Advanced Statistics

Erbil Technical Engineering College

PhD. Course of Technical Mechanical and Information Systems Engineering

2023-2024

Assist. prof. Dr. Paree khan A. Omer

Administration & Economics College / Statistics and Informatics Department

Lect. 2 & 3

SPSS: Statistical Package for the Social Sciences

SPSS Statistics is software for managing data and calculating a wide variety of statistical procedures.

Four types of windows are:

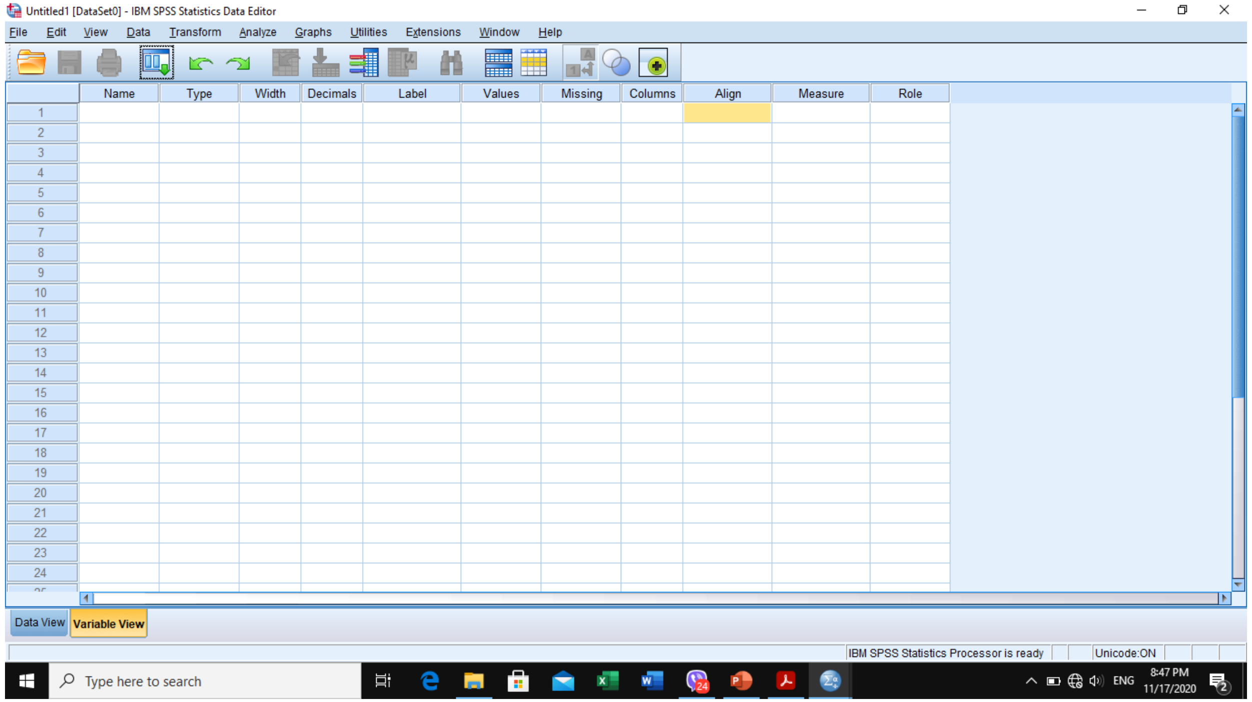
1. Data Editor Window.

