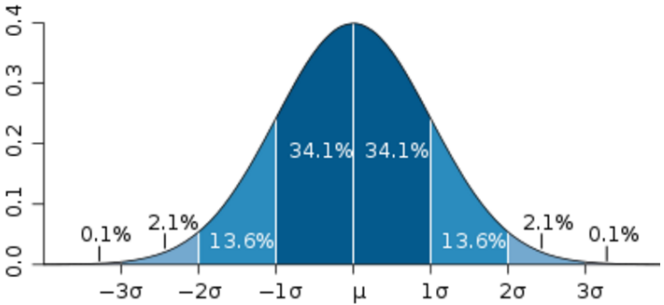
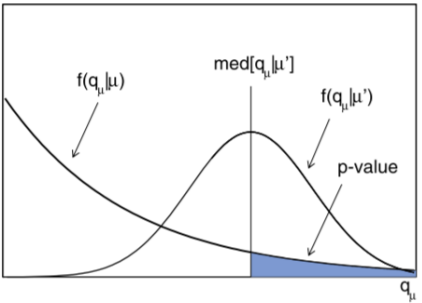




PH2282 part 1: Statistical models

Applied Multi-Messenger Astronomy 2:
Statistical and Machine Learning Methods in Particle and Astrophysics

Matteo Agostini

TUM - summer term 2019

Introduction to the course

Practical information:

- time schedule: every Friday from 10:00 to 13:45
- location: PH 1161
- responsible: Prof. Dr. Resconi (elisa.resconi@tum.de)
- continued assessment, more on this next week
- questions: elisa.resconi@tum.de, martin.wolf@tum.de

Structure of the course:

- Dr. Matteo Agostini: introduction to applied statistical methods
- Dr. Hans Niederhausen: more on statistical methods and applications to IceCube
- Dr. J. Michael Burgess: fitting a line, applications to astronomy and cosmology
- Dr. Patrick Vaudrevange: cluster analysis and neural networks

4 lecturers discussing statistics from 4 different points of view!
Peek into the analysis methods used in different fields

Introduction to the first block of lectures

About me:

- Postdoc researcher at TUM (E15)
- Office: 3063
- Research: Experimental Neutrino Phsyics (origin of neutrino masses, neutrino oscillations, solar neutrinos...)
- Office Hours: flexible, organized through email
- Email: matteo.agostini@ph.tum.de

About my lectures (first three Fridays):

1) Statistical models
2) Likelihoods and point estimation
3) Hypothesis testing
4) Test statistics and p-values
5) Interval estimation
6) Bonus topic TBD: chi-square or sensitivity

Lecture organization – 1

Hands-on session for each of the six parts:

- real-life problems and coding solutions
- gives the chance to use new statistical concepts!
- basic experience with linux and coding is needed
- macros in ROOT and python with solutions: github.com/mmatteo/TUM-lectures-PH2282
- I will stay around and try to help

Solution 1:

- 10:00 to 10:45: lecture
- 10:45 to 11:15: exercise
- 11:15 to 12:00: lecture
- 12:00 to 13:45: exercise / lunch

Solution 2:

- 10:00 to 11:30: lecture
- 11:30 to 13:45: exercise / lunch

Lecture organization – 2

IT survey :

- computers?
- linux environment? shell?
- programming experience?
- python? numpy? pylab? scipy?
- C++/ROOT?

Survey on statistic background:

- courses about statistics?
- Gaussian distribution, Poisson distribution?
- likelihoods?
- fitting?
- hypothesis testing?
- confidence intervals?

Motivations and Goals of this course

- statistics is becoming increasingly important in Physics as the analyses become more complicated and the experiments more expensive
- statistical methods can be viewed as tools to drive the intuition of the analyst and extract from the data results that are usable/comparable within the community
- statistics is a branch of mathematics but it is often taught as a collection of tools. The rationale might not be clear without studying the full mathematical framework
- we will approach statistics as an applied science and try to develop some intuition on how to handle data. This is enough for most of the physicists but it is only the tip of the iceberg

The scientific narrative (K. Cranmer, 1503.07622)

When your colleague asks you over lunch to explain your analysis, you tell a story. It is a story about the signal and the backgrounds:

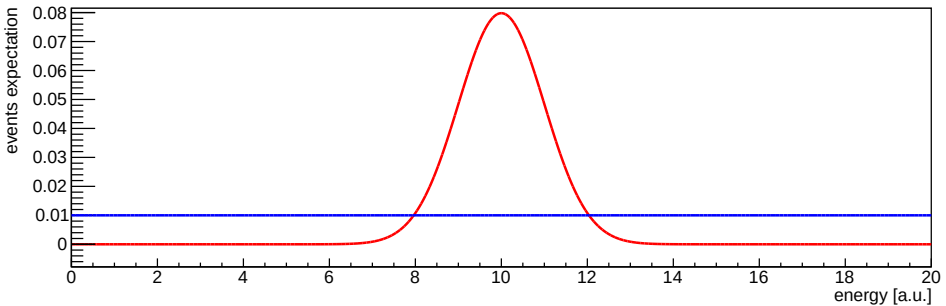
- *are they estimated using Monte Carlo simulations, a side-band, or some data driven technique?*
- *Is the analysis based on counting events or do you use some discriminating variable, like an invariant mass or perhaps the output of a multivariate discriminant?*
- *What are the dominant uncertainties in the rate of signal and background events and how do you estimate them?*
- *What are the dominant uncertainties in the shape of the distributions and how do you estimate them?*

Even a scientist working mostly on hardware should be able to handle such a conversation because this is what the result of the experiment is based on!

The reference analysis for this set of lectures – The Model

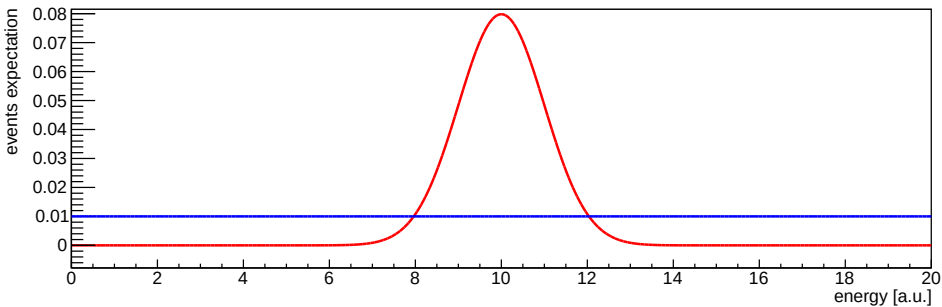
An experiment searches for an excess of events due to a signal in a energy region in which also background events are present. The experiment measures a number of events and for each its energy. The number of events expected from the signal and background is λ_s and λ_b. The energy distribution expected by signal and background events are:

- signal -> Gaussian distributed in energy (μ =10, σ =1)
- background $->$ flat distributed in energy



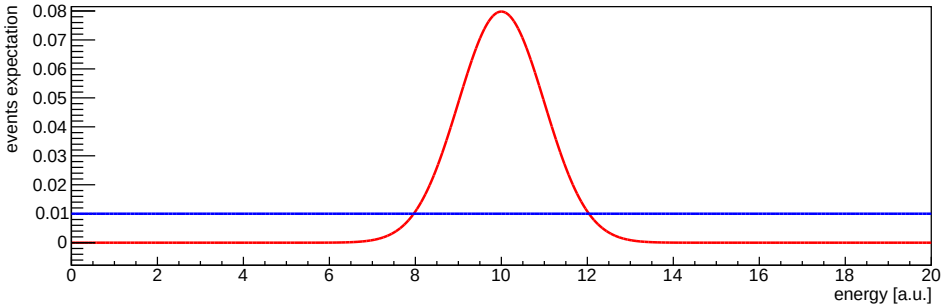
The reference analysis for this set of lectures – The Model

- assuming that the mean and sigma of the Gaussian are known, the only unknown parameters of the model are the expectation for the background λ_b and for the signal λ_s, i.e. the expected number of signal counts
- λ_s can be regarded as the "strength of the signal"
- The total number of events expected from background and signal follows a Poisson distribution: $N_s \sim \text{Poisson}(n_s; \lambda_s)$ and $N_b \sim \text{Poisson}(n_b; \lambda_b)$



