MAST7866 — Foundations of Data Science

Computing Session 4 — Numerical Descriptive Measures

Task 1 – Numerical summaries of data

Watch the video "Numerical summaries of data" on moodle.

Task 2 – Measures of Location

Data can be summarised using numerical measures such as the mean, median and mode.

Consider the following data of the number of letters in a sample of 11 words from a page of text:

3 2 10 5 9 4 2 6 3 4 3

Recall that the sample mean is the average of the observed numbers. To calculate, we add all of the observations up and divide by the total number of observations.

Mathematically, for a sample of size n, with observations {x_1,x_2,…,x_n} this is written as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean is a very common measure of location and one advantage of it is that every data value is taken into account. However, the mean is sensitive to outliers (extreme values) in the data.

To calculate this in R, input the data:

```
> Letters<-c(3,2,10,5,9,4,2,6,3,4,3)
```

Then, calculate the mean:

```
> mean(Letters)
```

```
[1] 4.636364
```

Note that if your data contains any missing values, denoted by NA, the function mean will return NA by default. To get R to ignore the NA values you use the command:

```
> mean(Letters,na.rm=TRUE)
```

Challenge 1

The following are the ages (in years) of all eight employees of a small company:

53 32 61 27 39 44 49 57

Calculate the population mean.

```
> Age<-c(53,32,61,27,39,44,49,57)
```

```
> mean(Age)
```

Next, consider sampling this population. Use R to take a random sample of size 3 from this population and calculate the sample mean:

```
> x1<-c(sample(Age,3,replace=FALSE))
```

```
> mean(x1)
```

Repeat these two commands. What do you notice about the sample means you calculate?

The **median** is the middle value in a data set that has been ordered from smallest to largest. When data have been ordered in increasing order, the median lies in the $\frac{(n+1)}{2}^{th}$ value.

To calculate this in R:

```
> median(Letters)
```

```
[1] 4
```

The median is a more robust measure than the mean, since it is not affected by outliers. It is often a better measure than the mean if the data are highly skewed (see later!) However, it can be a disadvantage that the median only uses position and does not consider the specific values of the data.

The **mode** of a data set is the value that occurs most frequently. Note that some data sets may have more than one mode, or may have none at all.

There is no function in R which computes the mode. Instead we can tabulate the data:

```
>y<-table(Letters)
```

And then report which of the tabulated values has the highest frequency:

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

Task 3 – Measures of Dispersion

Measures of location summarise a data set with one number, but they do not tell us everything about a distribution.

Consider the following data showing the overall percentages children from School A and School B achieved in several tests.

School A					School B				
31.1	71.3	91.7	41.6	54.7	60.9	62.1	31.8	75.2	51.3
62.3	61.2	89.9	23.7	20.3	66.8	44.2	78.1	44.0	70.6
10.8	42.9	86.7	51.8	62.1	58.8	46.5	57.9	65.9	53.7
64.4	75.3	52.4	83.6	81.1	60.3	79.8	42.7	80.4	59.6
70.2	60.2	72.4	73.5	85.6	63.2	54.1	58.1	71.5	62.2
80.1	73.3	83.9	39.0	2.9	53.1	47.9	60.9	57.4	81.0

The data are stored in an excel file called school on moodle. Load the data into R.

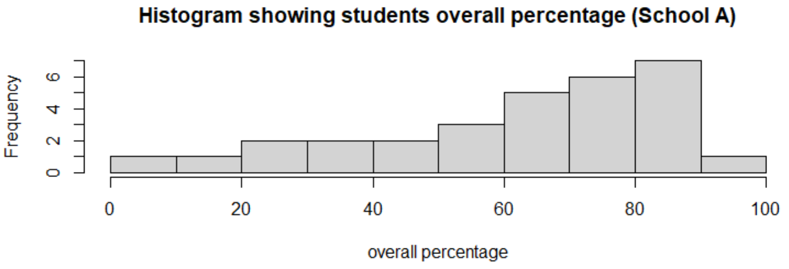
We can plot histograms of these data using R:

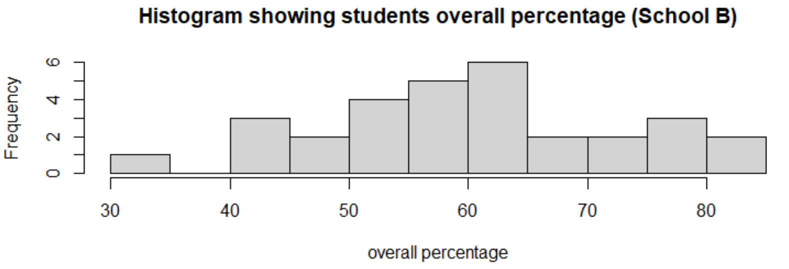
```
> attach(school)

> par(mfrow=c(2,1))

> hist(schoolA,breaks=10,main="Histogram showing students
overall percentage (School A)",xlab="overall percentage")

> hist(schoolB,breaks=10,main="Histogram showing students
overall percentage (School B)",xlab="overall percentage")
```





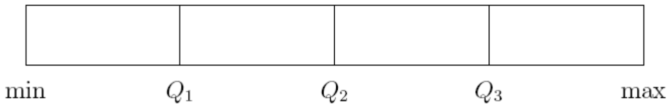
The mean overall percentage is 60 for both schools (check this in R), but the histograms clearly show the distributions are different; there is a much larger variation of values in School A compared to School B. We need measures of dispersion to provide information on the variation in a data set.

The range is the difference between the largest and smallest values in a dataset.

```
> range_A <- max(schoolA)-min(schoolA)

> range_B <- max(schoolB)-min(schoolB)
```

Before we look at interquartile range (IQR), we need to know how to compute *quartiles*, which divide an ordered set of data in to quarters.



After dividing the data into quarters, 25% of the data lies below the lower quartile Q_1, 50% below the median Q_2 and 75% below the upper quartile Q_3.

The quartiles of a data set can be calculated by

- ordering the data from smallest to largest
- find the median, Q_2 (we have already learnt how to do this)
- Q_1 is the median of the lower half of the data (of the values below Q_2)
- Q_3 is the median of the upper half of the data (of the values above Q_2)

The interquartile range is the range of the middle 50% of the data,

$$\text{IQR} = Q_3 - Q_1.$$

This is a more robust measure than the range as it is not affected by outliers.

The five number summary of a data set consists of the minimum value, Q_1, median, Q_3 and maximum value. A five number summary can then be used to create a **boxplot**.

```
> fivenum(schoolA)

> boxplot(schoolA,main="School A",ylab="overall percentage")
```

Challenge 2

We are going to see how to define multiple boxplots on the same plot. We are going to use inbuilt data set mtcars.

Type the following to find out more about the data set:

```
>?mtcars
```

To see the first few rows of the data set we can type

```
>head(mtcars)
```

Let us create a box plot for vehicle weight for each type of car:

```
>boxplot(wt~cyl,data=mtcars,main="Vehicle Weight",xlab="Number
of Cylinders",ylab="Weight")
```

Investigate other boxplots you can make of this data set.

The **sample variance** (s^2) and **sample standard deviation** (s) are the most common measures of spread, and they explain how much variation there is in the data.

The best way to interpret the standard deviation is to think of it as the average distance of each of the observations in the data from the mean.

If all the observations were the same, then the standard deviation would be zero.

The formula for the sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2, \qquad (1)$$

but it is often easier to use

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]. \qquad (2)$$

In R we can use the command

```
> sd(schoolA)
```

Challenge 3

Calculate the standard deviation of the School B data. What can you conclude about the data from the 2 schools?

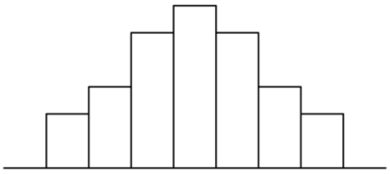
Task 4 – Shapes of Distributions

Watch the video "Shapes of Distributions" on moodle.

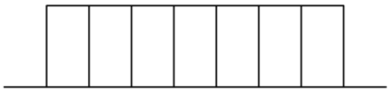
Task 5 – Shapes of Distributions

If a sample is representative of a population, then a histogram of the sample data should have a shape that is similar to the distribution of a population.

