Introductory Statistics for HEP

A. Sznajder

UERJ
Instituto de Fisica

January - 2019

Outline

1. Introduction
2. Probability
3. Monte Carlo Method
4. Point and Interval Estimation
5. Hypothesis Test
6. Goodness of the Fit

Introduction

HEP advances through the interplay of top-down (theory guided) and bottom-up (data driven) processes. Some of the important tools used by HEP are:

HEP Tools

- **Statistics:** measure physics parameters (point estimation) and estimate uncertainty(interval estimation) , decide between two theories (hypotehsis test) and decide on compatibility between data and a theory (GOF)
- **MC Event Generators:** extract theory predictions , study detector effects and understand data
- **Machine Learning:** modern tools for data analysis like Neural Networks for classification , regression and data generation

Course Contents

1) Introduction to Statistics for HEP Data Analysis

- Basic concepts of statistics
- Parameter and interval estimation
- Hypothesis testing and goodness of the fit

2) The Physics of Event Generators

- Physics processes , Feynman diagrams and cross sections
- Monte Carlo event generator
- Madgraph

3) Machine Learning

- Introduction to machine learning: classification and regression
- The multilayer perceptron (MLP)
- Universal approximation , vanishing gradient and deep learning
- Convolutional network, autoencoder and adversarial network

Bibliography

Bibliography

- Glen Cowan , http://www.pp.rhul.ac.uk/~cowan/stat_course.html
- Kyle Cranmer , https://indico.cern.ch/event/208901/contributions/1501070
- Diego Toneli , https://indico.cern.ch/event/598530/contributions/2547055
- Fred James , https://indico.desy.de/indico/event/11244/overview
- Luca Lista , http://people.na.infn.it/~lista/Statistics
- Books:
  - G. Cowan, Statistical Data Analysis, Oxford, 1998.
  - F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006
  - S. Brandt, Statistical and Computational Methods in Data Analysis, Springer, 1998
  - L. Lista, Statistical Methods for Data Analysis in Particle Physics, Springer, 2017

Introduction to Probabiity and Statistics for HEP Data Analysis

Probability

Probability theory is a branch of pure mathematics. It studies mathematical models of probability distributions and the description of its properties through parameters like mean, variance, skewness, correlations and etc.

Statistics

Statistics is essentially an inductive and empirical. If you have a model dependent on a parameter θ, it tries to answer what constrains the data impose on θ (Point and Interval Estimation). It also studies the decision of which probabilistic model better describes a population (Hypothesis Test).

Probabilidade and statistics are intimately related. In probability theory the object of study (population) is given a priori and we aim to describe its properties, while in statistics we infer the population properties from experimental data samples. In this way we can say that statistics studies the inverse problem of probability !

It is often said that mathematics is the language of science. It could well be said that statistics is the language of experimental science. It is through statistical concepts that we quantify the correspondence between theoretical predictions and experimental observations.

Axiomatic Probability Definition

Probability Axioms (Kolmogorov,1933)

Consider a set S with subsets A , B , ...

- $P(A) \geq 0$ for all $A \subset S$
- $P(S) = 1$
- If $A \cap B = \emptyset \ \rightarrow \ P(A \cup B) = P(A) + P(B)$

Theorems:

- $P(\overline{A}) = 1 - P(A)$
- $P(\overline{A} \cup A) = 1$
- $P(\emptyset) = 0$
- If $A \subset B$ then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Interpretation of Probability

Frequentist

Probability is defined *operationally* as a limit of relative frequency in a large number of trials. Considering a repeatable experiment with outcomes A,B,... , we have

$$P(A) = \lim_{N \to \infty} \frac{\#\text{ times outcome is } A}{N}$$

Bayesian

Probability is interpreted as an expectation, representing a state of knowledge or as quantification of a personal belief. A, B, ... are hypotheses or statements that have a value of true or false

$$P(A) = \text{ degree of belief that A is true}$$

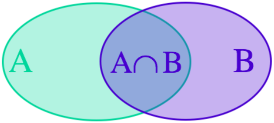
Both interpretations are consistent with the Probability Axioms. The Frequentist is usually more usefull in particle physics , but Bayesian provides an easier treatment of non-repeatable fenomena (syst. uncertainties , probability for Higgs existence ...)

Conditional Probability

The conditional probability of A given B (with $P(B) \neq 0$) is given by

Conditional Probability

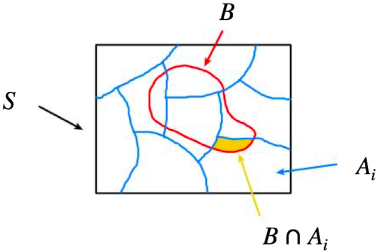
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



- A and B are statistically independent if $P(A \cap B) = P(A)P(B) \;\Rightarrow\; P(A|B) = P(A)$
- It's important to realize that in general $P(A|B) \neq P(B|A)$!

Law of Total Probability

Let's consider a subset B of the sample space S. If we partition S in disjoint subsets A_i such that $\cup_i A_i = S$ we have that



Law of Total Probability

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

Proof:

$$B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$$

$$\Rightarrow \; P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

$$\Rightarrow \; P(B) = \sum_i P(B|A_i)P(A_i)$$

Bayes Theorem (Bayes, 1763)

Bayes Theorem

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

, where $P(A)$ is a prior probability because it's independent of any information on B and vice-versa for $P(B)$

Proof:

$$\begin{cases} P(A|B) = \dfrac{P(A \cap B)}{P(B)} \\ P(B|A) = \dfrac{P(B \cap A)}{P(A)} \end{cases} \Rightarrow \quad P(A|B)P(B) = P(B|A)P(A)$$

From the law of total probability one can rewrite Bayes Theorem as

Bayes Theorem (Total Probability Form)

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

Probability Density Function

A random variable can be discrete or continuous. Considering the outcome of an experiment on a continuous variable x , we define the PDF $f(x)$ as the probability to observe x in the interval $[x, x + dx]$

Probability Density Function (PDF)

$$P(x \in [x, x + dx]) = f(x)dx$$

In Frequentist interpretation $f(x)dx$ is the fraction of times x is observed in $[x, x + dx]$!