The reference analysis for this set of lectures – The Model

- This statistical model might seem simple, but most of the statistical analysis in particle and nuclear physics can be traced back to it
- Large number of papers on this problem: sometimes called on/off problem or counting experiment
- on/off because an energy cut can be used to divide the data in a set containing only background events and a set containing both background and signal
- counting experiment because the data can be analyzed using the total number of counts in "off" data set and "on" data set



The reference analysis for this set of lectures – The Data

- Data can be in the most general case a set of events, each with a given energy: $\mathcal{D} = \{x_1, x_2, \ldots x_N\}$ where N is the total number of events

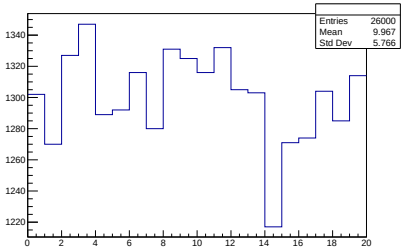
- Data can also be prepared in order to simplify the analysis without loosing information (i.e. data reduction)
  - data can be simplified using two numbers: the total number of events in the signal region and the number of events in the background region (effectively a two-bin analysis)
  - data can be divided in bins (e.g. the first bin grouping the events with energy between $x = 0$ and $x = 1$

- if the data are used considering each event separately, the resulting analysis is typically called "unbinned". If events are grouped, it is typically called binned analysis.

Tasks of statistical inference

Which questions can I try to address once I have a model and some data?

- Is there a signal? If so, how strong is the evidence for the signal?
- If there is an evidence for a signal:
  - what is the best estimate of the signal expectation λ_s?
  - What is a reasonable range of expectation values for λ_s?
- If there is no evidence for a signal:
  - Which range of expectation values for λ_s I can exclude?
  - Which range of values is still compatible with my data?
- are my data compatible with the model?
- what about the background?



Tasks of statistical inference

Which questions can I try to address once I have a model and some data?

- Is there a signal? If so, how strong is the evidence for the signal? — Hypothesis Testing
- If there is an evidence for a signal:
  - what is the best estimate of the signal expectation λ_s? — Point Estimation
  - What is a reasonable range of expectation values for λ_s? — Interval Estimation
- If there is no evidence for a signal:
  - Which range of expectation values for λ_s I can exclude? — Interval Estimation
  - Which range of values is still compatible with my data? — Interval Estimation
- are my data compatible with the model? — Hypothesis Testing: goodness of fit
- what about the background?

Tasks of statistical inference

Task name	Task description	Some Frequentist tools	Some Bayesian tools
Point Estimation	what is the best estimate for a parameter of the model?	Maximum likelihood estimator	Median or mode of posterior distribution
Hypothesis Testing	which (model) hypothesis can be accepted or rejected given the data?	likelihood ratios	Bayes factors
Interval Estimation	which range of values is plausible for a given parameter of the model?	inverse hypothesis test	intervals of the posterior distribution
Goodness of Fit	are my data compatible with a model?	chi-square test, likelihood ratio test	posterior predictive p-values

Frequentist and Bayesian tools

- For each tasks there are many tools that can be grouped in classes at different levels
- The two top groups are Frequentist vs Bayesian methods whose difference is very deep, at the level of the basic concept of probability
- In a nutshell:
  - Frequentist methods dial with the probability that the observed data are generated by a given model:

$$P(\text{Data}|\text{Model})$$

  - Bayesian methods dial with the probability of the model or of the model parameters given a prior and the observed data:

$$P(\text{Model}|\text{Data})$$

- This first set of lectures will focus on Frequentist models, Bayesian models will be introduced later in the course.

Ingredients of a Frequentist analysis

1) Concept of probability
2) Statistical model and data set
   - Random variables
   - Parameters of the model
   - Probability distribution functions (PDF's)
3) Likelihood function $\mathcal{L}$
4) Estimators [Point Estimation]
5) Test statistics [Hypothesis Testing]
   - power and size of a test
   - Likelihood ratios
6) Confidence interval [Interval Estimation]
   - coverage
   - inverting an hypothesis test

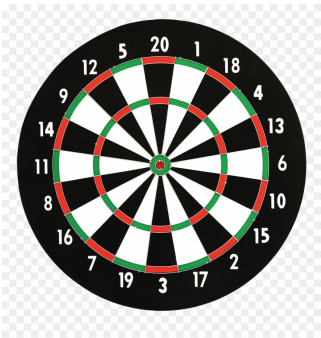
Questions?

