PH227: AI and Data Science

Tutorial Sheet 5: Python/Data-Science Part

Note: This tutorial sheet contains programming problems related to hypothesis testing discussed in the lectures. You have to submit the codes corresponding to the starred (*) problems to the TAs. The marks of the starred problems are indicated next to them.

1. * Consider the database file weight-height.csv which has three columns: (a) the first column denotes the gender (Male or Female), (b) the second column contains the height (in inches), and (c) the third one has the weight in the units of lbs (pounds). The first 5000 entries correspond to males, and the next 5000 for females; therefore the csv file has 10000 rows, excluding the header. Write a Python program which will:

   (a) based on the user input, make a histogram plot of heights of the males and females (5000 data points each) separately, and also of the combined data (10000 data points)

   (b) repeat the same for the weight

   (c) Calculate the standard deviation (σ) and mean (μ) for both height and weight separately for males and females, and also for the joint data. [4 marks]

2. * Write a Python program for performing z-test on the previous data. For this purpose:

   (a) randomly draw a sample of 40 weights from the male list and compute the sample mean $\bar{x}$, and the z-score

   (b) ask the user for the α value

   (c) perform left, right, and two-tailed z-tests on the data to ascertain whether or not H_0 stays valid

   (d) repeat the same procedure for females and the joint data

   (e) repeat (a)–(d) for the height data [8 marks]

3. * Pretend that for the population of problem 1, you don't know the population standard deviation σ. Therefore, by drawing n random samples (let n be user defined), calculate the sample mean $\bar{x}$, and the sample standard deviation s. Now repeat all the steps of problem 3, and perform t tests, instead of the z test. How do the results depend on the sample size n? [8 marks]