Université de Genève Autumn term 2022

# Statistical Methods in Physics (14P058)

Prof. Federico Sánchez (federico.sancheznieto@unige.ch)
Dr. Hepeng Yao (hepeng.yao@unige.ch)
Dr. Knut Zoch (knut.zoch@unige.ch)

## Exercise I – Basics of statistics and python

22 & 29 September 2022, 9:15, room: SCI–202

**Task 1:** *Drawing random numbers from a PDF*

a) Use the `numpy.random` module to initialise a default random number generator (hint: check the numpy documentation).

b) Draw a batch of 100 random values from a uniform distribution in the interval $[0, 1)$.

c) Use the `matplotlib.pyplot` module to fill a histogram with the drawn values. As an example, the histogram should have 20 equidistant bins in the range $[0, 1]$.

d) Draw a batch of 1000 random values from a Gaussian distribution with mean $\mu = 15$ and standard deviation $\sigma = 1.0$. Fill these values into another histogram and plot it.

**Task 2:** *Defining and plotting analytical functions*

a) Write a function in python for a Gaussian distribution including the norm factor $\frac{1}{\sigma\sqrt{2\pi}}$. Hint: the python function should take three arguments, i.e., the value $x$ at which it is evaluated, the mean $\mu$ and the standard deviation $\sigma$. It should return its evaluated value at $x$.

b) Create an array of 100 equidistantly spaced values in the range $[10, 20]$ (you can use the `numpy.linspace` method for that).

c) Create an array that contains the corresponding values of the evaluated Gaussian distribution for $\mu = 15$, $\sigma = 1.0$, such that you have 100 pairs of $x$ and $y$ values.

d) Plot the analytical function using `matplotlib.pyplot.plot` into the same figure as the random values drawn from a Gaussian distribution in the previous task.

**Task 3:** *Fitting functions to data*

As a last step, let's use the previously defined Gaussian function and perform a parameter estimate based on the drawn data points:

a) Histogramize the 1000 drawn values into a histogram with 50 bins in the range $[10, 20]$.

b) Extract the bin centres and bin values.

c) Perform a maximum likelihood estimate of the two function parameters, $\mu$ and $\sigma$, using the `scipy.optimize.curve_fit` method and the previously defined python function for Gaussian distributions.

**Task 4:** *Bonus question: the Monty Hall problem*

The Monty Hall problem is a probability puzzle[1], loosely based on the American television game show “Let’s Make a Deal” and named after its original host, Monty Hall. More information can be found, e.g., on wikipedia. One popular formulation of the problem is the following:

> *Suppose you’re on a game show, and you’re given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what’s behind the doors, opens another door, say #3, which has a goat. He says to you, “Do you want to pick door #2?” Is it to your advantage to switch your choice of doors?*

Contrary to popular belief, switching doors is actually beneficial and increases chances of winning the car to ⅔, compared to an unchanged probability of ⅓ with the originally chosen door.

Write a small function that mimics the above problem: you pick a random door, then the “host” opens once of the other two doors, always one with a goat. You are left with the choice of keeping your originally chosen door or switching to the other unopened door.

a) Simulate a large number of toy experiments where you pick a door at random, but stick with your original door choice. What is the probability of winning?

b) Simulate a large number of toy experiments where you pick a door at random, but then switch your choice after one goat is revealed. What is the probability of winning?

[1] Original formulation: Selvin, Steve. “A problem in probability (letter to the editor)”. The American Statistician. 29 (1): 67–71. doi:10.1080/00031305.1975.10479121

Université de Genève Autumn term 2021

# Statistical Methods in Physics (14P058)

Prof. Federico Sánchez (federico.sancheznieto@unige.ch)
Dr. Hepeng Yao (hepeng.yao@unige.ch)
Dr. Knut Zoch (knut.zoch@unige.ch)

## Exercise II – Maxwell-Boltzmann distribution

7 October 2021, 9.15 CEST, room: SCI–202

Consider a box system containing a large number of $^{85}$Rb atoms with mass $m = 85m_0$ at room temperature $T$, where $m_0$ is the atomic mass unit. Along each single direction $i = x, y, z$, we assume the particles follow a probability density distribution

$$f(v_i) = \left(\frac{m}{2\pi k_B T}\right)^{1/2} e^{\frac{mv_i^2}{2k_B T}} \tag{1}$$

where $v_i$ is the velocity along each direction and $k_B$ is the Boltzmann constant. Here, we assume the $f(v_i)$ on each direction is independent with each other.