As x must be somewhere we have the PDF normalization condition

PDF Normalization

$$\int_{-\infty}^{\infty} f(x)dx = P(x \in [-\infty, +\infty]) = 1$$

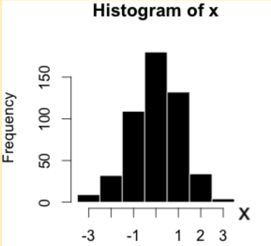
Histograms

A histogram is a function m_i that counts the number of observations (frequency) that fall into each of the disjoint categories (bins). Considering the outcome of an experiment a set of n observations $x_1, x_2, x_3, \ldots, x_n$, we can display in a histogram, giving a visualization of the shape, localization and dispersion of the data

Data

Bin	Content
-3.5 to -2.51	9
-2.5 to -1.51	32
-1.5 to -0.51	109
-0.5 to 0.49	180
0.5 to 1.49	132
1.5 to 2.49	34
2.5 to 3.49	4

Histogram



PDF is a limit case of a histogram with infinite data sample, zero bin width and normalized to a unit area

Joint , Marginal and Conditional PDF

A distribution can be characterized by more than one random variable. We define the joint PDF as the probability of observing the point $\vec{x} = (x_1, x_2, \ldots, x_n)$ to be in the volume element $d\vec{x}^n$

Joint PDF

$$P(x_i \in [x_i, x_i + dx_i]) = f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$

The marginal PDF is defined as the distribution projected on one axis

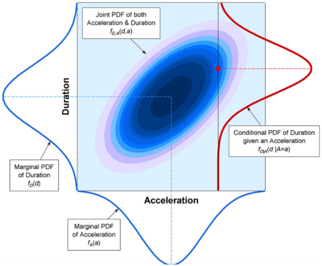
Marginal PDF

$$f_2(x_2) = \int f(x_1, x_2) dx_1$$

The conditional PDF is defined as a distribution for a slice of one variable

Conditional PDF

$$g(x_1 | x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$$



Two variables are independent if the PDF factorizes as $f(x_1, x_2) = f(x_1)f(x_2)$

Expectation Values , Mean and Variance

The expectation value $E[x] = \mu$ (mean) of a random variable x and $E[a(x)] = \mu_a$ of function $a(x)$ are defined as

Expectation Value x or Mean

$$\mu = E[x] = \int_{-\infty}^{+\infty} x f(x) dx$$

Expectation Value of $a(x)$

$$\mu_a = E[a] = \int_{-\infty}^{+\infty} a(x) f(x) dx$$

Of special interest is the expectation value $E[(x-\mu)^2] = \sigma^2$ (variance), which measures the spread around the mean μ (width of the PDF)

Expectation Value of $(x-\mu)^2$ or Variance

$$V = E[(x-\mu)^2] = \int_{-\infty}^{+\infty} a(x) f(x) dx \quad \Rightarrow \quad V = E[(x-\mu)^2] = E[x^2] - \mu^2$$

The standard deviation of x is the square root of the variance: $\sigma_x = \sqrt{E[(x-\mu)^2]}$

Covariance and Correlation

The covariance V_{xy} between two random variables x and y is defined as

Covariance V_{xy}

$$V_{xy} = E[(x-\mu_x)(y-\mu_y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_x)(y-\mu_y)f(x,y)dxdy$$

$$\Rightarrow \quad V_{xy} = E[xy] - \mu_x\mu_y$$

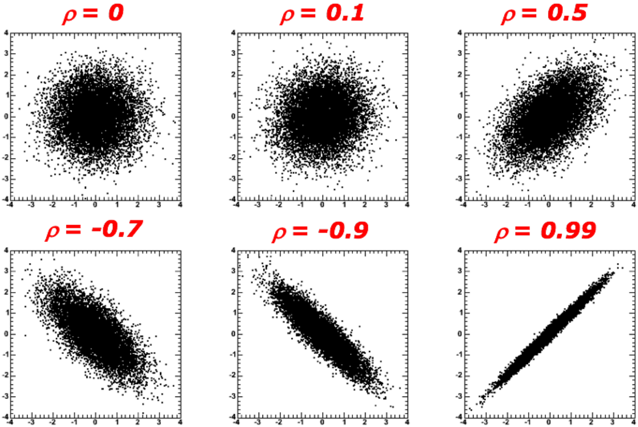
We can also define a correlation coefficient between the random variables x and y as

Correlation Coeficient

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x\sigma_y}$$

For independent variables x and y we have $E[xy] = E[x]E[y] = \mu_x\mu_y$ and $V_{xy} = 0$ (uncorrelated)

Correlation Coeficient

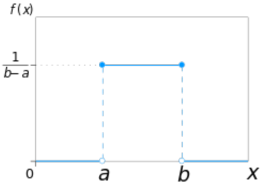


Probability Distributions

Uniform Probability Distribution

Uniform Probability Distribution

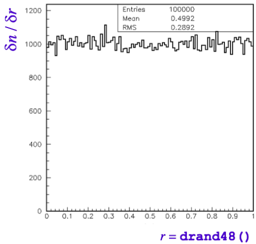
$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{, for } a \leq x \leq b \\ 0 & \text{, for } x < a \text{ or } x > b \end{cases}$$



Uniform random number generators have a central importance for random number generation (ex: Monte Carlo and simulation)

Pseudo-Random Number Generation

- pseudo-random number are numbers that look close to random, but were generated using a deterministic process (computer)
- programming languages come with implementations to generate uniform pseudo-random numbers (ex: drand48() in C , random() in Python , TRandom() in ROOT)
- generator uniformity and period depends on algorithm



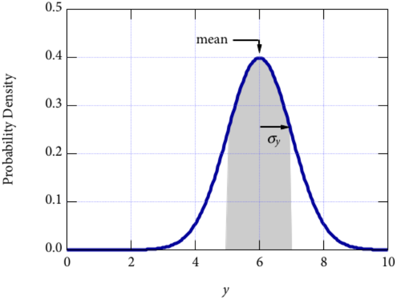
Gaussian Probability Distributions

Gaussian (Normal) distribution plays an important role due to central limit theorem

Gaussian Probability Distribution

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

- μ is the mean
- σ is the variance(width)



Measurement errors are often the sum of many contributions, so frequently measured values can be treated as Gaussian distributed

$$y = \sum_{i=1}^{N} x_i$$

In the limit $N \to \infty$ the random variable y is gaussian with $E[y] = \sum \mu_i$ and $V[y] = \sum \sigma_i$

