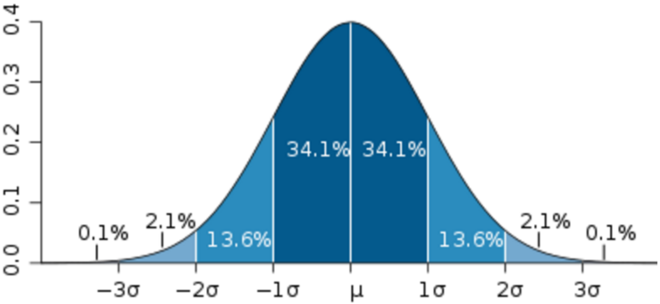
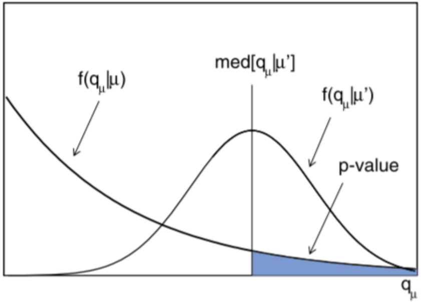




Introduction to Statistics

Applied Multi-Messenger Astronomy
Hans Niederhausen

TUM - winter term 2020/21

some general info

About me:

- Postdoc researcher at TUM
- Member of IceCube Collaboration since 2012
- Office: 2132
- Research: Experimental Astroparticle Physics with Neutrinos (diffuse flux of high energy astrophysical neutrinos, searches of neutrino sources, statistical methods ...)
- Office Hours: flexible, organized through email
- Email: hans.niederhausen@tum.de

some general info

About me:

- Postdoc researcher at TUM
- Member of IceCube Collaboration since 2012
- Office: 2132
- Research: Experimental Astroparticle Physics with Neutrinos (diffuse flux of high energy astrophysical neutrinos, searches of neutrino sources, statistical methods ...)
- Office Hours: flexible, organized through email
- Email: hans.niederhausen@tum.de

About the upcoming lectures:

1) Statistical models, Likelihoods and point estimation
2) Hypothesis testing, p-values, Interval estimation
3) Intro to Machine Learning
4) TBD.

some general info

About me:

- Postdoc researcher at TUM
- Member of IceCube Collaboration since 2012
- Office: 2132
- Research: Experimental Astroparticle Physics with Neutrinos (diffuse flux of high energy astrophysical neutrinos, searches of neutrino sources, statistical methods ...)
- Office Hours: flexible, organized through email
- Email: hans.niederhausen@tum.de

About the upcoming lectures:

1) Statistical models, Likelihoods and point estimation
2) Hypothesis testing, p-values, Interval estimation
3) Intro to Machine Learning
4) TBD.

credit to Dr. Matteo Agostini for preparing an earlier version of these slides.

Lecture organization

IT survey :

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

Lecture organization

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

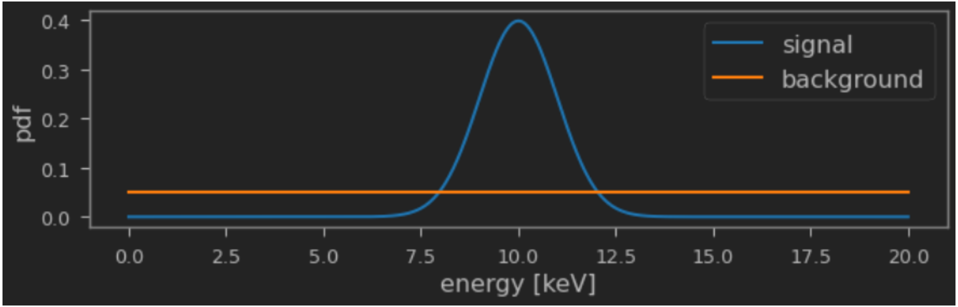
- statistics is a branch of mathematics but it is often taught as a collection of tools. The rationale might not be clear without studying the full mathematical framework

- 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 reference analysis for this set of lectures – The Model

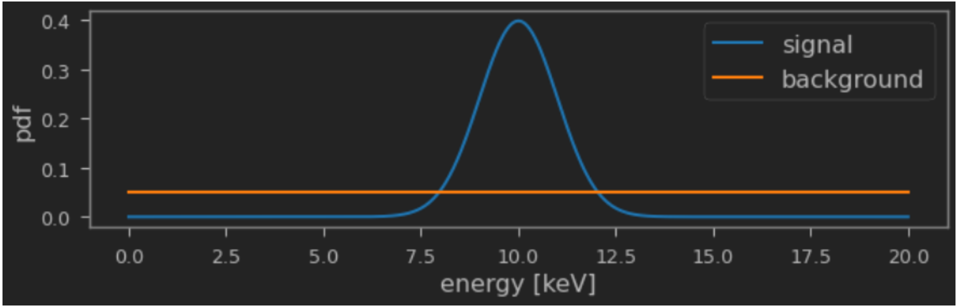
An experiment searches for an excess of events due to a signal in a energy region in which also background events are present. The experiment measures a number of events and for each its energy. The number of events expected from the signal and background is λ_s and λ_b. The energy distribution expected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

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



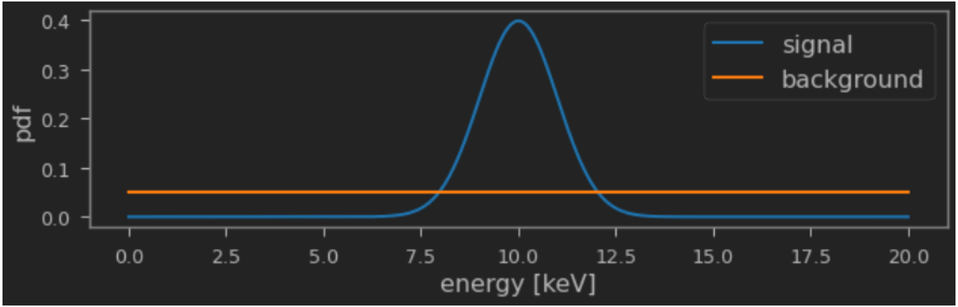
The reference analysis for this set of lectures – The Model

- 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 $N_b \sim \text{Poisson}(n_b; \lambda_b)$, thus $N \sim \text{Poisson}(n; \lambda = \lambda_s + \lambda_b)$

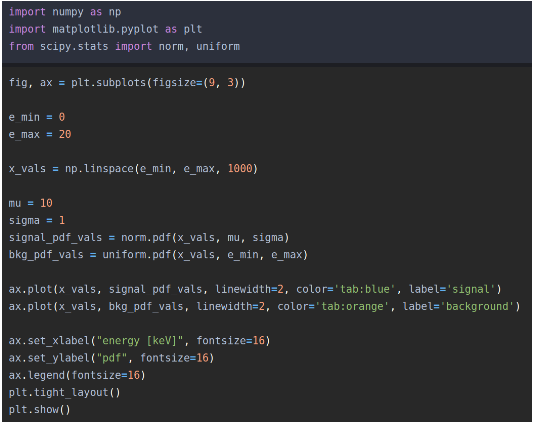


The reference analysis for this set of lectures – The Model

- This statistical model might seem simple, but most of the statistical analysis in nuclear, particle and astro 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

