Grades for assignment #1 will be released later today or tomorrow.

IMPORTANT:

- Some folks did not consider the uncertainty given in question 2! Remember that in data analysis we cannot simply ignore meaningful information. And even if a piece of data is not being used, an explanation needs to be given. **(Marks deducted)**

- Some folks did not plot or provide a visual inspection of the fit in question 2! In all curve fitting demos, we always provide a (first) inspection of the fit with a visual of the model across the data points. **(Marks deducted)** *"But you did not ask us to plot the fit"*…

That is the whole point of assignments at the graduate level; not everything will be super-prescribed or detailed (you will need to show a significant leap in solving the problem). You are entering or are doing research already. In exploratory research, you will have to solve problems with partial to minimum information. Developing independence, developing your own questions, and going beyond the "basics" are key in graduate studies.

Simulation of random variables

Any random variable can be generated using

$$X = F^{-1}(U)$$

where F^{-1} represents the inverse of the cumulative distribution associated with the target variable X, and U represents a uniform random variable with $U \in [0,1]$.

Method: *N* independent samples u_i are drawn uniformly and they are transformed into x_i according to the equation above. The transformed set of x_i values is a random sample set describing the variable of interest.

This method relies on the availability of $F(\dots)$ and its inverse!

Simulation of an exponential distribution

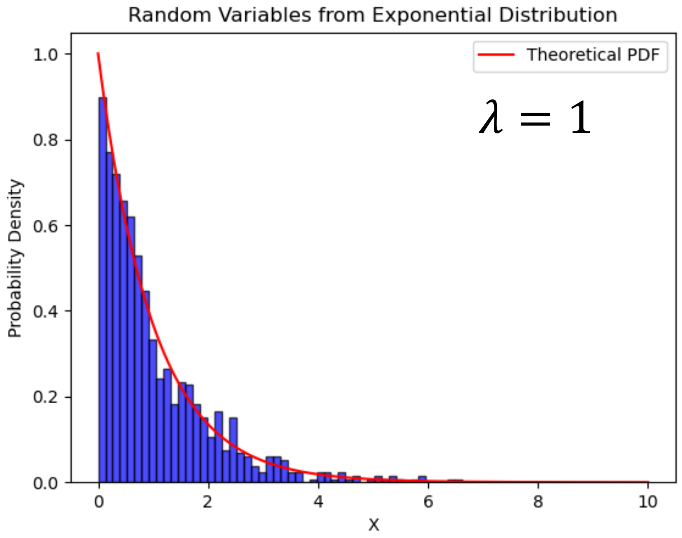
The cumulative distribution of an exponential distribution with parameter λ is

$$F(x) = 1 - e^{-\lambda x}$$

with $x \geq 0$ representing possible values of the exponential variable. Its inverse, also known as the quantile function, is

$$x = -\frac{\ln(1-u)}{\lambda}$$

where $0 \leq u \leq 1$ represents possible values of the standard uniform distribution. As an example, using 1000 drawn random values u_i yield the sample distribution function for the exponential distributed variable X



$PDF = \lambda e^{-\lambda x}$

Simulation of Gaussian variable

The cumulative distribution of the standard normal distribution is

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

which cannot be inverted analytically! This a special function (like an error function) which can be computed numerically with Maclaurin series. But with a simple change of variables, we can simplify this problem as shown below.

Consider a pair of variables (X, Y) and the functions $U = u(X, Y)$ and $V = v(X, Y)$ that transform them to the pair of variables (U, V).

$$g(u, v) = h(x, y)|J|$$

in which $|J|$ is the determinant of the Jacobian

$$J = \left[\frac{\partial(x, y)}{\partial(u, v)}\right] = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$

Simulation of Gaussian variable

Consider two random variables X and Y distributed as standard Gaussians as

$$h(x, y) = \frac{1}{2\pi} e^{-\left(x^2+y^2\right)/2}$$

Consider a transformation from Cartesian to polar coordinates

$$x = r\cos\theta$$
$$y = r\sin\theta$$

and the Jacobian of the transformation is

$$J = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

with determinant $|J| = r$. Then the distribution of (r, θ) is

$$g(r, \theta) = \frac{1}{2\pi} r e^{-r^2/2}$$

with $r \geq 0$, $0 \leq \theta \leq 2\pi$.

Simulation of Gaussian variable

The joint distribution $g(r,\theta)$ can be written as the product of two functions

$$g(r,\theta) = g_1(r)g_2(\theta)$$

with $g_1(r) = re^{-r^2/2}$ known as the *Rayleigh distribution* and $g_2(\theta) = 1/2\pi$ which is the uniform distribution for the angle θ. **These two distributions have a closed analytic form for their cumulative distributions** so the random variables R and Θ can be simulated to draw a pair of independent standard Gaussians.

Starting with the Rayleigh distribution, that has a CDF as

$$G_1(r) = 1 - e^{-r^2/2}$$

Its inverse will provide the quantile function as

$$r = \sqrt{-2\ln(1-u)} = G_1^{-1}(u)$$

This result shows that $R = \sqrt{-2\ln(1-U)} = G_1^{-1}(U)$ simulates a Rayleigh distribution from a standard uniform variable $U \in [0,1]$.

Simulation of Gaussian variable

For the angle, its CDF is

$$G_2(\theta) = \begin{cases} \theta/2\pi & 0 \leq \theta \leq 2\pi \\ 0 & \text{otherwise} \end{cases}$$

Its inverse will provide the quantile function as

$$\theta = 2\pi v = G_2^{-1}(v)$$

This result shows that $\Theta = 2\pi V = G_2^{-1}(V)$ simulates a uniform distribution from a standard uniform variable $V \in [0,1]$.