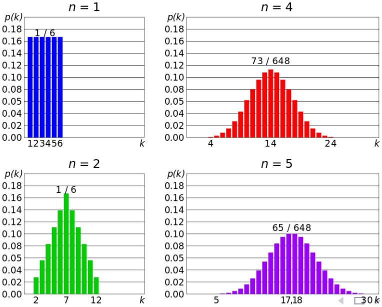
Central Limit Theorem

Central Limit Theorem

The Central Limit Theorem(CLT) states that the sum of N independent and identically distributed random variables, with means μ_i and variances σ_i^2, will tend to a Gaussian(Normal) as $N \to \infty$. It also states that:

$$\mu = \sum_i \mu_i \quad , \quad V = \sum_i \sigma_i^2 \quad \xrightarrow{\text{for identical } \sigma_i} \quad \sigma_\mu = \frac{\sigma}{\sqrt{N}}$$

The average of N random variables x (ex: fair dice) converges to a Gaussian, independently of the original distributions



Exponential Probability Distribution

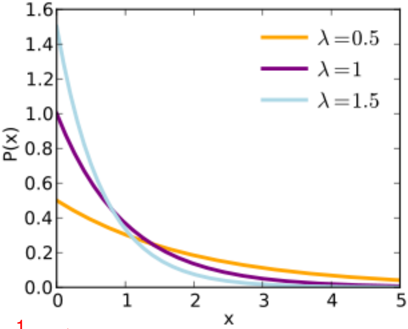
The exponential distribution describes describes random processes in which events occur independently and at a constant rate (memoryless) , like particle decays. Consider a decay process with a constant decay rate λ containing $N(t)$ particles

$$\frac{dN}{dt} = -N\lambda \quad \Rightarrow \quad N(t) = N_0 e^{-t\lambda}$$

Exponential Probability Distribution

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{, for } x \geq 0 \\ 0 & \text{, for } x < 0 \end{cases}$$

- $\frac{1}{\lambda}$ is the mean
- $\frac{1}{\lambda^2}$ is the variance(width)



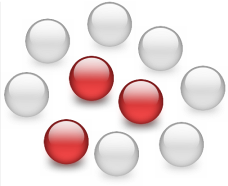
Ex: proper decay time t of an unstable particle $p(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$, where τ is the mean lifetime

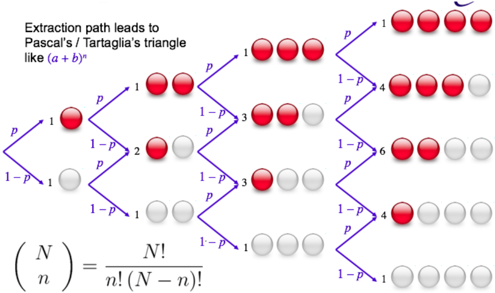
Bernoulli Trials

Bernoulli Trials

A Bernoulli trial is a random experiment with exactly two possible (binary) outcomes, "success"and "failure"[1] , in which the probability of success is the same every time the experiment is conducted.

- **RED:** success
- **WHITE:** failure




Extraction path leads to Pascal's / Tartaglia's triangle like $(a+b)^n$

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

Success could also be a track reconstructed by a detector, or event selected by a set of cuts !

[1] Success and failure are merely labels for the two outcomes, and should not be taken literally

Binomial Probability Distribution

Consider N independent experiments (Bernoulli trials). The outcome of each trial is a binary result: *Success* or *Failure* , where the success has a probability p and failure $(1-p)$. For a number n of successes $(0 \leq n \leq N)$ we have

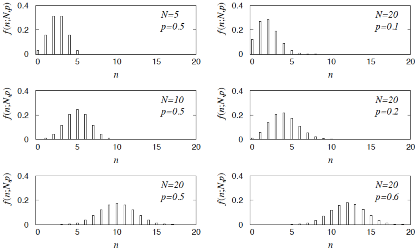
$$\underbrace{p\,p\,p\ldots p}_{n\ times}\ \underbrace{(1-p)\,(1-p)\,(1-p)\ldots(1-p)}_{(N-n)\ times} = p^n(1-p)^{(N-n)}$$

As the order of the outcomes (permutations) is not relevant, we have the following distribution

Binomial Distribution

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

- Mean is given by
  $\mu = E[n] = \sum_{n=0}^{N} nf(n|N,p) = Np$
- Variance is given by
  $V = E[n^2] - (E[n])^2 = Np(1-p)$



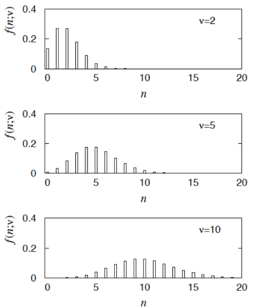
Ex: if we observe N decays of $W^{\pm}$, the number n of muonic decays is a binomial with probability $p = BR(W \rightarrow \mu\nu)$

Poisson Probability Distributions

Consider binomial distribution of n in the limit where $N \to \infty$ and $p \to 0$. Considering the mean as $E[n] = Np \to \nu$ we have

$$f(n|N, p = \frac{\nu}{N}) = \frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n (1 - \frac{\nu}{N})^{(N-n)}$$

$$= \frac{\nu^n}{n!} \underbrace{\frac{N(N-1)\ldots(N-n+1)}{N^n}}_{1} \underbrace{\left(1 - \frac{\nu}{N}\right)^N}_{e^{-\nu}} \underbrace{\left(1 - \frac{\nu}{N}\right)^{-n}}_{1}$$



Poisson Probability Distribution

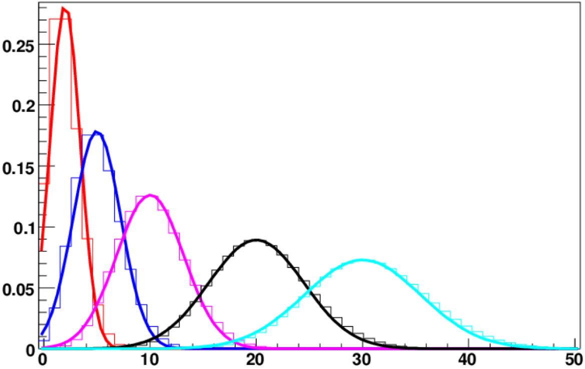
$$f(n|\nu) = \frac{\nu^n}{n!} e^{-\nu}$$

- Mean is given by $\mu = E[n] = \nu$
- Variance is given by $V = E[n^2] - (E[n])^2 = \nu$

Ex: number n of scattering events with cross section σ observed for a given luminosity, with $\nu = \sigma L$

From Poisson to Gaussian Distributions

In the limit of large ν the Poisson distribution is well aproximated by the Gaussian distribution