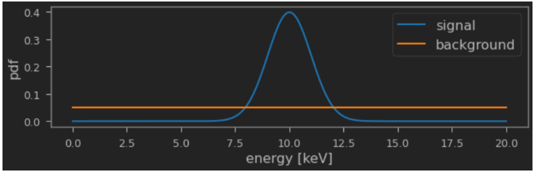


Aside: plotting code



```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm, uniform
```

```
fig, ax = plt.subplots(figsize=(9, 3))

e_min = 0
e_max = 20

x_vals = np.linspace(e_min, e_max, 1000)

mu = 10
sigma = 1
signal_pdf_vals = norm.pdf(x_vals, mu, sigma)
bkg_pdf_vals = uniform.pdf(x_vals, e_min, e_max)

ax.plot(x_vals, signal_pdf_vals, linewidth=2, color='tab:blue', label='signal')
ax.plot(x_vals, bkg_pdf_vals, linewidth=2, color='tab:orange', label='background')

ax.set_xlabel("energy [keV]", fontsize=16)
ax.set_ylabel("pdf", fontsize=16)
ax.legend(fontsize=16)
plt.tight_layout()
plt.show()
```



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: $\mathcal{D} = \{x_1, x_2, \dots x_N\}$ where N is the total number of events

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 ideally 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 $x = 1$

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

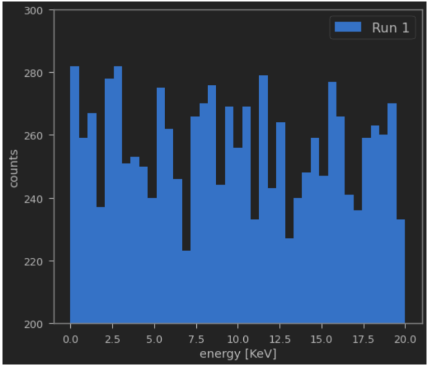
- Data can also be prepared in order to simplify the analysis ideally without loosing information (i.e. data reduction)
  - data can be simplified using two numbers: the total number of events in the signal region and the number of events in the background region (effectively a two-bin analysis)
  - 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.

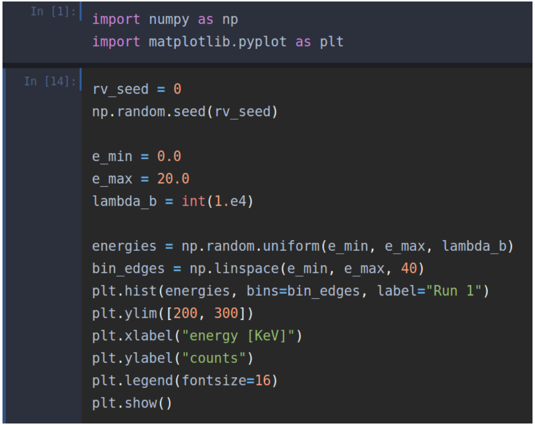
Tasks of statistical inference

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?

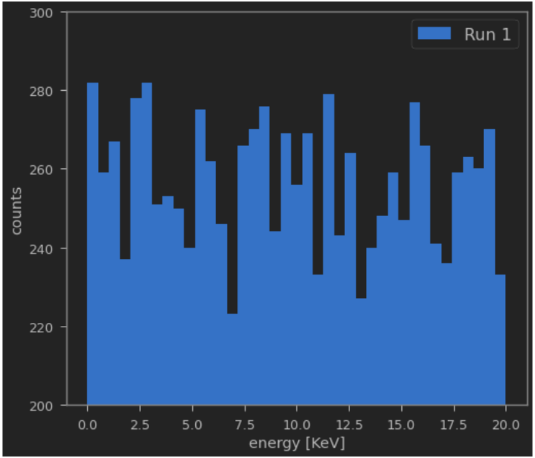


Aside: plotting code



```
In [1]:  import numpy as np
         import matplotlib.pyplot as plt
```

```
In [14]: rv_seed = 0
         np.random.seed(rv_seed)

         e_min = 0.0
         e_max = 20.0
         lambda_b = int(1.e4)

         energies = np.random.uniform(e_min, e_max, lambda_b)
         bin_edges = np.linspace(e_min, e_max, 40)
         plt.hist(energies, bins=bin_edges, label="Run 1")
         plt.ylim([200, 300])
         plt.xlabel("energy [KeV]")
         plt.ylabel("counts")
         plt.legend(fontsize=16)
         plt.show()
```



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
Point Estimation	what is the best estimate for a parameter of the model?	Maximum likelihood estimator
Hypothesis Testing	which (model) hypothesis can be accepted or rejected given the data?	likelihood ratios
Interval Estimation	which range of values is plausible for a given parameter of the model?	inverse hypothesis test
Goodness of Fit	are my data compatible with a model?	chi-square test, likelihood ratio test

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 use the probability that the observed data are generated by a given model

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

    to perform statistical inference.
    Inference is based on probability statements about procedures (e.g. point estimators) in comparison to the observed outcomes of these procedures.
  - Bayesian methods convert the data probabilities into probabilistic statements about the model or the model parameters given a prior and the observed data:

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

- Won't be able to cover Bayesian methods in this 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)

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)

- throw 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)
- throw 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

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)
- throw 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 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:

- 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:

- 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 (density) 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:

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

Probability distributions

The probability distribution of a random variable is a function that associates to each possible outcome a probability (density) 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:

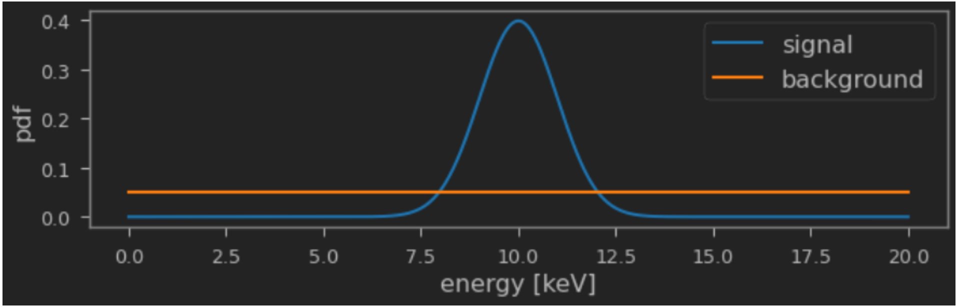
$$\text{f(N-points)} = \begin{cases} 1 & \text{fraction of area resulting in 1 point} \\ 2 & \text{fraction of area resulting in 2 points} \\ .. \end{cases}$$