2. Output Viewer Window.

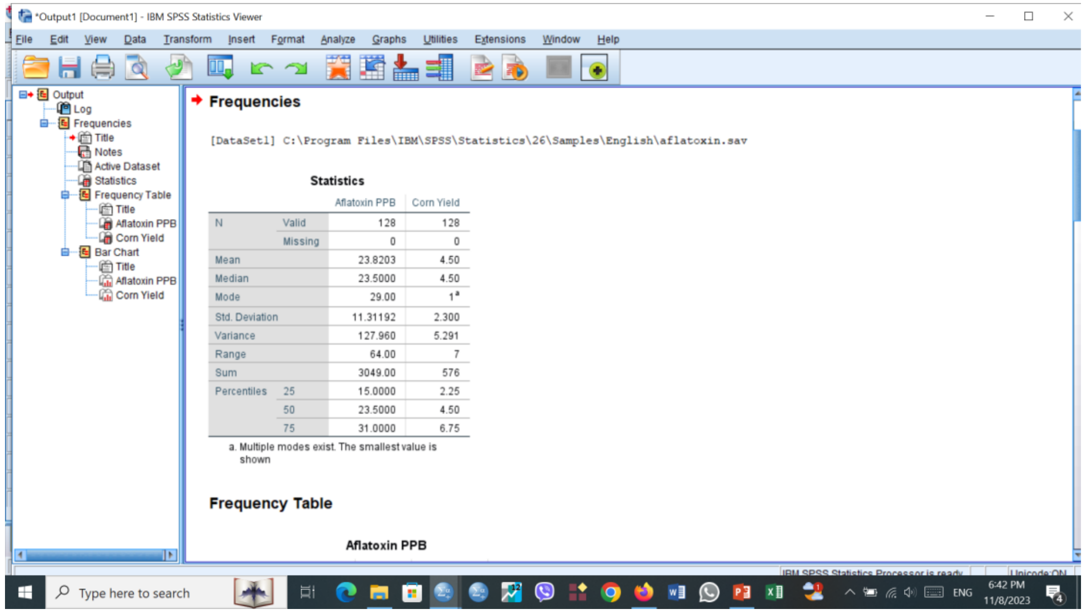
3. Chart Editor Window.

4. Syntax Editor Window

The Data Editor consists of two windows. By default the **Data View**, which allows the data to be entered and viewed, the other window is the **Variable View**, which allows the types of variables to be specified and viewed. The user can toggle between the windows by clicking on the appropriate tabs on the bottom left of the screen.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1											
2											
3											
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											



Frequencies

```
[DataSet1] C:\Program Files\IBM\SPSS\Statistics\26\Samples\English\aflatoxin.sav
```

