IB 120/201 - Lab 4

R Basics & Probability Distributions

Due Date: February 19, 2021 by 12:59pm PST

University of California, Berkeley *GSI: Ksenia Arzumanova*

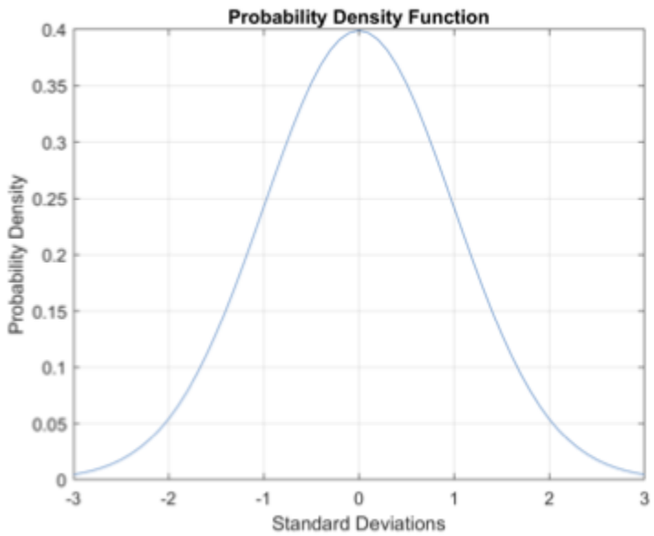
Background

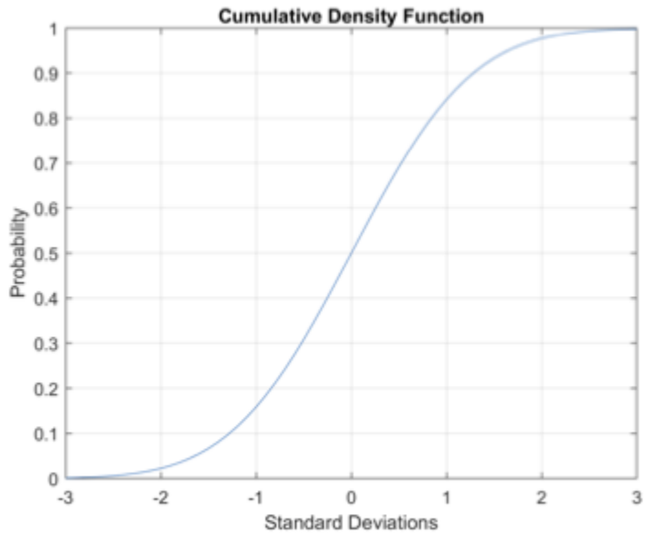
Random Variables

Informally, a random variable is a variable whose values depend on outcomes of a random event. The domain of a random variable is called the *sample space*, or the set of possible outcomes of event (such as heads or tails in a coin toss). Random variables have *probability distributions* associated with its sample space. Random variables can be discrete (finite or countable set of values), the probability distribution of which can be described by a probability mass function (PMF), or continuous (all values in an interval), the probability distribution of which can be described by a probability density function (PDF).

Probability Distributions

A probability distribution is a function which gives the probability of occurrence of an event confined within one experiment. The most famous and widely used probability distribution is the normal (or Gaussian) distribution. The central limit theorem states that the sum of independent random variables will resemble a normal distribution as the number of samples increases. For example, the distribution of students' heights at UC Berkeley resembles a bell curve, which is the normal distribution. A cumulative distribution function (CDF) shows the probability that a random variable will take a value less than or equal to x, which in the graphs below have units of standard deviations. Recall that standard deviations are a measure of the amount of variation in the data.





Assignment

1. Create a new script in RStudio (File -¿ New File -¿ R Script).
2. Draw 100 random samples from the normal distribution with a mean of 50, and standard deviation of 5. Plot the histogram of these samples.
3. Using a line of code, convince yourself that 50% of the density is to the left of the mean above.
4. Repeat part 2, but with standard deviations of 10, and then 100. Describe, in comments within the script, how the spread of the data (or domain/range) change with standard deviation.
5. Now draw again 100 random samples from the normal distribution with a mean of 50, and standard deviation of 5. Calculate the densities/probabilities of these random samples.
6. Plot the samples and densities calculated in part 4. This should look similar to the first figure in the Background section.
7. Calculate the *cumulative* densities of the same 100 samples from part 4. Plot the samples and their cumulative densities. This should look similar to the second figure in the Background section.
8. Under your plot function from part 7, type in the following code:

$$abline(h = 0.5, v = 50)$$

This should have created grid lines on your plot. In trying to understand the relationship between PDFs and CDFs, how does the point in the graph outlined by the gridlines correspond to your answer in part 3. Answer in comments within the script.

Make sure to save your script. It will save your script to the directory you choose within your JupyterHub. When you're done, manually go to that directory, find your .R file, check the box next to it, and up above under the 'Files' tab, click 'Download'. Upload your file to bCourses.