Math 525 – Statistics I, Fall 2020

Assignment 1

Due: Sunday, September 13, by 8:00PM

Problem 1: With this problem, you will improve your coding and data visualization skills.

(i) Develop a function that takes as input the radius of a sphere and returns its surface area. Similarly, develop a function that takes as input the radius of a sphere and returns its volume.

(ii) Use the functions you developed for (i) to develop a new function that takes as input the radius of a sphere and returns both its surface area and volume.

(iii) In a single figure with multiple panels, demonstrate how the surface area and volume of a sphere changes as well as how a sphere's surface area and volume are related to each other when the sphere's radius varies between 0 and 1 μm. The resulting figure should be graphically informative and scientifically accurate.

(iv) Briefly describe how you could modify the source code you developed for (i) and (ii) so they apply to cubes instead of spheres.

(v) The dataset provided captures experimentally estimated surface areas and volumes of individual *E. coli* cells. Based on these data, argue on whether *E. coli* cells attain a spherical or cubical shape. Explain in detail your reasoning.

Associated data: For step (v) use `e_coli_volume_area.mat`. This dataset contains surface area and volume measurements. The measurements are reported in units of μm^2 and μm^3, respectively.

Problem 2: With this problem, you will improve your simulation skills and test your understanding of random variables.

(i) Develop a function that applies the fundamental theorem of simulation and simulates draws from Categorical$_{\sigma_1,\cdots,\sigma_M}(\pi_{\sigma_1},\cdots,\pi_{\sigma_M})$. This function, should be able to generate variables with any number of categories M. To generate uniform random variables use `rand`.

(ii) Develop a function that applies the fundamental theorem of simulation and simulates draws from Cauchy(μ,σ). This distribution is supported over the entire real line and has a probability density given by

$$p(x) = \frac{1}{\pi\sigma} \frac{1}{1 + \left(\frac{x-\mu}{\sigma}\right)^2}$$

To generate uniform random variables use `rand`.

(iii) Use simulations to verify that the functions you developed in (i) and (ii) simulate variables with the correct statistics.

Problem 3: In this problem, you will investigate the Box-Muller algorithm.

The Box-Muller algorithm proceeds as following:

- Generate u and v by drawing uniform random variables over the interval $[0, 1]$.
- Set $x = \mu + \sigma\sqrt{-2\log u}\,\cos(2\pi v)$ and $y = \mu + \sigma\sqrt{-2\log u}\,\sin(2\pi v)$.

Upon completion, the values of x and y each follow a Normal(μ, σ^2) distribution.

(i) Show mathematically that the Box-Muller algorithm indeed produces values with the correct statistics.

(ii) Develop a function that takes as input μ, σ and implements the Box-Muller algorithm. To generate uniform random variables use `rand`.

(iii) Use simulations to verify that the function you developed in (ii) simulates variables with the correct statistics.

Problem 4: In this problem, you will use the maximum likelihood principle to analyze experimental data.

Spectroscopic experiments, held in TTTR mode, proceed as following: Pulses of light are sent to a chemical or biological sample. With each pulse, a molecule within the sample becomes excited. Following a short period of time, the excited molecule emits a photon which is collected and detected with appropriate equipment. In such an experiment, the measurements consist of the time elapsed between a pulse and the photon detection. Each measurement w encodes the time d that the molecule remained excited and some error r caused by the detection hardware. Both contributions are additive, i.e. each measurement is given by the sum $w = d + r$. By fundamental laws of physics, it is known that d is an exponential random variable with a rate λ characteristic of the sample under investigation. By detector engineering, it is ensured that the error r is normally distributed around zero with some variance v characteristic of each detector.

(i) The dataset `TTTR_calibration.mat` contains calibration measurements of two different detectors. These measurements are obtained under artificial conditions that ensure $d \approx 0$. Accordingly, they encode only the error of the detectors. Use these measurements to estimate the error variances of the two detectors. To do so, apply the maximum likelihood principle and carry out the calculations involved analytically.

(ii) Derive an analytic expression for the probability density $p(w)$ of each individual measurement of an actual experiment.

(iii) Given that the detectors are already calibrated from step (i), describe in detail the remaining steps for the estimation of λ from the measurements of an actual experiment.

(iv) Implement the method you developed in (iii) and use the measurements in the provided dataset `TTTR_experiment.mat` to find the value of λ. To carry out the involved optimization, you might use `fminsearch` or any other optimization strategy you prefer.

(v) Describe in detail an alternative method that relies on the maximum likelihood principle and estimates the values of v and λ using calibration and actual measurements simultaneously. In other words, describe a method that uses a single likelihood function for all parameters and all datasets. You do not need to implement this method.

Associated data: The datasets `TTTR_calibration.mat` and `TTTR_experiment.mat`, for steps (i) and (iv) contain measurements from the same two detectors. The detector are labeled with A and B. All measurements are reported in units of ns.