Practical Session 4 : two sample testing

Two sample testing for proportion

Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a a simple random sample of 100 women and 200 men from a population of 100,000 volunteers.

At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold.

1. State the null hypothesis H_0 and the alternative one H_A
2. Perform a z test to answer this question. Conclude

Two sample testing on the iris dataset

The data used in this example is from Kaggle.com and was posted by the user Web IR. The link to the data set is here `https://www.kaggle.com/webirlab/iris-data/data`. The data set contains the sepal and petal length and width of various floral species. We will be testing to see if there is a significant difference in the sepal width between the species Iris-setosa and Iris-versicolor which are variables "sepal_width" and "species" respectively.

1. Import the data using `read_csv`. Comment the following command line
   `df.groupby("species")['sepal_width'].describe()`
2. Create 2 data frames that are subsets of the original data where each data frame only contains data for a respective flower species
   ```
   setosa = df[(df['species'] == 'Iris-setosa')]
   versicolor = df[(df['species'] == 'Iris-versicolor')]
   ```
3. We now want to perform a t-test
   (i) Test equality of variances using the function `levene` of the library `scipy.stats`. Can we accept the hypothesis of equality of variances of the two samples?
   (ii) Is the sepal width a Gaussian variable. Display a QQ-plot using `probplot` of the library `scipy.stats` to answer this question.
   (iii) Perform a t-test using `ttest_ind`. What is your conclusion?