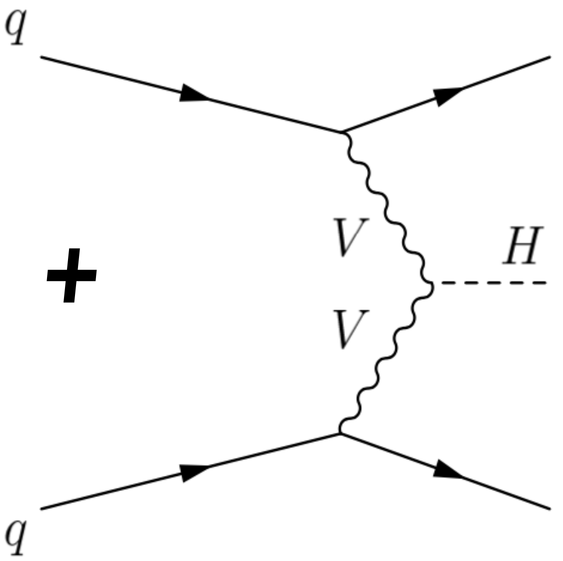
→***ideal scenario for applied statistics!***

# Statistical Analysis at the LHC

In particle physics,

1. Physical laws (and the observations we make) are ***probabilistic in nature*** due to quantum mechanics
2. The way we perform experiments (e.g collider experiments) means that ***events are statistically independent***

$\Psi \sim$ 

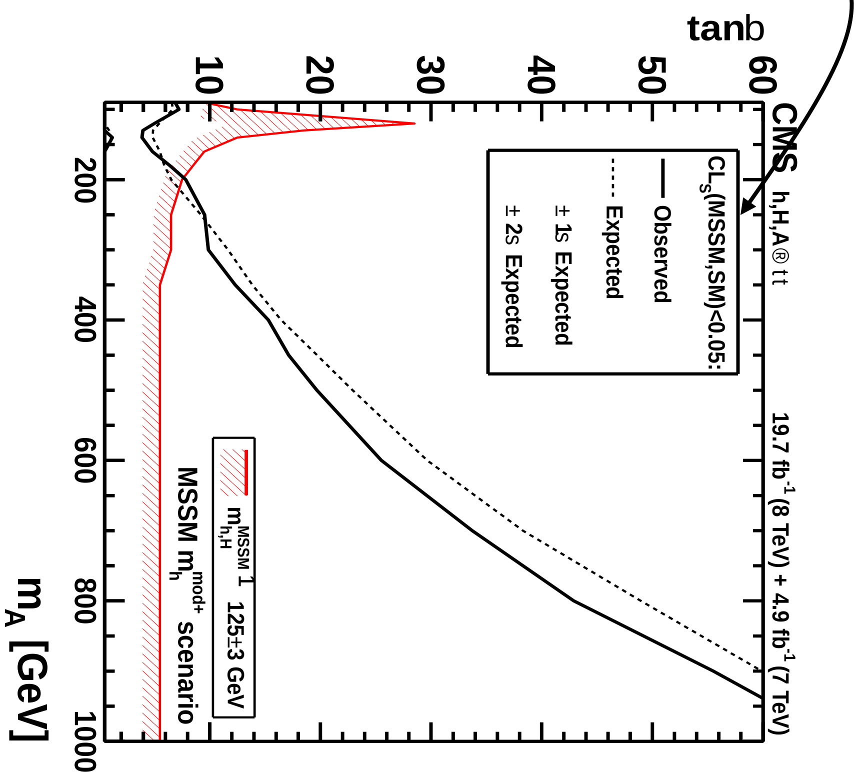
$+$ 

$+\ldots$

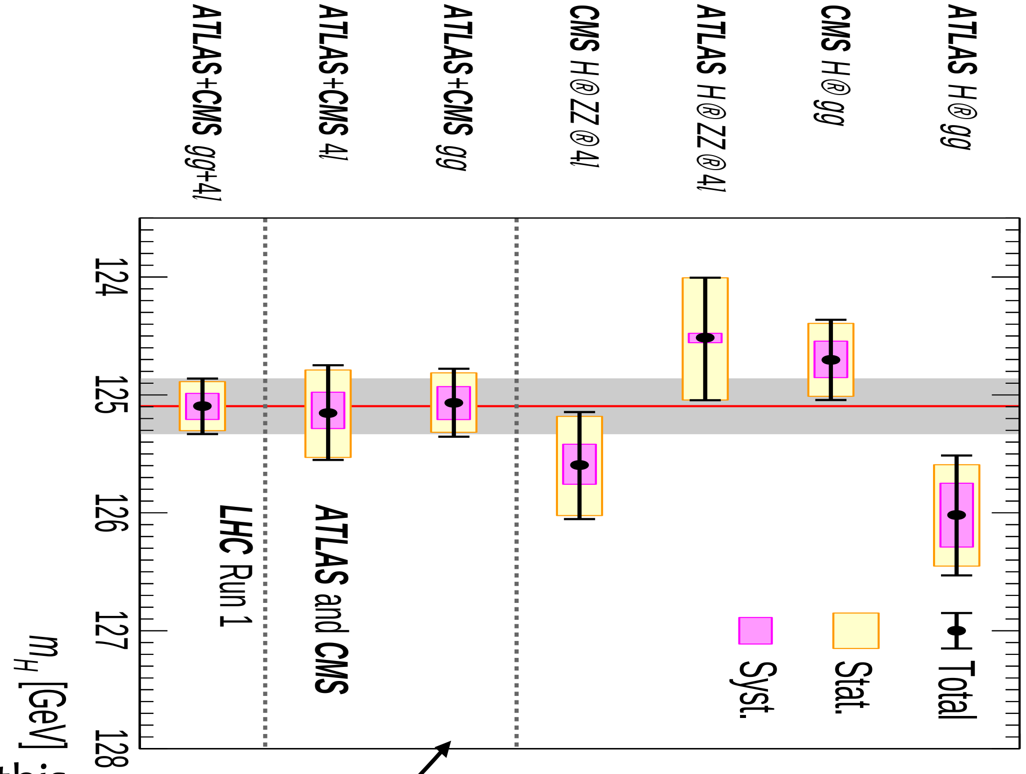
→***ideal scenario for applied statistics!***

# Statistical Analysis at the LHC

What does a plot like this



or this



actually tell us, and how are they made?

# 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,

- $P(X_i) \geq 0 \text{ for all } i$
- $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$ 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, Ω={1, 2, 3, 4, 5, 6}

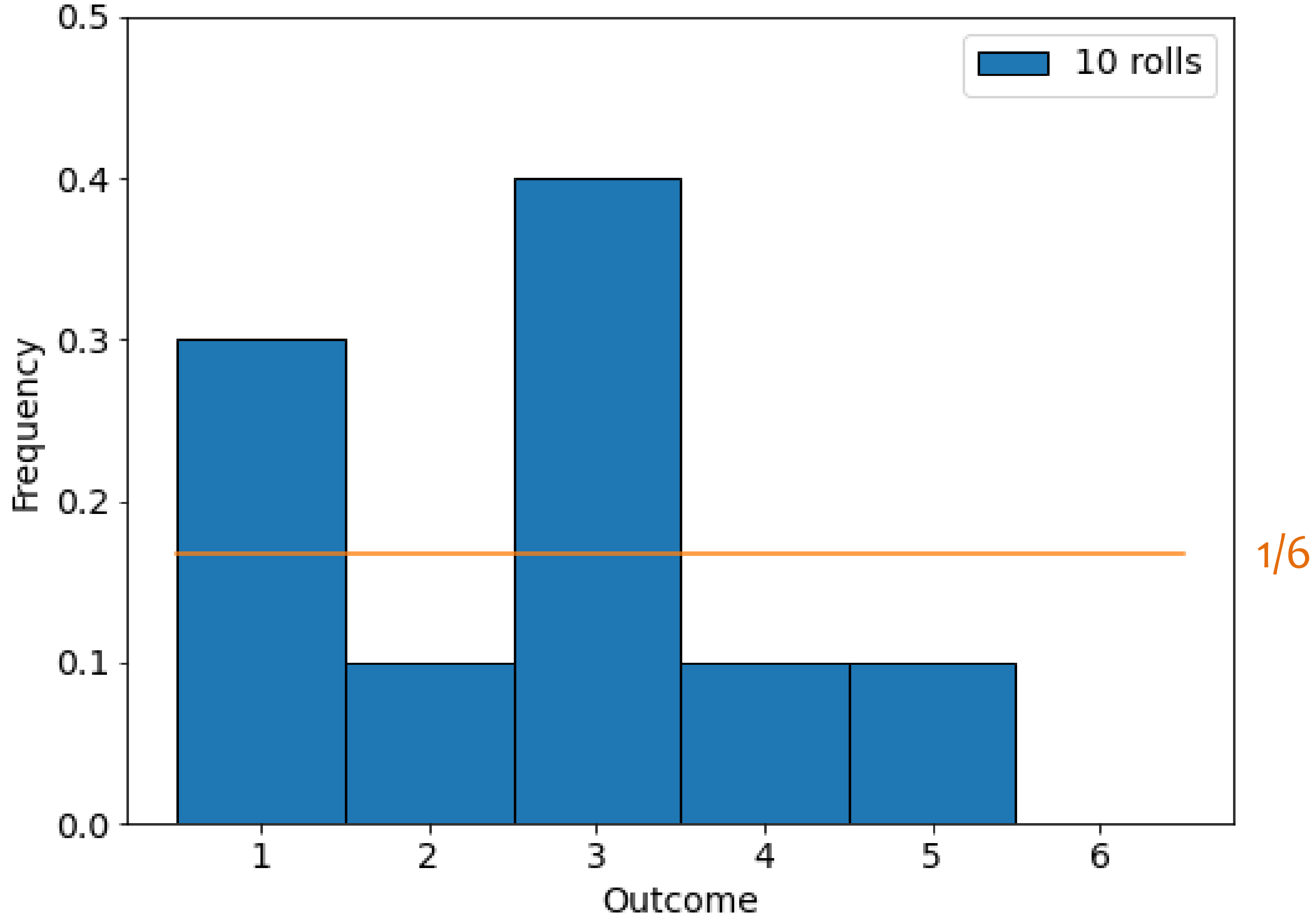
# Frequentist probability

Frequentist (or classical) probability is the definition 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

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



# Bayesian Probability

Bayesians think differently about the interpretation of probability.

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 very helpful!

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

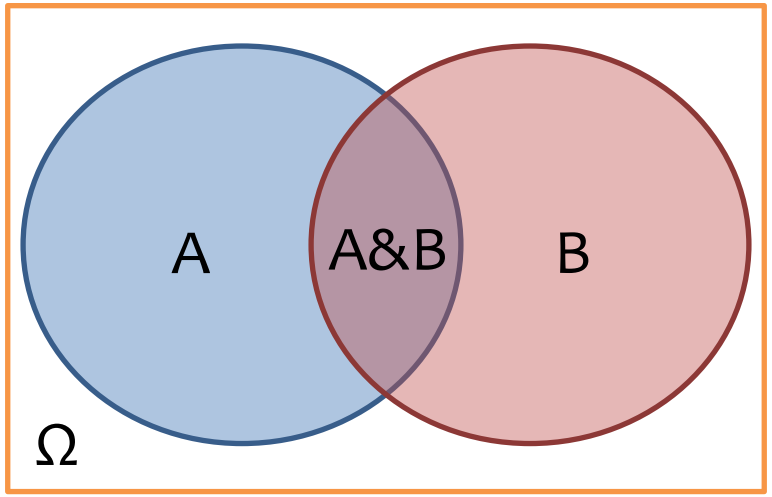
In these lectures, I'll avoid discussing Bayes' vs Frequentist interpretation of probability too much → usually for particle physicists we **choose the one which is most useful** for a given analysis

# Bayes Rule

The probability that two events ***A and B occur*** is given by

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

Probability that A occurs given that B has occurred



Rearranging gives us Bayes' rule!

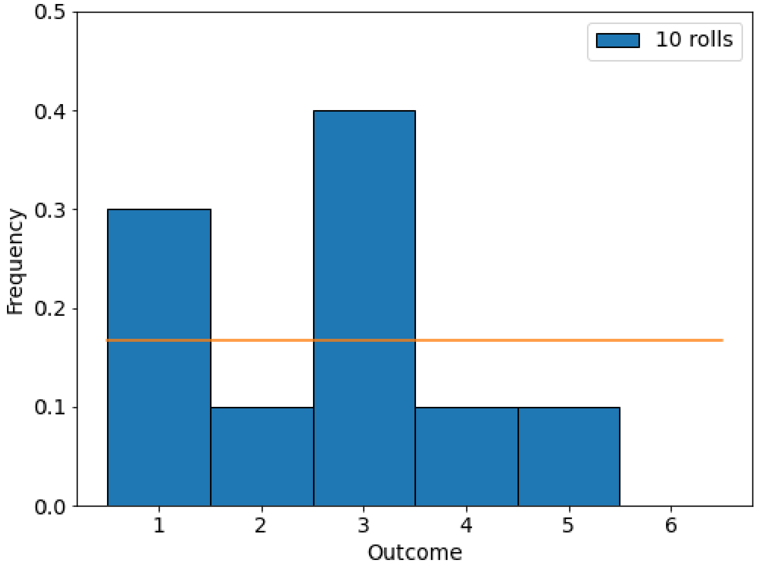
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# Probability Distributions

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

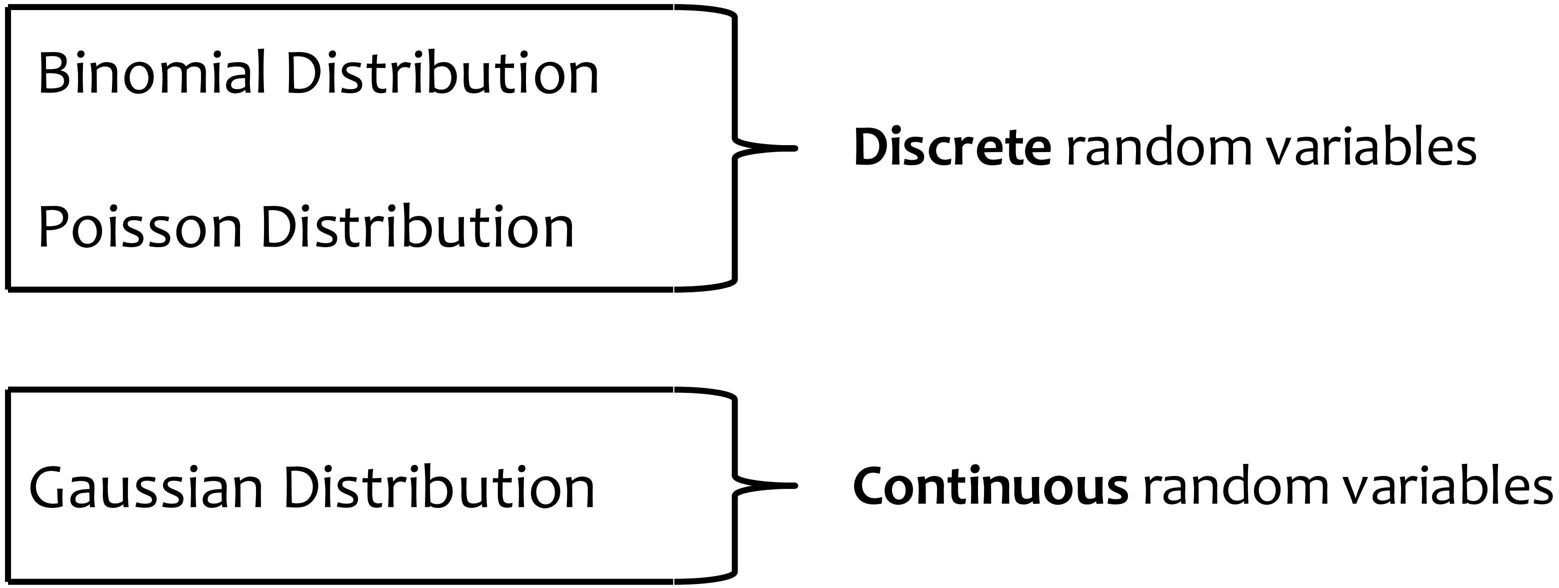
The corresponding probabilities P(X)Ω over all possible values of X ∈ Ω form a probability distribution. As an example, if X is the outcome of a single die roll, then Ω = {X = 1, X = 2,X = 3,X = 4,X = 5} and P(X) =1/6 for every possible value of X, meaning the **probability distribution is uniform**.