Properties:

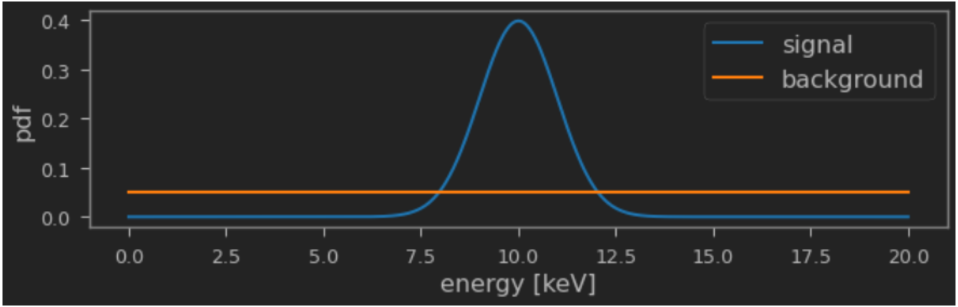
- probability distributions for discrete variables are called "probability mass functions (PMF)" while for continuous variables are called "probability density functions (PDF)"
- both PMF and PDF are positive function (probability between 0 and 1)
- both PMF and PDF are normalized (via sum, integral)

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

What are possible random variables in our reference analysis?

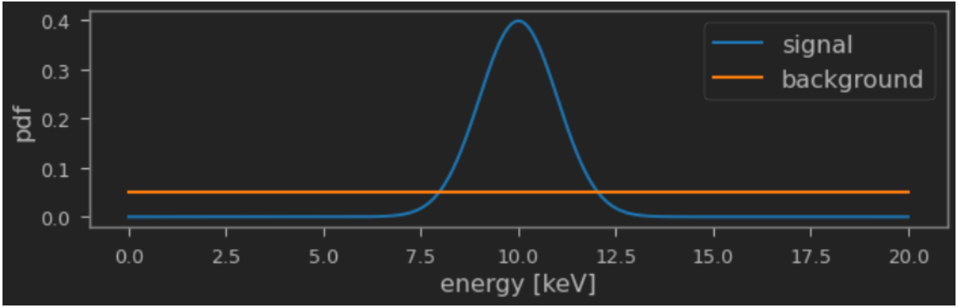


What are possible random variables in our reference analysis?



- 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?

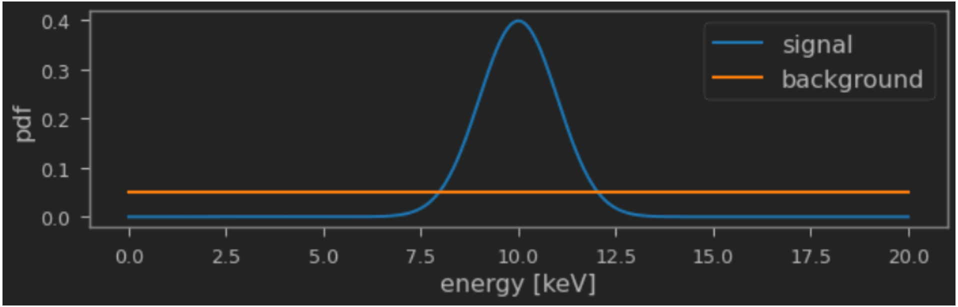


Random Variable	PDF
X the energy for a signal only scenario ($\lambda_b = 0$)	$X \sim f_X^s(x; \mu, \sigma) = \text{Gauss}(x; \mu, \sigma)$
X the energy for a background only scenario ($\lambda_s = 0$)	$X \sim f_X^b(x) = \text{constant}$
N, total event number for a signal only scenario ($\lambda_b = 0$)	$N \sim f_N^s(n; \lambda_s) = \text{Poisson}(n; \lambda_s)$
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:
$$\int_{x_{\min}}^{x_{\max}} f(x)dx = 1$$
- 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 $\lambda_b > 0$?

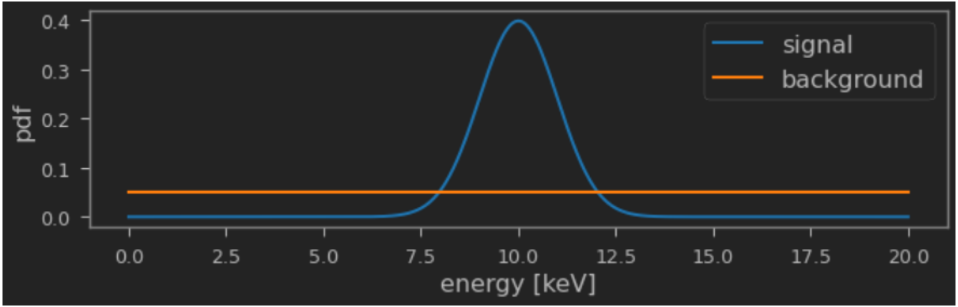
A bit more on the PDF's



The PDF for a generic *mixture 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 *mixture 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 $f_N(n;\lambda) = \dfrac{e^{-\lambda}\lambda^n}{n!}$
- 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 counts 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:

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

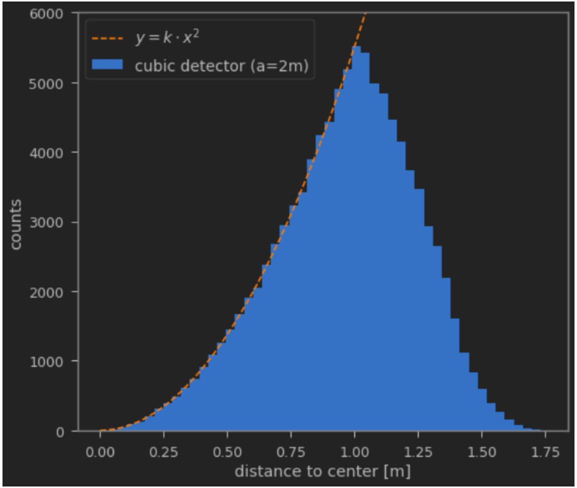
Role of the model in statistical inference

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

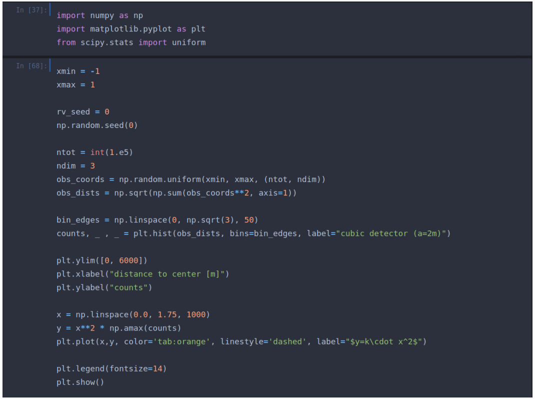
2) construction of the joint probability function for the data
   - used as tool for estimating the model parameters, confidence intervals and perform hypothesis testing

Role of the model in statistical inference

Consider a cubic detector ($a = 2m$) able to reconstruct for each event 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?