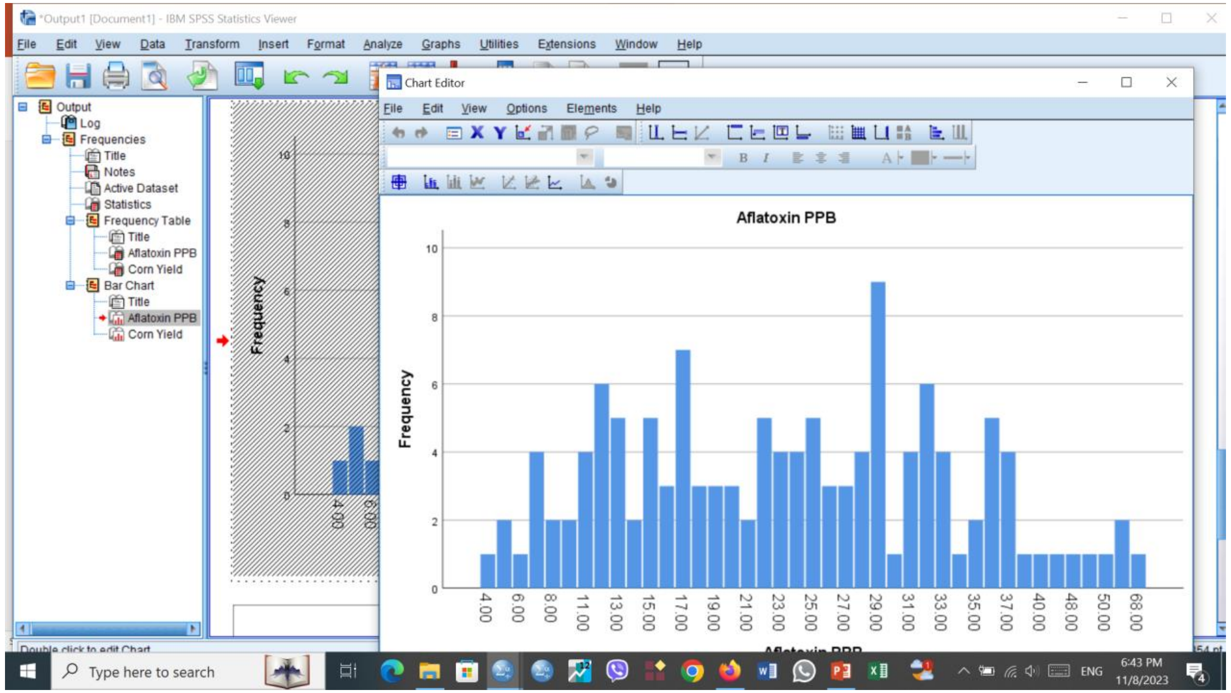
Statistics

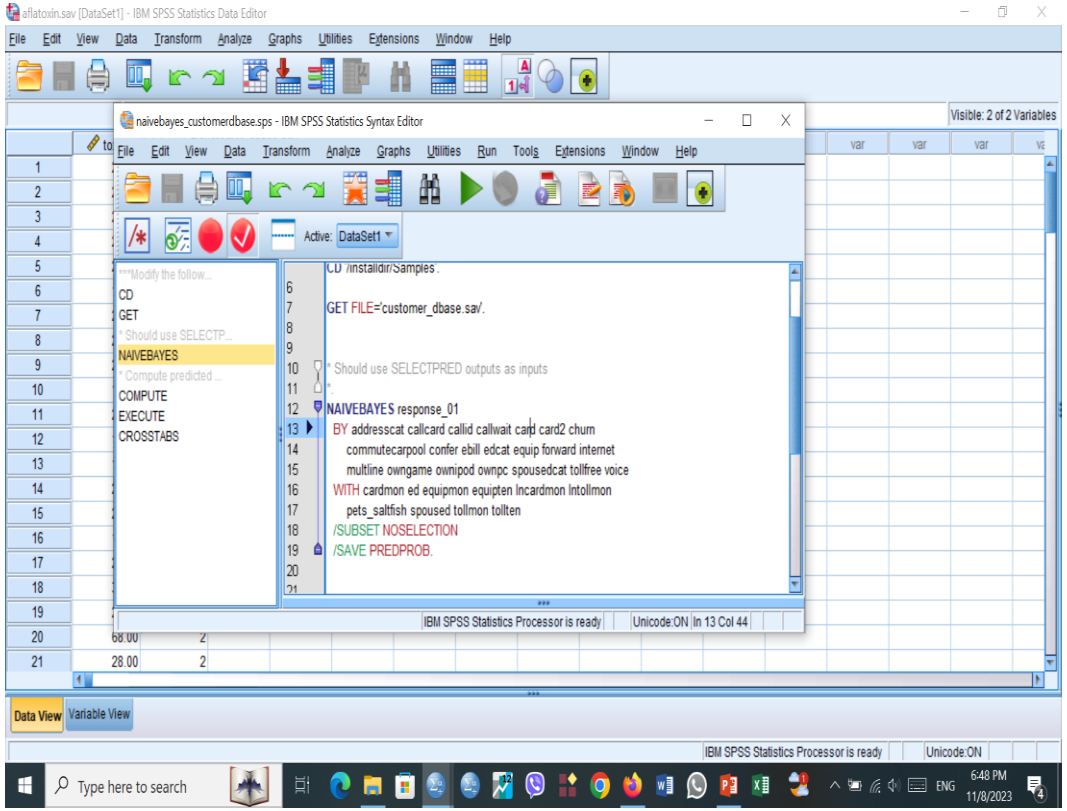
		Aflatoxin PPB	Corn Yield
N	Valid	128	128
	Missing	0	0
Mean		23.8203	4.50
Median		23.5000	4.50
Mode		29.00	1[a]
Std. Deviation		11.31192	2.300
Variance		127.960	5.291
Range		64.00	7
Sum		3049.00	576
Percentiles	25	15.0000	2.25
	50	23.5000	4.50
	75	31.0000	6.75

a. Multiple modes exist. The smallest value is shown

Frequency Table

Aflatoxin PPB




aflatoxin.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Visible: 2 of 2 Variables
var
naivebayes_customerdbase.sps - IBM SPSS Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Run Tools Extensions Window Help
Active: DataSet1
***Modify the follow...
CD
GET
* Should use SELECTP...
NAIVEBAYES
* Compute predicted ...
COMPUTE
EXECUTE
CROSSTABS
IBM SPSS Statistics Processor is ready
Unicode:ON
In 13 Col 44
Data View
Variable View
Type here to search
ENG
6:48 PM
11/8/2023

```
CD '/installdir/Samples'.

GET FILE='customer_dbase.sav'.


* Should use SELECTPRED outputs as inputs
*.
NAIVEBAYES response_01
  BY addresscat callcard callid callwait card card2 churn
     commutecarpool confer ebill edcat equip forward internet
     multline owngame ownipod ownpc spousedcat tollfree voice
  WITH cardmon ed equipmon equipten lncardmon lntollmon
     pets_saltfish spoused tollmon tollten
  /SUBSET NOSELECTION
  /SAVE PREDPROB.
```