1. What is the average and variance of the distribution $f(v_i)$? Generate on your computer an array which contains the velocity information on the $x$-axis $v_x$ for $N = 20\,000$ particles and plot a histogram for it. Compute its skewness and kurtosis numerically, and comment.
2. Now, we fix $N = 20\,000$. Based on the array you generated, generate a histogram for the speed of the particle $v = |\vec{v}|$, and prove that it fits the shape
$$f(v) = \left(\frac{m}{2\pi k_B T}\right)^{3/2} 4\pi v^2 e^{-\frac{mv^2}{2k_B T}} \tag{2}$$
3. Derive analytically the probability density function for the kinetic energy $E_k = \frac{1}{2}mv^2$. And then prove it with your numerical data.
4. Write a function on your computer: it should generate an array which contains the information of $E_k$ for a given particle number $N$ and a given temperature $T$. Then, plot the histogram of $f(E_k)$ for three different temperatures, $T = 100\,\mathrm{K}, 300\,\mathrm{K}, 600\,\mathrm{K}$, on the same plot.
5. At five given temperatures, $T = 10\,\mathrm{K}, 50\,\mathrm{K}, 100\,\mathrm{K}, 300\,\mathrm{K}, 600\,\mathrm{K}$, and fixed $N = 20000$ compute with your code the expectation value $\langle E_k \rangle$. We suppose it should follow the shape $\langle E_k \rangle = \alpha k_B T$. Estimate the value of $\alpha$.
6. Redo the procedure from question 5 with $N = 500$ particles. Comment.
7. Compute the variance $\sigma^2$ of the distribution $f(E_k)$ for the case $N = 20\,000$ and $T = 300\,\mathrm{K}$ and check how it compares with $(k_B T)^2$. Compute also the skewness and kurtosis, and comment on your results.

*Hints:*

a) One can draw random numbers from a Gaussian distribution in python with the function "np.random.normal($\mu$, $\sigma$)".

b) Other possibly useful functions: plt.hist(), np.var(), scipy.stats.skew(), scipy.stats.kurtosis()

c) Universal constants: $m_0 = 1.66 \times 10^{-27}\,\mathrm{kg}$, $k_B = 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$.

Université de Genève Autumn term 2022

# Statistical Methods in Physics (14P058)

Prof. Federico Sánchez (federico.sancheznieto@unige.ch)
Dr. Hepeng Yao (hepeng.yao@unige.ch)
Dr. Knut Zoch (knut.zoch@unige.ch)

# Exercise III – Monte Carlo integration I

27 October & 3 November 2022, 9:15, room: SCI–202

Today we look at Monte Carlo integration techniques and a small example from particle physics where these techniques could be used. Let's look at the top quark, the heaviest known elementary particle. The top quark is unstable and decays. Thus, according to the Heisenberg uncertainty principle, its mass distribution is not a delta peak, but follows a Cauchy distribution (in HEP often referred to as Breit-Wigner distribution).

The top quark mass is measured to be approximately $m_t = 173\,\text{GeV}$ with a decay width calculated to be $\Gamma_t \approx 1.33\,\text{GeV}$. The decay width corresponds to the "full width at half maximum" value of the Cauchy distribution (i.e. $\Gamma = 2\gamma$).

### Task 1: *Accept–reject method*

a) Implement two functions in python that evaluate the Cauchy and Gauss distributions according to the equations below.

$$\text{Cauchy distribution:} \quad f_C(x; x_0, \gamma) = \frac{1}{\pi\gamma} \cdot \frac{\gamma^2}{(x - x_0)^2 + \gamma^2} \tag{1}$$

$$\text{Gauss distribution:} \quad f_G(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}$$

*Hint:* Choose the definitions of the python functions to have the following signatures:

```
def Cauchy(x, x_0, gamma):
    """Implementation here."""
def Gauss(x, mu, sigma):
    """Implementation here."""
```

b) Using your definition, plot the mass distribution of the top quark in the interval $[170, 176]$. Remember that $\Gamma_t$ is the *full* width at half maximum, but $\gamma$ is the half width.

c) Now use the accept–reject method to sample from a Cauchy distribution. Initialise a random number generator (`np.random.default_rng()`) and generate a set of 2000 pairs of random $x, y$ values. Choose $x \in [170, 176]$. For $y$, use a uniform distribution $h(x)$ such that $\forall x \in [170, 176] : h(x) \geq f_C(x)$. Remember: the closer $h(x)$ is to the Cauchy distribution, the higher your acceptance rate! Plot the generated pairs of $x, y$ values using the `plt.scatter(x, y)` function, together with the Cauchy distribution.

d) Implement the acceptance criterion: accept all $x, y$ pairs where $y <= f_C(x)$. Then make the same plot as before, but with accepted and rejected points in different colours.