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 ~ U(1, 6)



# Most important Probability distributions

We'll take a look at a few common probability distributions (they're common at least to me but let me know if any of them sound completely off-topic)



Binomial Distribution

Poisson Distribution

**Discrete** random variables

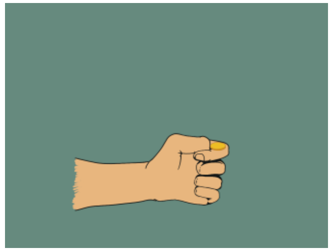
Gaussian Distribution

**Continuous** random variables

# 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

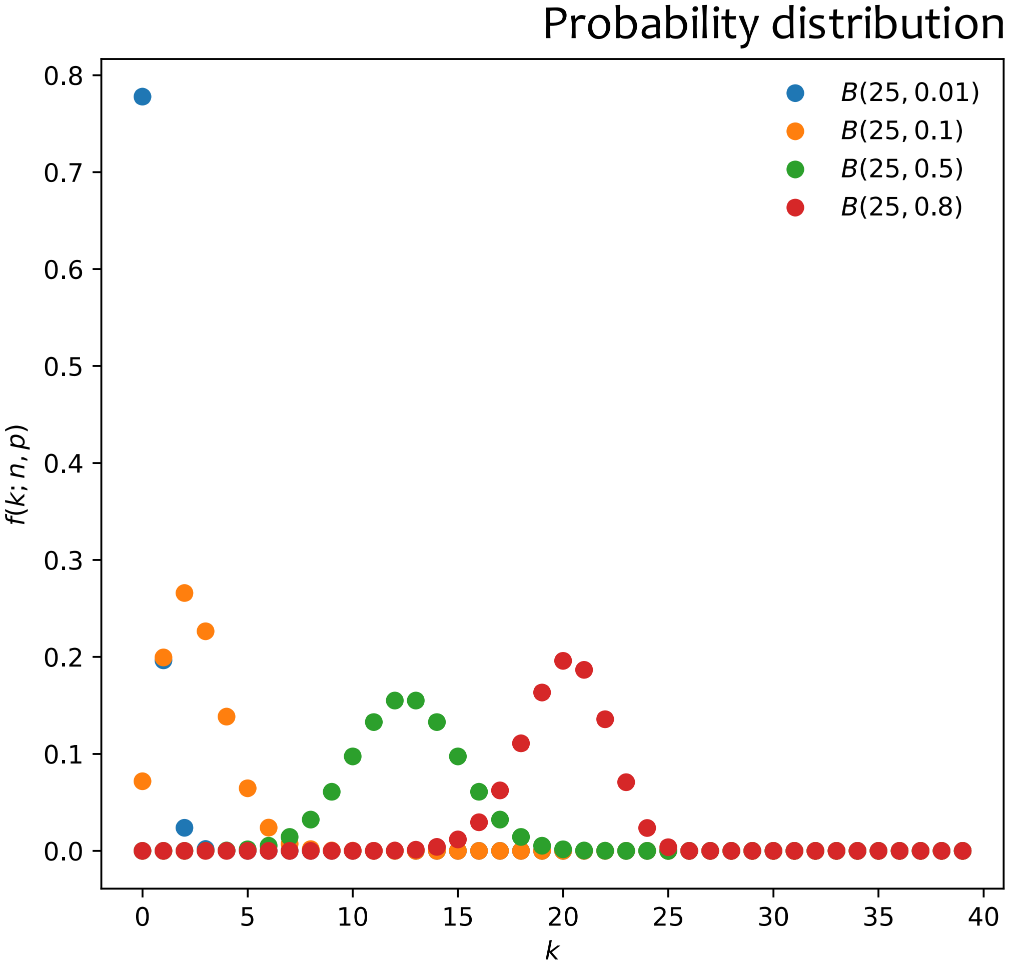
$$f(k;n,p) = \binom{n}{k} p^k q^{n-k}$$

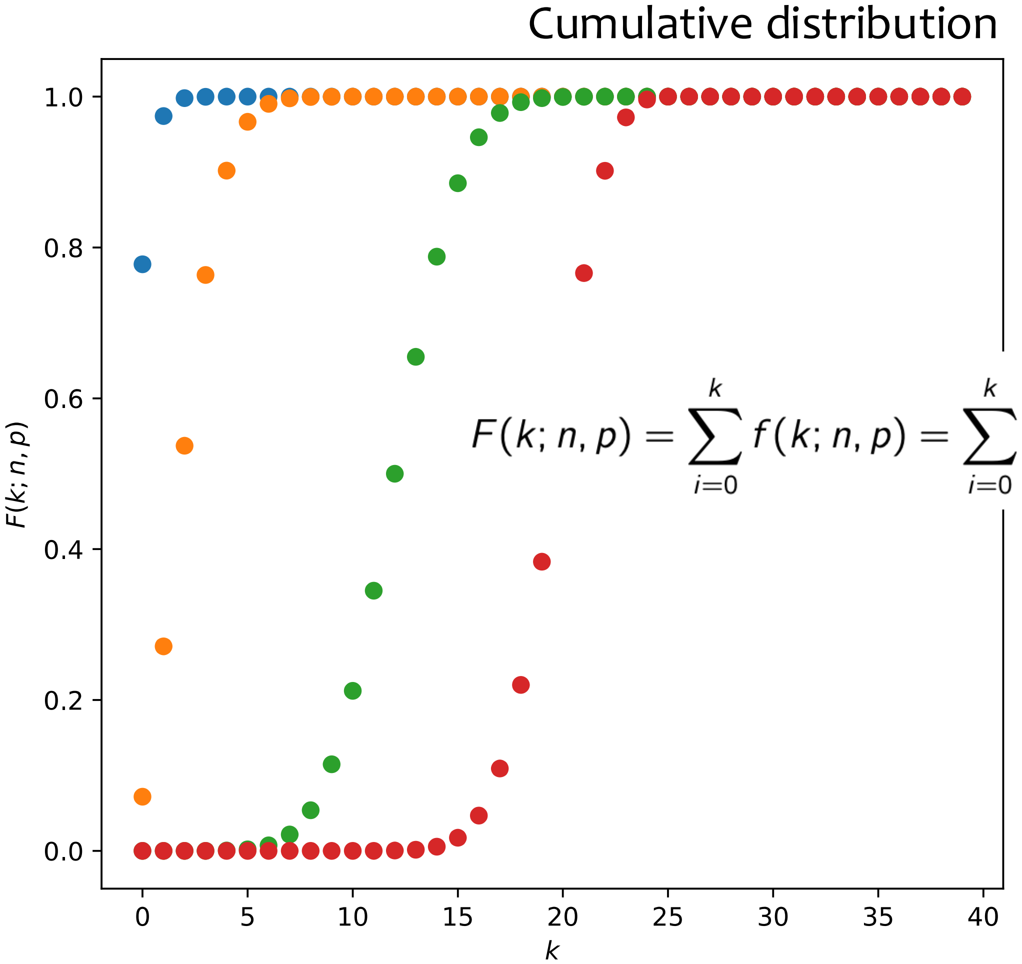
Distribution f

Parameters of "model" (n, p)

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

# Binomial Distribution





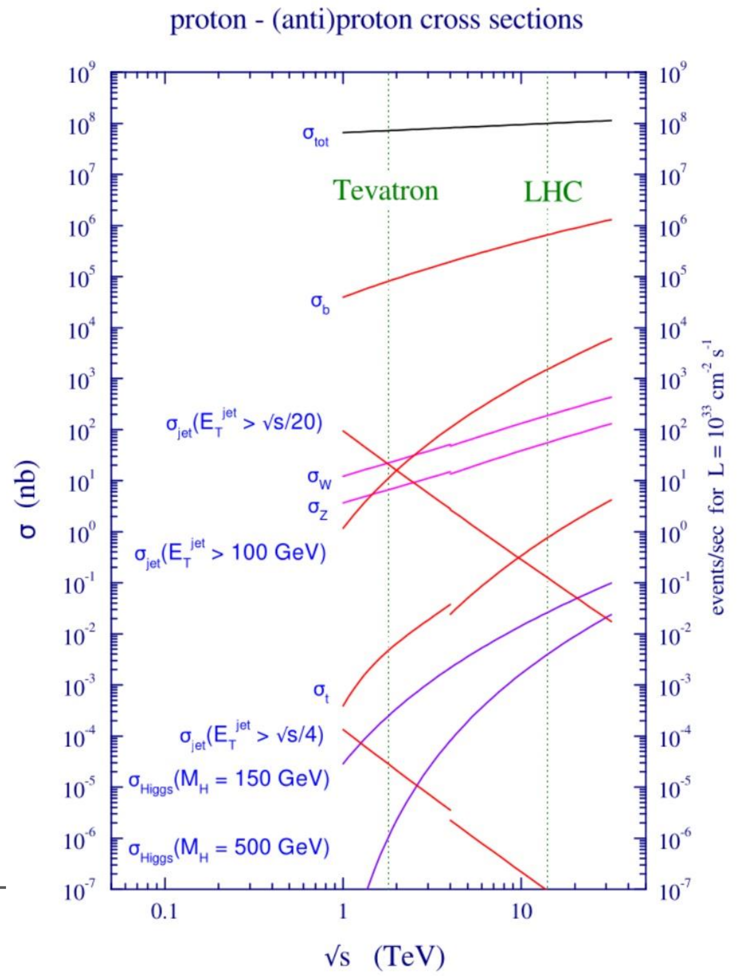
$$F(k;n,p)=\sum_{i=0}^{k} f(k;n,p)=\sum_{i=0}^{k}\binom{n}{k}p^{k}q^{n-k}$$

# Poisson Distribution

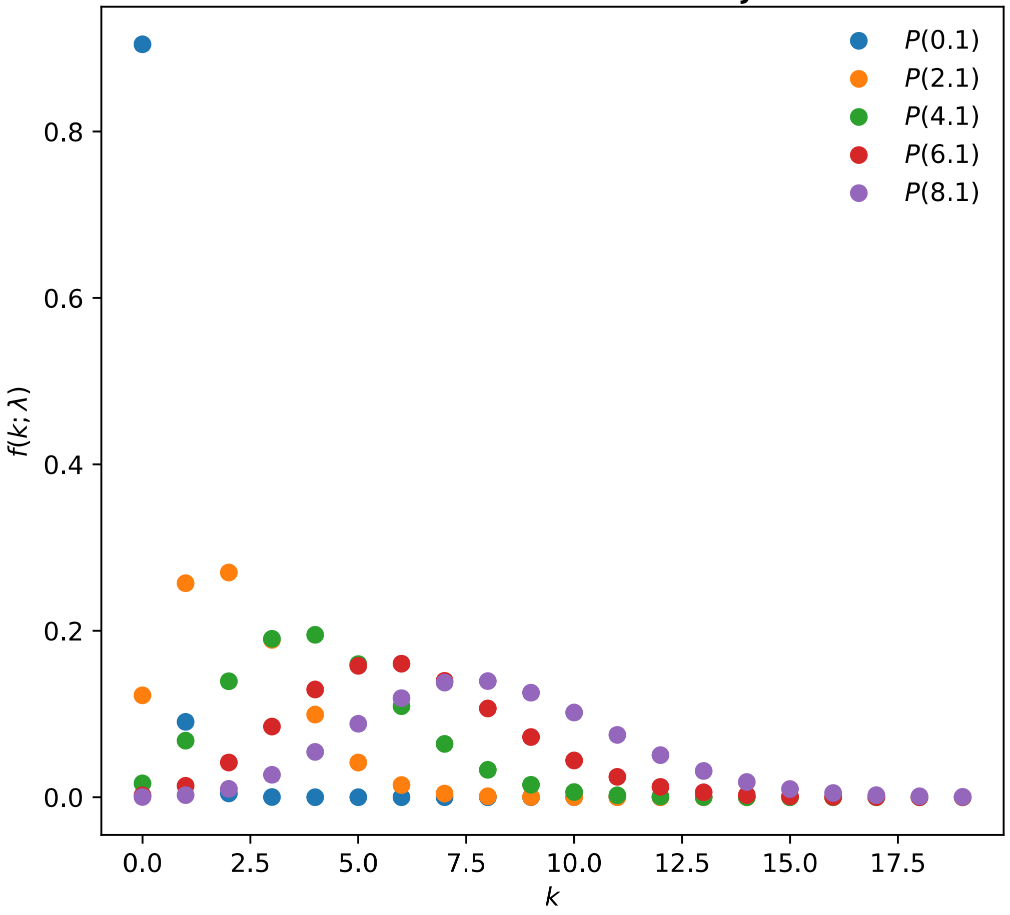
The Poisson distribution is a limit case of the binomial distribution where n → ∞, and p → 0 such that the product **λ = np** is constant.