20	68.00	2
21	28.00	2

There are 10 characteristics to be specified under the columns of the Variable View:

1. Name, the chosen variable name.
2. Type, the type of data.
3. Width, the width of the actual data entries.
4. Decimals, the number of digits to the right of the decimal place to be displayed for data entries.
5. Label, a label attached to the variable name.
6. Values, labels attached to category codes, for categorical variables, an integer code should be assigned to each category and the variable defined to be of type “numeric.”
7. Missing, missing value codes. SPSS recognizes the period symbol as indicating a missing value.
8. Columns, width of the variable column in the Data View.
9. Align, alignment of variable entries.
10. Measure, measurement scale of the variable. The default chosen by SPSS depends on the data type.

To Obtain Frequency Tables

From the menus of choose:

SPSS : Analyze → Descriptive Statistics → Frequencies

From frequencies main dialog box select one or more categorical or quantitative variables.

Optionally, you can:

- Click Statistics for descriptive statistics for quantitative variables.
- Click Charts for bar charts, pie charts, and histograms.
- Click Format for the order in which results are displayed.

Frequencies is indicated as a description for a single qualitative (categorical) variable, while for a scale variable frequency table becomes too long and full of single cases. However, it is always a good idea to start any statistical research with frequencies for every variable.

Central Tendency

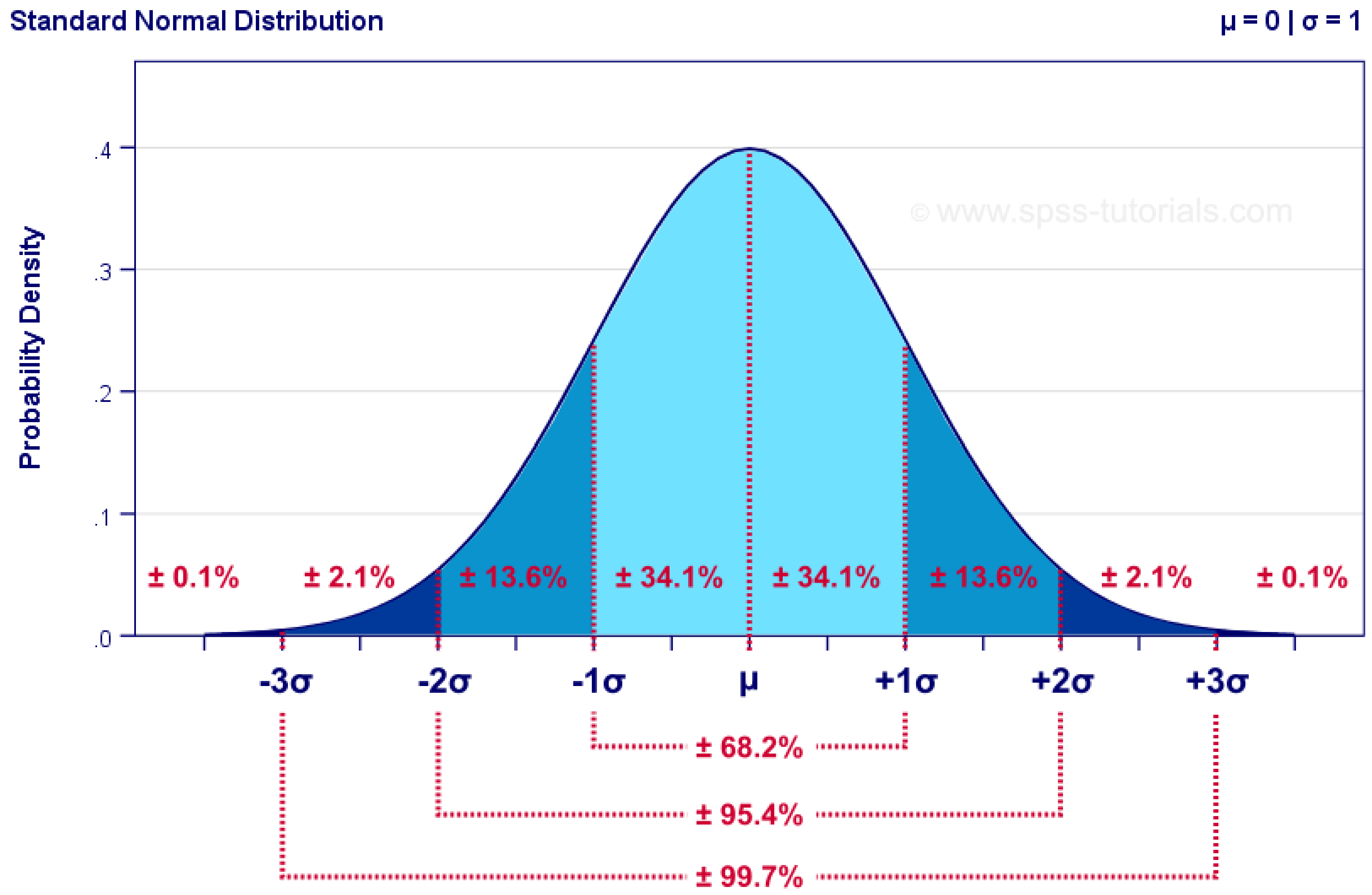
Statistics that describe the location of the distribution include the mean, median, mode, and sum of all the values.