*Hint:* comparison operators in python can also work with lists. For example, in the expression `a = (b <= c)`, `b` and `c` can be lists. `a` will then be a list of booleans.

e) Calculate the acceptance rate and comment.

f) Repeat the accept–reject method with a Gaussian distribution as an envelope. The Gauss distribution $f_G(x)$ should serve as an *envelope* to the Cauchy distribution, so again $\forall x \in [170, 176] : f_G(x) \geq f_C(x)$. Choose $\mu = m_{top}$, then pick an appropriate value for $\sigma$ and a scaling value for the entire function. Remember: the closer you are to the Cauchy distribution, the higher the acceptance rate. Plot them together.

g) Now, for the already generated $x \in [170, 176]$, sample 2000 new $y$ values based on the Gaussian envelope that you picked. Create a plot that shows the Cauchy distribution of the top quark mass, the Gaussian envelope, as well as the accepted and rejected sample points (in different colours).

h) Calculate the acceptance rate again and compare it to that of the uniform envelope. Is it what you expected? Comment.

**Task 2:** *Bonus: equal-density sampling*

When trying to fill the sampled values into a histogram with `plt.hist()`, you will notice that their density is not flat in $x$ (because we did not use a uniform envelope).

a) What event weights do you need to add such that the filled histogram follows the shape of the top quark mass distribution? Plot the final histogram and overlay it with the Cauchy distribution.

*Hint:* remember which function you used to sample the $y$ values. You need to use that function and scale with the number of bins / number of samples.

b) Once you have added the event weights correctly, get the return values of `plt.hist()` like in the following code:

```
contents, bin_edges, _ = plt.hist(x, weights=weights, bins=bins)
```

Calculate the sum of all bin contents times the bin width. This should be an estimator of the integral of the Cauchy distribution for $x \in [170, 176]$!

c) Use the following snippet to calculate the actual integral of the Cauchy distribution and compare with the estimate from the previous task. Play with the number of drawn sample points. Can you get a better estimate?

```
import scipy.integrate as integrate
integrate.quad(lambda x: Cauchy(x, m_top, gamma_top), 170, 176)[0]
```

Université de Genève Autumn term 2022

# Statistical Methods in Physics (14P058)

Prof. Federico Sánchez (federico.sancheznieto@unige.ch)
Dr. Hepeng Yao (hepeng.yao@unige.ch)
Dr. Knut Zoch (knut.zoch@unige.ch)

# Exercise IV – Monte Carlo integration II

10 & 17 November 2022, 9:15, room: SCI–202

In this exercise, we will firstly revisit the Maxwell-Boltzmann distribution for the kinetic energy $E_k$ of atoms in a three-dimensional (3D) classical system at room temperature. We will start directly from the energy distribution and compute the physical properties with Monte Carlo techniques. Then, we will turn to the case of 3D quantum Fermi gases at low temperature, where a Fermi-Dirac distribution should be applied to describe the system.

**Task 1:** *Revisit of the Maxwell-Boltzmann Distribution*

Consider a box system containing a large number of $^{40}$K with mass $m = 40m_0$ at room temperature $T$, where $m_0$ is the atomic mass unit. The kinetic energy of the system $E_k$ follows the probability density function

$$f(E_k) = 2\left(\frac{1}{k_B T}\right)^{3/2}\sqrt{\frac{E_k}{\pi}}e^{-\frac{E_k}{k_B T}} \tag{1}$$

In Exericse II, we computed the expectation value and the variance of $E_k$ based on the facts that $f(v_i)$ follows a Gaussian (which can be directly generated with the *numpy.random* function) and that $f(E_k)$ can be obtained from $f(v_i)$ with a change of variables. In this exercise, we assume that we do not have any knowledge of $f(v_i)$ and we try to compute the statistical properties of $E_k$ directly from Eq. (1) with Monte Carlo methods.

1. Prove analytically that the function $f(E_k)$ has its maximum at

$$f\left(\frac{k_B T}{2}\right) = \sqrt{\frac{2}{\pi}}\left(\frac{1}{k_B T}\right)e^{-1/2} \tag{2}$$

2. Use the accept-reject method to generate $E_k$ for $N = 10\,000$ particles in the range $[0, 10k_B T]$. Fill the values into a histogram, plot it and fit it with Eq. (1).

   *Hint:* We suggest to define python functions to generate the Maxwell-Boltzmann distribution. This will simplify the process of solving the following questions.

3. Based on the results of question 2, compute the expectation value for $E_k$

$$\langle E_k \rangle = \int E_k f(E_k) dE_k \tag{3}$$

   and show that it satisfies $\langle E_k \rangle = 1.5\, k_B T$.