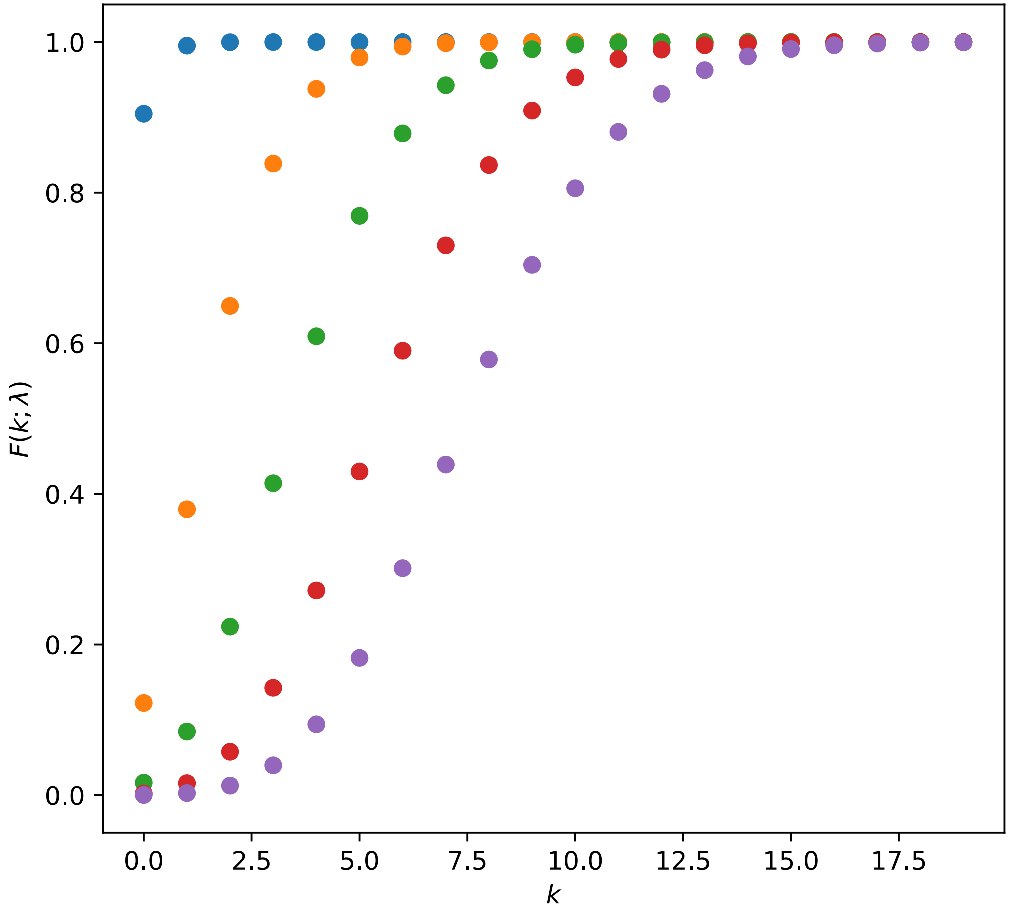
In particle physics, this is a very common distribution since it models processes that are rare (p → 0) in data sets that are very large (n → ∞).

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



# Poisson Distribution


Probability distribution


Cumulative distribution

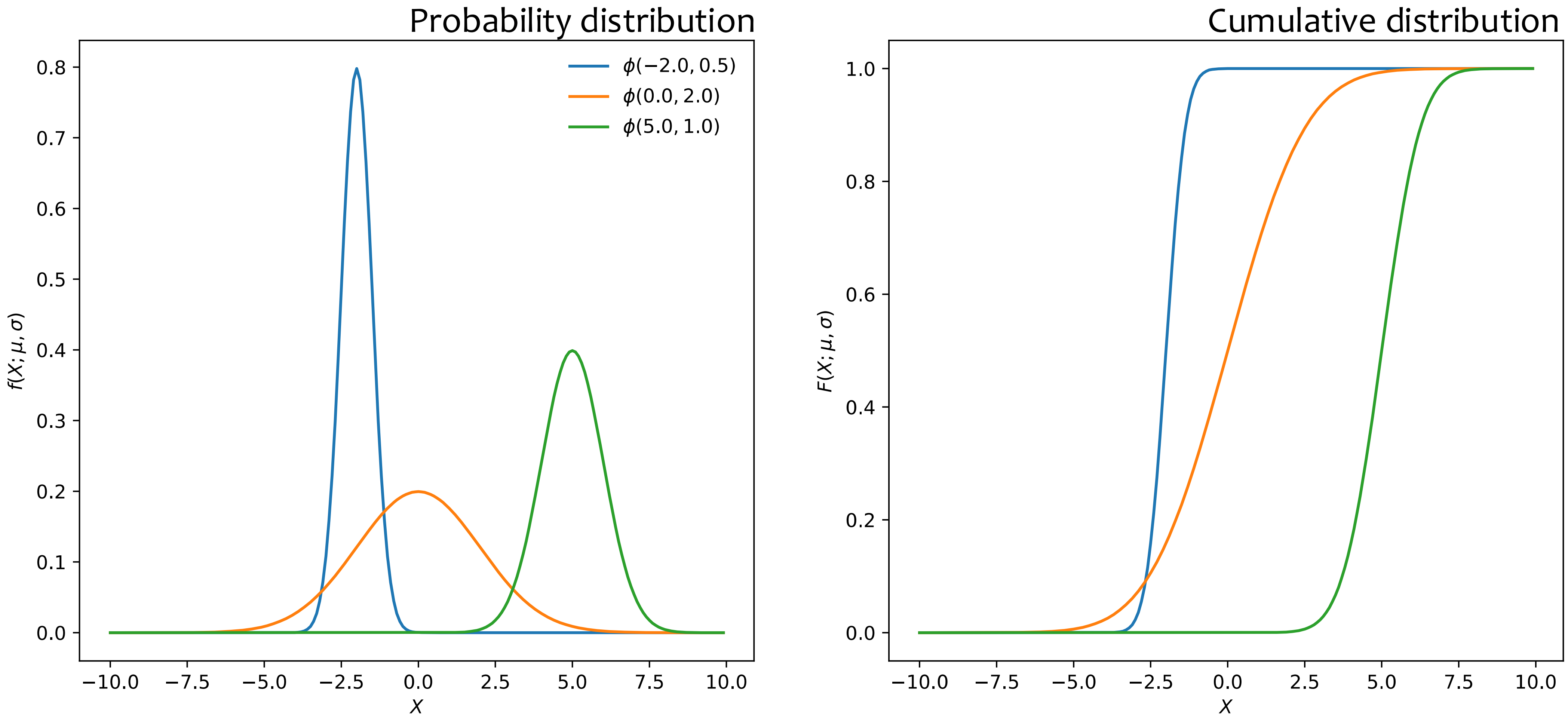
# Gaussian Distribution

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 distribution then,

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

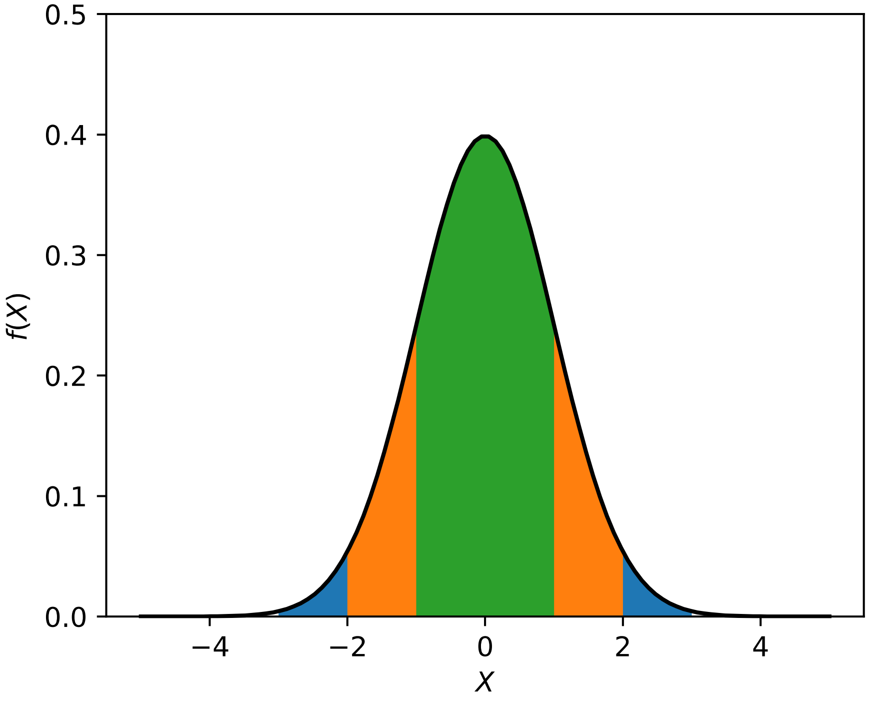
# Gaussian Distribution

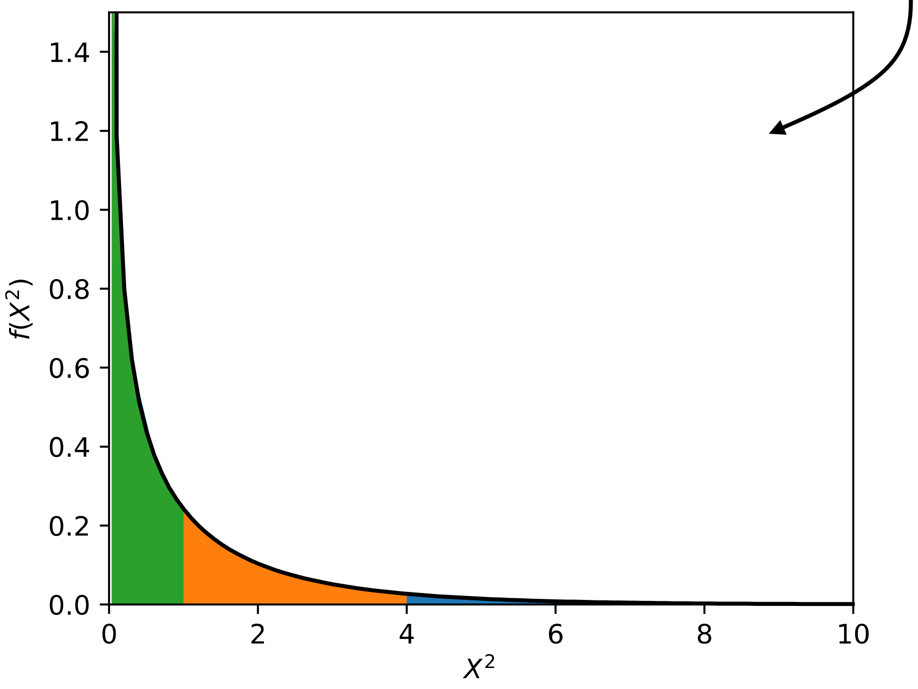


# $\chi^2$ distribution

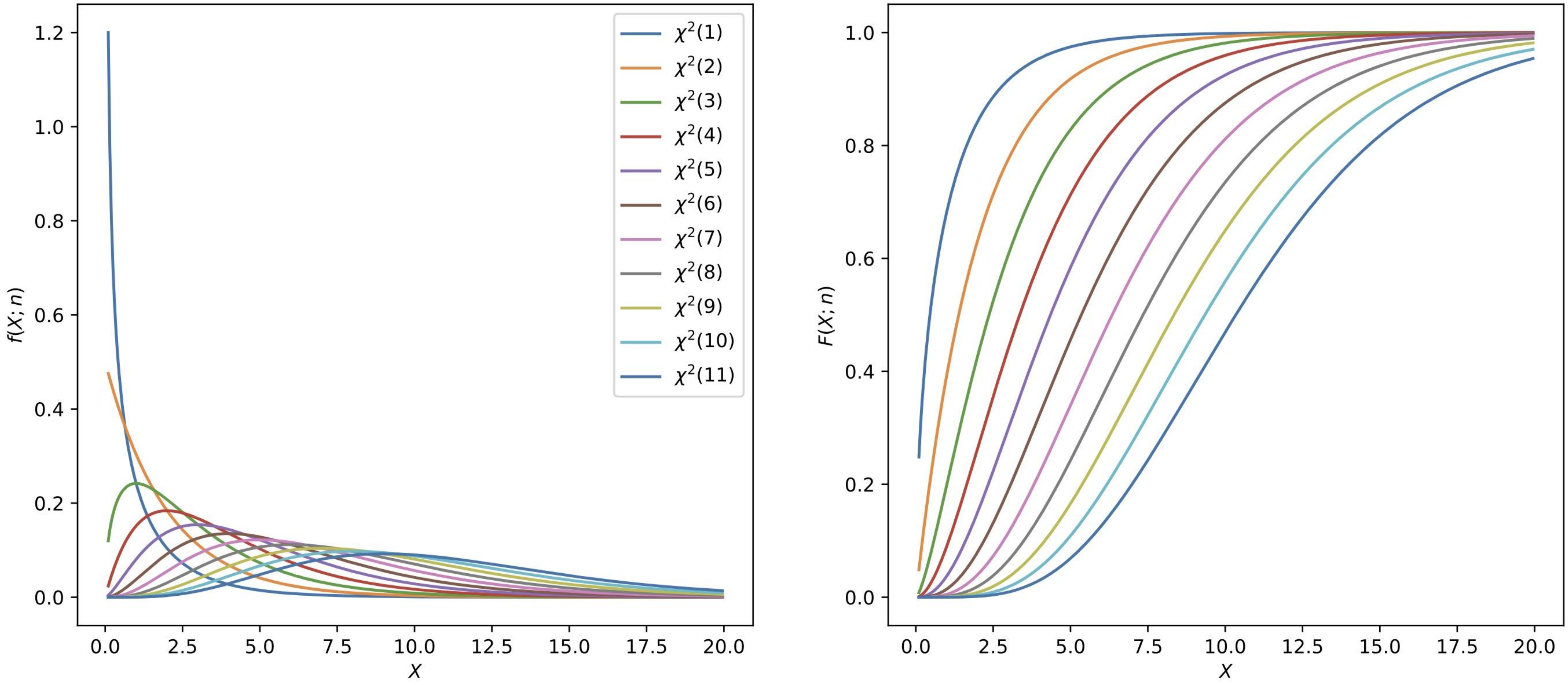
The $\chi^2$ distribution is related to the normal distribution. If $T \sim \varphi(T; 0, 1)$ then $X = T^2$ will be distributed as a $\chi^2(1)$ - "chi-square" distribution with 1 degree of freedom,

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





# $\chi^2$ distribution



# Multi-variate distributions

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

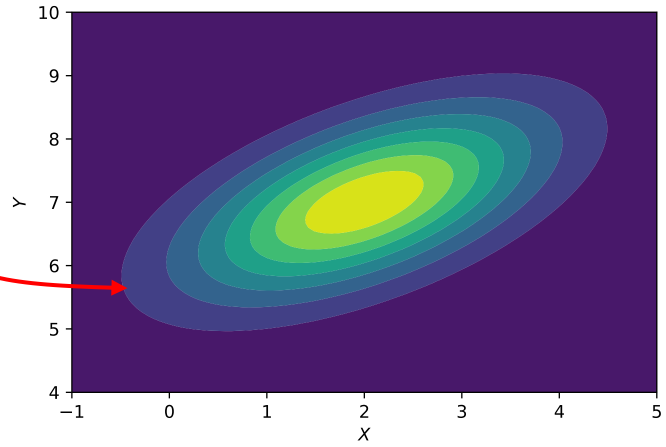
$$f(\boldsymbol{X};\boldsymbol{\mu},\mathbf{V}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{V})}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T (\mathrm{V}^{-1})(\boldsymbol{X}-\boldsymbol{\mu})}$$

