Descriptive Statistics

Process of Descriptive Statistics in R

- The measure of central tendency
- Measure of variability

Measure of central tendency

It represents the whole set of data by a single value. It gives us the location of central points. There are three main measures of central tendency:

Mean: It is the sum of observations divided by the total number of observations. It is also defined as average which is the sum divided by count.

```
# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                stringsAsFactors = F)

# Compute the mean value
mean = mean(myData$Age)
print(mean)
```

Mode: It is the value that has the highest frequency in the given data set. The data set may have no mode if the frequency of all data points is the same. Also, we can have more than one mode if we encounter two or more data points having the same frequency.

```
# Import the library
library(modeest)

# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                stringsAsFactors = F)

# Compute the mode value
mode = mfv(myData$Age)
print(mode)
```

Median:I t is the middle value of the data set. It splits the data into two halves. If the number of elements in the data set is odd then the center element is median and if it is even then the median would be the average of two central elements.

```
# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                stringsAsFactors = F)

# Compute the median value
median = median(myData$Age)
print(median)
```

Measure of variability

In Descriptive statistics in R measure of variability is known as the spread of data or how well is our data is distributed. The most common variability measures are:

Range: The range describes the difference between the largest and smallest data point in our data set. The bigger the range, the more is the spread of data and vice versa.

```r
# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                  stringsAsFactors = F)

# Calculate the maximum
max = max(myData$Age)
# Calculate the minimum
min = min(myData$Age)
# Calculate the range
range = max - min

cat("Range is:\n")
print(range)

# Alternate method to get min and max
r = range(myData$Age)
print(r)
```

Variance: It is defined as an average squared deviation from the mean. It is being calculated by finding the difference between every data point and the average which is also known as the mean, squaring them, adding all of them, and then dividing by the number of data points present in our data set.

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

```r
# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv",
                  stringsAsFactors = F)

# Calculating variance
variance = var(myData$Age)
print(variance)
```

Standard deviation: It is defined as the square root of the variance.

$$\sigma = \sqrt{\frac{\Sigma (x - u)^2}{N}}$$

```
# Import the data using read.csv()
myData = read.csv("CardioGoodFitness.csv", stringsAsFactors = F)

# Calculating Standard deviation
std = sd(myData$Age)
print(std)
```

Other measures of central tendancy are moments:

```
install.packages("moments")
install.packages("e1071")
install.packages("actuar")
```

```
all.moments(x, order.max=4)

raw2central()

central2raw()
```

Example

```
install.packages("moments")

library(moments)

x=c(0,1,2,3)

p=c(0.1,0.2,0.3,0.4)

m0=1

m1=sum(x*p)

m2=sum(x*x*p)

m3=sum(x*x*x*p)

m4=sum(x*x*x*x*p)

m=c(m0,m1,m2,m3,m4)

m

raw2central(m)
```

Problem

Find the mean, median and mode of the eruption duration in the data set faithful.

Solution

We apply the median function to compute the median value of eruptions.

```
duration = faithful$eruptions

mean(duration)

median(duration)

# Import the library

library(modeest)

mode = mfv(duration)

print(mode)

#Other method

y <- table(duration)

names(y)[which(y==max(y))]
```

Problem

Find the third central moment of eruption duration in the data set faithful.

Solution

We apply the function moment from the e1071 package. As it is not in the core R library, the package has to be installed and loaded into the R workspace.

```
> library(e1071)

> duration = faithful$eruptions

> moment(duration, order=3, center=TRUE)
```

Problem

Find the skewness of eruption duration in the data set faithful.

Solution

We apply the function skewness from the e1071 package to compute the skewness coefficient of eruptions. As the package is not in the core R library, it has to be installed and loaded into the R workspace.

```
> library(e1071)

> duration = faithful$eruptions

> skewness(duration)
```

Note: The normal distribution has zero excess kurtosis and thus the standard tail shape. It is said to be mesokurtic. Negative excess kurtosis would indicate a thin-tailed data distribution, and is said to be platykurtic. Positive excess kurtosis would indicate a fat-tailed distribution, and is said to be leptokurtic.

Problem

Find the excess kurtosis of eruption duration in the data set faithful.

Solution

We apply the function kurtosis from the e1071 package to compute the excess kurtosis of eruptions. As the package is not in the core R library, it has to be installed and loaded into the R workspace.

```
> library(e1071)

> duration = faithful$eruptions

> kurtosis(duration)
```

Exercise

1. A random variable X has the following probability distribution

X :	0	1	2
P(X=x):	1/3	1/3	1/3,

   Find the moment generating function, first four raw moments and the first four central moment.
   Write a R program for above problem.
2. The first three moments of the distribution about the value 3 of the random variable are 2, 10, -30 respectively. Find mean variance and skewness.
   Write a R program for above problem.
3. A random variable X has the probability distribution
   $P(X=x)=\frac{1}{8}\,{}^{3}C_{X}$, $X=0,1,2,3$, Find the moment generating function of X and then find mean and variance.
   Write a R program for above problem.
4. Find the first four moments about mean of the random variable X whose probability mass function is given by

X:	-2	3	1
P(X):	1/3	1/2	1/6

   Write a R program for above problem.