# 142.351, 260014: Statistical Methods of Data Analysis

W. Waltenberger[1] (Lecture)
L. Einfalt[1,2] (Tutorials TU Wien and Universität Wien)
F. Reindl[1,2] (Tutorials TU Wien and Universität Wien)

[1] Institut für Hochenergiephysik der Österreichischen Akademie der Wissenschaften, A-1050 Wien, Nikolsdorfer Gasse 18
[2] Technische Universität Wien, Atominstitut, A-1020 Wien, Stadionallee 2

**Winter term 2023/2024**

## Exercise sheet 3

**Hand-in by:** Friday, 10. November 2023 - 11:00
**Class:** Friday, 10. November 2023 - 14:15

### Example 3.1

Let $X$ be distributed according to the gamma distribution $\mathrm{Ga}(a, b), a > 2$ and $Y = 1/X$. Determine the density, expectation value $\mu$ and variance $\sigma^2$ of $Y = 1/X$. Compare the exact values of $\mu$ and $\sigma^2$ with those that follow from first and second order linear error propagation, respectively.

**Example 3.2 (Prog)**

You measure the energy spectrum of a radioactive $^{55}$Fe source. The decay process of the source produces X-ray radiation at 5.9 keV and 6.49 keV, in a ratio of about 9:1. You observe these lines as Gaussian peaks due to their natural uncertainties $\sigma_0$ and $\sigma_1$. The expected energy spectrum is therefore proportional to the term

$$9 \cdot \mathcal{N}(E|\mu = 5.9, \sigma_0) + \mathcal{N}(E|\mu = 6.49, \sigma_1).$$

Your detector is subject to an additional measurement uncertainty (resolution) which can be approximated by $\mathcal{N}(\mu = 0, \sigma_2)$ and smears the expected spectrum.

a) Calculate the density function of the measured spectrum depending on $\sigma_0, \sigma_1$ and $\sigma_2$ by convolving the expected spectrum with the resolution.

b) Your detector has a resolution of $\sigma_2 = 0.1$ keV. You identify two Gaussian peaks in your measured spectrum with $\tilde{\sigma_0} = \tilde{\sigma_1} = 0.4$ keV. Calculate the natural uncertainties $\sigma_0$ and $\sigma_1$.

c) Simulate and plot the measured spectrum by generating N random numbers in Python, according to the following scheme:

- Draw N random numbers $X_0$ from $\mathcal{U}(0, 1)$.
- Draw N random numbers $X_1$: For each $X_0$ smaller than 0.9, draw a random number from $\mathcal{N}(5.9, 0.3)$, for each $X_0$ larger than 0.9, draw a random number from $\mathcal{N}(6.49, 0.3)$.
- Draw N random numbers $X_2$ from $\mathcal{N}(0, 0.1)$.
- $X_1 + X_2$ is the simulation of your measured spectrum.

**Example 3.3**

Let $X \sim \text{Norm}(\mu, \sigma^2)$. The distribution of $Y = e^X$ is called *lognormal distribution.*

a) Determine the density (PDF) of the distribution.

b) Calculate mean, variance, median, and mode of density.

c) Show that the product of two lognormally distributed random variables is again lognormally distributed.

**Example 3.4 (Prog)**

The so-called *inversion method* can be used to generate random numbers following a specific distribution from equally distributed random numbers (usually in the interval 0-1). This concept is introduced step by step in the following example with reference to the article arXiv:2003.09172.
Already during the Mariner 2 mission in the 1960s it was found that solar winds in this region of our solar system are composed of two main components: a fast component originating from coronal holes and a slower component of still mostly unknown origin. In arXiv:2003.09172 these two components are described by a bi-Gaussian distribution of the following form:

$$f(x) = h_1 \exp\left(-\frac{(x-p_1)^2}{2w_1^2}\right) + h_2 \exp\left(-\frac{(x-p_2)^2}{2w_2^2}\right)$$

In the case of the proton velocity $x = v_p$ of the solar wind the following values for the parameters were determined:

| | $h_1$ | $p_1$ | $w_1$ | $h_2$ | $p_2$ | $w_2$ |
|---|---|---|---|---|---|---|
| values | 3.00 | 373 km/s | 57 km/s | 1.09 | 510 km/s | 103 km/s |

a) Construct a probability density (PDF, $\rho(v_p)$) and then calculate the cumulative density (CDF) which is given as $F(v_p) = \int_0^{v_p} \rho(v_p')dv_p'$. (Note: The lower limit of integration is 0, since negative velocities make no sense and thus $\rho(v_p < 0) = 0$. Also use the approximation $\mathrm{erf}(x) = 1$ for $x > 3$). Then plot $\rho(v_p)$ and $F(v_p)$ in the range (200-800) km/s.

b) Invert the CDF ($F^{-1}$) numerically (hints in the skeleton). Now draw 10000 random numbers in the interval (0-1) and apply the CDF$^{-1}$ to these (this might take some time). Histogram the result - it should have the same shape as the original distribution.

**Example 3.5 (Prog)**

Compulsory for TU (142.351), optional for University (260014)!
Continuation of Example 3.4:
The inversion method can also be used to generate random numbers which follow an arbitrary binned distribution.

c) Use the histogram created in subtask 3.5 b) to create a CDF and simulate 10000 events from it.

d) Briefly discuss the advantages and disadvantages of *simulating* an analytical function and a histogram.

e) Fit the histograms created in subtasks b) and c) with the function $f(v_p)$ in $h_1$ und $h_2$ (i.e. $p_1$, $p_2$, $w_1$ and $w_2$ are fixed in the fit) and compare the result with your expectation. Note: fits can be made in Python easily using e.g. curve_fit().

**Example 3.6 (Prog)**

We want to generate random uniformly distributed points on the unit sphere. A tempting way to do this is to generate a uniform distribution of the two spherical coordinates $\theta \in [0, 2\pi)$ and $\varphi \in [0, \pi]$.

a) Generate $N$ points with coordinates $(r = 1, \theta, \varphi)$ in the way described above. Transform them to Cartesian coordinates and plot them in 3D (help in the skeleton). What do you observe for increasing $N$?

b) The method from a) is not working, why? How is this related to the area element (solid angle) $d\Omega = \sin(\varphi)\, d\theta\, d\varphi$?
Use the probability that a point lies within an infinitesimal area element $P(\Omega)d\Omega$ to find the joint distribution $P(\theta, \varphi)$. Then calculate the marginal distributions $P(\theta)$, $P(\varphi)$ and use the inverse sampling method from 3.4 (you can do all steps analytically this time) to correctly sample values for $\theta$ and $\varphi$. Plot your results as in a).