Statistics Project

Haiyan Cheng
Department of Computer Science
Willamette University
Salem, OR 97301
hcheng@willamette.edu

CS-435 Computational Science and Applications

Fall 2016

> *If you torture the data long enough, it will confess.*
>
> — RONALD COASE

1 Introduction

In this project you will learn how to do the following things in MATLAB :

1. Import data into MATLAB.
2. Reorganize data and assign them to named variables.
3. Perform simple statistical analysis on the data.
4. Generate a scatter plot and a histogram.
5. Display and save results.

You will learn those MATLAB programming techniques in the context of a simple statistical analysis project: Compute mean and standard deviation for specific groups of data, and generate a plot of the data.

To get you started, I will introduce the basic concepts of sample mean, sample variance, sample standard deviation, a commonly used probability distribution called normal distribution, and the concept of a histogram.

1.1 Measure of location

Sample mean describes a central tendency of a group of sampled data. In mathematics and statistics, it is the arithmetic mean, and often denoted by a bar, for example $\overline{x}$ is the mean of a set of observational data x. For a statistical sample $x = \{x_1, x_2, \ldots, x_n\}$, the sample mean for x is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \tag{1}$$

There are also several alternatives to the arithmetic mean. The *median* is the midpoint of a group of data. To calculate median, we first sort the data in ascending order, if the number of the data is odd, the middle value is the median, if the number of the data is even, the arithmetic mean of the middle two values is the median. Sometimes, the median is also called "the 50th percentile."

The *mode* is the value that occurs most frequently.

Here is an example: for $x = \{2, 4, 1, 5, 8, 10, 10\}$, the mean is

$$\frac{2+4+1+5+8+10+10}{7} = 5.7143$$

The median of x is 5, since it separates the seven numbers into two halves, three less than five and three greater than five.

The mode is 10, since it occurs more often than any other number.

In figure 1, these three quantities: mode, median and mean are represented with the colors: red, green, and blue, respectively. The figure shows that for a non-symmetric distribution, these values may not coincide with each other.

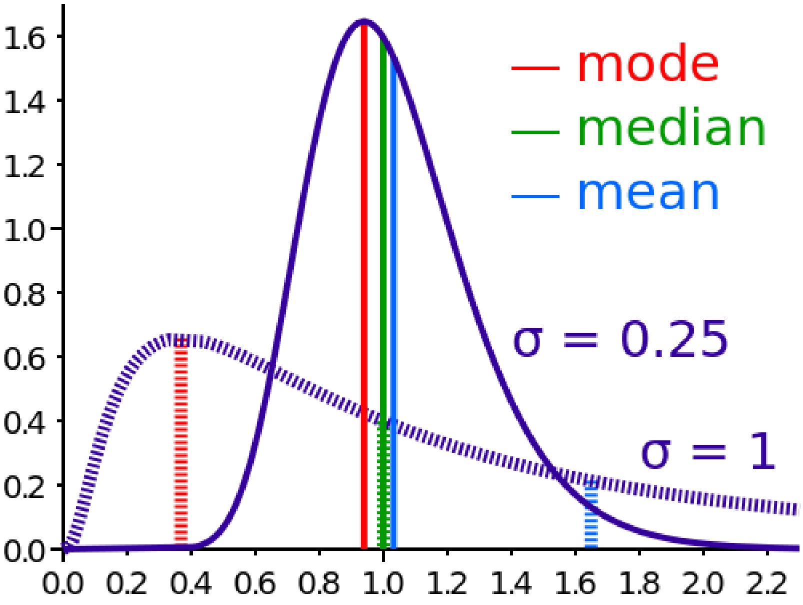

Figure 1. Comparison of mean, median and mode of two log-normal distributions with different standard deviation, from Wikipedia [1].

1.2 Measure of spread

The simplest measure of spread is the *range*, computed by the difference of the largest and the smallest value. Range is easy to calculate, but is not a reliable method for spread measurement. It is highly sensitive to the sample size and it is very sensitive to extreme values.

The concept of variance describes how far the sample is spread. Or, in the other word, how far the sample data is spread around the mean. If the individual samples are spread out widely around the mean, the variance will be large.