What's a probability?

Probability can be intuitive:

- toss a coin, what's the probability that it lands heads up?

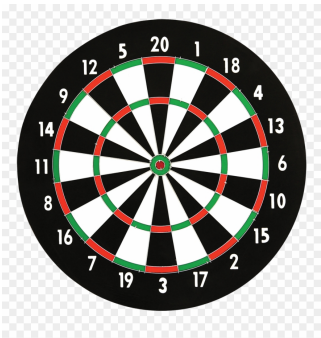
What's a probability?



Probability can be intuitive:

- toss a coin, what's the probability that it lands heads up?
  - The set of possible outcomes is $\{H, T\}$
  - $P(\{H\}) + P(\{T\}) = 100\%$
  - $P(\{H\}) = P(\{T\}) = 50\%$ (if the coin is fair)
- through a dart, what's the probability of a 20?

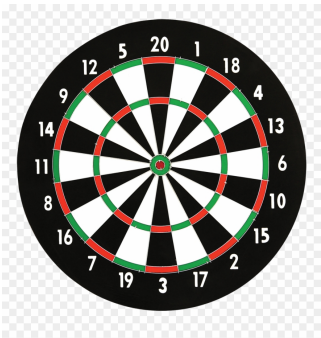
What's a probability?



Probability can be intuitive:

- toss a coin, what's the probability that it lands heads up?
  - The set of possible outcomes is $\{H, T\}$
  - $P(\{H\}) + P(\{T\}) = 100\%$
  - $P(\{H\}) = P(\{T\}) = 50\%$ (if the coin is fair)
- through a dart, what's the probability of a 20?
  - The set of possible outcomes is larger $\{1, 2, \ldots, 20, 21, \ldots\}$
  - $P(\{1\}) + P(\{2\}) + \cdots + P(\{\text{the wall}\}) = 100\%$
  - $P(\{20\}) \propto$ area

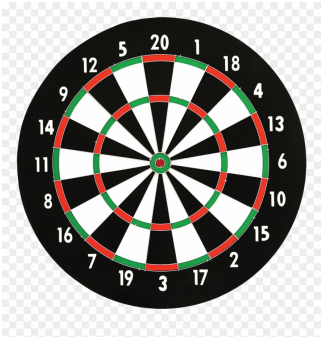
What's a probability?



Probability can be intuitive:

- toss a coin, what's the probability that it lands heads up?
  - The set of possible outcomes is $\{H, T\}$
  - $P(\{H\}) + P(\{T\}) = 100\%$
  - $P(\{H\}) = P(\{T\}) = 50\%$ (if the coin is fair)
- through a dart, what's the probability of a 20?
  - The set of possible outcomes is larger $\{1, 2, \ldots, 20, 21, \ldots\}$
  - $P(\{1\}) + P(\{2\}) + \cdots + P(\{\text{the wall}\}) = 100\%$
  - $P(\{20\}) \propto$ area
- How can check if a coin is fair? How can I estimate the probability for a not-fair coin to land heads up?

What's a probability?



Probability can be intuitive:

- toss a coin, what's the probability that it lands heads up?
  - The set of possible outcomes is $\{H, T\}$
  - $P(\{H\}) + P(\{T\}) = 100\%$
  - $P(\{H\}) = P(\{T\}) = 50\%$ (if the coin is fair)
- through a dart, what's the probability of a 20?
  - The set of possible outcomes is larger $\{1, 2, \dots, 20, 21, \dots\}$
  - $P(\{1\}) + P(\{2\}) + \cdots + P(\{\text{the wall}\}) = 100\%$
  - $P(\{20\}) \propto$ area
- How can check if a coin is fair? How can I estimate the probability for a not-fair coin to land heads up?
- What is a possible definition for probability?