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

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

$\mu_X = 2$
$\mu_Y = 7$
$\sigma_X = 1.22$
$\sigma_Y = 1$
$\rho\ = 0.57$

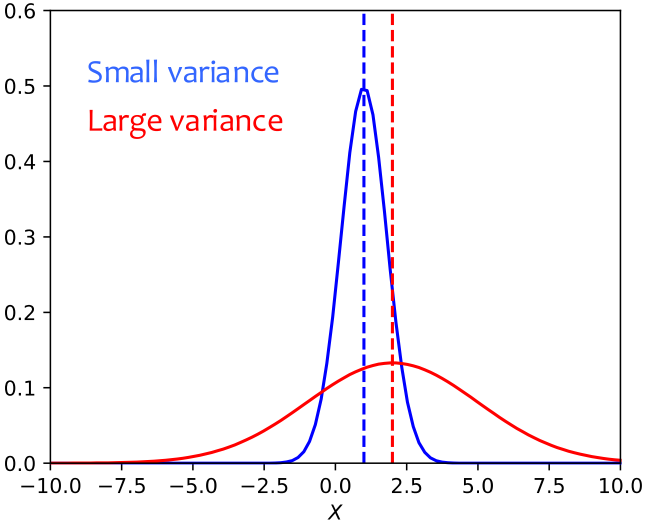


Multivariate distributions quickly become complicated and difficult to estimate (more about this later)

# Moments

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 ⟶ $E[a \cdot g(X) + b \cdot h(Y)] = a \cdot E[g(X)] + b \cdot E[h(Y)]$

The expectation value (mean) is often referred to as the first moment of the distribution of X, but of course we can define higher moments too. For example, the expectation of the function g(X) = (X – E[X])$^2$ is called the *variance** of X under f (X),

$$V(X) = E[(X - E[X])^2] = \int_{\Omega} (X - E[X])^2 f(X) dX$$

With the properties $V(X + a) = V(X)$ and $V(aX) = a^2 V(X)$ for a constant value *a*

# Moments

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 ($X_1, X_2, \ldots, X_N$ ) of size N, we can define the n–th sample moment as,

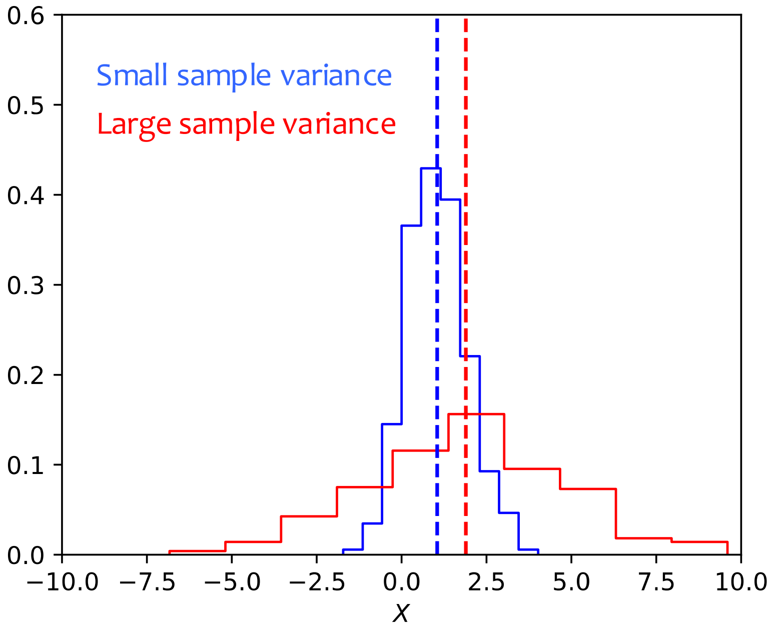
$$m_n = \frac{1}{n}\sum_i^N X_i^n.$$

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,

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

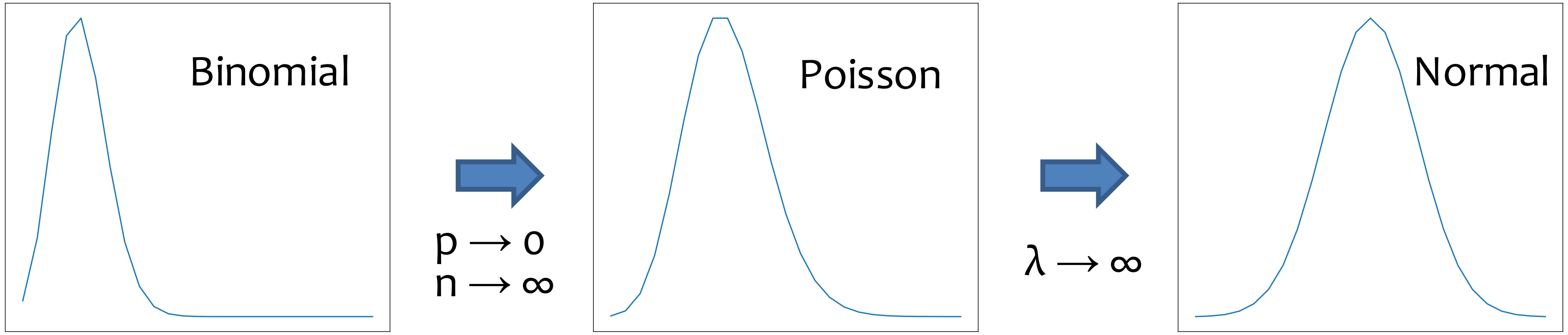


# Moments

| **Probability distribution** | $E[\cdot]$ | $\mathrm{Var}[\cdot]$ |
|---|---|---|
| $f(k;n,p) = \binom{n}{k} p^k q^{n-k}$ | $np$ | $np(1-p)$ |
| $f(k;\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$ | $k$ | $\sqrt{k}$ |
| $f(X;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$ | $\mu$ | $\sigma$ |

# The Central Limit Theorem

In certain limiting cases, distributions of random variables can be well approximated by other (often more convenient) distributions.
E.G …



Binomial

$p \rightarrow 0$
$n \rightarrow \infty$

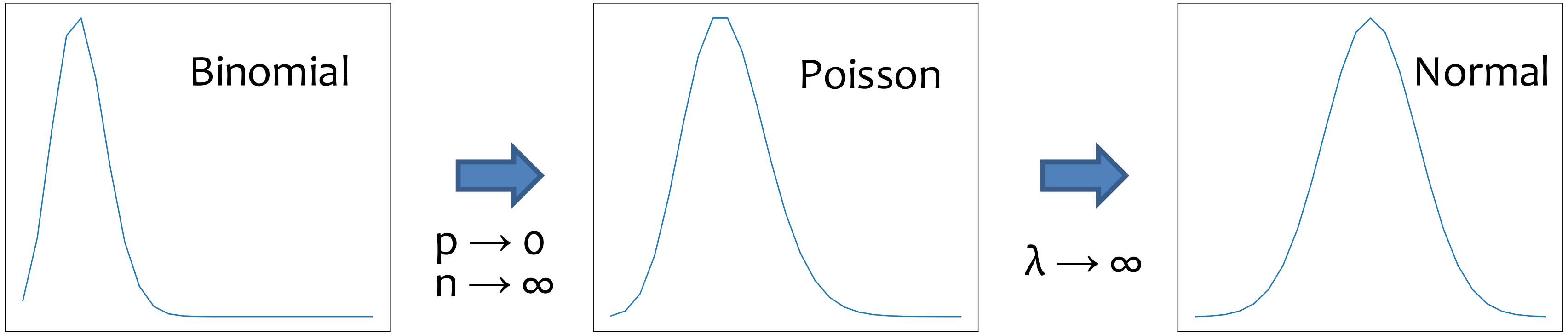
Poisson

$\lambda \rightarrow \infty$

Normal

# The Central Limit Theorem

In certain limiting cases, distributions of random variables can be well approximated by other (often more convenient) distributions.
E.G …

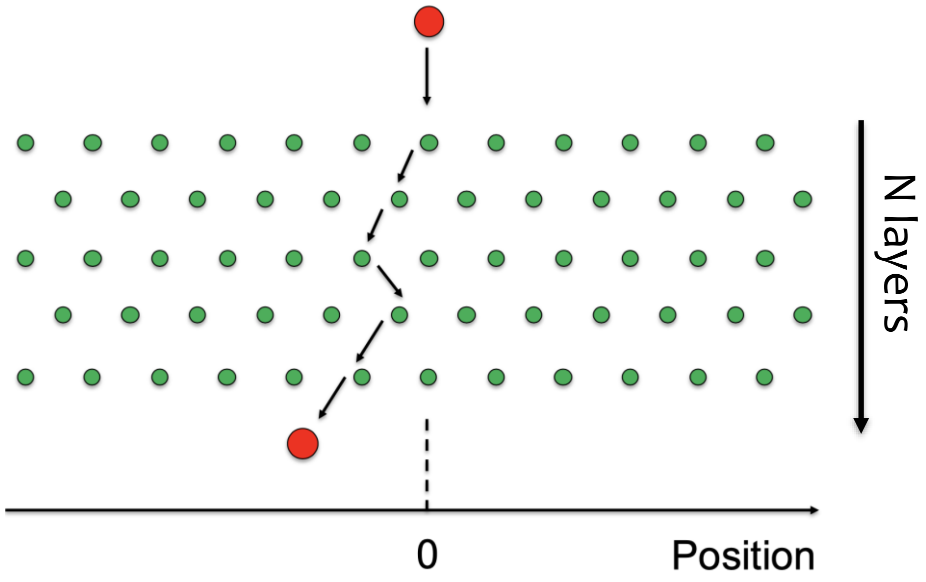


**Theorem**: Suppose we have a sequence of independent random variables $X_i$ , each from a distribution with mean $\mu^i_1$ and variance $\nu^i_2$. Define,

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

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

# The Central Limit Theorem

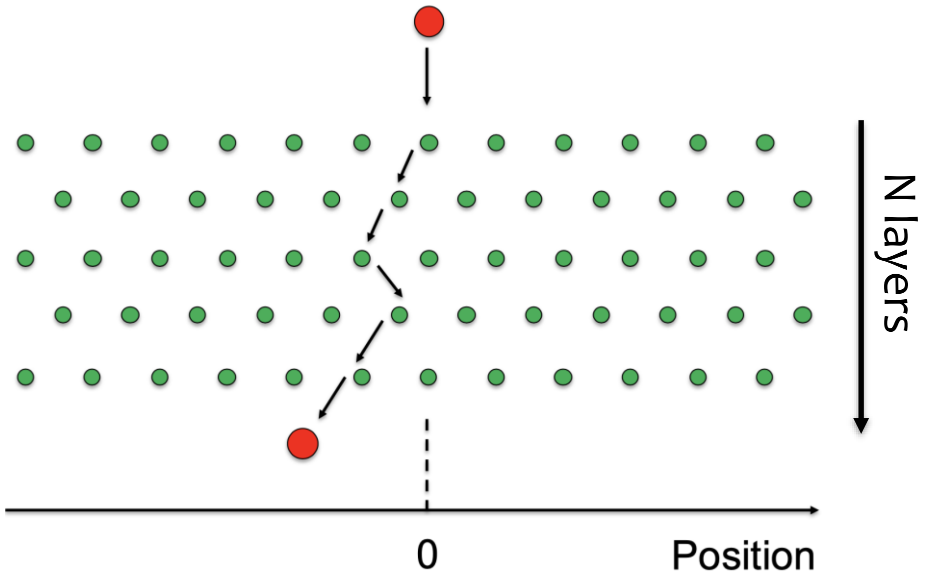


Example: a Galton board with N layers

The change in position after each layer $X_i$ can be thought of as a uniform random variable $X_i \in [-1, 1]$

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

# The Central Limit Theorem



Example: a Galton board with N layers

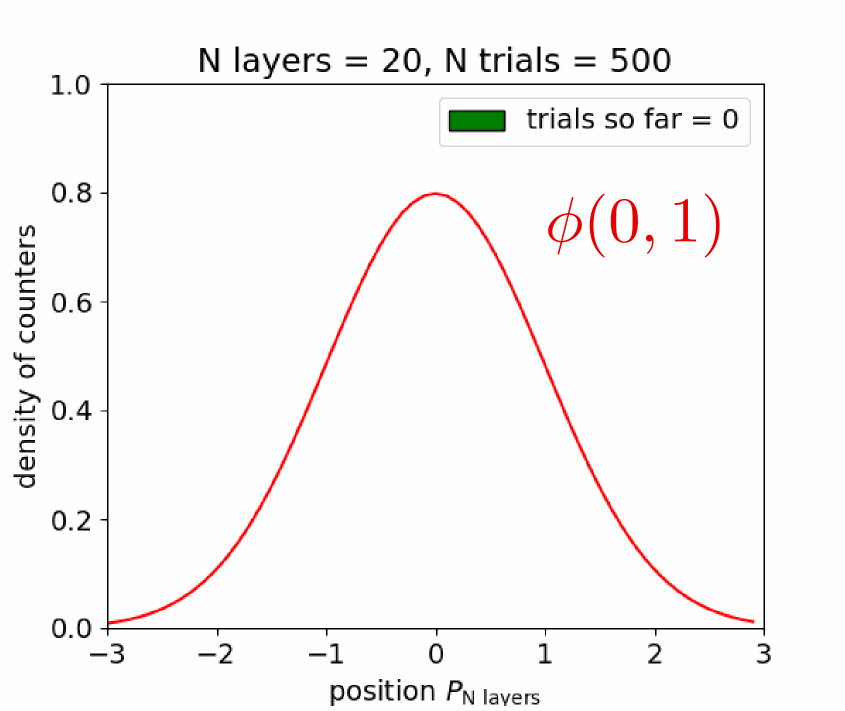
The change in position after each layer $X_i$ can be thought of as a uniform random variable $X_i \in [-1, 1]$