Since the deviation from the mean can be in either direction, greater than or less than the mean, we sum up the square of the distance from the mean and divided by the total number of samples minus one. For example, the variance of the above data example $x = \{2, 4, 1, 5, 8, 10, 10\}$ can be computed as

$$\begin{aligned}\text{Unbiased variance of } x &= \frac{1}{7-1} \times \Big[(2-5.7143)^2 + (4-5.7143)^2 + (1-5.7143)^2 \\ &+ (5-5.7143)^2 + (8-5.7143)^2 + (10-5.7143)^2 + (10-5.7143)^2\Big] \\ &= 13.5714\end{aligned} \tag{2}$$

In general, for a set of statistical sample $x = \{x_1, x_2, \ldots, x_n\}$, the unbiased sample variance is defined as:

$$\begin{aligned}\sigma^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}\end{aligned} \tag{3}$$

Note that we divide by $n-1$, this term $n-1$ is often called *the degree of freedom*. The name is derived from the fact that the sum $\sum_{i=1}^{n} \overline{x} - x_i$ is zero. So if $\overline{x}$ is known and $n-1$ of the values are specified, the remaining value is fixed. Thus, only $n-1$ of the values are said to be freely determined. Another reason we divide by $n-1$ is because it does not make sense to define spread of a single data point when $n = 1$.

The *standard deviation* is defined as the square root of the variance.

$$\begin{aligned}\text{Standard deviation of } x &= \sqrt{13.571} \\ &= 3.6839\end{aligned} \tag{4}$$

1.3 Data distribution and histogram

Often times, we are interested in how the sample data is distributed. The histogram is a graph tool that shows the count of the data point falling in various ranges, or how frequently the data shows up in a specific range.

We can graph groups of number according to how often they appear. For example, $x = \{1, 2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6\}$, we can graph them as shown in figure 2:

This example is a very simple case. In real data set, almost all numbers will be unique, for example: $\{3, 11, 12, 19, 22, 23, 24, 25, 27, 29, 35, 36, 37, 45, 49\}$, in that case we *bin* the data into different ranges: $0-10, 10-20, 20-30, 30-40, 40-50$, and count the occurrence to plot.

A histogram can be used to approximate a probability distribution. The most common probability distribution is the normal distribution, sometimes called Gaussian distribution in honor of Carl Friedrich Gauss, a famous mathematician. The normal distribution is a bell-shaped curve with the most likely event occurs in the center and the data has less tendency to produce unusually extreme values. A normal distribution is completely described by its mean and standard deviation. Notice that since the normal distribution is symmetric, its mean, mode and median value are the same.

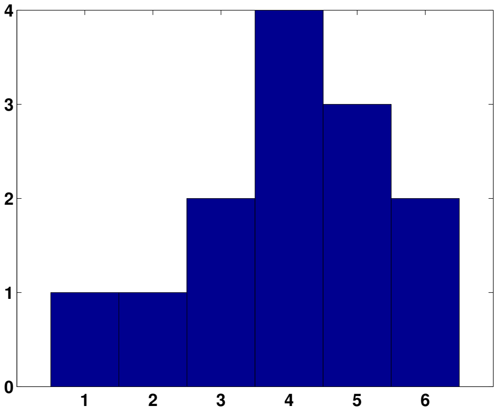

Figure 2. A simple histogram.

Figure 3 shows four different normal distributions that are completely described by their respective means and variances. The red bell-shaped normal distribution is called the standard normal with mean 0 and variance 1.

With normal distribution, three standard deviations account for 99.7% of the total samples. The standard deviation is often used as a measure of uncertainty in many subjects. We will investigate how risk is reflected in the standard deviation in finance applications in the next project.

The normal distribution is described by the following function:

$$y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

2 In-class project

2.1 Basic statistical analysis

The normally accepted human body temperature is $98.6 \pm 0.9°F$, but is it true population mean is really $98.6°F$? Recent medical research has discovered that the mean normal temperature is really $98.2°F$ degrees! A data sample is taken by measuring a group of male and a group of female's body temperature and their heart rates. This information is stored in the data file `normtemp.dat`. The first column contains body temperature (degree Fahrenheit), the second column indicates gender (1 = male, 2 = female), and the third column is the heart rate (beats per minutes).

Write a MATLAB program to do the following:

1. Read the data file into MATLAB.
2. Extract and organize the required information by assigning them to named variables.
3. Calculate the sample mean and standard deviation for body temperature and heart rate.
4. Display the corresponding information in the console and write the output to a file.
5. Calculate the sample mean and standard deviation for males' and females' body temperature and heart rate.