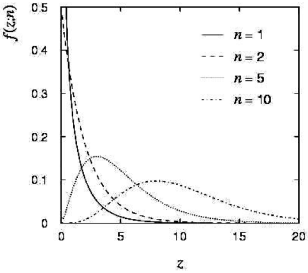
Chisquare Probability Distribution

The chisquare probability distribution for the continuous random variable z for n degrees of freedom is defined by

Chisquare (χ^2) Distribution

$$f(z|n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

- Mean is given by $E[z] = n$
- Variance is given by $V[z] = E[z^2] - (E[z])^2 = 2n$



Consider N independent Gaussian random variables x_i , with means μ_i and variances σ_i^2 . The chisquare estimator defined bellow follows the chisquare distribution

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

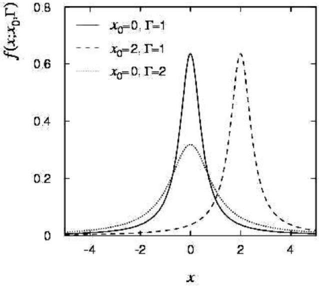
Breit-Wigner Probability Distribution

The Breit-Wigner probability distribution for the continuous random variable x , with a full width at half maximum Γ and mode x_0 is defined by

Breit-Wigner Distribution

$$f(x|\Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

- Mean is not defined
- Variance is ∞



Change of variable of a PDF

Consider a a 1-1 transformation of variable $y = t(x)$. This maps the interval $x_a \leq x \leq x_b$ into $y_a \leq y \leq y_b$, so the probabilities bellow are equal

$$Prob(x_a < x < x_b) = Prob(y_a < y < y_b)$$
$$\Rightarrow \int_{x_a}^{x_b} f(x)dx = \int_{y_a}^{y_b} g(y)dy$$
$$\Rightarrow \int_{x_a}^{x_b} f(x)dx = \int_{y_a \to x_a}^{y_b \to x_b} \underbrace{g\left(t(x)\right) \frac{dy}{dx}}_{\text{function of x}} dx$$

A PDF is positive definite, so it's transformation law must include the Jacobian absolute value

PDF Transformation Law

$$f(x) \;\rightarrow\; g(y) = f\left(t^{-1}(y)\right) \left|\frac{dx}{dy}\right|$$

The PDF properties constraints the transformation function $x = t(y)$, so it must be single valued and monotonic , otherwise it would have no inverse !

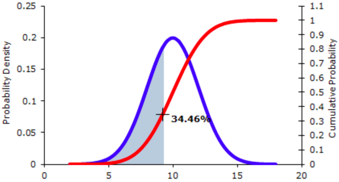
Cumulative Distribution Function

Given a random variable x and it's PDF $f(x)$, the cumulative distribution function CDF is defined a the probability for the random variable to be smaller than x

Cumulative Distribution Function(CDF)

$$F(x) = \int_{-\infty}^{x} f(x')dx' \;\Rightarrow\; f(x) = \frac{F(x)}{dx}$$

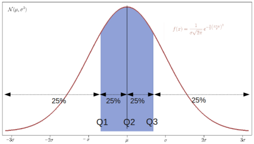
For well behaved functions we can specify the PDF by its CDF !



We define the quantile as the value of x which gives $F(x_\alpha) = \alpha$. Quantiles are cut points dividing the range of a PDF into equal probabilities intervals and it's given by the inverse CDF

Quantile of Order α

$$x_\alpha = F^{-1}(\alpha)$$



Uniform Distribution from CDF

Consider a change o variable $u = F(x)$ defined by the CDF of $f(x)$. The random variable u defined by this integral transform has the following properties

CDF Properties

- $u = F(x)$ is a random variable in the interval $[0, 1]$
- u obeys a uniform distribution , so it's PDF is $g(u) = 1$

So , there always exists a metric in which the PDF is uniform !

Proof:

$$u = F(x) = \int_{-\infty}^{x} f(x')dx' = \int_{-\infty}^{u} \underbrace{f\left(F^{-1}(u')\right)\left|\frac{dx'}{du'}\right|}_{\text{function of u'}} du'$$

$$= \int_{0}^{u} g(u')du' \quad \Rightarrow \quad g(u) = 1 \qquad (1)$$

Monte Carlo Method

Monte Carlo - Inverse Transform Method

The random variable $u = F(x)$ defined by the CDF of $f(x)$ has a uniform distribution. By inverting the CDF one can generate a sample that follows the $f(x)$ distribution using a uniform random number generator.

Inverse Transform Method

$$u = F(x) = \int_{-\infty}^{x} f(x')dx' \;\Rightarrow\; x = F^{-1}(u)$$

As an example , consider the exponential PDF given by $f(x) = \frac{1}{\xi} e^{-x/\xi}$

Example:

$$u = \int_{-\infty}^{x} \frac{1}{\xi} e^{-x'/\xi} dx' \;\Rightarrow\; x = -\xi \, log(1 - u)$$

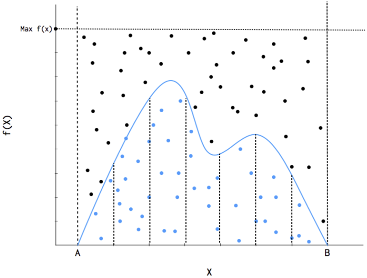
The inversion method only works for functions that $F^{-1}(u)$ exists !

Monte Carlo - Acceptance and Rejection Method

Procedure to generate a sample of random numbers following a distribution $f(X)$ using the Acceptance-Rejection method

Acceptance-Rejection Method

1. Find the maximum f_{max} of $f(X)$ in $[A, B]$
2. Generate a uniform random number X in $[A, B]$
3. For each X generate a uniform random number Y in $[0, f_{max}]$
4. If $Y < f(X)$ accept the point (X, Y) , otherwise reject
5. Return to 2



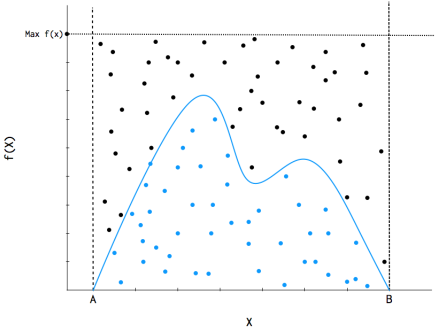
The efficiency in Acceptance-Rejection is given by the fraction of accepted points and for peaked functions it can be very low !