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

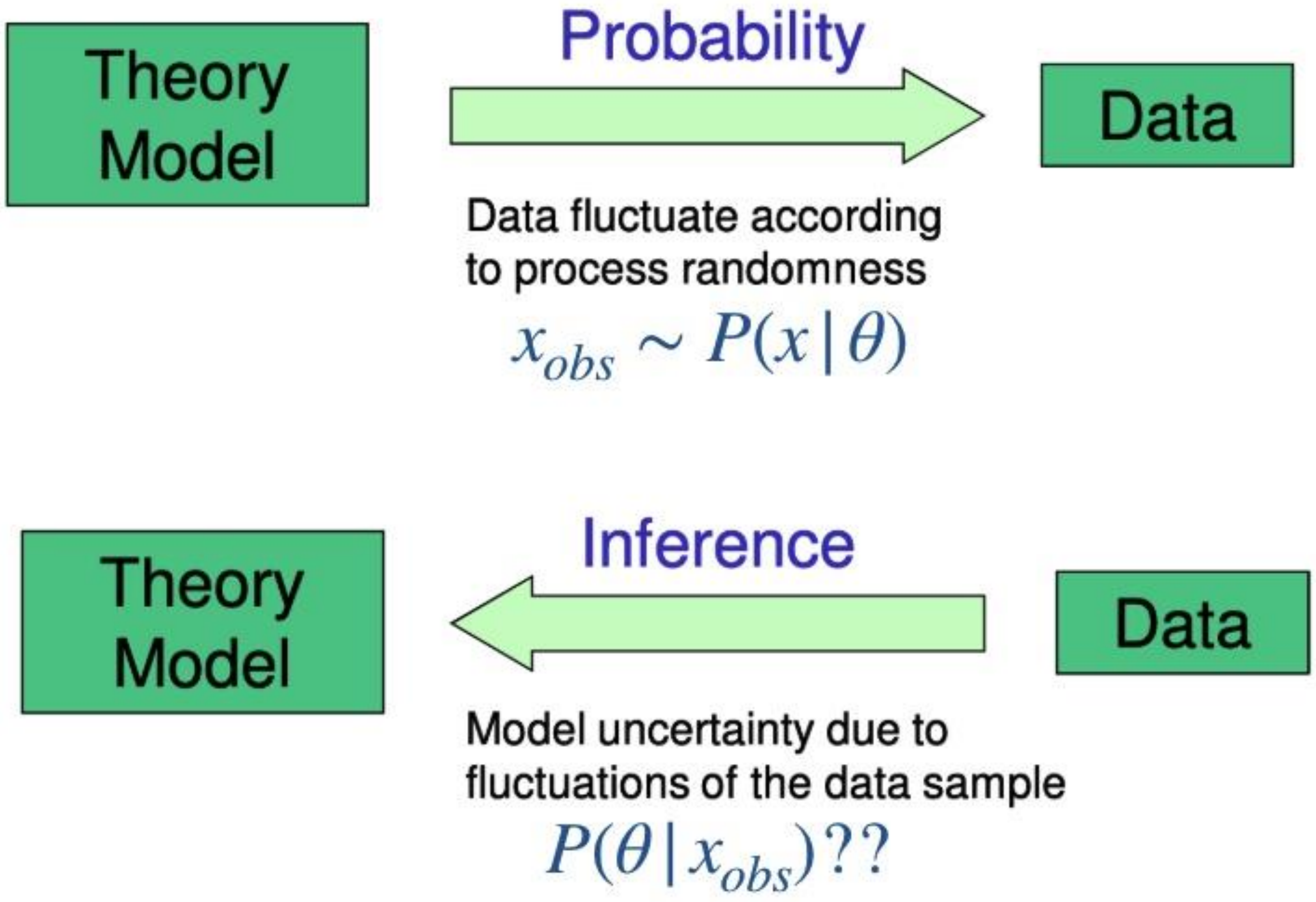
$$T_N = \frac{\bar{X}N - \sum_i \mu_1^i}{\sqrt{\sum_i v_2^i}} = \frac{\frac{N}{N}\sum_i X_i - 0}{\sqrt{N}} = P_N$$

Since E[X] = 0, Var[X] = 1

From CLT $$P_N \rightarrow \phi(0,1)$$



# Statistical Inference



Data fluctuate according to process randomness

$$x_{obs} \sim P(x|\theta)$$

Model uncertainty due to fluctuations of the data sample

$$P(\theta|x_{obs})??$$

N. Smith

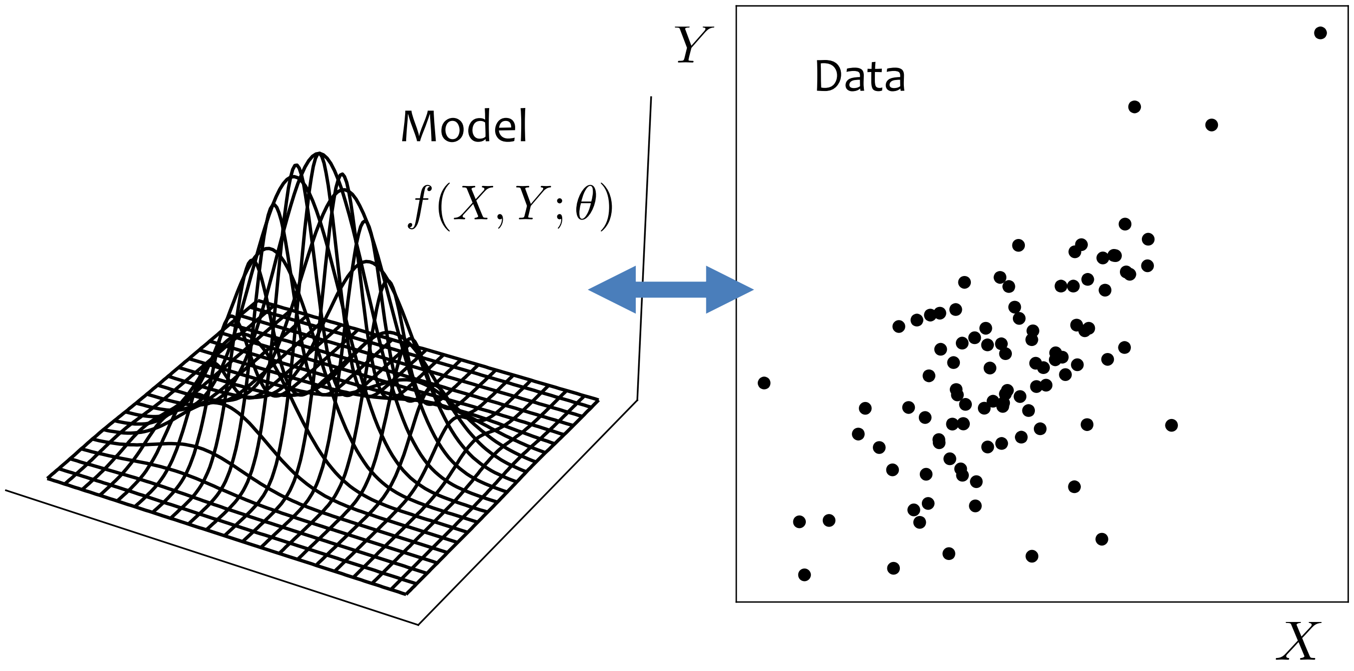
# Estimators

So far, we only talked about probability distributions → Where does the data come in?

Typically, we want to learn something about our model using the data → estimators is the first thing we can use the data for

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

- $\theta$ usually a parameter of a **model**
- $\hat{\theta}$ is a property of a specific set of observations → its value depends on the **data** observed and **it is itself a random variable**



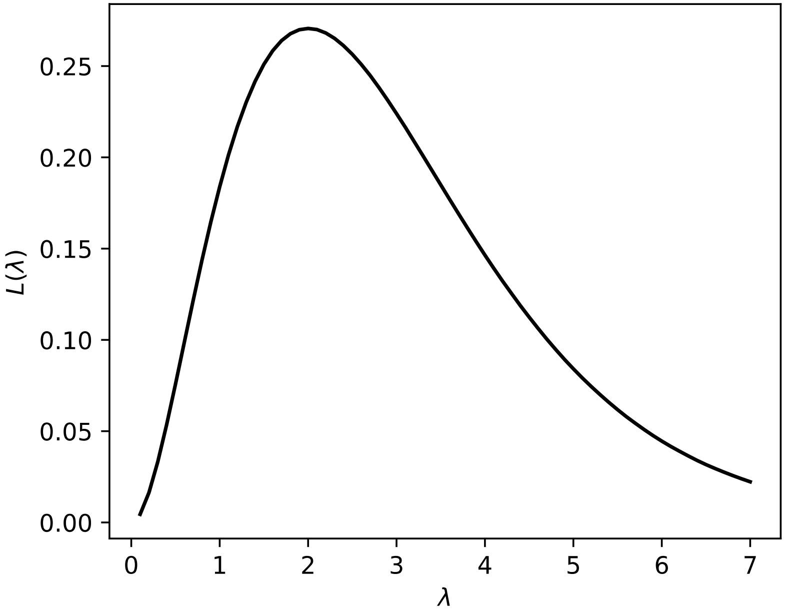
There are many different estimators, but we will focus on the most common one used at the LHC → the **maximum likelihood estimator**

# Likelihoods

The likelihood is a function of the model parameters **θ** at a fixed value for the data **X**

The likelihood function is proportional to the probability density P $L(\theta) \propto P(X|\theta).$

**Example:** For the Poisson distribution, let's suppose we observe the number of events k=2

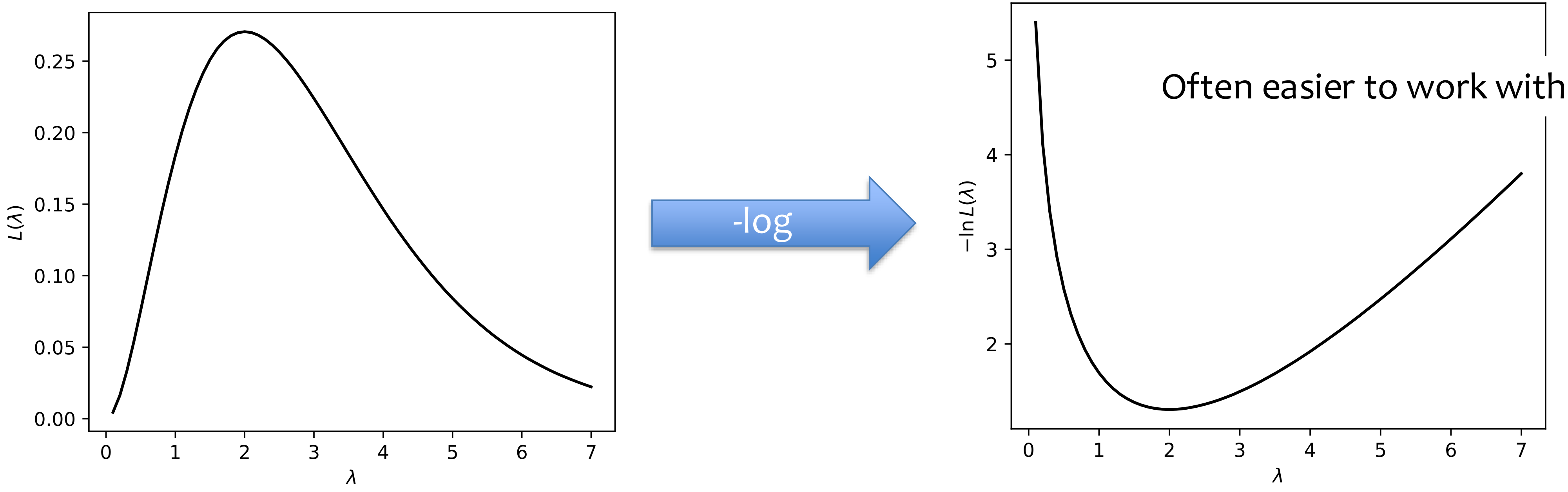


# Likelihoods

The likelihood is a function of the model parameters **θ** at a fixed value for the data **X**

The likelihood function is proportional to the probability density P $$L(\theta) \propto P(X|\theta).$$

**Example:** For the Poisson distribution, let's suppose we observe the number of events k=2



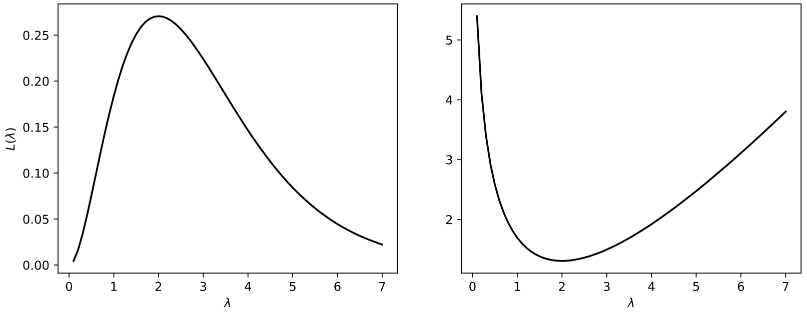
Often easier to work with $q = -\log(L)$

# Maximum likelihood estimator

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

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

Maximum likelihood ←→ Minimum –ve log likelihood



**Example:** For the Poisson distribution

$$q = \lambda - k\ln\lambda - \ln(k!)$$

$$\frac{dq}{d\lambda} = 1 - \frac{k}{\lambda} \implies \hat{\lambda} = k$$

# Numerical minimization (Gradient Descent)

iminuit

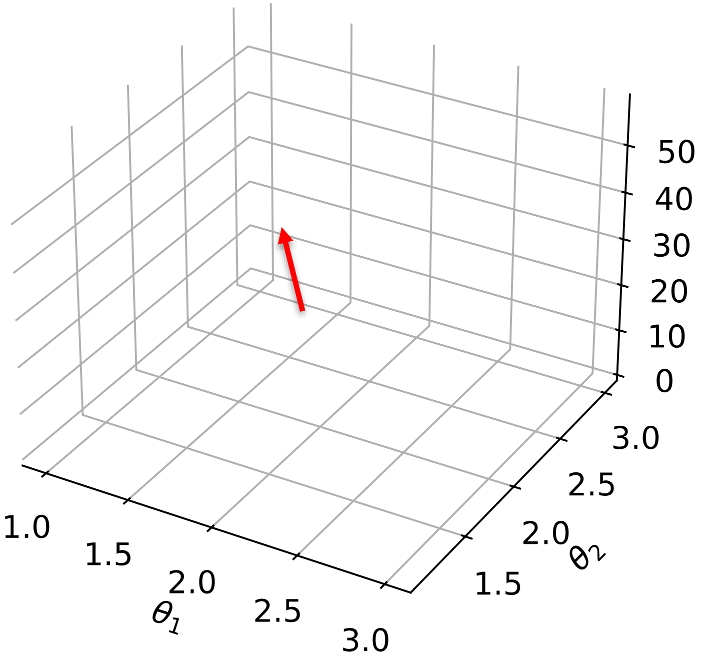
TensorFlow SciPy

ROOT Data Analysis Framework

...