4. Compute the variance of $E_k$ and compare it with $(k_B T)^2$.

5. Compute $\langle E_k \rangle$ defined in Eq. (3) with the inverse transform method.

   *Hint:* One should take advantage of the term $\exp\left(-\frac{E_k}{k_B T}\right)$ in $f(E_k)$.

6. Assume the generated values of $E_k$ are measurement from an actual experiment. From this data, we can obtain the corresponded velocity $v = \sqrt{2mE_k}$. What's the standard deviation for the velocity $\sigma_v$? Compute it both numerically and analytically.

**Task 2:** *Fermi-Dirac distribution*

Now, we move to the Monte Carlo generator of three-dimensional Fermi gas at low temperature. Consider a 3D box system which contains a large number of $^{40}$K atoms with mass $m = 40\,m_0$, where $m_0$ is the atomic mass unit. At low enough temperatures and a particle density of $n = 1$, the energy $\epsilon$ of the atoms follows the Fermi-Dirac distribution

$$f(\epsilon) = A\frac{\sqrt{\epsilon}}{\exp\left(\beta(\epsilon - \mu)\right) + 1} \tag{4}$$

where $\beta = 1/(k_B T)$ denotes the inverse energy scale related to the temperature $T$, $\mu$ is the chemical potential of the system, and $A$ is a pre-factor related to the properties of the system. Let's consider $T = 3\,\mathrm{K}$ and $\mu = 30\,k_B T$. Under these conditions, we have $A \simeq 3.5 \times 10^{31}$.

1. Build a MC process to generate this probability density function for a large number of data points $N = 10\,000$. Fill your results in a histogram, plot them and verify their correctness by comparing them with the curve of the analytical form in Eq. (4).

2. At low temperatures, the fermions show quantum behavior, which means that the mean energy per particle follows

   $$\mathbb{E}\left[\epsilon\right] \simeq \frac{3}{5}n\mu. \tag{5}$$

   Compute the sample mean $\langle \epsilon \rangle$ and compare with the expectation value. Also, compute $\langle \epsilon \rangle / k_B T$ and comment.

Université de Genève Autumn term 2022

# Statistical Methods in Physics (14P058)

Prof. Federico Sánchez (federico.sancheznieto@unige.ch)
Dr. Hepeng Yao (hepeng.yao@unige.ch)
Dr. Knut Zoch (knut.zoch@unige.ch)

## Exercise V – Basics of parameter estimation

24 November & 1 December 2022, 9:15, room: SCI–202

In this exercise, we are going to look at some basics of parameter estimation. Let's first set the foundation and implement random sampling from a Gaussian distribution, before looking at the $\chi^2$ distribution and the $t$-distribution in more detail.

**Task 1:** *Gaussian sampling & calculation of the $\chi^2$ value*

a) Implement a Gaussian distribution as you did for Exercise sheet 3. Plot a Gaussian distribution for $\mu = 2$, $\sigma = 0.5$ to test your implementation.

*Hint:* choose the signature of the function to have the following form:

```
def Gauss(x, mu=0, sigma=1):
```

b) Implement a function that samples $x, y$ pairs from a Gaussian distribution. Ideally your function should have a signature such as this one:

```
def sample_from_Gaussian(n_samples, mu=0, sigma=1):
```

*Hint:* initialise a random number generator (`np.random.default_rng()`), which you can use to draw random numbers from a uniform distribution, $r \in [0, 1)$. Based on the inversion trick, you can then use the percent point function (i.e. the inverse of the c.d.f.!) of the Gaussian distribution as implemented in `scipy.stats.norm.ppf(r, loc=0, scale=1)` to project $r$ on the Gaussian envelope. The location parameter corresponds to the population mean $\mu$, the scale to the standard deviation $\sigma$ of the population. For sampling in $y$, you can simply use your previously defined `Gauss(x, mu, sigma)` function.

c) Now sample a low number of $x, y$ pairs, e.g. $n = 6$, from the Gaussian distribution. Plot them together with the Gaussian envelope as a scatter plot. You can play around with the sample size to get a better idea if your random sampling is working.

d) Implement a function to calculate the $\chi^2$ value of the sample according to:

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

**Task 2:** *Comparison with the $\chi^2$ distribution*

Now we fix $n = 6$, but perform the same experiment a total of $n_{\mathrm{exp}} = 5000$ times.

a) Calculate the $\chi^2$ values of each of the experiments and fill them into a histogram. Compare the histogram with the p.d.f. of the corresponding $\chi^2$ distribution.

b) What does it mean if a variable, such as the one defined in eq. (1), is distributed according to a $\chi^2$ distribution?