Therefore, the use of two independent uniform distributions U and V can be used to simulate a Rayleigh and a uniform angular distribution according to

$$\begin{cases} R = \sqrt{-2\ln(1-U)} \\ \Theta = 2\pi V \end{cases}$$

Simulation of Gaussian variable

Therefore, the use of two independent uniform distributions U and V can be used to simulate a Rayleigh and a uniform angular distribution according to

$$\begin{cases} R = \sqrt{-2\ln(1-U)} \\ \Theta = 2\pi V \end{cases}$$

Using the Cartesian to polar transformation, we get

$$\begin{cases} X = R\cos\Theta = \sqrt{-2\ln(1-U)}\ \cos(2\pi V) \\ Y = R\sin\Theta = \sqrt{-2\ln(1-U)}\ \sin(2\pi V) \end{cases}$$

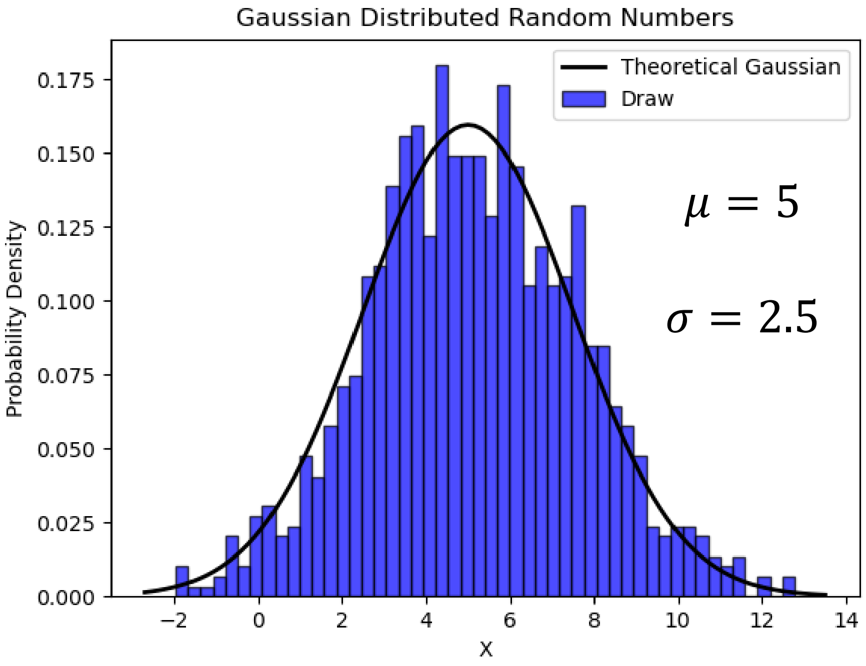
The equations above can be easily implemented using two independent standard uniform variables drawn between 0 and 1. The equations above are for a standard Gaussian simulation, but they can be further transformed to Gaussians of any mean and variance via simple rescalings. For example, a Gaussian X' of mean μ and variance σ^2 is related to the standard Gaussian X by the transformation

$$X = \frac{X' - \mu}{\sigma}$$

Simulation of Gaussian variable

Therefore, X' can be simulated as

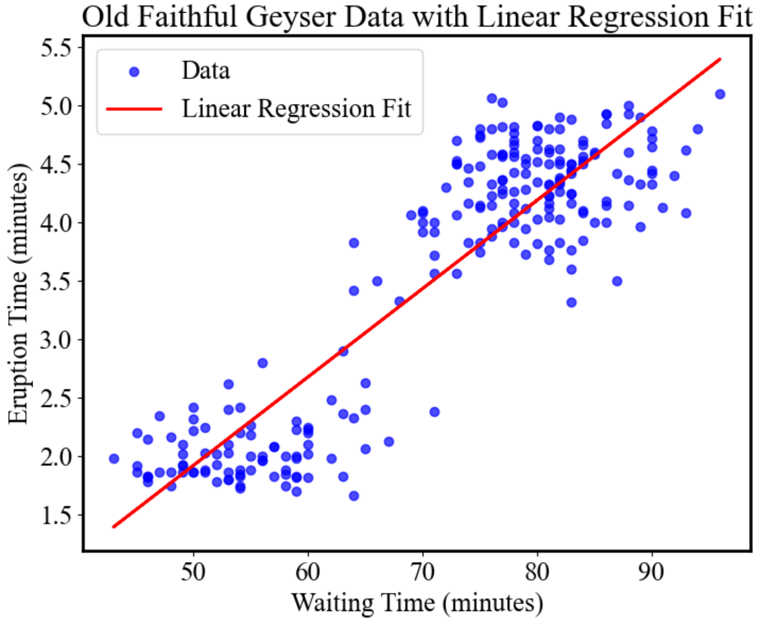
$$X' = \sqrt{-2\ln(1-U)}\ \cos(2\pi V)\,\sigma + \mu$$



Re-sampling methods

Re-sampling methods are useful to estimate best-fit parameters and their uncertainties in the fit when the best-fit values and their uncertainties cannot be estimated with adequate accuracy. Moreover, certain methods of estimation relying on statistics may result in estimators that are *biased*.

A way of removing bias in the estimators, while at the same time providing uncertainties on the estimate, is to **re-sample** the original measurements or data, for example, by using a *subset* of the data or using randomly drawn samples from the original data.



Old Faithful Geyser Data

Description: waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

Format: a data frame with 272 observations on 2 variables.
1. eruptions (numeric) Eruption time in mins
2. Waiting (numeric) Waiting time to next eruption in mins

Härdle, W. (1991). Smoothing Techniques with Implementation in S. New York: Springer.

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. Applied Statistics, 39, 357–365.