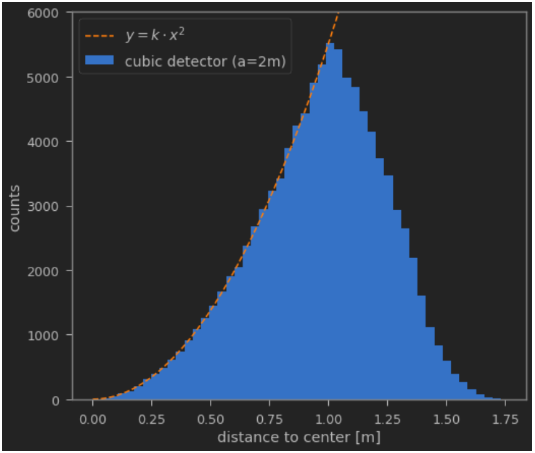


Aside: plotting code



```
In [37]: import numpy as np
         import matplotlib.pyplot as plt
         from scipy.stats import uniform
```

```
In [68]: xmin = -1
         xmax = 1

         rv_seed = 0
         np.random.seed(0)

         ntot = int(1.e5)
         ndim = 3
         obs_coords = np.random.uniform(xmin, xmax, (ntot, ndim))
         obs_dists = np.sqrt(np.sum(obs_coords**2, axis=1))

         bin_edges = np.linspace(0, np.sqrt(3), 50)
         counts, _ , _ = plt.hist(obs_dists, bins=bin_edges, label="cubic detector (a=2m)")

         plt.ylim([0, 6000])
         plt.xlabel("distance to center [m]")
         plt.ylabel("counts")

         x = np.linspace(0.0, 1.75, 1000)
         y = x**2 * np.amax(counts)
         plt.plot(x,y, color='tab:orange', linestyle='dashed', label="$y=k\cdot x^2$")

         plt.legend(fontsize=14)
         plt.show()
```



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)

Questions?

Parameters of the model

The parameters of a model can be divided into three groups:

- parameters of interest (POI): the parameters on which we want to do statistical inference (e.g. λ_s)
- nuisance parameters: parameters that are not known but whose value is not of interest (e.g. λ_b)
- known/fixed parameters: parameters of the model whose value is known and fixed (e.g. μ and σ)

In our case λ_s is the parameter of interest, λ_b is a nuisance parameter and μ/σ are fixed. Which parameters are nuisance or fixed depends on the theory behind the model. The value of fixed parameters is considered as known with infinite accuracy, while the value of nuisance parameters is typically related to a measurement and is constrained with some uncertainties.

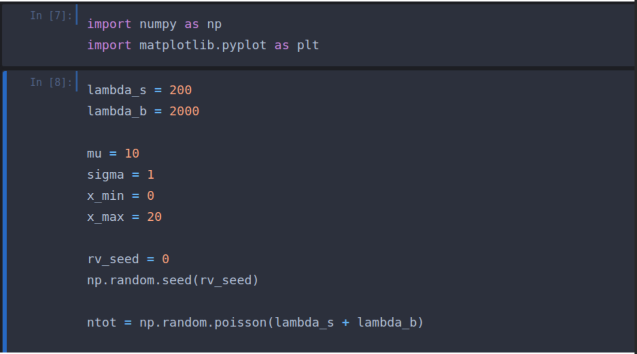
Role of the model in statistical inference

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

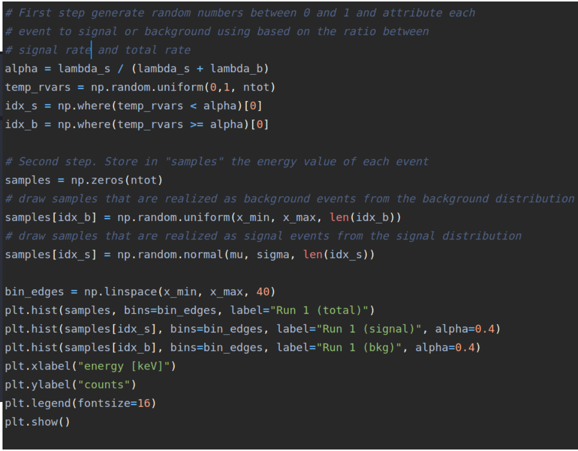
Random sampling

- Random sampling means to generate values $\{x_1, \ldots, x_N\}$ for a random variable $X \sim f_X(x)$
- generate random values from a uniform distribution $U \sim$uniform(0,1) (and other standard distributions) is relatively easy, the problem is to sample from non-standard distributions.
- Sampling from *mixture-models* can be done in two steps:
  - generate total counts from poisson distribution with expectation $\lambda_{tot} = \lambda_s + \lambda_b$
  - step 1: for each count, radomly assign type (signal or bkg) by flipping a coin with probability of success given by probability to find signal type ($p_s = \lambda_s / \lambda_{tot}$)
  - step 2: for each count generate energy from either gaussian (if type is signal) or uniform (if type is background)
  - this procedure guarantees that the sample is equivalent to being produced directly from mixture model