Monte Carlo Integration - Acceptance-Rejection

Acceptance-Rejection

Consider a function $f(x)$ limited by f_{max} in the interval $[A, B]$. To integrate it using acceptance-rejection method one samples points uniformly distributed within the box

$$\int_A^B f(x)dx = A_{box}\left(\frac{\#\text{ points under curve}}{\#\text{ points generated}}\right) = A_{box}\ Eff$$

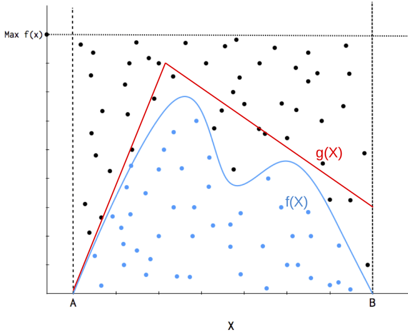


Monte Carlo - Importance Sampling

To improve efficiency one combines the Acceptance-Rejection and Invertion methods in a hybrid one. Instead of sampling X uniformly, it's sampled from an envelope function $g(X)$ that aproximates $f(X)$. We use the Invertion method to sample $g(X)$ and then Accept-Reject to sample $f(X)$.

Importance Sampling

1. Find a function $g(X)$ that is invertible and $g(X) \geq f(X)$ for all X in $[A, B]$
2. Generate a random number X from $g(X)$ using inversion method
3. Generate a random number Y in $[0, 1]$
4. If $g(X).Y < f(X)$ keep the point (X, Y), otherwise reject
5. Return to step 2



MC Integration as Averages

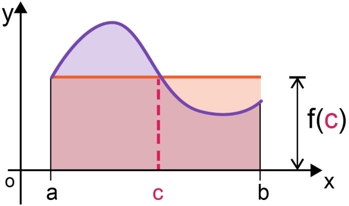
Another way of understanding the MC integration as an average is through the mean value theorem

Mean Value Theorem for Integrals

Let $f : [a, b] \to \mathbb{R}$ be a continuous function. Then there exists c in (a, b) such that

$$\int_a^b f(x)dx = (b-a)f(c)$$

The value of $f(c)$ is the mean value of f in $[a, b]$



We can estimate the mean value of $f(x)$ in $[a, b]$ directly by sampling the function uniformly in the interval and taking the average $f(c) \simeq \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} f(x_i)$

Integrals as an Average

The integral of $f(x)$ and its variance can be estimated by sampling N points $\{x_i\}$ in the interval $[a, b]$

$$\begin{cases} I_N = \int_a^b f(x)dx \simeq (b-a)\left[\frac{1}{N}\sum_{i=1}^N f(x_i)\right] \\ V_N \simeq \frac{(b-a)^2}{N}\sum_{i=1}^N [f(x_i)]^2 - I^2 \end{cases} \quad \Rightarrow \quad I = I_N \pm \sqrt{\frac{V_N}{N}} \quad (\text{ CLT Theorem })$$

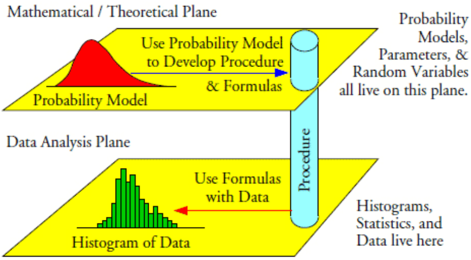
Point and Interval Estimation

Statistical Description of Data

Parent Distribution and Sample

Whenever we are measuring a quantify we are sampling a parent distribution PDF

- **Point Estimation:** parameter estimation from data sample
- **Interval Estimation:** uncertainty estimation from data sample



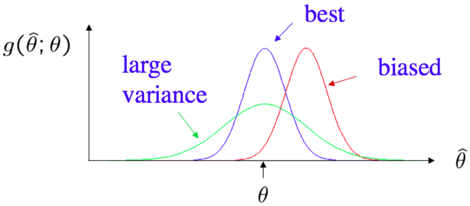
Point and Interval Estimation

Point estimation involves the use of a data sample to estimate a single value for an unknown population parameter (ex: mean, variance ...)

Point Estimator Properties

Estimators are functions of data [2], hence estimators are random variables with their own probability distributions. If we repeat the measurement many times, the estimates would follow a PDF $g(\hat{\theta}, \theta)$

- We want small bias (systematic error): $b = E[\hat{\theta}] - \theta$
- We want a small variance (statistical error): $V = E[(\theta - \hat{\theta})^2]$



[2] An estimator is a test statistic, i.e. a data function, that depends on a model parameter

Point and Interval Estimation

Given a data sample we can describe its properties directly using the point estimators for the mean and variance

The Estimator for the Mean

The mean μ is the expectation value $E[x]$ of the parent distribution. For a data sample $\{x_1, x_2, \ldots, x_n\}$ we define the estimator $\hat{\mu}$ of the mean as the sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \overline{x}$$

If we were to use other data samples, the estimates from each sample would follow a distribution. The variance of the mean estimator $\hat{\mu}$ over these samples is given by

Variance of $\hat{\mu}$

$$V_{\hat{\mu}} = \frac{V}{n} \quad \Rightarrow \quad \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

Point and Interval Estimation

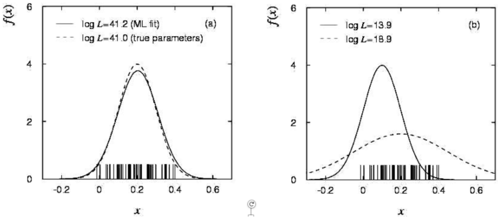
Suppose we perform an experiment modeled by a PDF $f(x,\theta)$, whose outcome is $\{x_1, x_2, \dots x_n\}$

Maximum Likelihood Estimator(MLE)

The Maximum Likelihood (ML) estimator of the model parameter θ is defined as the value $\hat{\theta}$ for which the likelihood $L(\theta)$ evaluated on data sample $\{x_1, x_2, \dots x_n\}$ is a maximum

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

If the hypothesized θ is close to the true value, there's a high probability to get data like we actually observe !