We use many different computational tools for finding the MLE. Many of these different tools use a common algorithm → **gradient descent**

$$\nabla(q)|_{\boldsymbol{\theta}_{init}} = \begin{bmatrix} \frac{\partial q}{\partial \theta_1} \\ \frac{\partial q}{\partial \theta_2} \\ \vdots \\ \frac{\partial q}{\partial \theta_n} \end{bmatrix}_{\boldsymbol{\theta}=\boldsymbol{\theta}_{init}}$$



Iterative process

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{init} - k \times h \times \nabla(q)|_{\boldsymbol{\theta}_{init}}$$

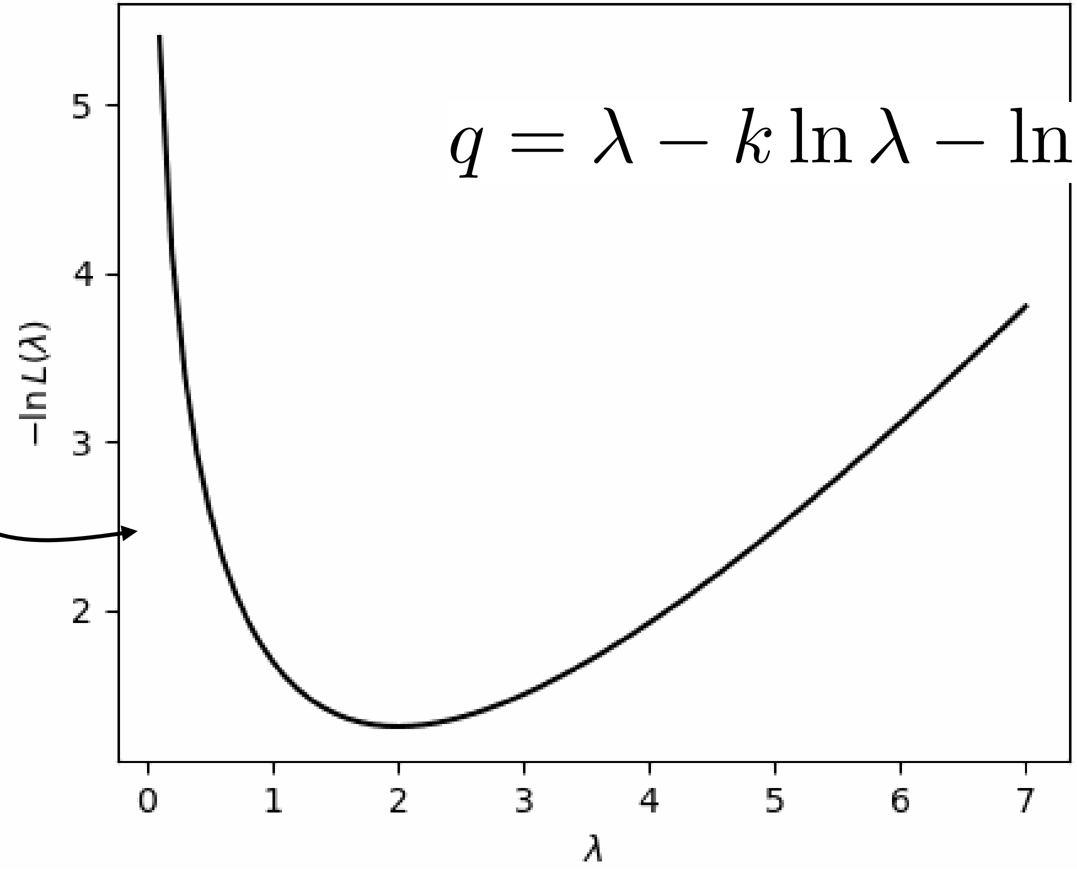
$$\nabla(q)|_{\boldsymbol{\theta}_{k*}}$$

Stop when $|\nabla(q)|_{\boldsymbol{\theta}} = \sqrt{\left(\frac{\partial q}{\partial \theta_1}\right)^2 + \left(\frac{\partial q}{\partial \theta_2}\right)^2 + \ldots + \left(\frac{\partial q}{\partial \theta_n}\right)^2} < \epsilon$

# Numerical minimization (Gradient Descent)

We typically use numerical minimization algorithms since often calculating these analytically is not possible → there's lots of these tools available for you to use in common HEP software

**Example:** For the Poisson distribution with **k =2**



$$q = \lambda - k\ln\lambda - \ln(k!)$$

Simple gradient descent algorithm finds same analytic solution

# Properties of MLE

MLEs have **very nice properties**, which is why they are common in LHC analyses

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

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

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**!

# Properties of MLE

MLEs have **very nice properties**, which is why they are common in LHC analyses

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

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

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**!

2. Relationships between parameter estimates are **preserved**

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

This is obvious since

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

# Properties of MLE

MLEs have **very nice properties**, which is why they are common in LHC analyses

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

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

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**!

2. Relationships between parameter estimates are **preserved**

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

This is obvious since

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

3. MLEs are **consistent estimators** i.e the value of the **estimator converges to the true value** as more data are included (n → ∞)

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \qquad \text{for any } \varepsilon > 0$$

# Variance of the Poisson estimator

From the 3rd property, we know that the variance of the MLE* converges to zero as k → ∞

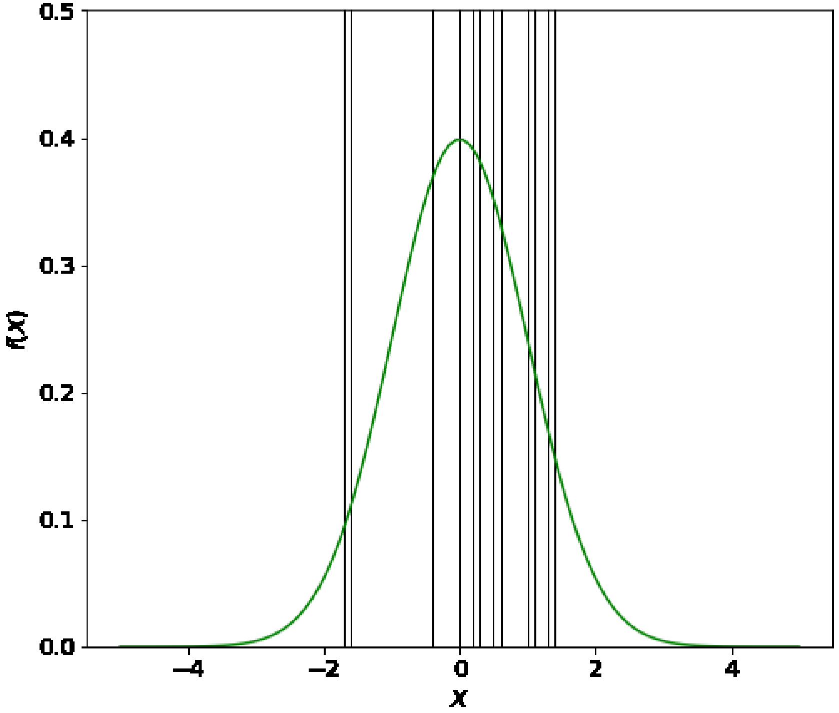
**Example,** Poisson random variable

$$\mathrm{Var}\left(\frac{\hat{\lambda}}{\lambda}\right) = \frac{1}{\lambda^2}\mathrm{Var}(k) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

Which means for larger samples, this will decrease to zero

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

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



* Generally, the law of large numbers tells us this is true for other estimators too

# (Co)variance of Gaussian estimators

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,

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

But if instead we Taylor expand

$$2\left(q(\theta)-q(\hat{\theta})\right) = 2\left(\cancelto{=0}{q'}(\hat{\theta})\cdot(\theta-\hat{\theta}) + \frac{1}{2}q''(\hat{\theta})(\theta-\hat{\theta})^2\right)$$

We must have $$\frac{1}{\sigma^2} = q''(\hat{\theta})$$

So the variance can be estimated from the second derivative of q!

# (Co)variance of Gaussian estimators

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,

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

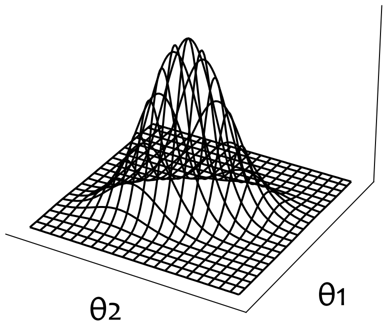
But if instead we Taylor expand

$$2\left(q(\theta)-q(\hat{\theta})\right) = 2\left(q'(\hat{\theta})\cdot(\theta-\hat{\theta}) + \frac{1}{2}q''(\hat{\theta})(\theta-\hat{\theta})^2\right)$$

We must have $$\frac{1}{\sigma^2} = q''(\hat{\theta})$$

So the variance can be estimated from the second derivative of q!

More generally



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

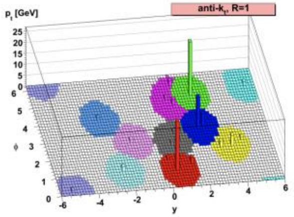
# Combined likelihoods

Our data at the LHC is much richer than a single Poisson random variable

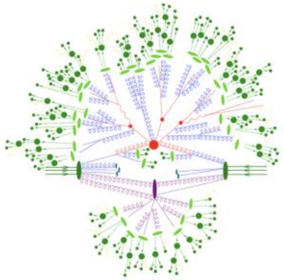
*(Example from K. Cranmer)*

Observables | Detector interactions | Shower splittings | Parton-level momenta | Theory parameters

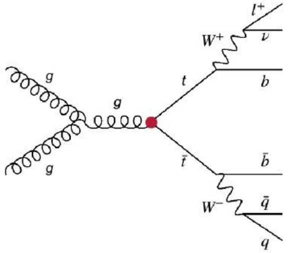
$$x \longleftarrow z_d \longleftarrow z_s \longleftarrow z_p \longleftarrow \theta$$



[Source: M. Cacciari, G. Salam, G. Soyez 0802.1189]



[Source: F. Krauss]



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

Evaluating this integral is essentially impossible → reduce the dataset x → x’ and use simulation to estimate p(x’|θ)

# Combined likelihoods

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

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

where $f_i$ are the p.d.f for each observation $X_i$

# Combined likelihoods

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

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

where $f_i$ are the p.d.f for each observation $X_i$

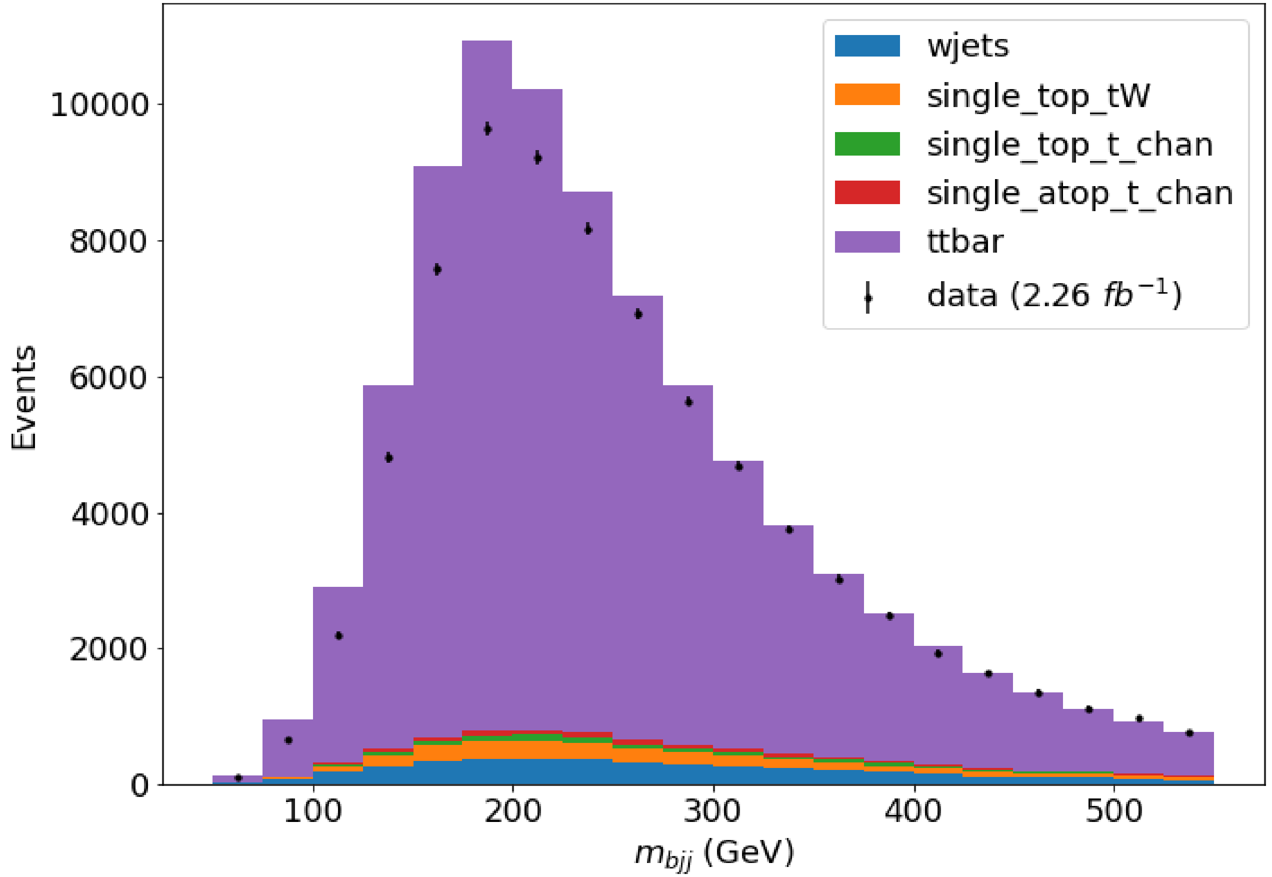
A histogram can be treated as a set of N independent Poisson random variables where N = number of bins!

$$f_i(k_i;\lambda_i) = \frac{\lambda_i^{k_i}}{k_i!}e^{-\lambda_i}$$

At the LHC, we split the contributions into **signal (S)** and **backgrounds (B)**

$$\lambda_i = rS_i + B_i$$

And L → L(*r*) is a function of the **signal strength**

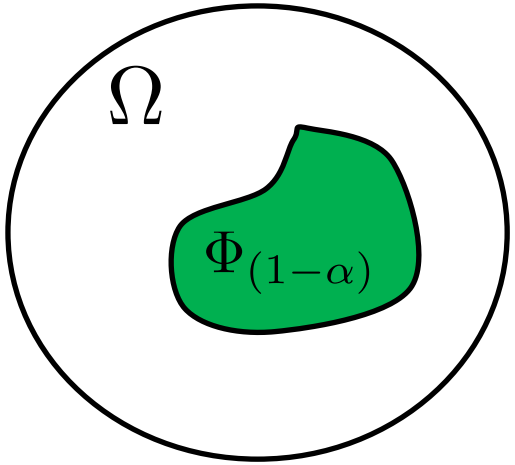


# Confidence Intervals

An estimator is only part of making a measurement at the LHC → we also report ***uncertainties.*** In **frequentist statistics,** we use the concept of a confidence interval

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

The confidence region at a confidence level 100 x (1-$\alpha$) is a region which is constructed such that the true values of the true parameters $\theta_0$ is contained in the region with a probability (1-$\alpha$)



Example: a **68% confidence interval** in θ will contain the true value $\theta_0$ in **68% of the outcomes in data**

