Simulation (Computer Exercise)

1. Assume that the daily temperature in Cairo in Summer time is normally distributed with $\mu = 37$ and $\sigma = 2$.

   - Using mathematics, find the probability that, in a particular day, the temperature exceeds 42?
   - Using simulation, answer the question above by generating 10^5 observations of that random variable and find the fraction of them exceeding the value of 42. Repeat the experiment 10 times with different samples of the same size; is there any change in the answer? Why?
   - Generate 10 observations of the same random variable and plot their values on a horizontal line; this gives you the feeling of how separated they are. plot their histogram as well and compare to the pdf of the random variable. Repeat the whole procedure 4 times, every time with multiplying the sample size by 10, i.e., generate 100 observations for the second time and 1000 for the third time,...etc. This example gives you the feeling of two things. First, the dense region of a pdf is translated to many observations in sampling. Second, the more observations you have the closer you are to the "true" pdf.

2. Assume that $X \sim N(0,1)$. Find and plot the pdf of Y, μ_Y, and σ_Y^2 where $Y = X^2$ (a χ^2 distribution). Re-solve the problem using simulation techniques, and plot the histogram of Y instead of the pdf. Compare the pdf and the histogram.

3. We will start this problem with the following quotation:

   > The term *regression* has an interesting history, dating back to the work of Sir Francis Galton in the 1800s. Galton investigated the relationship between heights of fathers and heights of sons. He found, not surprisingly, that tall fathers tend to have tall sons and short fathers tend to have short sons. However, he also found that very tall fathers tend to have short sons and very short fathers tend to have taller sons. (Think about it—it makes sense.) Galton called this phenomenon *regression toward the mean* (employing the usual meaning of *regression*, "to go back"), and from this usage we get the present use of the word *regression*.[1]

   In regression we get the data and try to build the model that expresses this data set. In this problem we will do the opposite for better understanding. We assume that the height of a son at the age 10 is proportional to the height of his father by a factor of 0.7. Of course, this is in average, i.e., if we have different sons for the same father it is unlikely to have their heights equal at the age of 10. Hence, the model of this problem can be

   $$Y = 0.7X + \varepsilon, \tag{1}$$

   where ε is a random variable. Then, $\mathrm{E}[Y|X = x] = 0.7x$. For the time being assume that $\varepsilon \sim N(0,1)$. Generate and plot a data set that satisfies the following. The height of fathers, X, ranges from 150 cm to 190 cm in steps of 0.4 cm (this generates 100 different values for X). For each value of X generate the heights of 100 sons. Look at the data; you will see the pattern (or the trend).

   Notice: in this problem the conditional distribution of Y given X is normal with $\mu_{Y|X} = 0.7x$ and $\sigma^2_{Y|X} = 1$. However, if X is a random variable, where X and ε are independent, then $\sigma_Y^2 = 1 + 0.7^2\ \sigma_X^2$ (why?)

4. In the previous problem we have generated 100 values for X uniformly from 150 to 190, in steps of 0.4. However, the actual distribution of X, the hight of fathers, is almost normal with $\mu_X = 170$, and $\sigma_X^2 = 1.96$. Then, from (3), it can be shown that the joint distribution of X and Y is binormal (try to prove it).

   - Now, generate 100 observations from the distribution of X, and for each value generate 100 observations from the conditional distribution of Y; plot the 10,000 observations and compare to the data set above.
   - What is the covariance matrix and the mean vector of the vector $(X, Y)'$? Directly, generate 10,000 observations from that binormal distribution and compare to the two data sets above. The patterns of the two data sets of this problem should agree.

[1] Quoted from Casella, G. and Berger, R. L. (2002), *Statistical inference*, Duxbury advanced series, Australia ; Pacific Grove, CA: Duxbury/Thomson Learning, 2nd ed

- For fathers and sons living in Asia—where people are shorter than anywhere else—$\mu_X = 150$, and $\sigma_X^2 = 1.5$. Generate and plot 10,000 observations for Asians and plot this data set in different color, on the same figure, along with the same data set above. Watch the effect of the mean and the variance on the pattern. Using free hand, and your sense as well, try to draw a curve that best separates the two populations.