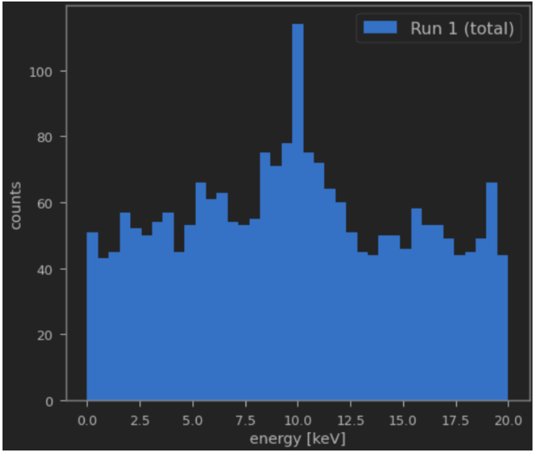


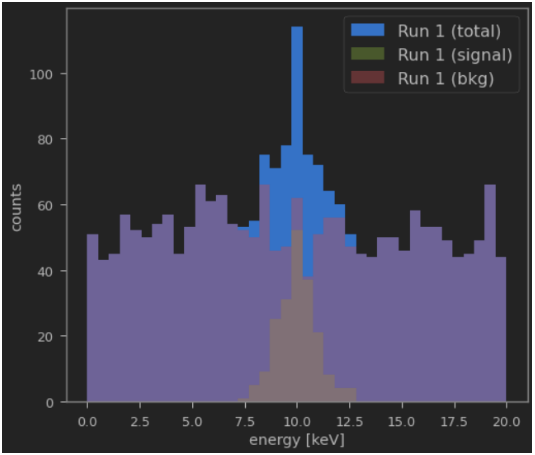
```python
import numpy as np
import matplotlib.pyplot as plt
```

```python
lambda_s = 200
lambda_b = 2000

mu = 10
sigma = 1
x_min = 0
x_max = 20

rv_seed = 0
np.random.seed(rv_seed)

ntot = np.random.poisson(lambda_s + lambda_b)
```



```python
# First step generate random numbers between 0 and 1 and attribute each
# event to signal or background using based on the ratio between
# signal rate and total rate
alpha = lambda_s / (lambda_s + lambda_b)
temp_rvars = np.random.uniform(0,1, ntot)
idx_s = np.where(temp_rvars < alpha)[0]
idx_b = np.where(temp_rvars >= alpha)[0]

# Second step. Store in "samples" the energy value of each event
samples = np.zeros(ntot)
# draw samples that are realized as background events from the background distribution
samples[idx_b] = np.random.uniform(x_min, x_max, len(idx_b))
# draw samples that are realized as signal events from the signal distribution
samples[idx_s] = np.random.normal(mu, sigma, len(idx_s))

bin_edges = np.linspace(x_min, x_max, 40)
plt.hist(samples, bins=bin_edges, label="Run 1 (total)")
plt.hist(samples[idx_s], bins=bin_edges, label="Run 1 (signal)", alpha=0.4)
plt.hist(samples[idx_b], bins=bin_edges, label="Run 1 (bkg)", alpha=0.4)
plt.xlabel("energy [keV]")
plt.ylabel("counts")
plt.legend(fontsize=16)
plt.show()
```


Run 1 (total)
counts
0
20
40
60
80
100
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
energy [keV]


Run 1 (total)
Run 1 (signal)
Run 1 (bkg)
counts
0
20
40
60
80
100
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
energy [keV]

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**

Joint probabilities

Up to now we have been considering only probabilities for a single random variable. Now we want to define a joint probability for the overall outcome of a set of random variables

What is the probability of two heads up if we toss two coins?

Joint probabilities

Up to now we have been considering only probabilities for a single random variable. Now we want to define a joint probability for the overall outcome of a set of random variables

What is the probability of two heads up if we toss two coins?

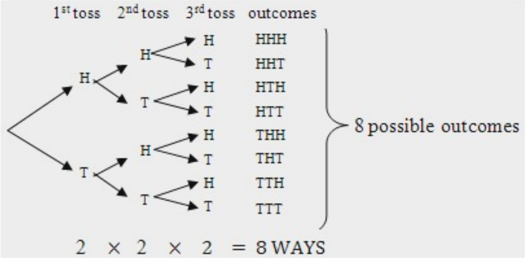
- X is the variable associated to the side of the first coin
- Y is the variable associated to the side of the second coin
- the outcome of a trial is (x, y)
- the possible outcomes are $\{\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}\}$
- what is $P(\{H, H\})$?

Joint probabilities

What is the probability of three heads up if we toss three coins?

Joint probabilities

What is the probability of three heads up if we toss three coins?



The possible outcomes are

$$\{\{H, H, H\}, \{H, H, T\}, \{H, T, H\}, \{H, T, T\}\}, \{\{T, T, T\}, \{T, T, H\}, \{T, H, T\}, \{T, H, H\}\}$$

Joint probability distributions

If two random variables are uncorrelated, the probability of a composite outcome is given by the product of the probability of each outcome:

$$P(X, Y) = P(X) \cdot P(Y)$$

Similarly, the probability distribution function for a composite outcome is

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

Joint probability distributions

If two random variables are uncorrelated, the probability of a composite outcome is given by the product of the probability of each outcome:

$$P(X, Y) = P(X) \cdot P(Y)$$

Similarly, the probability distribution function for a composite outcome is

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

In general, if we have a vector of N random variables $\vec{X} = \{X_1, X_2, \ldots, X_N\}$, their joint probability distribution is:

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{N} f_{X_i}(x_i)$$

Joint probability distributions for our reference example

The joint PDF for a vector of N uncorrelated random variables $\vec{X} = \{X_1, X_2, \ldots, X_N\}$ is:

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{N} f_{X_i}(x_i)$$

In our reference example, the probability distribution for the energy of an event is:

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

The joint PDF for N events will thus be:

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

Joint probability distributions for our reference example

Our model has two random variables, X the energy of each event and N the number of events. The joint PDF will hence be:

$$\begin{aligned} f_{N,\vec{X}}(n,\vec{x};\mu,\sigma,\lambda_s,\lambda_b) &= f_N(n;\lambda_s+\lambda_b)\cdot f_{\vec{X}}(\vec{x};\mu,\sigma,\lambda_s,\lambda_b) \\ &= \frac{e^{-(\lambda_s+\lambda_b)}(\lambda_s+\lambda_b)^n}{n!}\cdot\prod_{i=1}^{N}\left\{\frac{1}{\lambda_s+\lambda_b}\left[\lambda_s\cdot\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}+\lambda_b\cdot\frac{1}{20}\right]\right\} \end{aligned}$$

The likelihood function

Let $f_{\vec{X}}\left(\vec{x};\vec{\theta}\right)$ denote the joint probability distribution of the random variable $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ for a given set of parameters $\vec{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$. Given an observed value of $\vec{X}$ denoted with $\vec{x}$, the function of $\vec{\theta}$ defined by:

$$\mathcal{L}(\vec{\theta}; \vec{x}) = f_{\vec{X}}(\vec{x}, \vec{\theta})$$

is called the **likelihood function**.

The likelihood function

Let $f_{\vec{X}}\left(\vec{x};\vec{\theta}\right)$ denote the joint probability distribution of the random variable $\vec{X} = \{X_1, X_2, \ldots, X_n\}$ for a given set of parameters $\vec{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$. Given an observed value of $\vec{X}$ denoted with $\vec{x}$, the function of $\vec{\theta}$ defined by:

$$\mathcal{L}(\vec{\theta};\vec{x}) = f_{\vec{X}}(\vec{x},\vec{\theta})$$

is called the **likelihood function**.