# Confidence Intervals

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

We use the ratio of likelihoods

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

**Note the extra factor of 2 now**

Let's look at what this looks like for our Poisson random variable where n = 4

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

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

This is large when

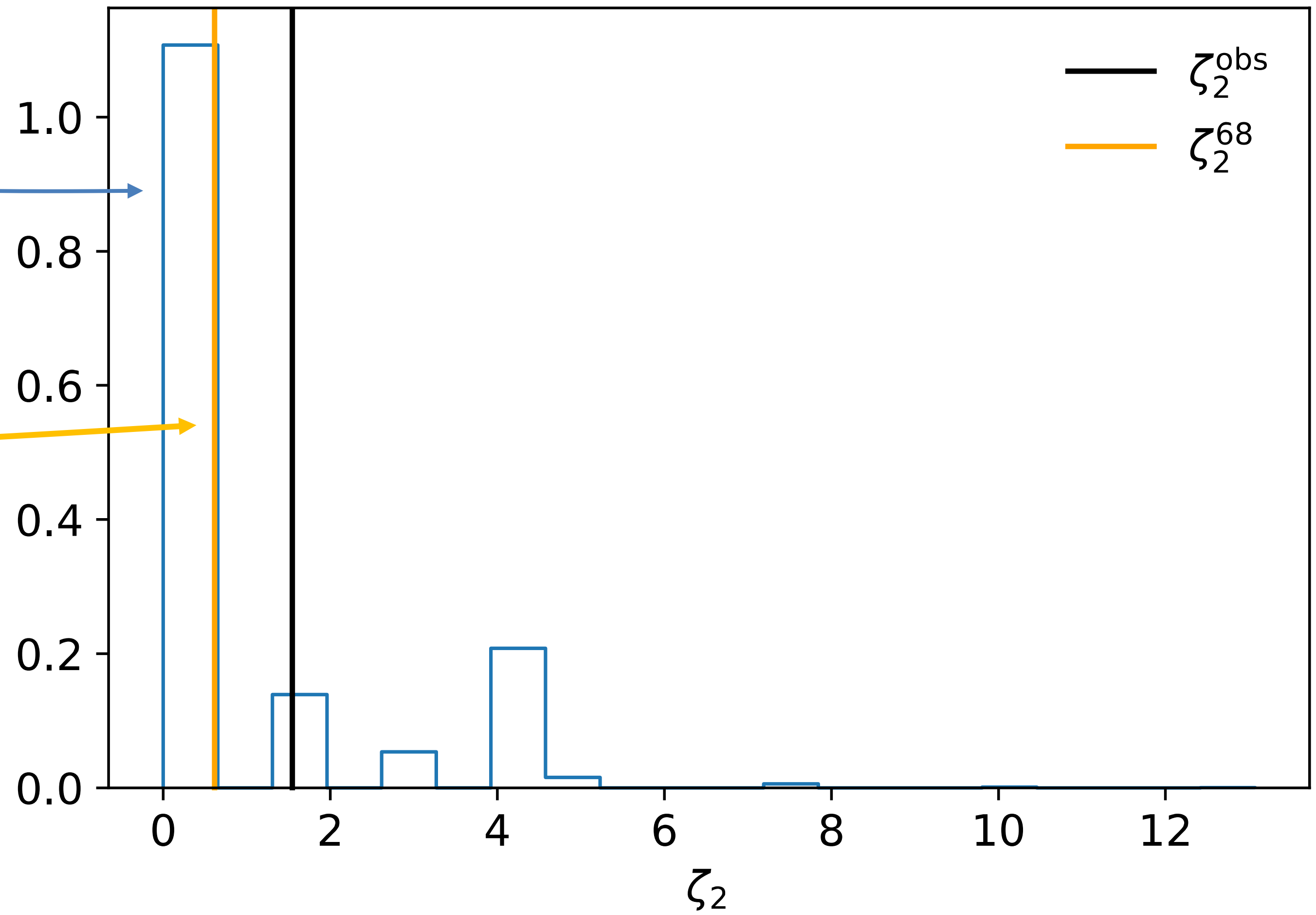
$$|\hat{\lambda} - \lambda| >> 0$$

Remember the MLE for λ is given by $\frac{dq}{d\lambda}|_{\lambda=\hat{\lambda}} = 0 \implies \hat{\lambda} = n$

# Confidence Intervals

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

Look at the distribution of $\zeta_\lambda$ when $\lambda$=2



From this distribution, we can calculate the value of $\zeta_2^{68}$

Defined as

$$\int_{\zeta_2^{68}}^{+\infty} f(\zeta_\lambda; \lambda = 2)d\zeta_\lambda = 1 - 0.68$$

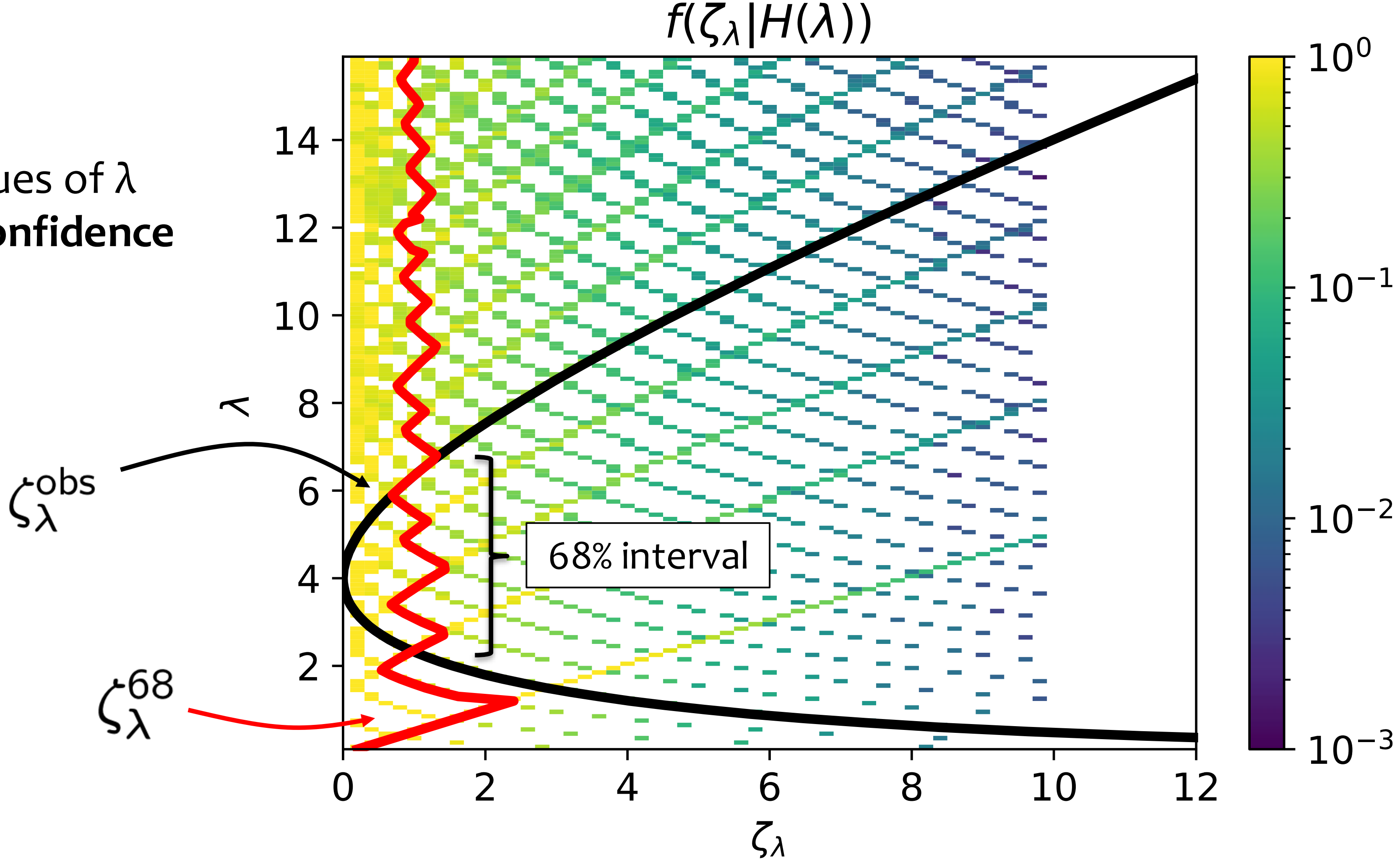
Finally, we include this value in our interval if $\zeta_2^{\rm obs} < \zeta_2^{68}$

# Confidence Intervals

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

We include any value of λ for which

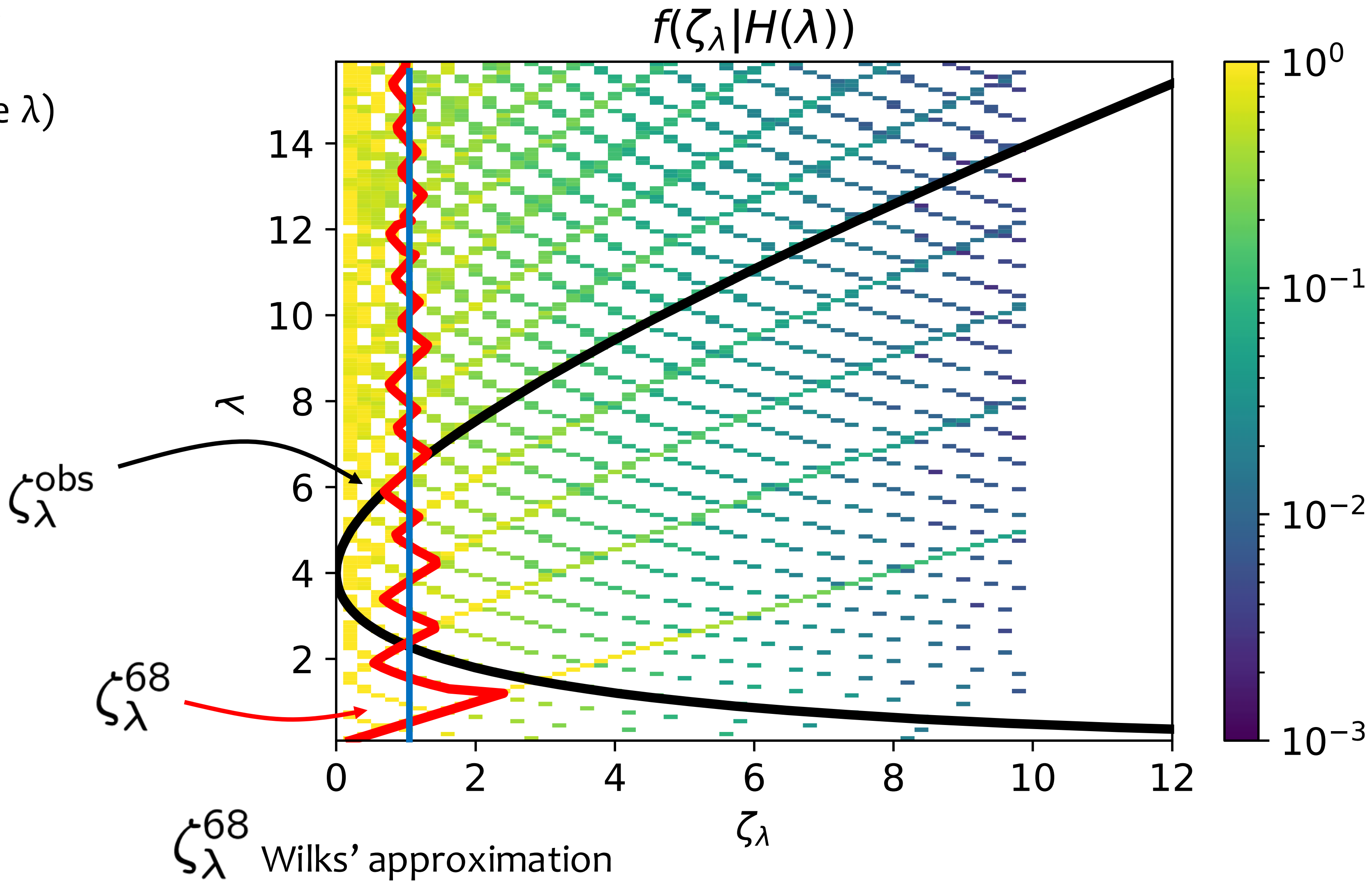
$$\zeta_\lambda^{\mathrm{obs}} \leq \zeta_\lambda^{68}$$



# Wilks' theorem

Wilks' theorem tells us that the distribution of $\zeta_\lambda$ is a $\chi^2(1)$ for large n (in this case for large λ)

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



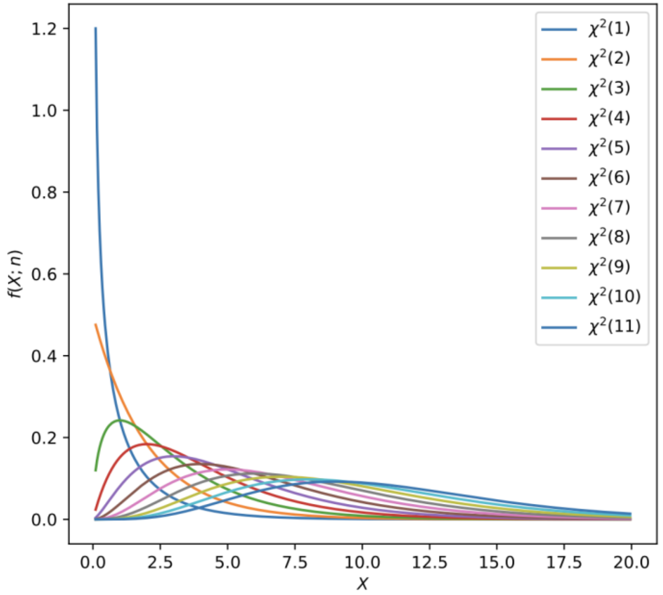
*the proof of this uses the central limit theorem

# Wilks' theorem

In general, Wilks' theorem gives us the result for any number of degrees of freedom.