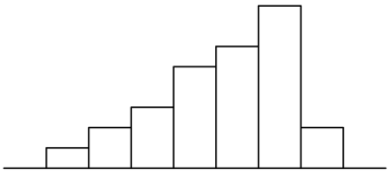
Symmetric



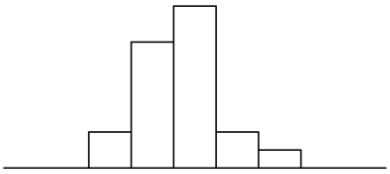
Uniform



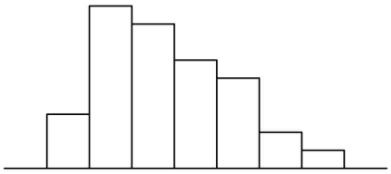
Negatively skewed



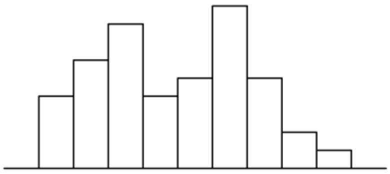
Unimodal

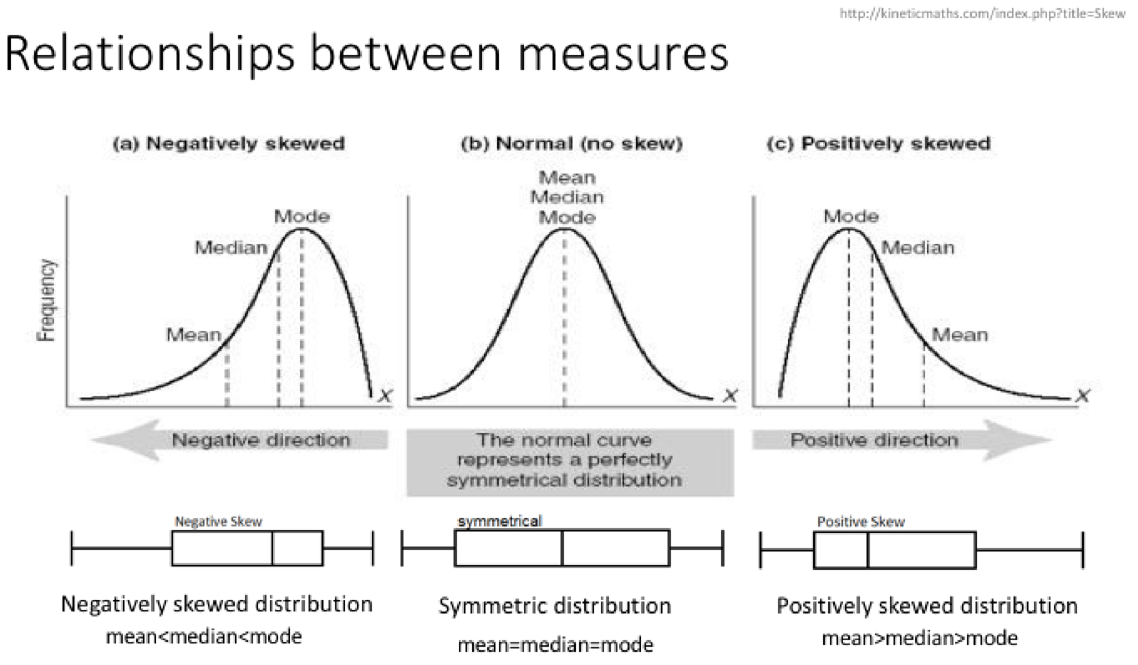


Positively skewed



Bimodal





http://kineticmaths.com/index.php?title=Skew

Relationships between measures

Negatively skewed distribution
mean<median<mode

Symmetric distribution
mean=median=mode

Positively skewed distribution
mean>median>mode

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point.

Suppose we have a sample of size n, with observations {x_1,…,x_n}, the Fisher-Pearson coefficient of skewness is defined by:

$$g_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

Note that many software packages actually compute an adjusted skewness statistic which adjusts for sample size.

How to interpret skewness?

- The skewness for a symmetric distribution is zero.
- Negative values indicate data that are skewed left – i.e. have left tails which are long relative to the right tail (negatively skewed)
- Positive values indicate data that are skewed right – i.e. have right tails which are long relative to the left tail (positively skewed)