The frequentist idea of probability

Frequentist probability *or* **frequentism** *is an interpretation of probability; it defines an event's probability as the limit of its relative frequency in a large number of trials. This interpretation supports the statistical needs of experimental scientists and pollsters; probabilities can be found (in principle) by a repeatable objective process (and are thus ideally devoid of opinion).*
[en.wikipedia.org/wiki/Frequentist_probability]

The frequentist idea of probability

Frequentist probability *or* **frequentism** *is an interpretation of probability; it defines an event's probability as the limit of its relative frequency in a large number of trials. This interpretation supports the statistical needs of experimental scientists and pollsters; probabilities can be found (in principle) by a repeatable objective process (and are thus ideally devoid of opinion).*
[en.wikipedia.org/wiki/Frequentist_probability]

Going back to the example of tossing a coin, the frequency of times it will land heads up will converge by increasing the number of trials towards the probability for heads up.

The frequentist idea of probability

Frequentist probability *or* **frequentism** *is an interpretation of probability; it defines an event's probability as the limit of its relative frequency in a large number of trials. This interpretation supports the statistical needs of experimental scientists and pollsters; probabilities can be found (in principle) by a repeatable objective process (and are thus ideally devoid of opinion).*
[en.wikipedia.org/wiki/Frequentist_probability]

Going back to the example of tossing a coin, the frequency of times it will land heads up will converge by increasing the number of trials towards the probability for heads up.

Warning: not all kinds of probabilities can be described with the frequentist definition.....

Random Variables

A **random variable**:

- is a variable whose possible values are outcomes of a random phenomenon
- is a variable in the sense that the frequency of its outcomes depends on the properties of the phenomenon (aka the parameters of a models)
- is random in the sense that the outcome of the process is random, ergo unpredictable

Random Variables

A **random variable**:

- is a variable whose possible values are outcomes of a random phenomenon
- is a variable in the sense that the frequency of its outcomes depends on the properties of the phenomenon (aka the parameters of a models)
- is random in the sense that the outcome of the process is random, ergo unpredictable

In the coin tossing example, random variables can be:

Random Variables

A **random variable**:

- is a variable whose possible values are outcomes of a random phenomenon
- is a variable in the sense that the frequency of its outcomes depends on the properties of the phenomenon (aka the parameters of a models)
- is random in the sense that the outcome of the process is random, ergo unpredictable

In the coin tossing example, random variables can be:

- the total number of tails or heads
- the results of a sequence of trials, e.g. T, H, T, H

Random Variables

A **random variable**:

- is a variable whose possible values are outcomes of a random phenomenon
- is a variable in the sense that the frequency of its outcomes depends on the properties of the phenomenon (aka the parameters of a models)
- is random in the sense that the outcome of the process is random, ergo unpredictable

In the coin tossing example, random variables can be:

- the total number of tails or heads
- the results of a sequence of trials, e.g. T, H, T, H

In the dart example, random variables can be:

Random Variables

A **random variable**:

- is a variable whose possible values are outcomes of a random phenomenon
- is a variable in the sense that the frequency of its outcomes depends on the properties of the phenomenon (aka the parameters of a models)
- is random in the sense that the outcome of the process is random, ergo unpredictable

In the coin tossing example, random variables can be:

- the total number of tails or heads
- the results of a sequence of trials, e.g. T, H, T, H

In the dart example, random variables can be:

- the number of points scored with a dart
- the number of points scored with 3 darts

Probability distributions

The probability distribution of a random variable is a function that associates to each possible outcome a probability value for it to occur.

In the coin tossing example, the probability distribution is:

$$\text{f(side)} = \begin{cases} 0.5 & \text{if side is head} \\ 0.5 & \text{if side is tail} \end{cases}$$

In the dart example, the probability distribution is:

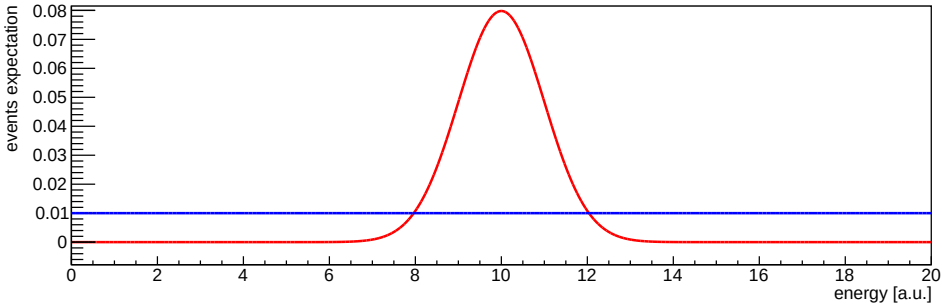
$$\text{f(N-points)} = \begin{cases} 1 & \text{fraction of area resulting in 1 point} \\ 2 & \text{fraction of area resulting in 2 points} \\ .. & \end{cases}$$

Properties:

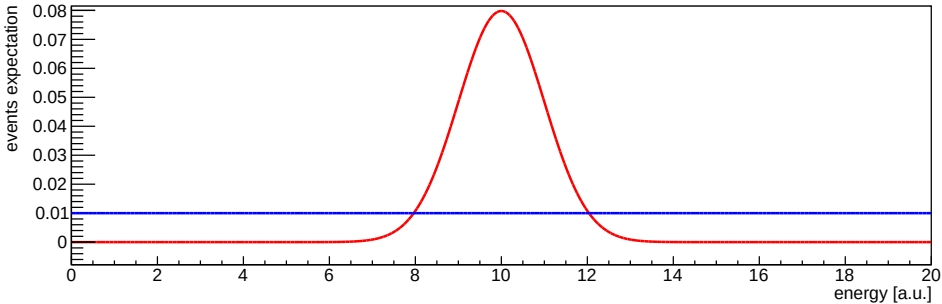
- probability distributions for discrete variables are called "probability mass functions (PMF)" while for continuous variables are called "probability density functions (PDF)"
- both PMF and PDF are positive function (probability between 0 and 1)
- both PMF and PDF are normalized

random variable	outcome of random variable	PDF
X	x	$X \sim f_X(x)$

What are possible random variables in our reference analysis?

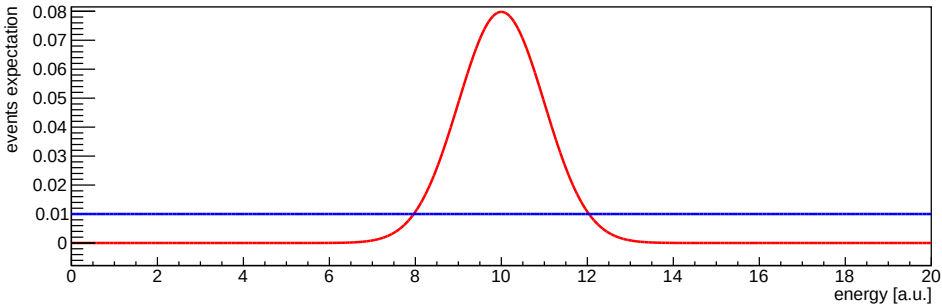


What are possible random variables in our reference analysis?



- X the energy of an event
- N_{tot} the total number of events with energy in the range $[0, 20]$
- $N_{\text{side-bands}}$ the total number of events with energy in the range $[0, 5] \cup [15, 20]$
- $N_{\text{signal-region}}$ the total number of events with energy in the range $[5, 15]$
- $\{N_1, N_2, \ldots, N_{\text{M}}\}$ the numbers of events in a set of M bins

What are the PDF's?



Random Variable	PDF
X the energy for a signal only scenario ($\lambda_b = 0$)	$X \sim f_X^s(x; \mu, \sigma) = \text{Gauss}(x; \mu, \sigma)$
X the energy for a background only scenario ($\lambda_s = 0$)	$X \sim f_X^b(x) = \text{constant}$
N, total event number for a signal only scenario ($\lambda_b = 0$)	$N \sim f_N^s(n; \lambda_s) = \text{Poisson}(n; \lambda_s)$
N, total event number for a background only scenario ($\lambda_s = 0$)	$N \sim f_N^b(n; \lambda_b) = \text{Poisson}(n; \lambda_b)$