The likelihood function is possibly the most important tool for statistical inference:

- think about it as the joint probability of a model, but rather than focusing on the outcome as a variable, consider the outcome as fixed by the experiment and the model parameters as variables
- the likelihood should however not be considered as a probability distribution as the model parameters are not random variables
- if we compare the likelihood function at two points and find that

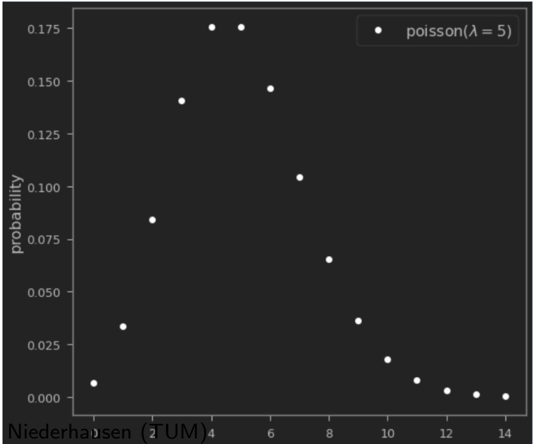
  $$P_{\vec{\theta}_1}\left(\vec{X} = \vec{x}\right) = L\left(\vec{\theta}_1;\vec{x}\right) > L\left(\vec{\theta}_2;\vec{x}\right) = P_{\vec{\theta}_2}\left(\vec{X} = \vec{x}\right)$$

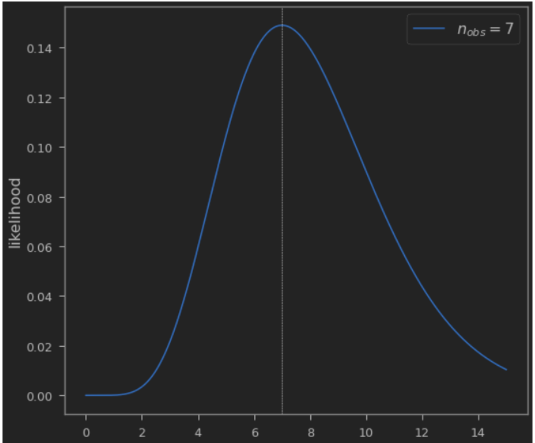
  then the observed data are more likely to have occurred if $\vec{\theta} = \vec{\theta}_1$ than if $\vec{\theta} = \vec{\theta}_1$ than if $\vec{\theta}_1$ is a more plausible value for the true value of $\vec{\theta}$ than $\vec{\theta}_2$.

Likelihood function for a Poisson process

Consider a Poisson process and a counting experiment (random variable is N). The expected number of counts in the data set is $\lambda = 5$ (which the experimenter doesn't know). Assume $n_{obs} = 7$ was measured. The (true, unknown) PDF and likelihood functions follow from

$$f_N(n; \lambda) = \mathcal{L}(\lambda; n) = \frac{e^{-\lambda}\lambda^n}{n!}$$

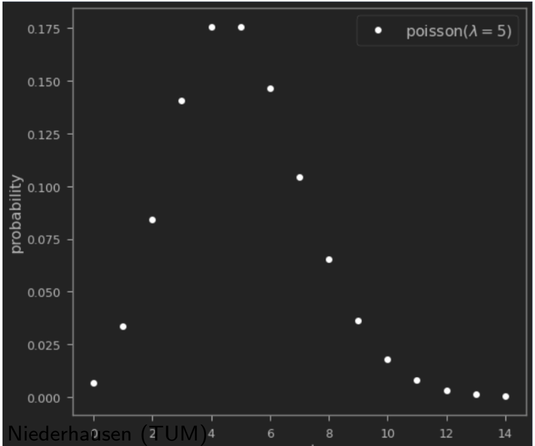


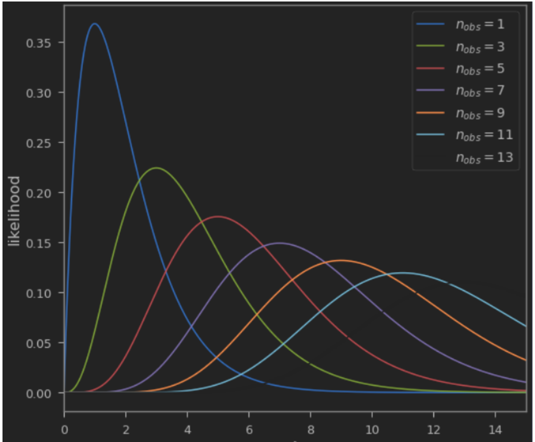


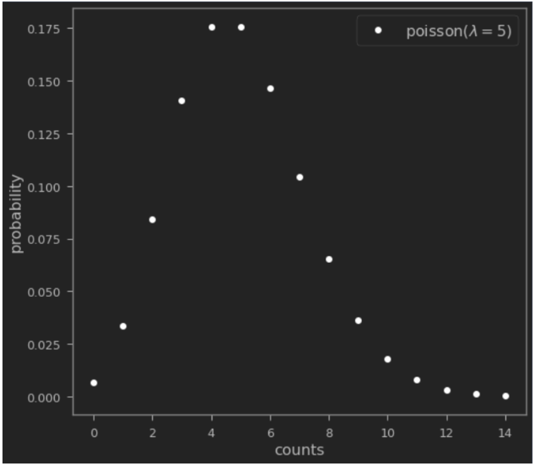
Likelihood function for a Poisson process

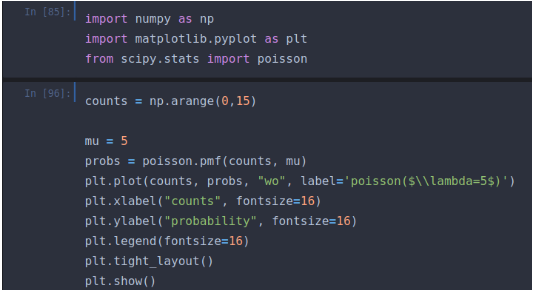
Consider a Poisson process and a counting experiment (random variable is N). The expected number of counts in the data set is $\lambda = 5$ (which the experimenter doesn't know). Assume $n_{obs} = 7$ was measured. The (true, unknown) PDF and likelihood functions follow from

