IST772 Notes Template: Chapter 1 – Statistical Vocabulary
Originality Assertion: By submitting this file you affirm that this writing is your own.

Name: Hendi Kushta
Date: 01/14/2023

****Important** Copying and/or pasting anything from the textbook will not be acceptable for your chapter notes submissions. You must write your notes in your own words and generate your own code, results, and graphs in R. This is what forces your brain to process the material that you read.**

MEASURES OF CENTRAL TENDENCY

The mean, median, and mode are the three most popular ways to determine central tendency.

The Mean

Is used by many people. It is also known as arithmetic mean or average. Found by adding all numbers together and dividing them with the total number of elements.

Ex. If we have grades = 6,7,8,9,10,4,4
mean = sum(a) / total_number(a) which is 48/7 = 6.85 approximately

The Median

Is the value that is found in the middle of the given elements, after the elements are sorted.

Ex. If we have grades = 6,7,8,9,10,4,4
We firstly order the grades, 4,4,6,7,8,9,10
Since there are just few numbers, we can see that 7 is the number in the middle, in this case is the median.

The Mode

Is the value with the highest frequency in a group of elements.

Ex. If we have grades = 6,7,8,9,10,4,4
From what we see, mode is 4 since it has the highest frequency from the given numbers.

IST772 Notes Template: Chapter 1 – Statistical Vocabulary
Originality Assertion: By submitting this file you affirm that this writing is your own.

MEASURES OF DISPERSION

The most used measure of dispersion is standard deviation.

The Range

It is very simple. It finds the minimum value and the maximum value in a given set of elements.

Ex. If we have grades = 6,7,8,9,10,4,4
Range will be 4 since it is the lowest number and 10 which is the highest number.

Deviations from the Mean

It is the distance of a point from the center.

Ex. If we have grades = 6,7,8,9,10,4,4
The average of the grades as I also found above is 6.85 approximately.
The distance of 6 from the center or mean is 0.85 and so on.

Sum of Squares

It helps to find the dispersion of data and deviation from the mean. It helps to find the variation of a group of elements. It is difficult to be understood or interpreted when more variables are added.

Ex. If we have grades = 6,7,8,9,10,4,4
sum of squares will be (6-6.85)^2 + (7-6.85)^2 + ... (4-6.85)^2

Variance

It is the sum of squares deviation from the mean divided from the total number of observations.

Ex. If we have grades = 6,7,8,9,10,4,4
variance = sum((6-6.85)^2 + (7-6.85)^2 + ... (4-6.85)^2) / 7

Standard Deviation

To have results not in square, so that it will be easier to understand, we find the standard deviation just by taking the square root of variance. Lower the standard deviation is, means that the data are distributed around the mean, higher, the more distant they are from the mean.

IST772 Notes Template: Chapter 1 – Statistical Vocabulary
Originality Assertion: By submitting this file you affirm that this writing is your own.

Ex. If we have grades = 6,7,8,9,10,4,4
standard deviation = sqrt(sum((6-6.85)^2 + (7-6.85)^2 + ….. (4-6.85)^2) / 7)

Box on p.14: Mean and Standard Deviation Formulas

Mean and Standard Deviation Formulas.

Mean => $\mu = \frac{\Sigma x}{N}$

μ -> small Greek letter <u>mu</u> which stands for the population mean.

Σ -> Capital Greek letter <u>sigma</u> which is a summation symbol.

x -> all data points.

Σx => summation of all data points.

N -> number of data points in observation.

<u>Ex</u> x => <u>2, 4, 0, 6</u>
-> Leads to $N=4$

substitute the numbers in the equation. -> $\mu = \frac{(2+4+0+6)}{4} = 3$.

Standard Deviation => $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

σ -> the standard deviation for the population.

nr. of observations

$\Sigma(x-\mu)^2$ -> sum of the squared deviations from the mean

-> when we divide sum of squares by nr of observations, we get variance, and as we said before the square root of variance gives standard deviation.

<u>Ex</u> x => <u>2, 4, 0, 6</u>
-> $N=4$

=> $\sigma = \sqrt{\frac{(2-3)^2 + (4-3)^2 + (0-3)^2 + (6-3)^2}{4}} = \sqrt{\frac{20}{4}} = \sqrt{5}$

IST772 Notes Template: Chapter 1 – Statistical Vocabulary
Originality Assertion: By submitting this file you affirm that this writing is your own.

DISTRIBUTIONS AND THEIR SHAPES

The Normal Distribution

Normal distribution is known also as bell curve distribution since it looks like a bell when its drawn. It is a distribution that is symmetric and shows that the closer the values are to the mean the more likely they are to occur, the further, the less chances they have to appear.

The Poisson Distribution

Poisson Distribution is mostly used related with time. It shows the number of times for an event to occur in a specific time.
Ex. The number of people ordering online at a specific period of time, lets say during holidays, winter break, black Friday etc.