The result is that for a log-likelihood difference with n parameters $\theta_1, \theta_2, \ldots, \theta_n$, the test statistic $\zeta_{\theta_1,\ldots,\theta_n}$ will be distributed under $H(\theta_1, \ldots, \theta_n)$ (the null hypothesis) as,

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



Read off values for where to apply threshold for different confidence intervals, for different number of parameters

| n | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 68.3% | 1.00 | 2.30 | 3.53 | 4.72 | 5.89 |
| 95.5% | 4.02 | 6.20 | 8.05 | 9.74 | 11.34 |
| 99.74% | 9.07 | 11.90 | 14.24 | 16.34 | 18.29 |

# Coverage

Frequentists care about the **coverage of confidence intervals** → check the fraction of intervals that contain the true value of λ ($\lambda_0$) for different values of $\lambda_0$

We can check for our Poisson case using

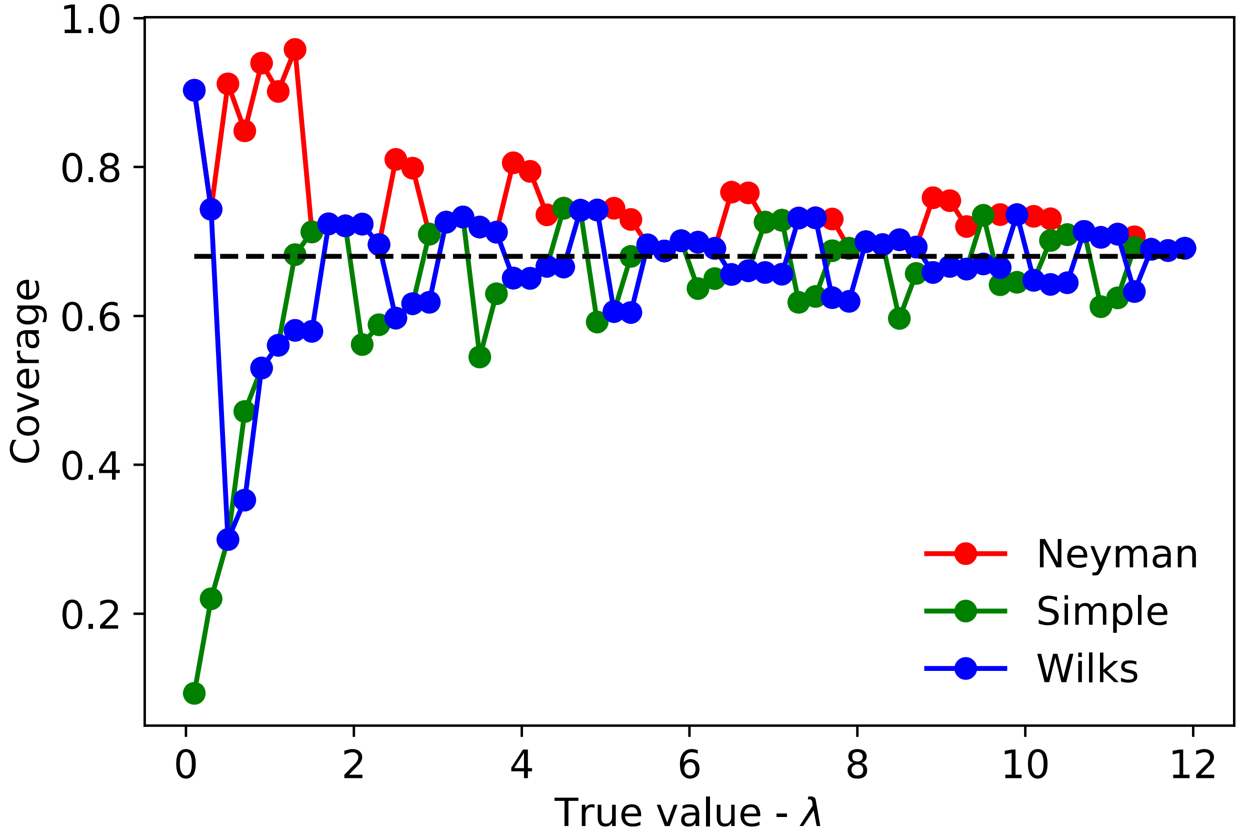
Full frequentist construction slow but only over covers

$$n_{\text{obs}} - \sqrt{n_{\text{obs}}} < \lambda < n_{\text{obs}} + \sqrt{n_{\text{obs}}}$$

Variance on n

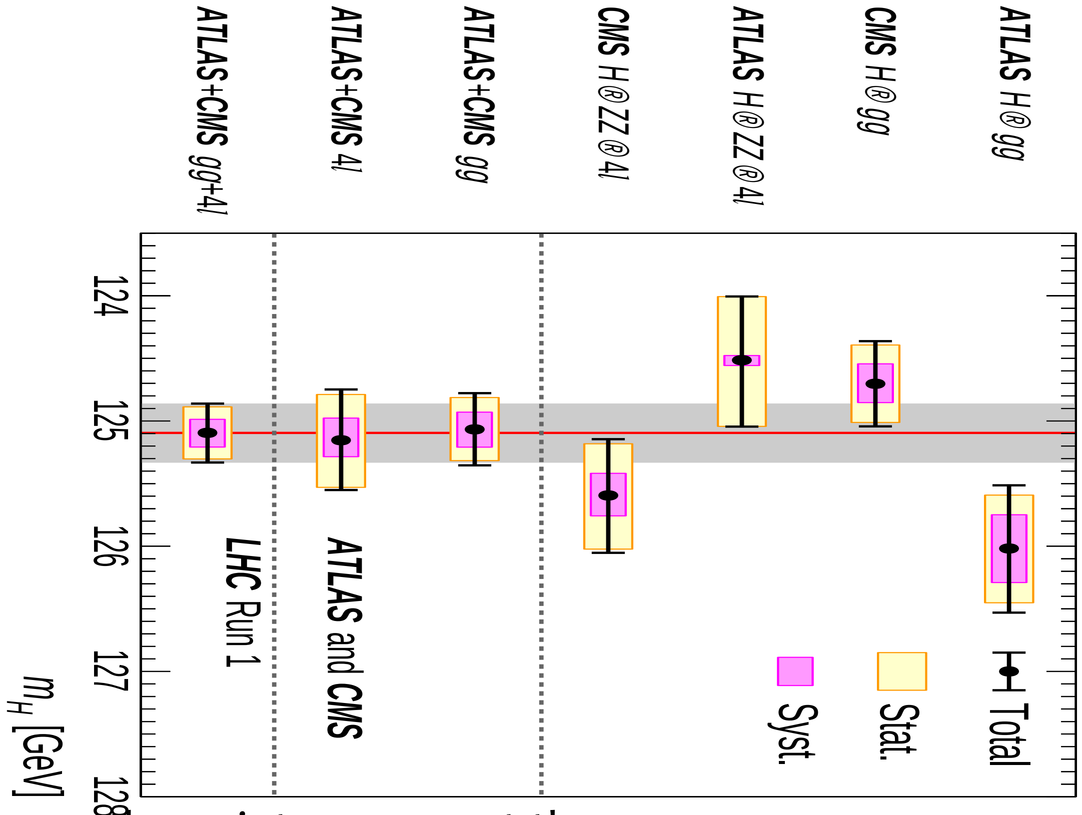
Very fast but can seriously undercover

Use Wilks' theorem - somewhere in between the others



# Uncertainties @ the LHC

When we see a plot like this…

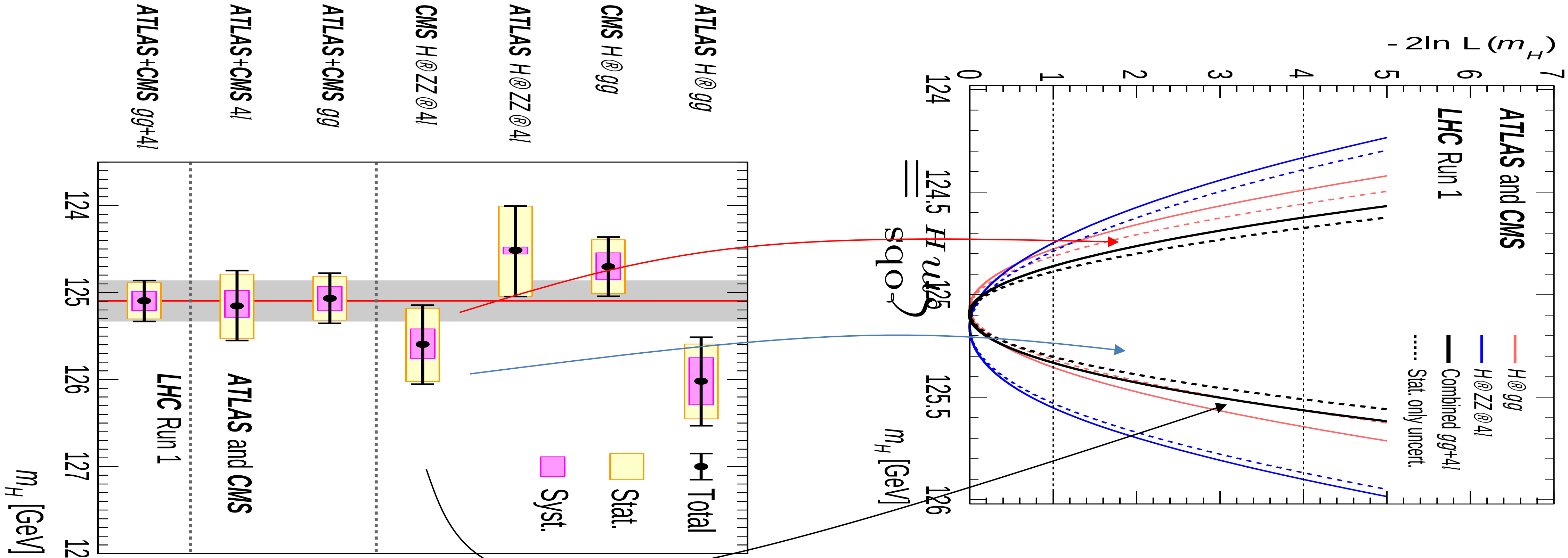


We know the points represent the Maximum Likelihood estimates

# Uncertainties @ the LHC

When we see a plot like this…

And we know that the error bars are derived from the region for which $\zeta^{\text{obs}}_{m_H} < 1$



We know the points represent the Maximum Likelihood estimates

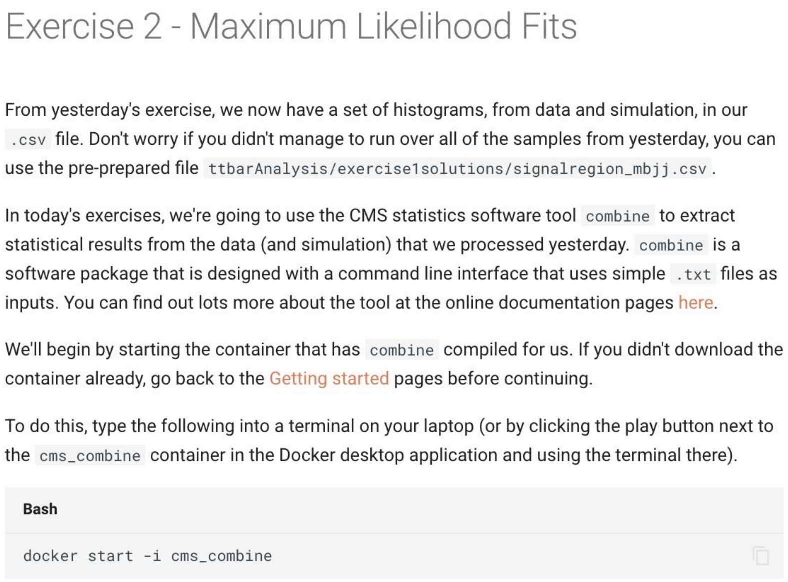
Tomorrow, we'll see what the dashed vs solid lines means

# Now it’s your turn!

In this afternoon’s exercises, you will use the histograms that you produced yesterday and calculate the MLE for the signal strength for the *tt* process – in HEP we call this “***fitting***” to the data

For this exercise, you will be using the software package `Combine`, that is used in almost every CMS statistical analysis

Make sure you have obtained the `cms_combine` container



Exercise 2 - Maximum Likelihood Fits

From yesterday's exercise, we now have a set of histograms, from data and simulation, in our `.csv` file. Don't worry if you didn't manage to run over all of the samples from yesterday, you can use the pre-prepared file `ttbarAnalysis/exercise1solutions/signalregion_mbjj.csv`.

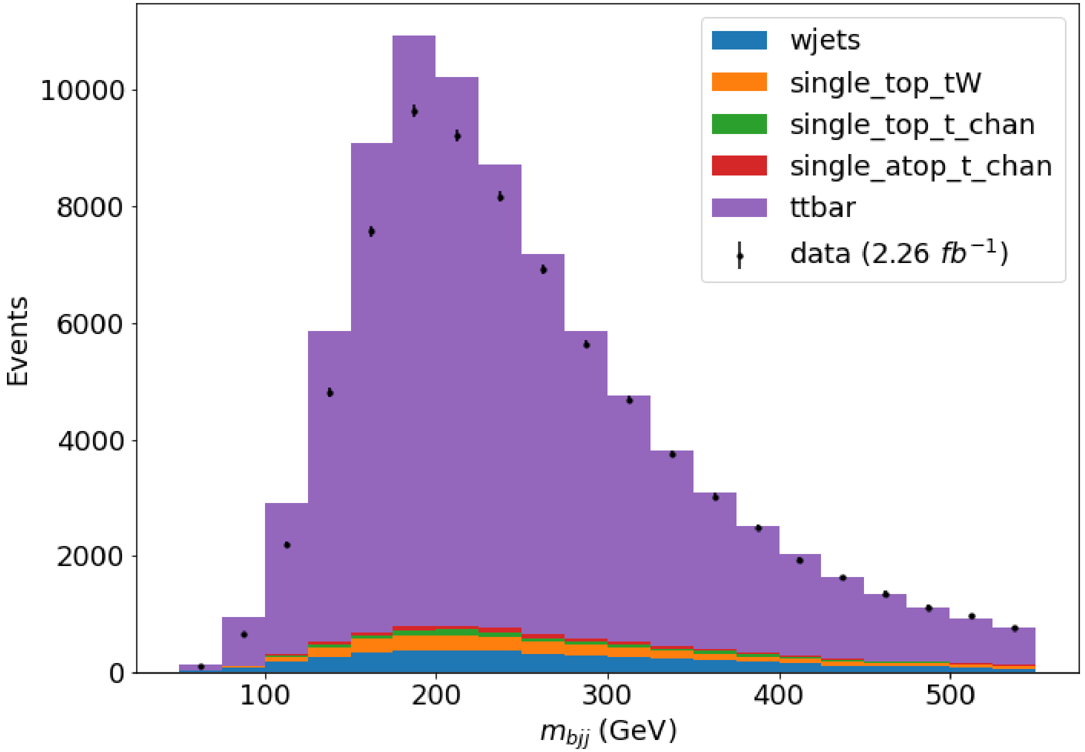
In today's exercises, we're going to use the CMS statistics software tool `combine` to extract statistical results from the data (and simulation) that we processed yesterday. `combine` is a software package that is designed with a command line interface that uses simple `.txt` files as inputs. You can find out lots more about the tool at the online documentation pages here.

We'll begin by starting the container that has `combine` compiled for us. If you didn't download the container already, go back to the Getting started pages before continuing.

To do this, type the following into a terminal on your laptop (or by clicking the play button next to the `cms_combine` container in the Docker desktop application and using the terminal there).

Bash

```
docker start -i cms_combine
```



Don’t worry if you didn’t complete the previous exercise, all of the solutions can be found in `ttbarAnalysis/exercise1solutions`