*Hint:* the p.d.f. of the $\chi^2$ distribution is accessible via `stats.chi2.pdf(x, df)`.

**Task 3:** *Student's t-distribution*

Last but not least we want to calculate the $t$-test variable according to

$$t = \sqrt{n}\,\frac{\bar{x} - \mu}{\hat{\sigma}} \tag{2}$$

for a set of $n_{\text{exp}} = 5000$ experiments. Here, $\hat{\sigma} = s$ is the (Bessel-corrected) estimate of the standard deviation of the population, i.e., the sample standard deviation.

a) Implement a function to calculate the $t$-test variable according to eq. (2) for a list of sampled values. The function could for example have the following signature:

```
def calculate_t(x, mu):
```

b) Again, use $n = 6$ and $n_{\text{exp}} = 5000$ and calculate the $t$-test variable for each of the experiments. Fill them into a histogram. Compare the histogram with the p.d.f. of the corresponding Student's $t$-distribution. What is the number of degrees of freedom to choose here? Why?

c) What does it mean if a variable, such as the one defined in eq. (2), is distributed according to a $t$-distribution?

*Hint:* the p.d.f. of the $t$-distribution is accessible via `stats.t.pdf(x, df)`.

Université de Genève Autumn term 2022

# Statistical Methods in Physics (14P058)

Prof. Federico Sánchez (federico.sancheznieto@unige.ch)
Dr. Hepeng Yao (hepeng.yao@unige.ch)
Dr. Knut Zoch (knut.zoch@unige.ch)

## Exercise VI – Maximum likelihood principle

13 December 2022, 15:15, room: SCI–222

In this exercise, we will use the maximum likelihood principle to perform parameter estimation. We will first generate the energy for a large number of atoms which follows the Boltzmann distribution. Then, we assume this to be the experimental data from an actual measurement and estimate the temperature of the system based on this data.

Similar to the previous exercise, we consider a box system containing a large number of $^{85}$Rb atoms with mass $m = 85\,m_0$ at room temperature $T$, where $m_0$ is the atomic mass unit. The kinetic energy of the system $E$ follows the probability density function

$$f(E) = 2\left(\frac{1}{k_B T}\right)^{3/2}\sqrt{\frac{E_k}{\pi}}e^{-\frac{E_k}{k_B T}} \tag{1}$$

As a reminder: the expectation value of this distribution is $1.5\,k_B T$.

1. Write a Monte Carlo process which generates the energy distribution for $N_{par} = 400$ particles at temperature $T = 300\,\mathrm{K}$. In the following, we will assume this is the data from a single measurement in an actual thermal gas experiment.

2. Write a function which computes the likelihood $L(E_1, E_2, ..., E_{N_{par}}\,|\,T)$ of your data at a given temperature $T$. As a reminder: the definition of the likelihood follows

$$L(\vec{X}|\theta) = \prod_i f(x_i|\theta). \tag{2}$$

3. In actual experiment, we may know in advance a wide range of temperatures where the system's temperature locates. In our case, assume that we know the system has a temperature in the range $[250\,\mathrm{K}, 350\,\mathrm{K}]$. Taking a list of $T$ in the range $[250\,\mathrm{K}, 350\,\mathrm{K}]$ with a distance of $1\,\mathrm{K}$, plot $-\ln L$ as a function of $T$ and locate the minimum of $-\ln L$. This will be your estimation of $T_e$.

4. Now, we would like to perform two different methods to estimate the uncertainty of the obtained $T_e$. We start with repeating the same estimation process and study the distribution of $T_e$. To do so, you should repeat the same process of question 1-3 with $N_{exp} = 200$. (In an actual experiment, this means you perform the same measurement 200 times on the same system.) Then, check that the distribution of $T_e$ follows a Gaussian. Find the standard variance $\sigma_{T_e}$ of this Gaussian.

5. Another estimation of the uncertainty focuses only on a single set of measurement. Taking your data from question 3, find the two values $T_1$ and $T_2$ which satisfy: (a) the difference between its corresponding $-\ln L$ and the one of $T_e$ is less than 0.5, (b) the value $T_2 - T_1$

is maximized. Then, the obtained value $(T_2 - T_1)/2$ is the estimation of the uncertainty. Compare it with the result from question 4.

6. Fix $N_{exp} = 200$, estimate the Fisher information $I(\hat{T})$ with particle numbers $N_{par} = 100, 200, 400$. Comment on your results.

7. Fix $N_{par} = 200$, estimate the Fisher information $I(\hat{T})$ with experiment numbers $N_{exp} = 100, 200, 400$. Comment on your results.