R Code Fragments and Explanation

The Mean

```
grades <- c(6,7,8,9,10,4,4) # create a vector of numbers with some grades
mean(grades) # calculate the mean of the grades.
```

The Median

```
grades <- c(6,7,8,9,10,4,4) # create a vector of numbers with some grades
median(grades) # orders and find the median value of the vector of numbers.
```

The Mode

There si no mode function in R, so we need to install a package named modeest which stands for mode estimation.

```
install.packages("modeest") # install the necessary package to find mode
library(modeest) # in order to use the library that we have installed
grades <- c(6,7,8,9,10,4,4) # create the grades vector
mfv(grades) # find the mode using mfv function found in the package we installed.
```

The Range

```
grades <- c(6,7,8,9,10,4,4) # create the grades vector
range(grades) # gives the minimum and maximum values in the grades vector.
```

IST772 Notes Template: Chapter 1 – Statistical Vocabulary

Originality Assertion: By submitting this file you affirm that this writing is your own.

Variance

We can find variance in 2 different ways by going in a long way, by finding firstly the sum of squares and then dividing by the number of observations, but the easiest way is by using var() function

```
grades <- c(6,7,8,9,10,4,4) # create the grades vector
var(grades) # find the variance for the given vector of numbers
```

Standard Deviation

Just like with the variance, we can find standard deviation in a long way, by doing many calculations, like finding firstly sum of squares, then finding variance, then getting the square root of variance to find standard deviation. In R it is just a single line of code.

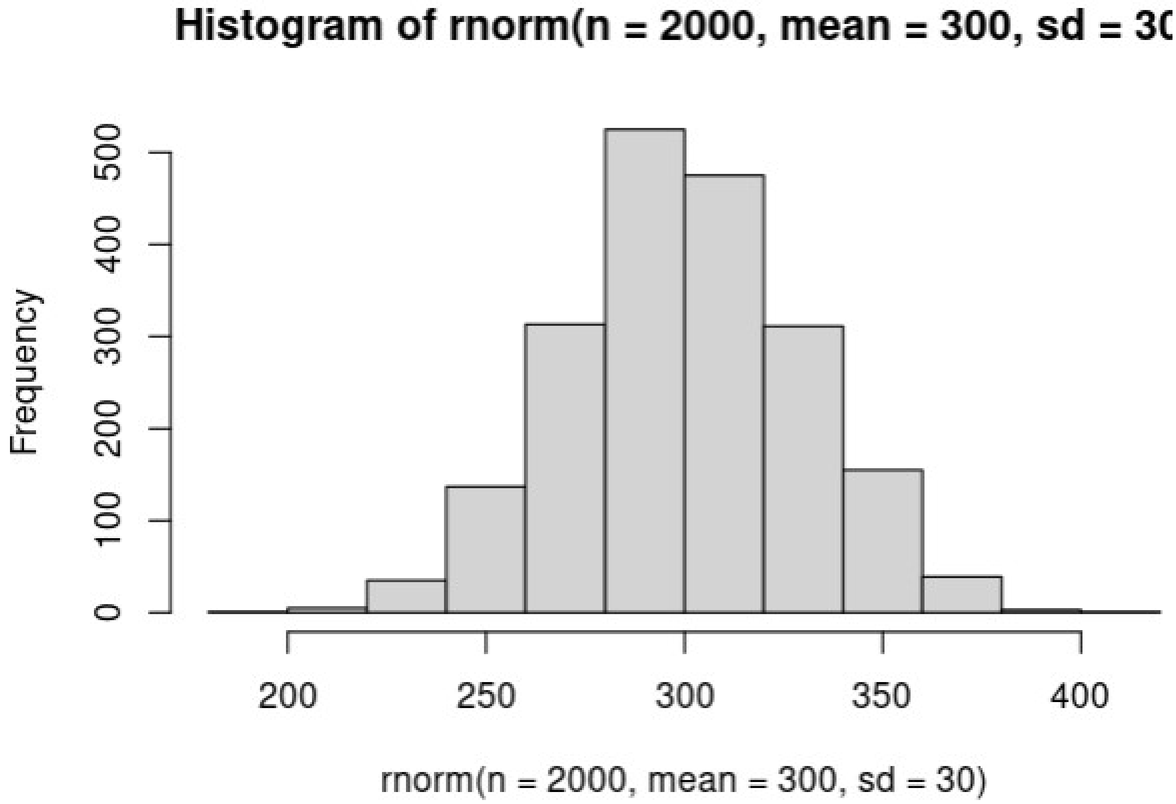
```
grades <- c(6,7,8,9,10,4,4) # create the grades vector
sd(grades) # find the standard deviation through sd() function
```

The Normal Distribution

I will draw a histogram in R to show the normal distribution.
To draw a histogram in R we use hist() function, and to give a normal distribution, we use rnorm() function.

hist(rnorm(n=2000, mean=300, sd=30)) # give 3 parameters to the normal distribution, n which is the number of 2000 random numbers with a mean of 300 and standard deviation of 30. To give more bars, we just add breaks = (number of bars that we want to show).

Originality Assertion: By submitting this file you affirm that this writing is your own.



The

Poisson Distribution

Same as in normal distribution, I will show the Poisson Distribution by building a histogram, but in this case instead of using rnorm() function, I am using rpois() function. This function takes only 2 parameters. Number of observations, or data, and the second parameter is lambda which shows the mean number of an event to happen within a given period of time.

hist(rpois(n=2000, lambda = 0.8))

IST772 Notes Template: Chapter 1 – Statistical Vocabulary
Originality Assertion: By submitting this file you affirm that this writing is your own.