- **Mean,** A measure of central tendency. The arithmetic average, the sum divided by the number of cases.
- **Median,** If there is an even number of cases, the median is the average of the two middle cases when they are sorted in ascending or descending order. The median is a measure of central tendency not sensitive to outlying values (unlike the mean, which can be affected by a few extremely high or low values).
- **Mode,** The most frequently occurring value. If several values share the greatest frequency of occurrence, each of them is a mode. The Frequencies procedure reports only the smallest of such multiple modes.
- **Sum,** The sum or total of the values, across all cases with non missing values.

Dispersion

Statistics that measure the amount of variation or spread in the data include the standard deviation, variance, range, minimum, maximum, and standard error of the mean.

- **Std. deviation,** A measure of dispersion around the mean. In a normal distribution, 68% of cases fall within one standard deviation of the mean and 95% of cases fall within two standard deviations. For example, if the mean age is 45, with a standard deviation of 10, 95% of the cases would be between 25 and 65 in a normal distribution.
- **Variance,** A measure of dispersion around the mean, equal to the sum of squared deviations from the mean divided by one less than the number of cases. The variance is measured in units that are the square of those of the variable itself.


Standard Normal Distribution
μ = 0 | σ = 1
Probability Density
.4
.3
.2
.1
.0
© www.spss-tutorials.com
± 0.1%
± 2.1%
± 13.6%
± 34.1%
± 34.1%
± 13.6%
± 2.1%
± 0.1%
-3σ
-2σ
-1σ
μ
+1σ
+2σ
+3σ
± 68.2%
± 95.4%
± 99.7%

- **Range,** The difference between the largest and smallest values of a numeric variable, the maximum minus the minimum.
- **Minimum,** The smallest value of a numeric variable.
- **Maximum,** The largest value of a numeric variable.
- **S. E. mean.** A measure of how much the value of the mean may vary from sample to sample taken from the same distribution.

Distribution

Skewness and kurtosis are statistics that describe the shape and symmetry of the distribution.

- **Skewness.** A measure of the asymmetry of a distribution. The normal distribution is symmetric and has a skewness value of 0. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail.
- **Kurtosis.** A measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero.

To Obtain Descriptive Statistics

From the menus choose:

$SPSS : Analyze \rightarrow Descriptive\ Statistics \rightarrow Descriptive$

From Descriptive main dialog box select one or more variables.

Optionally, you can:

- Select Save standardized values as variables to save (z) scores as new variables.
- Click Options for optional statistics and display order.

Descriptive statistics (mean, median, standard deviation, minimum, maximum, range, skewness, kurtosis) is indicated as a description for a single scale variable and usually it does not make sense for categorical variables.

