1940223\_shapiro-wilk\_ttest.R

pranav

2020-08-07

#PART 1: Checking if a data set seems to follow normal distribution

#The Shapiro-Wilk test is used to check how closely a data set follows normal distribution.

#H\_0: data follows normal distribution  
#H\_1: data does not follow normal distribution  
#If p > 0.05, H\_0 is accepted.  
#If p < 0.05, H\_0 is rejected.

#DATA SET 1  
data1 = rbinom(100, 100, 0.4)  
shapiro.test(data1)

##   
## Shapiro-Wilk normality test  
##   
## data: data1  
## W = 0.98844, p-value = 0.541

#p > 0.05  
#Hence, data1 can be said to follow normal distribution.  
#Hence, H\_0 is accepted.

#DATA SET 2  
data2 = c(100, 99, 67, 1, 2, -2)  
shapiro.test(data2)

##   
## Shapiro-Wilk normality test  
##   
## data: data2  
## W = 0.78657, p-value = 0.04435

#p = 0.04435 < 0.05  
#Hence, data2 cannot be said to follow normal distribution.  
#Hence, H\_0 is rejected.  
#========================  
#PART 2: Estimating the mean of a population from which the data set is assumed to be derived

#The t-test is used to determine whether the means of two groups are equal to each other.  
#It is also used to judge if a sample is derived from a normal distribution, which is the purpose for which it is used here.  
#Hence, it can also be used to judge if a sample is representative of a normal population or not.  
#The assumption for the test is the data is sampled from normal distributions with equal variances.

#H\_0: sample mean - population mean is 0  
#H\_1: sample mean - population mean is not 0  
#If p > 0.05, H\_0 is accepted.  
#If p < 0.05, H\_0 is rejected.

#DATA SET 3  
data3 = rpois(100, 4)  
val = mean(data3)  
t.test(data3)

##   
## One Sample t-test  
##   
## data: data3  
## t = 19.151, df = 99, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 3.657264 4.502736  
## sample estimates:  
## mean of x   
## 4.08

#p < 0.05  
#Hence, data3's mean is very different from population mean.  
#Hence, data3 is not a representative sample.  
#Hence, H\_0 is rejected.

#DATA SET 4  
data4 = c(12, 23, 4, -1, -32, 122, -232, 32)  
t.test(data4)

##   
## One Sample t-test  
##   
## data: data4  
## t = -0.25312, df = 7, p-value = 0.8074  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -93.07727 75.07727  
## sample estimates:  
## mean of x   
## -9

#p = 0.8074 > 0.05  
#Hence, data4's mean may be said to be close to the population mean.  
#Hence, data3 may be considered as a representative sample.  
#Hence, H\_0 may be accepted.