$$f_N(n; \lambda) = \mathcal{L}(\lambda; n) = \frac{e^{-\lambda}\lambda^n}{n!}$$



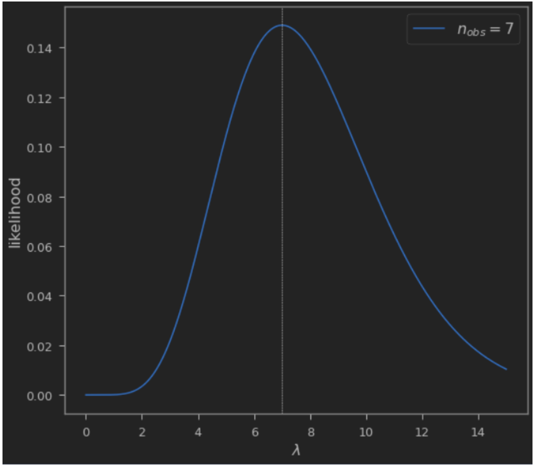


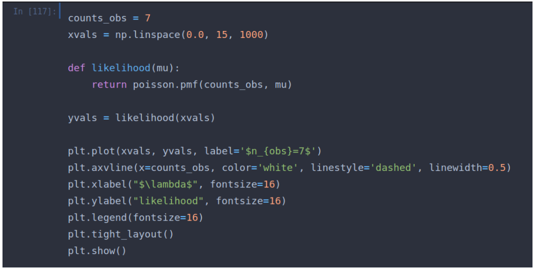




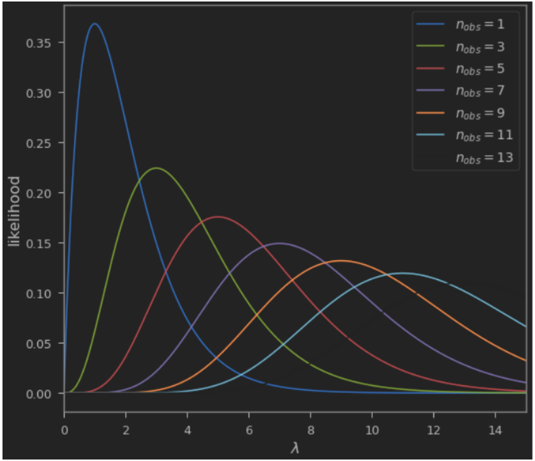
```
In [85]: import numpy as np
         import matplotlib.pyplot as plt
         from scipy.stats import poisson
```

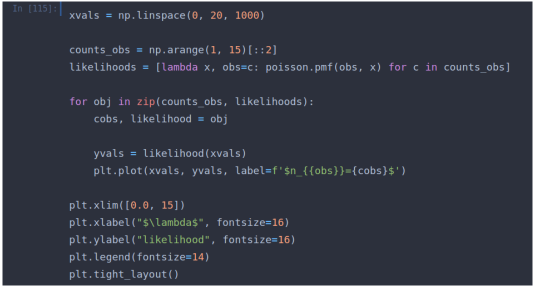
```
In [96]: counts = np.arange(0,15)

         mu = 5
         probs = poisson.pmf(counts, mu)
         plt.plot(counts, probs, "wo", label='poisson($\\lambda=5$)')
         plt.xlabel("counts", fontsize=16)
         plt.ylabel("probability", fontsize=16)
         plt.legend(fontsize=16)
         plt.tight_layout()
         plt.show()
```





```python
counts_obs = 7
xvals = np.linspace(0.0, 15, 1000)

def likelihood(mu):
    return poisson.pmf(counts_obs, mu)

yvals = likelihood(xvals)

plt.plot(xvals, yvals, label='$n_{obs}=7$')
plt.axvline(x=counts_obs, color='white', linestyle='dashed', linewidth=0.5)
plt.xlabel("$\lambda$", fontsize=16)
plt.ylabel("likelihood", fontsize=16)
plt.legend(fontsize=16)
plt.tight_layout()
plt.show()
```





```
In [115]: xvals = np.linspace(0, 20, 1000)

          counts_obs = np.arange(1, 15)[::2]
          likelihoods = [lambda x, obs=c: poisson.pmf(obs, x) for c in counts_obs]

          for obj in zip(counts_obs, likelihoods):
              cobs, likelihood = obj

              yvals = likelihood(xvals)
              plt.plot(xvals, yvals, label=f'$n_{{obs}}={cobs}$')

          plt.xlim([0.0, 15])
          plt.xlabel("$\lambda$", fontsize=16)
          plt.ylabel("likelihood", fontsize=16)
          plt.legend(fontsize=14)
          plt.tight_layout()
```

The likelihood principle

The likelihood principle states that all relevant information for inference about $\vec{\theta}$ is contained in the likelihood function for the observed data given the assumed statistical model.

A discussion about the likelihood principle and its statistical/philosophical implications is beyond the scope of this course, but it should not be hard to believe that something connected to the joint probability is all we need for statistical inference

Likelihood function for our reference example

For a given data set in which $N = n$ and $\vec{X} = \vec{x}$:

$$\begin{aligned}
\mathcal{L}(\lambda_s, \lambda_b; n, \vec{x}) &= f_{N,\vec{X}}(n, \vec{x}; \mu, \sigma, \lambda_s, \lambda_b) \\
&= \frac{e^{-(\lambda_s+\lambda_b)}(\lambda_s + \lambda_b)^n}{n!} \cdot \prod_{i=1}^{N} \left\{ \frac{1}{\lambda_s + \lambda_b} \left[\lambda_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} + \lambda_b \cdot \frac{1}{20} \right] \right\}
\end{aligned}$$

this is called an "extended unbinned" likelihood function

- it is unbinned in the sense that the energy of each events is used
- it is extended in the sense that the total number of events observed is not fixed and it is described by the Poisson term

Questions?

Point Estimators

- An estimator is a rule for calculating an estimate of a given quantity based on observed data:
- Point Estimators are rules for calculating a single value which is to serve as a "best guess" or "best estimate" of an unknown parameter

Point Estimators

- An estimator is a rule for calculating an estimate of a given quantity based on observed data:
- Point Estimators are rules for calculating a single value which is to serve as a "best guess" or "best estimate" of an unknown parameter
- Examples:
  - Let X be a random variable with a Gaussian PDF with median μ and variance σ^2. Let $x_1, x_2, \ldots, x_n$ be a set of outcomes from independent measurements of X. Estimators for the median and variance of the Gaussian are:

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

$$\hat{\sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Point Estimators

- An estimator is a rule for calculating an estimate of a given quantity based on observed data:
- Point Estimators are rules for calculating a single value which is to serve as a "best guess" or "best estimate" of an unknown parameter
- Examples:
  - Let X be a random variable with a Gaussian PDF with median μ and variance σ^2. Let $x_1, x_2, \ldots, x_n$ be a set of outcomes from independent measurements of X. Estimators for the median and variance of the Gaussian are:

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

$$\hat{\sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

- Some of the properties of an estimator:
  - mean squared error: the average squared difference between the estimator and the parameter it estimates (precision)
  - bias: the difference between the expectation value of the estimator and the true value of the parameter (accuracy)

Maximum likelihood estimators

For a given outcome $\vec{x}$, the maximum likelihood estimator (MLE) of a parameter θ is the parameter value that maximizes $\mathcal{L}(\theta; \vec{x})$ considered as a function of θ with $\vec{x}$ held fixed. We will denote the MLE with $\hat{\theta}(\vec{x})$

- intuitively, the MLE is a reasonable (and the most popular) choice for an estimator.
- The MLE is the parameter point for which the observed sample is more frequent.
- if the likelihood function is differentiable (in $\vec{\theta}$), possible candidates for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that solve
$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\vec{\theta}; \vec{x}) = 0, \qquad i = 1, \ldots, k.$$
- if the likelihood is hard to differentiate, the maximum can be found with computational methods. To this purpose, the problem of finding the maximum of the likelihood is converted into finding the minimum of the negative log-likelihood:

$$\hat{\theta} = \left\{ \max_{\theta \in \Theta} \mathcal{L}(\theta; x) \right\} = \left\{ \min_{\theta \in \Theta} -2 \log \mathcal{L}(\theta; x) \right\}$$