Measures of central tendency

1. **mode**

The mode of a data set is the value that occurs most frequently.

2. **Median**

The median is the central value of an ordered distribution. To find it,

1. Order the data from smallest to largest.

2. For an odd number of data values in the distribution, Median = Middle data value

For an even number of data values in the distribution:

$$median = \frac{sum\ of\ middle\ two\ values}{2}$$

3.The Mean

The most commonly used measure of center for quantitative variable is the sample mean. The sample mean of the variable is the sum of observed values in a data divided by the number of observations. If the sample size is n, then the mean of the variable X is:

$$mean = \frac{Sum\ of\ all\ entries}{number\ of\ entries} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} \quad ,$$

The sum $x_1 + x_2 + x_3 + \cdots + x_n$ is denoted as $\sum_{i=1}^{n} x_i$

The sample mean of the variable is the sum of observed values $x_1, x_2, x_3, \ldots, x_n$ in a data divided by the number of observations n. The sample mean (statistic) is denoted by $\bar{x}$, and expressed operationally,

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

- **Measures of variation**

we will examine three of the most frequently used measures of variation; the sample range, the sample interquartile range and the sample standard deviation, Measures of variation are used mostly only for quantitative variables.

1. Range

The sample range of the variable is the difference between its maximum and minimum values in a data set:

$$\text{Range} = \text{Max} - \text{Min}.$$

Quartiles

the quartiles of the variable divide the observed values into quarters, or 4 equal parts, the variable has three quartiles, denoted by Q1, Q2 and Q3. the first quartile, Q1, is the number that divides the bottom 25% of the observed values from the top 75%; second quartile, Q2, is the median, which is the number that divides the bottom 50% of the observed values from the top 50%; and the third quartile, Q3, is the number that divides the bottom 75% of the observed values from the top 25%.

1. The first quartile Q1 is at position $\frac{n+1}{4}$
2. The second quartile Q2 (the median) is at position $\frac{n+1}{2}$
3. The third quartile Q3 is at position $\frac{3(n+1)}{4}$

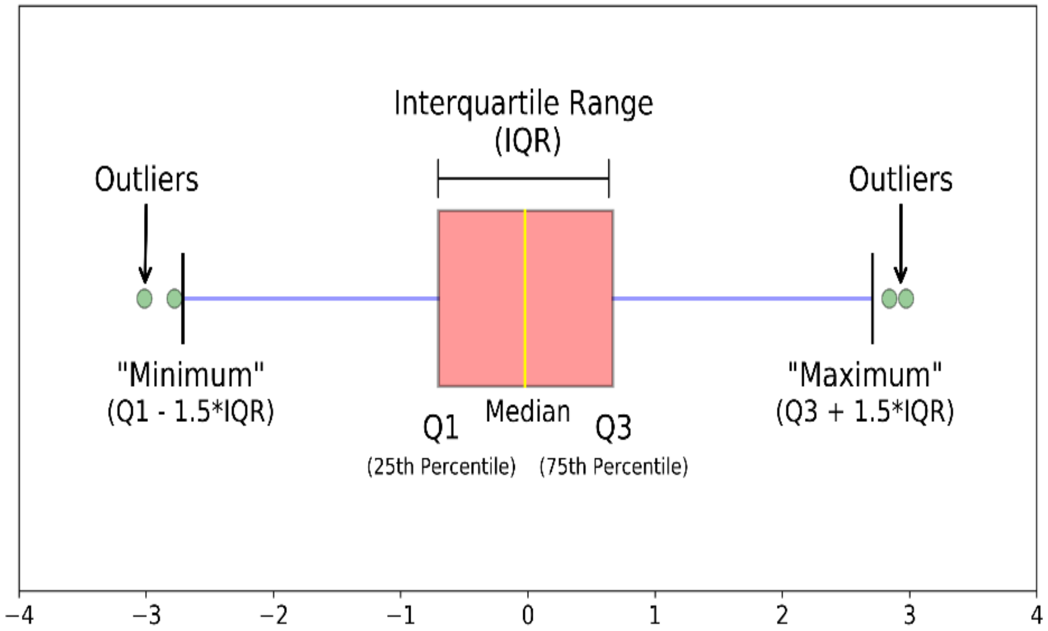
Interquartile range:

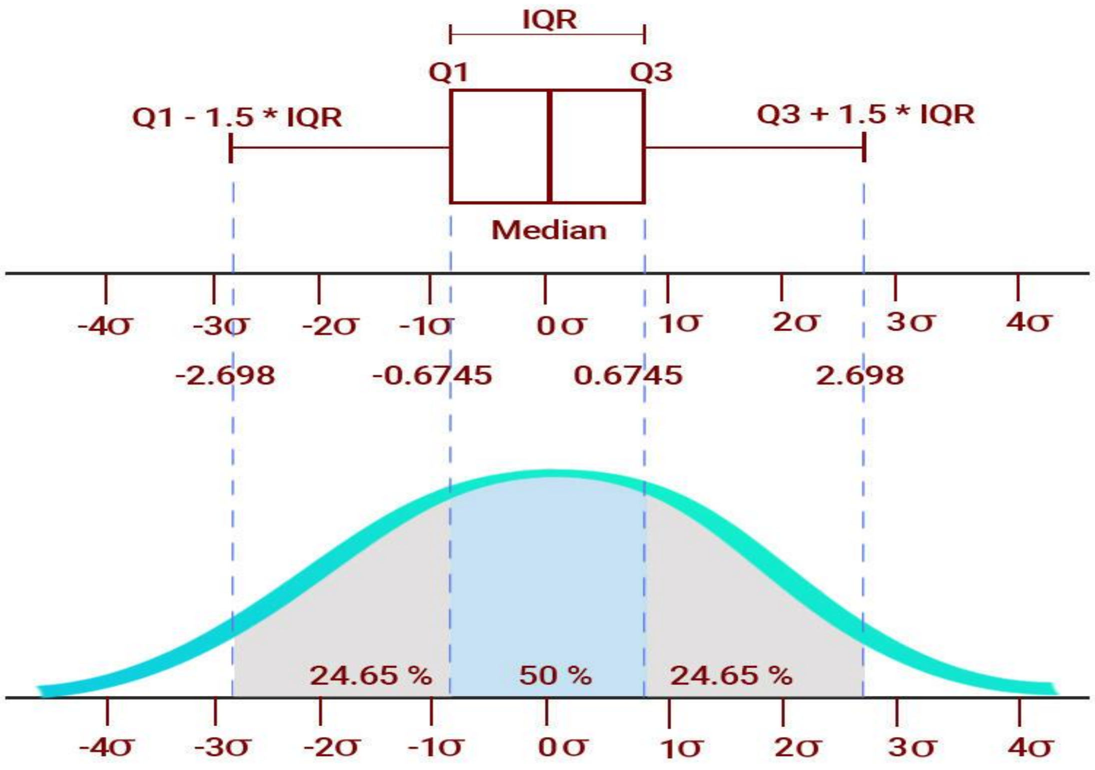
The sample interquartile range of the variable, denoted as IQR, is the difference between the first and third quartiles of the variable, that is, IQR = Q3 − Q1.

Roughly speaking, the IQR gives the range of the middle 50% of the observed values.

Note//The five-number summary of the variable consists of minimum, maximum, and quartiles written in increasing order: Min, Q1, Q2, Q3, Max.

A boxplot is based on the five-number summary and can be used to provide a graphical display of the center and variation of the observed values of variable in a data set.


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
"Maximum"
(Q3 + 1.5*IQR)
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4


IQR
Q1
Q3
Q1 - 1.5 * IQR
Q3 + 1.5 * IQR
Median
-4σ
-3σ
-2σ
-1σ
0σ
1σ
2σ
3σ
4σ
-2.698
-0.6745
0.6745
2.698
24.65 %
50 %
24.65 %
-4σ
-3σ
-2σ
-1σ
0σ
1σ
2σ
3σ
4σ

3. Variance

Variance is a measure of dispersion of data points from the mean. Low variance indicates that data points are generally similar and do not vary widely from the mean. High variance indicates that data values have greater variability and are more widely dispersed from the mean. The value is provided by S^2 given by

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

4. Standard deviation

The sample standard deviation is the most frequently used measure of variability, although it is not as easily understood as ranges. The formula is the square root of the Variance.

$$S_{dx} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$