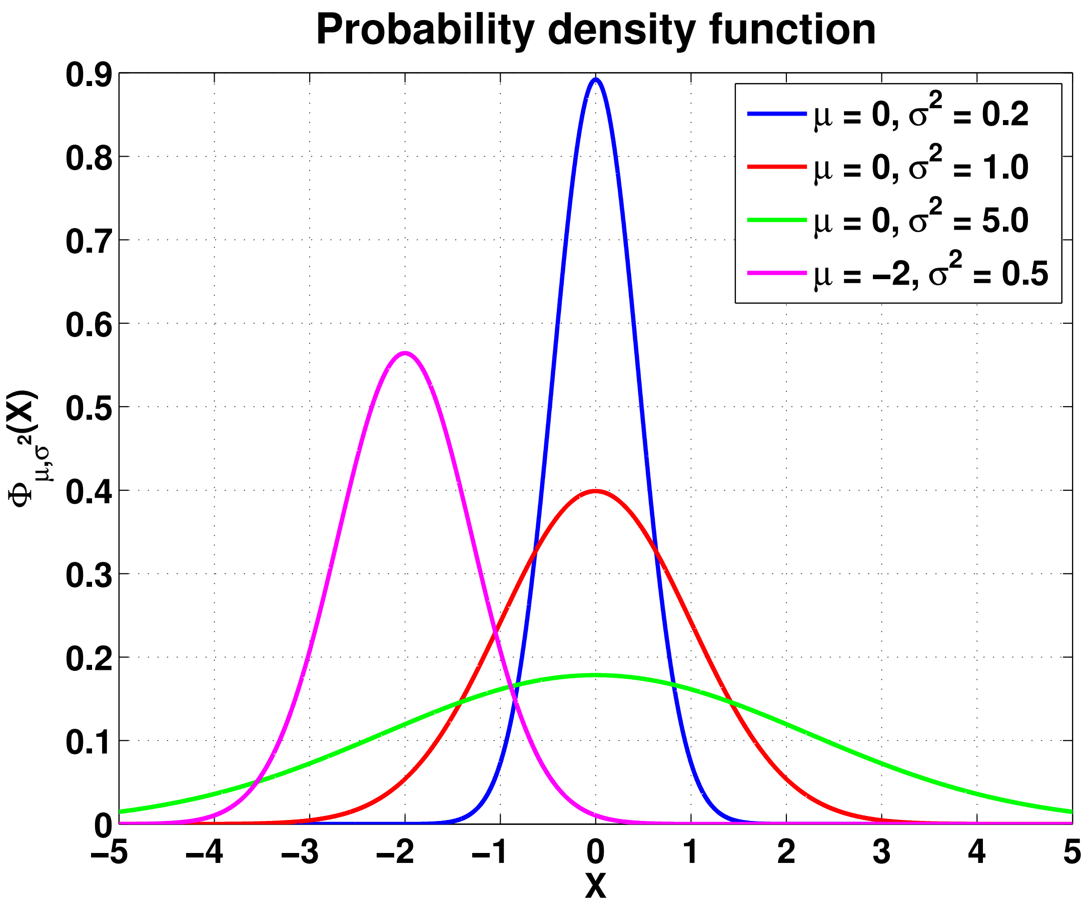

Figure 3. Four normal distributions with red being the standard normal distribution.

6. Generate a scatter plot and histogram of the body temperature and heart rate.

Before we start to deal with 2-dimensional data, let's work on a simpler problem first. The data file `bodytemp.dat` contains only one column of data, 65 male and 65 female body temperature. We'll perform the following with MATLAB :

1. Use load command to load the data into MATLAB.
2. Calculate the length of the data.
3. Generate a scatter plot
4. Calculate the mean body temperature
5. Calculate the variance of the body temperature

2.2 MATLAB plotting

MATLAB has a powerful plotting function. We will explore some of the plotting techniques by replicating figure 3. In this part, you will explore user-defined functions, MATLAB built-in functions, and various ways to customize your plot.

Your task is to replicate figure 3 using the normal PDF formula 5 for different means and variances.

3 Lab project

3.1 Statistical analysis

Researchers for Consumer Reports analyzed three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). The calories and sodium contents are listed for each given type. The information is stored in a data file `hotdog.dat` with three columns.

The first column uses three numbers to represent different kinds of hotdogs (1 = beef, 2 = meat, 3 = poultry), second column lists the calories for each hotdog, and the third column is the sodium (mg) content.

Decide which meat type contain the most calorie, the most sodium, the least calorie and least sodium. Decide the overall mean calorie, standard deviation of calorie, mean sodium and standard deviation of sodium. Display all the information on the console.

