1940223 MSE (R Programming)

Pranav

2020-09-25

data = USArrests  
head(data, 10)

## Murder Assault UrbanPop Rape  
## Alabama 13.2 236 58 21.2  
## Alaska 10.0 263 48 44.5  
## Arizona 8.1 294 80 31.0  
## Arkansas 8.8 190 50 19.5  
## California 9.0 276 91 40.6  
## Colorado 7.9 204 78 38.7  
## Connecticut 3.3 110 77 11.1  
## Delaware 5.9 238 72 15.8  
## Florida 15.4 335 80 31.9  
## Georgia 17.4 211 60 25.8

rape = data$Rape  
murder = data$Murder

DATASET DESCRIPTION

#

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. A data frame with 50 observations on 4 variables. The variable “rape” contains the rapes per state, while the variable “murder” contains the murders per state.

#

QUESTION 1

#

Consistency is the property of an estimator wherein the value of the estimator indefinitely approaches the parameter as the sample size gets indefinitely large. In other words, a consistent estimator is an estimator wherein the probability of the estimator equalling the parameter approaches 1 as the sample size approaches infinity. In other words, the estimator becomes indefinitely more accurate with larger and larger sample sizes.

#

Choose parameter as population mean and choose estimator as sample mean.

#

Let the population be the data of the column “Murder”

population\_mean = mean(murder)  
population\_mean

## [1] 7.788

for(i in c(50, 100, 500, 1000, 2500, 5000))  
 {  
 sample\_means = replicate(10, mean(sample(murder, size = i, replace = TRUE)))  
 print(abs(population\_mean - mean(sample\_means)))  
 }

## [1] 0.0748  
## [1] 0.3119  
## [1] 0.0266  
## [1] 0.08901  
## [1] 0.017404  
## [1] 0.001564

As can be seen, the difference between the population mean and the sample mean generally gets closer to 0 as the sample size gets larger (note that this is an overall tendency, and may not be true for a particular pair of samples, where one is smaller) This indicates that sample mean is a consistent estimator of population mean.

#

QUESTION 2

#

We only have the murders per 100,000 people for 1973. Take the population as the number of murders per 100,000 people in USA states for all time. We will try to test whether the average murders per 100,000 people in USA states for all time equals 8, the closest integer to the sample mean.

#

Now, we do not knw the standard deviation of the population. Hence, we use the t-test on the sample mean to estimate the population mean. The assumptions of a one sample t-test is as follows…

# 1. The population follows normal distribution  
# 2. The sample values are independent (i.e. one does not affect the other)  
# 3. There are no outliers.

The hypotheses of one sample t-test are…

# H\_0: Population mean is 8  
# H\_1: Population mean is not 8

If p > 0.05, we accept H\_0.

#

For testing normality of the population using the sample, we use the Shapiro-Wilk test. The hypotheses of the Shapiro-Wilk test are as follows…

# H-0: The sample's population follows normal distribution  
# H\_1: The sample's population does not follow normal distribution

If p > 0.05, we accept H\_0.

shapiro.test(murder)

##   
## Shapiro-Wilk normality test  
##   
## data: murder  
## W = 0.95703, p-value = 0.06674

p = 0.06674 > 0.05. Hence, we may say that the sample’s population i.e. the murders per 100,000 people for all time follows normal distribution.

#

Now, performing the t-test…

t.test(murder, mu = 8)

##   
## One Sample t-test  
##   
## data: murder  
## t = -0.34418, df = 49, p-value = 0.7322  
## alternative hypothesis: true mean is not equal to 8  
## 95 percent confidence interval:  
## 6.550178 9.025822  
## sample estimates:  
## mean of x   
## 7.788

p = 0.7322 > 0.05. Hence, we accept H\_0. Hence, we may say that the average murders per 100,000 people in USA states for all time equals 8. Note that it is not wise to estimate values with data from only one year, because different years may have very different conditions that could affect crime rates.

#

QUESTION 3

#  
# Let P\_1 denote the set of rapes per 100,000 people in USA states for all time  
# Let P\_2 denote the set of murders per 100,000 people in USA states for all time

We want to find out whether P\_1 is similar to P\_2 (note that these two sets of measures are the populations). To do this, we use the data of 1973 as our sample. Using the rates of murder and rape per 100,000 people given in the dataset, we estimate the difference of the means of P\_1 and P\_2, and test whether the difference is significant or not.

#

Since we don’t know the population standard deviations, we use two sample t-test. Also, since murder and rape are different crimes, committed by different people in various conditions, we cannot consider them paired.

#

To check whether the corresponding populations of the samples can be said to have equal variances, we use the F-test. The F-test has the following hypotheses…

# 1. The variances of the two populations are equal  
# 2. The variances of the two populations are not equal

If p > 0.05, we accept H\_0.

var.test(murder, rape)

##   
## F test to compare two variances  
##   
## data: murder and rape  
## F = 0.21624, num df = 49, denom df = 49, p-value = 3.347e-07  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.1227105 0.3810539  
## sample estimates:  
## ratio of variances   
## 0.216239

p = 3.347e-07 = 0.0003347 < 0.05. Hence we reject H\_0 i.e. we take the two populations to have unequal variances.

#

Now, two sample t-test for unequal variances has the following assumptions…

# 1. The populations follow normal distribution  
# 2. The samples are independent and random  
# 3. There are no outliers

The hypotheses of the two sample t-test are…

# H\_0: The difference of the two populations' means is 0  
# H\_1: The difference of the two populations' means is not 0

i.e. H\_0 suggests the populations are similar enough to be equal, while H\_1 is the contrary. If p > 0.05, we accept H\_0.

#

First checking for normality using the Shapiro-Wilk test (discussed in question 2)…

shapiro.test(murder)

##   
## Shapiro-Wilk normality test  
##   
## data: murder  
## W = 0.95703, p-value = 0.06674

p = 0.06674 > 0.05

shapiro.test(rape)

##   
## Shapiro-Wilk normality test  
##   
## data: rape  
## W = 0.94674, p-value = 0.0251

p = 0.0251 < 0.05

#

Hence, while the population for murders may be said to follow normal distribution, the population for rapes may not be said to follow normal distribution. We proceed with the test anyway, since it is a robust test, keeping in mind that the results may be inaccurate due to all the assumptions not being fulfilled.

#

The two sample t-test is as follows…

t.test(murder, rape, var.equal = FALSE, paired = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: murder and rape  
## t = -9.2031, df = 69.245, p-value = 1.237e-13  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -16.35807 -10.52993  
## sample estimates:  
## mean of x mean of y   
## 7.788 21.232

p = 1.237e-13 < 0.05. Hence, we reject H\_0. Hence, we may say that P\_1 and P\_2 are significantly different i.e. we may say that the no. of murders for each state is significantly different from the no. of rapes for each state, for all time.