A bit more on the PDF's

- probability distribution functions are normalized, in the sense that the sum of the probability of each possible outcome should be 1:

$$\int_{x_{\text{min}}}^{x_{\text{max}}} f(x)dx = 1$$

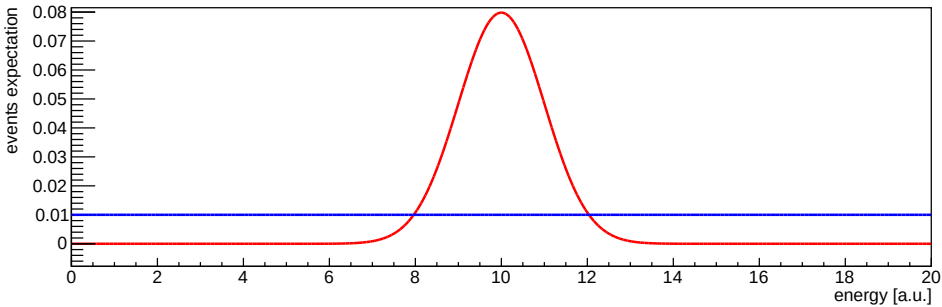
- in a background only scenario, the PDF for the energy an event is $f_X^b(x) = 1/20$:

$$\int_0^{20} f_X^b(x)dx = 1 \xrightarrow{f_X^b(x)=k} \int_0^{20} k\, dx = 1 \rightarrow k = 1/20$$

  and $f_X^s(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

- f_X^s is to a first approximation fully contained between 0 and 20. Truncated Gaussian distributions must be carefully renormalized.
- in the notation $f_X^s(x;\mu,\sigma)$ the first argument is the outcome of the random variable, the arguments after the ";" are fixed parameters of the model
- but what is the PDF for models for which both $\lambda_s > 0$ and $\lambda_b > 0$?

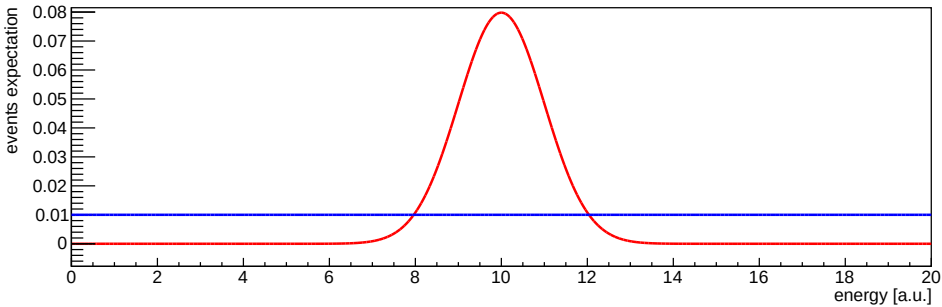
A bit more on the PDF's



The PDF for a generic model in which both $\lambda_s > 0$ and $\lambda_b > 0$ can be written as:

$$f_X(x;\mu,\sigma,\lambda_s,\lambda_b) = \frac{1}{\lambda_s+\lambda_b}\left[\lambda_s \cdot f_X^s(x;\mu,\sigma) + \lambda_b \cdot f_X^b(x)\right]$$

A bit more on the PDF's



The PDF for a generic model in which both $\lambda_s > 0$ and $\lambda_b > 0$ can be written as:

$$f_X(x;\mu,\sigma,\lambda_s,\lambda_b) = \frac{1}{\lambda_s+\lambda_b}\left[\lambda_s \cdot f_X^s(x;\mu,\sigma) + \lambda_b \cdot f_X^b(x)\right] = \frac{1}{\lambda_s+\lambda_b}\left[\lambda_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \lambda_b \cdot \frac{1}{20}\right]$$

where the coefficient $1/(\lambda_s+\lambda_b)$ serves to normalize the PDF.