Write a MATLAB program to do the following:

1. Load the data file `hotdog.dat` into MATLAB.
2. Extract required information and assign them to named variables.
3. For each kind of meat, calculate the mean, variances, standard deviation for calorie and sodium respectively.
4. Find out which kind of hotdog has the most and the least amount of calorie and sodium. Display the result in the console with text information and mean values. Write the result into a file `Project1_result.txt`.
5. Plot the histogram for calories contained in the meat hotdogs.
6. Plot the scatter plot of the sodium contained in the beef hotdogs i(f possible, use plot features to make your plot more visually attractive instead of using the default plot options.)

3.2 Explore and plot lognormal distribution

Write a MATLAB program to plot the lognormal distribution for different μ and σ values, try your best to replicate figure 4, explore how to add texts to the plot and use Latex format for labels. The log-normal probability density function is describe by (6):

$$y(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}} \tag{6}$$

4 Submission method and grading criteria

The 3 Lab project is the part you need to submit for grading. If there're sub problems, please create folders for each problem. If there are source data required to run your program, you need to put the data under the same folder, so that as soon as I unzip your code, I can test your code.

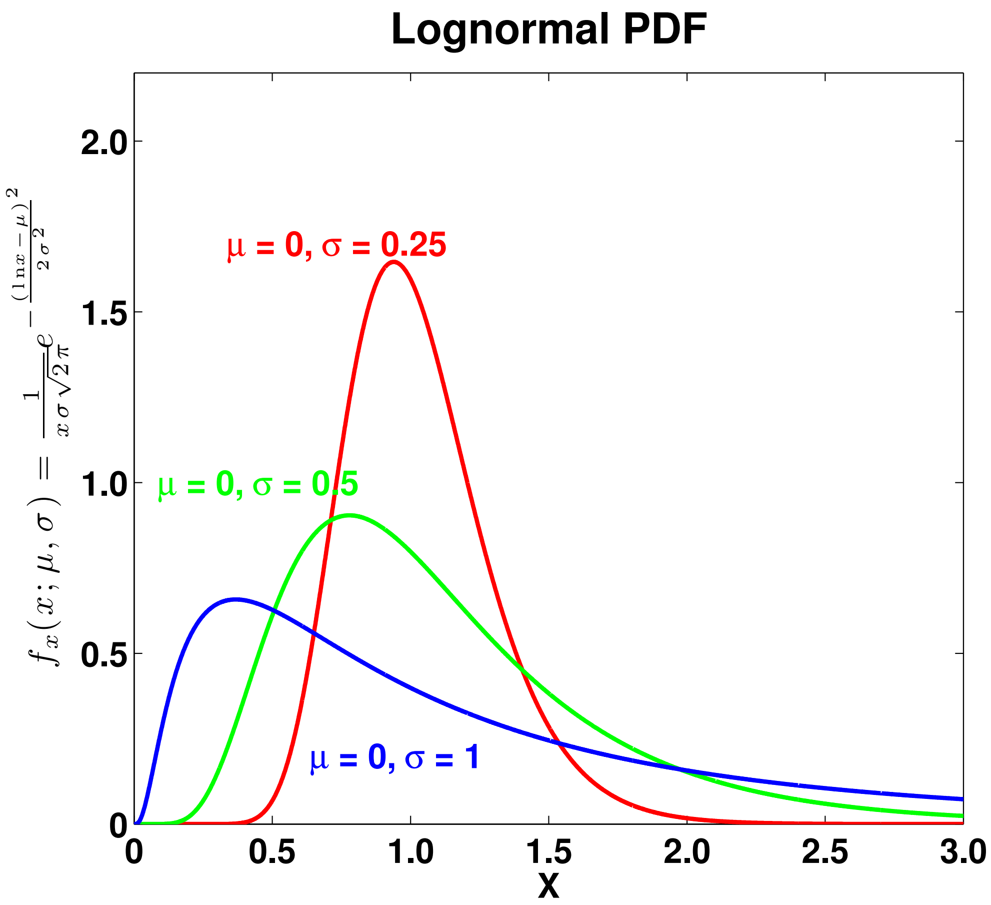

Figure 4. Three log-normal distributions.

Please zip all the folders that contain programs you want to submit and name it `YourFirstname_P1.zip` to the WISE dropbox before 11:59PM on Sep 14.

Your submission will be graded based on the following criteria:

- Your program runs without any error and requirements of additional data.
- Your program generates expected results.
- Your programs are property documented.
- You have answered all the questions as described in the lab project, and divided your program into different sections.

References

[1] Wikipedia. Arithmetic mean. Wikipedia, the free encyclopedia, 2012. Available at `http://en.wikipedia.org/wiki/Arithmetic_mean`.