The value of $\hat{\theta}$ that maximizes the likelihood is not the most likely value of θ. It's the value of θ that makes your data most likely

Point and Interval Estimation

Maximum Likelihood (ML) Estimator

- Instead of maximizing L we minimize $-ln(L)$, which is equivalent and easier: $\dfrac{\partial ln\,L}{\partial\theta} = 0$
- This minimization is usually solved numerically
- The likelihood and its maximum are invariant against re-parametrization: $L(\theta) = L(f(\theta))$

Asymptotically, for large data samples, the ML estimator has optimal properties

ML Asymptotic Properties

- ML estimator is asymptotically unbiased
- The estimates of $\hat{\theta}$ follows a normal distribution
- The variance of the ML estimator may be infered from: $\hat{V}(\hat{\theta}) \simeq -\left(\dfrac{\partial^2 ln\,L}{\partial\theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$

These asymptotic properties are not met with finite size samples, although often well approximated for large samples !

Point and Interval Estimation

From Maximum Likelihood to Least Squares Method

Suppose the outcome of n measurement of a quantity $\lambda(x_i,\theta)$ are $\{y_1, y_2, \ldots, y_n\}$ and they are known to be gaussian distributed with variances σ_i . The quantity $\lambda(x_i,\theta)$ depends on the control variable x_i and the parameter θ

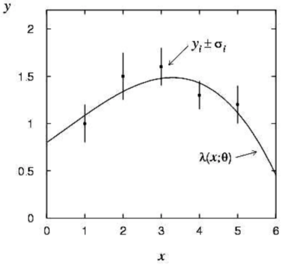
$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[-\frac{(y_i-\lambda(x_i,\theta))^2}{\sigma_i^2}\right] \Rightarrow -ln\,L(\theta) = \sum_{i=1}^{n} \frac{[y_i-\lambda(x_i,\theta)]^2}{\sigma_i^2} + \theta \text{ indep. term}$$

Hence, maximizing the likelihood is equivalent to minimizing the least-square estimator

Least Squares Estimator (Chisquare)

The Least Squares (LS) estimator is obtained from the minimization condition of the ML formula

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{[y_i-\lambda(x_i,\theta)]^2}{\sigma_i^2}$$



Point and Interval Estimation

Having estimated our parameter we now need to report its statistical error, i.e., how widely distributed estimates would be if we repeat the entire measurement many times

Variance of Estimators

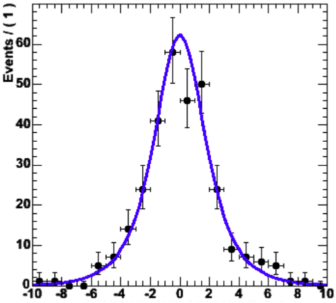
- **Simulation:** Monte Carlo the method can be used to simulate the experiment many times and one can use the ML estimator for each simulation to obtain a distribution of estimates.
- **Analytical:** if we know the PDF of the data and its integral, we can determine the variance directly from $V = E[\theta^2] - (E[\theta])^2$
- **Graphical:** approximation to the minimum variance bound

Point and Interval Estimation

We can use simplified simulated experiments (“toy Monte Carlo”) to understand the distribution of the ML estimators.

Variance of Estimators - MC Simulation

1. Choose a plausible true value of the parameter θ
2. Generate several sets of simulated data $\{x_i\}$ (experiments) by random sampling the model PDF $f(x, \theta)$
3. Maximize the likelihood in each set to estimate the parameter θ
4. Look at the distribution of the estimator $\hat{\theta}$
5. Repeat for all relevant choices of the parameter θ values



As is true for ML estimator in large sample limit, the distribution of estimates is roughly Gaussian !

Point and Interval Estimation

Expanding the likelihood in a Taylor series around its maximum $\hat{\theta}$ we have

$$log\ L(\theta) = log\ L(\hat{\theta}) + \left[\frac{\partial log\ L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 log\ L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

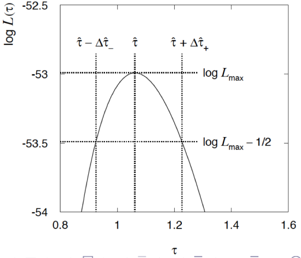
First term is L_{max} , the second is zero and for the third we use information inequality

$$log\ L(\theta) = log\ L_{max} - \frac{(\theta - \hat{\theta})^2}{2\sigma^2_{\hat{\theta}}}$$

Variance of Estimators - Graphical

So , displacing θ by one standard deviation , $\theta \rightarrow \hat{\theta} \pm \sigma_{\hat{\theta}}$ we have that *log L* decreases by $1/2$ from its maximum $log\ L_{max}$

$$log\ L(\hat{\theta} \pm \sigma_{\hat{\theta}}) = log\ L_{max} - \frac{1}{2}$$



Point and Interval Estimation

The standard deviation (variance) definition of measured uncertainty is usually applied for quantities with a gaussian PDF. For other distributions we use confidence intervals, which can lead to asymmetric error bars.

Confidence Interval

A interval on μ at $x\%$ confidence level is defined such that the true of value of μ is contained $x\%$ of the time in the interval.

- The output is not a probabilistic statement on the true value of μ.
- The true μ is fixed but unknown and each experiment will result in an estimated interval $[\mu_l, \mu_u]$. A fraction of $x\%$ of those intervals will contain the true value of μ
- Coverage is the guarantee that probabilistic statements is true (i.e. repeated future experiments do reproduce results in $x\%$ of cases)
- Definition of confidence intervals does not make assumptions on interval shape and we can choose two sided intervals (measurements) or one-sided intervals (limits)

Confidence intervals can also be applied to composite hypothesis to give an interval statement on a observation μ , instead of quoting just a *P-values* for a hypothesis with a fixed μ

Hypothesis Test

Hypothesis Testing

We use hypothesis testing to decide on the agreement between data and models. The null hypothesis H_0 is the one subjected to the test and expected to be true. Usually it's complemented with another hypothesis H_1 representing an alternative model.

Considering the null hypothesis H_0 as background (BKG) and the alternate hypothesis H_1 as signal plus background (SIG) we have the possible scenarios

Confusion Matrix

		Truth	
		BKG	SIG
Decision	BKG	*True Negative*	*False Negative (Type II Err.)*
	SIG	*False Positive (Type I Err.)*	*True Positive*

Test Statistic

To quantify the level of agreement between data and hypotheses we define a test statistic $t(\vec{x})$ as a scalar function of the data sample $\vec{x} = (x_1, x_2, \dots, x_n)$.

The usefullness of a test $t(\vec{x})$ depends on its discriminating power between the hypotheses !