MLE for the mean of a Gaussian distribution

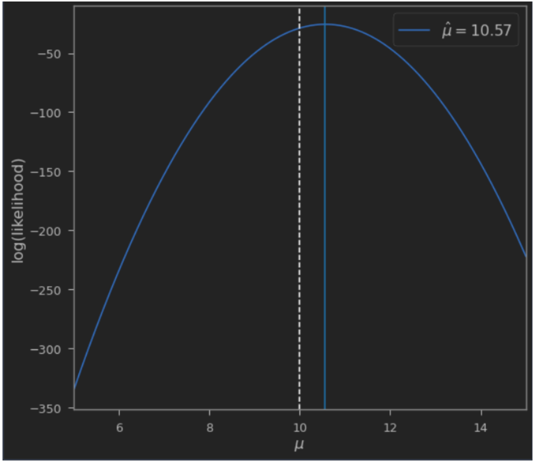
$$\mathcal{L}(\mu, \sigma; \vec{x}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

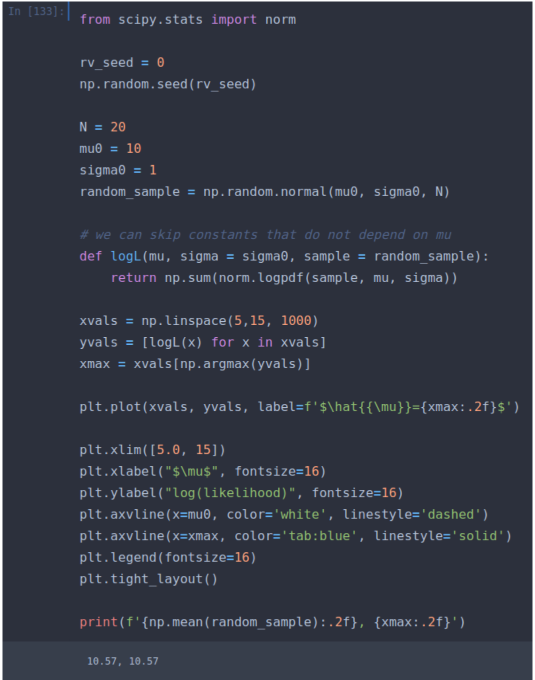
$$\log\left(\mathcal{L}(\mu, \sigma; \vec{x})\right) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}\right) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\mu}\log\mathcal{L}(\mu, \sigma; \vec{x}) = 0 \longrightarrow \frac{2}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu) = 0 \longrightarrow \sum_{i=1}^{N} x_i - N\mu = 0 \longrightarrow \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

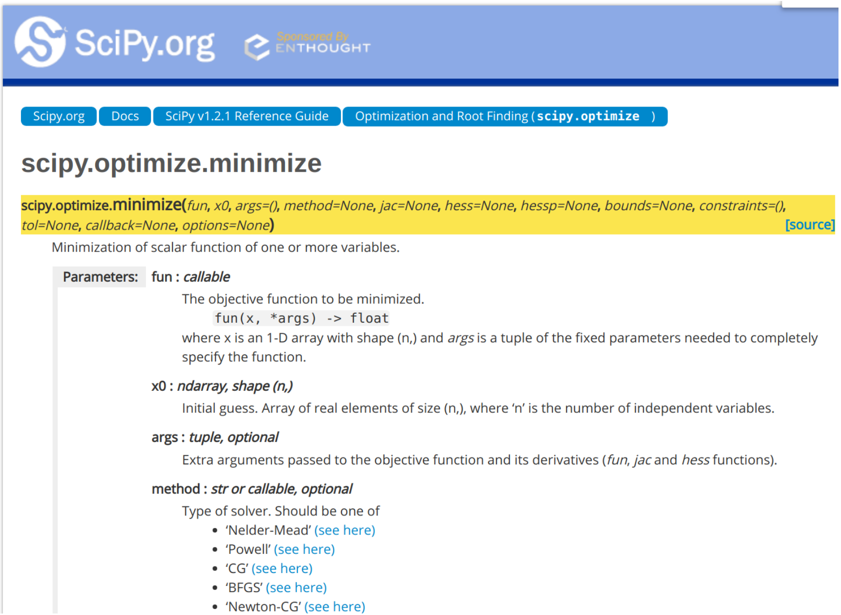
The sample mean is the maximum likelihood estimator for the parameter μ of the gaussian distribution.





```
from scipy.stats import norm

rv_seed = 0
np.random.seed(rv_seed)

N = 20
mu0 = 10
sigma0 = 1
random_sample = np.random.normal(mu0, sigma0, N)

# we can skip constants that do not depend on mu
def logL(mu, sigma = sigma0, sample = random_sample):
    return np.sum(norm.logpdf(sample, mu, sigma))

xvals = np.linspace(5,15, 1000)
yvals = [logL(x) for x in xvals]
xmax = xvals[np.argmax(yvals)]

plt.plot(xvals, yvals, label=f'$\hat{{\mu}}={xmax:.2f}$')

plt.xlim([5.0, 15])
plt.xlabel("$\mu$", fontsize=16)
plt.ylabel("log(likelihood)", fontsize=16)
plt.axvline(x=mu0, color='white', linestyle='dashed')
plt.axvline(x=xmax, color='tab:blue', linestyle='solid')
plt.legend(fontsize=16)
plt.tight_layout()

print(f'{np.mean(random_sample):.2f}, {xmax:.2f}')
```

```
10.57, 10.57
```


SciPy.org
Sponsored By ENTHOUGHT
Scipy.org
Docs
SciPy v1.2.1 Reference Guide
Optimization and Root Finding (scipy.optimize)

scipy.optimize.minimize

scipy.optimize.minimize(*fun, x0, args=(), method=None, jac=None, hess=None, hessp=None, bounds=None, constraints=(), tol=None, callback=None, options=None*) [source]

Minimization of scalar function of one or more variables.

Parameters: **fun** : *callable*

The objective function to be minimized.

```
fun(x, *args) -> float
```

where x is an 1-D array with shape (n,) and *args* is a tuple of the fixed parameters needed to completely specify the function.

x0 : *ndarray, shape (n,)*

Initial guess. Array of real elements of size (n,), where 'n' is the number of independent variables.

args : *tuple, optional*

Extra arguments passed to the objective function and its derivatives (*fun*, *jac* and *hess* functions).

method : *str or callable, optional*

Type of solver. Should be one of

- 'Nelder-Mead' (see here)
- 'Powell' (see here)
- 'CG' (see here)
- 'BFGS' (see here)
- 'Newton-CG' (see here)

Many minimizer algorithms available in scipy.
If the function is differentiable, methods based on its gradient are more performing.
(where gradients are defined w.r.t to model parameters)

You will be able to practice these concepts in the tutorial session/excercises.