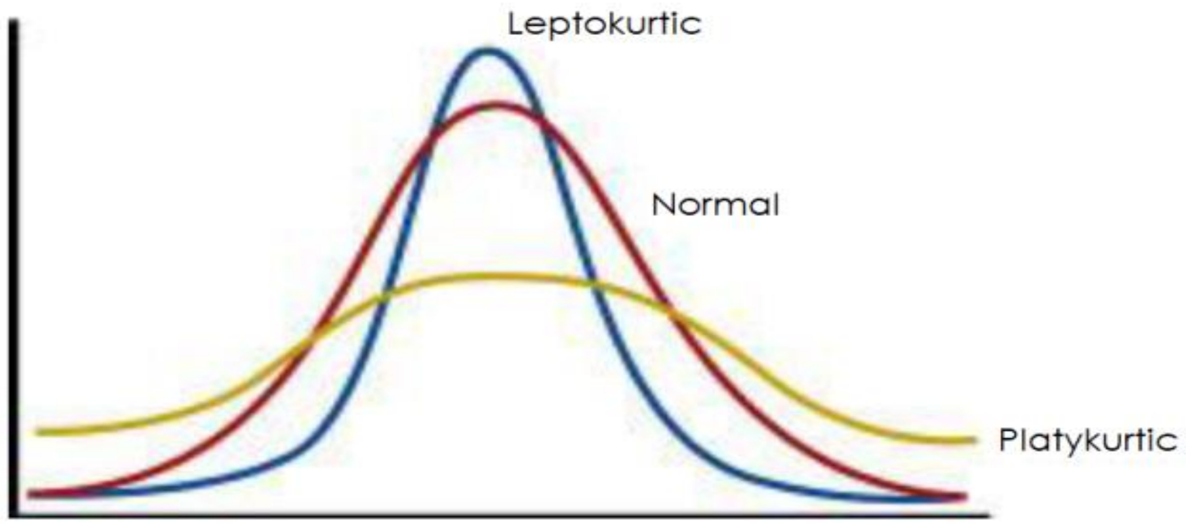
Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

Suppose we have a sample of size n, with observations $\{x_1,\ldots,x_n\}$, the kurtosis is defined by:

$$kurtosis = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4}$$

The kurtosis for a standard normal distribution is 3. For this reason, some define excess kurtosis by subtracting 3 from the formula above.

- A positive excess kurtosis indicates a “heavy-tailed” distribution – also referred to as Platykurtic
- A negative excess kurtosis indicates a “light-tailed” distribution – also referred to as Leptokurtic



https://javashine.wordpress.com/2017/08/21/measures-of-shape-skew-kurtosis/

Challenge 4

Calculate the skewness and kurtosis of school A and B. What can you conclude?

In R you will need to install package moments:

```
> install.packages(“moments”)

> library(moments)

> skewness(schoolA)

> kurtosis(schoolB)
```

Task 6 – Association between variables

Watch the video "Association between variables" on moodle.

Let us consider looking at the relationship between two of the variables in the iris data set. We have previously used this data set but if you would like a reminder about what it contains please type ?iris.

Plot a scatter plot of Petal.Length versus Petal.width:

```
> attach(iris)
```

```
> plot(Petal.Length,Petal.Width)
```

Now let us calculate the product moment correlation coefficient of these two variables.

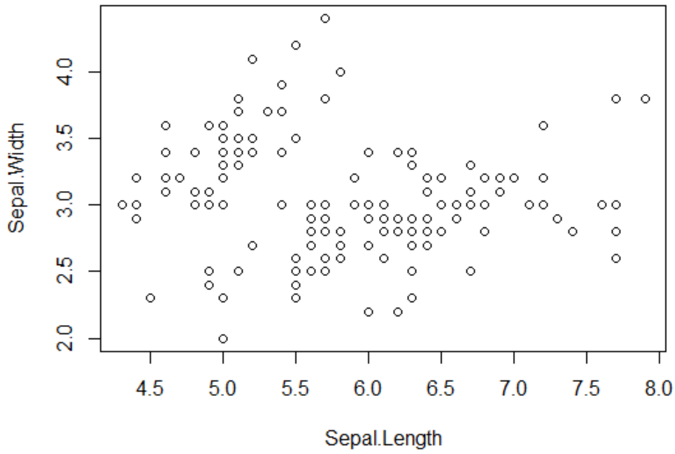
```
> cor.test(Petal.Length,Petal.Width)
```

This gives us a correlation of 0.962, suggesting a high positive correlation. This agrees with the strong linear relationship we can see in the scatter plot.

Note that in a later worksheet we will revisit this function to discuss the hypothesis test the function is performing, but for now we are just interested in obtaining the correlation statistic.

Challenge 5

Produce a scatter plot of Sepal.Length versus Sepal.Width and calculate the correlation. What do you conclude from the value of the correlation?


Sepal.Width
Sepal.Length