Hypothesis Testing

Consider a data sample of n measured values $\vec{x} = (x_1, x_2, \ldots, x_n)$. As the test statistic $t(\vec{x})$ is a random variable it has a PDFs $g(t|H_0)$ and $g(t|H_1)$ under the hypotheses

Type I Error

Probability to reject H_0 if true (false discovery)

$$\alpha = \int_{t_{cut}}^{\infty} g(t|H_0)dt$$

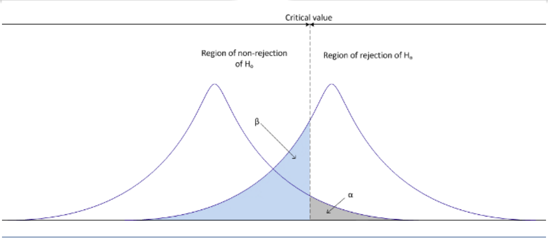
(Size or Significance Level=α)

Type II Error

Probability to accept H_0 if false (missed discovery)

$$\beta = \int_{-\infty}^{t_{cut}} g(t|H_1)dt$$

(Power=1-β)


Critical value
Region of non-rejection of H_0
Region of rejection of H_0
β
α

To perform a test one chooses a value of α which sets t_{cut} and evaluate $t_{obs} = t(\vec{x})$ on data. If $t_{obs} < t_{cut}$ one accepts null hypothesis H_0 , otherwise if $t_{obs} > t_{cut}$ one rejects it.

Neyman-Pearson Test Statistic

There are an infinity of possible choices for the test statistics $t(\vec{x})$ and we qualify it by its discrimination potential

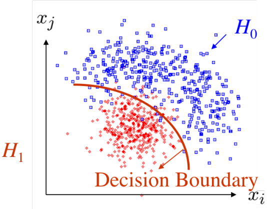
Neyman-Pearson Test Statistic

The most *Powerfull* (smaller β) test statistic for a given *Significance Level* α is the likelihood ratio

$$t(\vec{x}) = \frac{f(\vec{x} \mid H_1)}{f(\vec{x} \mid H_0)}$$

Decision Boundary

After choosing the size α and the test statistic, the critical value $t_{cut} = t(\vec{x})$ associated to α defines a hypersurface (decision boundary) in the data space $\mathbb{R}^n$. This boundary specifies acceptance and the critical (rejection) region.



Neyman-Pearson Hypothesis Test

In case one knows the analytic expression of the PDFs $f(\vec{x} \mid H_0)$ and $f(\vec{x} \mid H_1)$, the procedure for the Neyman-Pearson test is straightforward

Neyman-Pearson Hypothesis Test

1. Determine the expected distribution of t for the null hypothesis, $g(t|H_0)$
2. Define the size α of the test taking into account the cost of both type I and type II errors and obtain the critical region.
3. Determine the observed value of $t = t(\vec{x})$ from the measured data sample.
4. Check if the observed value of t lies in the critical region and make a decision: if t is within the critical region, reject H_0, otherwise, there is not enough evidence to reject H_0

Hypotheses are treated asymmetrically. Null hypothesis H_0 is special because we fix α and choose the test which maximize the power(minimize β) for the given α.

Constructing the Test Statistic

If we don't know $f(\vec{x} \mid H_0)$ and $f(\vec{x} \mid H_1)$ analytically we have the following options:

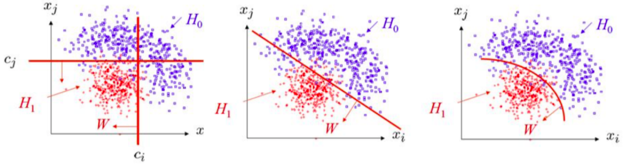
Monte Carlo

Generate a Monte Carlo simulation of signal and background events and construct multidimensional histograms corresponding to H_0 and H_1 hypotheses.

Obs: A histogram with M bins for each of the N components of $\vec{x}$, gives a total of M^N bins. So, for large number N the number of events necessary is prohibitive large (dimensional curse) !

Multivariate Analysis

For large number of observables, instead of trying to approximate the PDFs of $f(\vec{x} \mid H_0)$ and $f(\vec{x} \mid H_1)$, one can parametrize directly the test statistic (decision boundary). For linear tests one can use Fisher discriminant, while for non-linear test we need a BDT or Neural Networks.



Model Testing - Monte Carlo Method

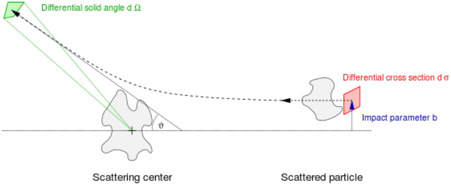
Differential Cross Section as Statistical Model

The differential cross section normalized by the total cross section can be seen as the model PDF $p(x|\mu)$. It's the probability to observe an event x, given parameter(s) μ.

For an example , let's consider the following differential cross section representing the probability of observing an event x in angle Ω with energy E

Normalized Differential Cross Section

$$p(x|\mu) = \frac{1}{\sigma}\frac{d\sigma}{d\Omega dE}$$



Obtaining a realistic model PDF in HEP is very complicated because our fundamental theories describes the physics process at parton level and that is very far from a real event observed in a detector ! One needs to process the partonic event through shower, hadronization, decay, detector simulation and reconstruction.

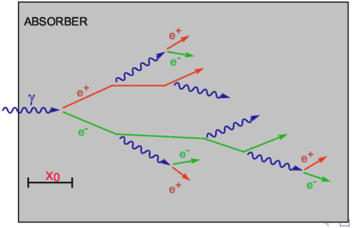
Model Testing - Monte Carlo Method

Although we know how to describe each step that takes from parton to detector level at an elementary level (QCD,QED,...), it's impossible to obtain an analytical expression in the form of a PDF. It involves multidimensional integrals that depends on stochastic processes, leading to variable number of integrands !

PDF for the testing a model (theory) at CMS Experiment

$$p(X|\theta) = \int dy_{part} \int dz_{shower} \int du_{hadr} \int dv_{detect} \int dw_{reco}\ p(X, y, z, u, v, w|\theta)$$

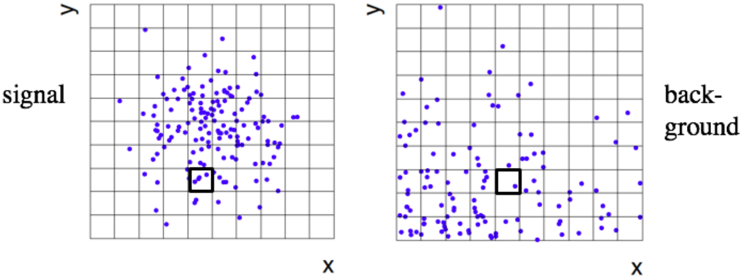
Example: an electron showers when it goes through the detector material creating a variable number of photons. This is a stochastic process leading to a fluctuating number of particles, making the integral to have a variable integration dimension



Model Testing - Monte Carlo Method

Aproximating the Test Statistic with Histograms

Suppose a given problem has only 2 variables. We can try using 2-D histograms to aproximate the test statistic PDFs using $N_{sig}(x, y)$ and $N_{bkg}(x, y)$ in corresponding cells.