A bit more on the PDF's

- The Poisson distribution for an observed number of counts given an expectation is $f_N(n;\lambda) = \frac{e^{-\lambda}\lambda^n}{n!}$
- The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event
- examples are radioactive decays, particle interactions, etc...
- for large expectations the Poisson distribution converges to a Gaussian distribution
- the PDF for the total number of coutns observed is a Poisson with expectation given by the same of the expectations:

$$f_N(n;\lambda_s+\lambda_b) = \frac{e^{-(\lambda_s+\lambda_b)}(\lambda_s+\lambda_b)^n}{n!}$$

Summary of our statistical model

Parameters of models:
λ_s and λ_b, i.e. the expectation for the total numbers of signal and background events

Elementary random variables:
N number of events (signal and background); X energy of an event

Probability distribution function for a single event energy:

$$X \sim f_X(x; \mu, \sigma, \lambda_s, \lambda_b) = \frac{1}{\lambda_s + \lambda_b} \left[\lambda_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \lambda_b \cdot \frac{1}{20} \right]$$

Probability distribution function for the number of events: background events:

$$N \sim f_N(n; \lambda_s + \lambda_b) = \frac{e^{-(\lambda_s+\lambda_b)}(\lambda_s + \lambda_b)^n}{n!}$$

Role of the model in statistical inference

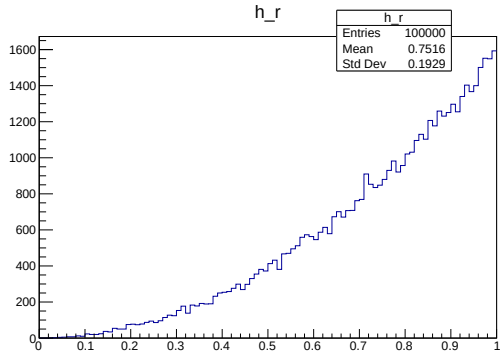
1) generation of ensemble of pseudo data (toy Monte Carlo):
   - to understand which data are expected by a model and with which frequency
   - to study the impact of the model parameters on the data
   - to test the analysis pipeline
   - to built probability distributions for complex random variables of the data

2) construction of the joint probability function for the data
   - used as tool for estimating the model parameters, confidence intervals and perform hypothesis testing

Role of the model in statistical inference

Consider a cubic detector able to reconstruct for each even the distance from the center. Assuming a homogeneous distribution of events inside the detector volume as a function of the x,y,z coordinates, which distribution is expected as a function of the radius?

```
1  TH1D h_r ("h_r","h_r",100,0,1);
2  TRandom3 rnd(0);
3  for (int i = 0; i < 1e5; i++) {
4     double x = rnd.Uniform(-1,1);
5     double y = rnd.Uniform(-1,1);
6     double z = rnd.Uniform(-1,1);
7     h_r.Fill( sqrt(x*x + y*y + z*z) );
8  }
9  h_r.Draw();
```



Summary of key concepts

- A random phenomenon is a phenomenon whose outcome is unpredictable
- The Frequentist probability is the frequency of an outcome for a large number of trials
- A random variable X is a variable whose possible values x are the outcomes of a random phenomenon. It is a variable in the sense that the frequency of its outcomes depends on the properties of the phenomenon (aka the parameters of a models). It is random in the sense that the outcome of the process is random, ergo unpredictable;
- The probability distribution of a random variable $X \sim f(x)$ is a function that associates a probability value for each possible outcome to occur
- A statistical model is defined by a set of parameters, a set of basic random variables, and their probability functions
- The probability functions of complex random variables can be constructed as a function of the basic random variables of the model (e.g. the number of counts within a energy bin)

Exercise 1

1) implement reference model assuming $\lambda_s = 1000$ counts and $\lambda_b = 1000$ counts (ergo $\lambda_{s+b} = 2000$ counts)
2) plot the PDF for the energy of an event
3) use model to generate pseudo-data:
   A) sample a random number of events froh the PDF

$$f_N(n; \lambda_s + \lambda_b) = \frac{e^{-(\lambda_s+\lambda_b)}(\lambda_s + \lambda_b)^n}{n!}$$

   B) for each events sample a random energy betwen 0 and 20 using as PDF:

$$f_X(x; \mu, \sigma, \lambda_s, \lambda_b) = \frac{1}{\lambda_s + \lambda_b}\left[\lambda_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \lambda_b \cdot \frac{1}{20}\right]$$

4) store events into a histogram with 20 bins in the energy range $[0, 20]$
5) inspect by eye which data set can be generated given this model. Repeate inspection using different model parameters. Reduce λ_s till the point the signal is not visible by eye.

macros in ROOT and python with solutions: github.com/mmatteo/TUM-lectures-PH2282