For M bins in each variable in N-dimensions we have M^N cells ! ⇒ It's NOT feasible to generate enough training data to populate all bins, so only works for low dimensional problems !

Model Testing - Matrix Element Method

We can obtain an analytical approximation of the model PDF by writing it in the following form

PDF for the model CMS Experiment

$$p(x|\theta) = \int dz_{parton} q(x|z_{parton}) \underbrace{p(z_{parton}|\theta)}_{|M|^2}$$

- For each θ (model hypothesis) we have a the partonic differential cross section, that depends on the matrix element $|M|^2 = p(x|\theta)$
- $q(x|z_{parton})$ is the probability of the partonic event, characterized by z_{parton}, to be observed at the detector as x (reco level). It's the transfer function that parametrizes the physics of the parton shower, hadronization, decay, detector and reco.
- The matrix element method is slow because one needs to integrate a multidimensional integral for each event and it usually needs to use Monte Carlo simulations for determining the transfer functions and also to perform the multidimensional integration over phase space.

Goodness of the Fit

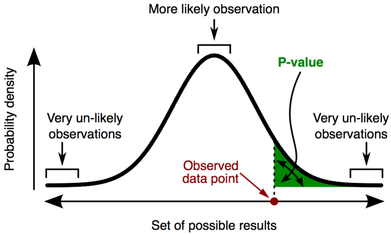
Goodness of the Fit Test (GOF)

We can quantify the agreement between data and a null hypotheses H_0 , without an alternative hypothesis. This *Goodness of the Fit* (GOF) test is quantified by the *P-value*

P-value Definition

It's the probability under the null hypothesis H_0 to obtain data as far away (or more) from the null hypothesis as the observed data

$$P = \int_{t_{obs}}^{\infty} g(t|H_0)dt$$



- t_{obs} is the value of the test statistic evaluated on data, so it's a random variable
- The smaller the *P-value* , the stronger the evidence against H_0 (harder to be bkg. fluctuation)
- The *P-value* is also known as Observed Significance Level
- In frequentist view the *P-value* is not the *H* probability because *H* not a random variable !

Parallel Between Hypothesis and GOF Tests

- If one had defined a critical region, the significance level α of the hypothesis test would correspond to the *P-value*
- In hypothesis test α is a constant chosen a priori, before looking into data, while in GOF the observed *P-value* is a random variable
- In hypothesis test the critical value (boundary) t_c is a function of the chosen α, while in GOF it's the value of the test statistic evaluated on data t_{obs}
- Although *P-value* are not confidence levels, sometimes one uses the notations $CL_b = 1 - P_0$ and $CL_{s+b} = P_1$ in analyses, where P_0 and P_1 are *P-values* for $H_0(background)$ and $H_1(signal + background)$ hypotheses

GOF: Observed Signal Significance (Poisson)

A GOF test can be applied in a counting experiment using the number n_{obs} as test statistic. If discrepancy between data and background only hypothesis is significant enough, one can claim a new discovery. The probability of observing $n = n_s + n_b$ events with mean $\nu = \nu_s + \nu_b$ is

Poisson Probability to Observe *n* Events

$$f(n, \nu_s, \nu_b) = \frac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}$$

The probability to observe n_{obs} or more events , considering it as a background only fluctuation $(\nu_s = 0)$ is

P-value for Observing n_{obs} Events

$$\begin{aligned} P(n \geq n_{obs}) &= \sum_{n=n_{obs}}^{\infty} f(n, \nu_s = 0, \nu_b) \\ &= 1 - \sum_{n=0}^{n_{obs}-1} f(n, \nu_s = 0, \nu_b) = 1 - \sum_{n=0}^{n_{obs}-1} \frac{(\nu_b)^n}{n!} e^{-(\nu_b)} \end{aligned} \quad (2)$$

GOF: Observed Signal Significance

Instead of quoting the *P-value* , publications often quote the observed signal significance Z. It's defined as the equivalent number of standard deviations and geometrically it corresponds to the equivalent area under the rightmost tail of a unit normal distribution

Signal Significance

$$\text{P-value} = \frac{1}{\sqrt{2\pi}} \int_{Z}^{\infty} e^{-x^2/2} dx = \frac{1}{2}\left[1 - Erf\left(\frac{Z}{\sqrt{2}}\right)\right]$$

By convention one claim observation or discovery according to the following criteria:

- OBSERVATION: $Z \geq 3\ (3\sigma)\ \Rightarrow P\text{-}value = 1.35x10^{-3}$
- DISCOVERY: $Z \geq 5\ (5\sigma)\ \Rightarrow P\text{-}value = 2.87x10^{-7}$

GOF: Pearson χ^2 Test

A GOF can be used to evaluate the χ^2 fit. Suppose we have a data histogram with N bins, containing n_i entries in bin i , where the expected number of entries ir ν_i. A test statistic that quantifies the level of agreement between data and expected histograms is the Pearson χ^2

Pearson χ^2

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i + \nu_i)^2}{\nu_i}$$

If data n_i is Poisson distributed with mean ν_i and $n_i > 5$ it can be shown that χ^2 will follow the distribution

χ^2 Distribution

$$f(z,N) = \frac{1}{2^{(N/2)}\Gamma(N/2)} z^{N/2-1} e^{-z/2} \text{ , where } \Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \qquad (3)$$

Pearson χ^2 Test (GOF)

The *P-value* is given by the integral of the χ^2 distribution

P-value for the χ^2

$$P(z \geq \chi^2) = \int_{\chi^2}^{\infty} f(z,N) = 1 - \int_{0}^{\chi^2} f(z,N)$$

An ambiguity of the χ^2 test is that it can depend on the histogram bining for small samples. One should require at least 5 entries per bin to use χ^2 distribution to evaluate significance. On the other hand too large bins loose information on the position of x within a bin.