Statistical analysis allows a better understanding of the basic structure of the data and carry out a number of procedures to identify possible correlations between variables, recognize trends and, as an integral part of the data mining tool and enabling predictions. Below we present the main two main features of a statistical analysis

- **CROSSTABULATION.**
- **DESCRIPTIVE STATISTICS.**

1 Types of Variables

The variables typically used in statistical analysis fall into one of the following three basic categories.

1.1 Arithmetic (or Quantitative or Scale) Variables

These variables take values in an interval of the real line and include the height, weight or income of an individual, the distance traveled by some automobile, the life-span of a machine, etc.

1.2 Categorical (or Qualitative or Nominal) Variables

These variables record qualitative attributes of the objects under consideration. Usually the possible categories are called the levels of the nominal variable. Examples of categorical variables include the political preference (Right, Center, Left), the preferred kind of music (Rock, Jazz, Classical, Country, Folk, etc.), the grade of a student in some exam (A, B, C, D, F), etc.

1.3 Ordinal Variables

These are nominal variables whose levels can be ordered in some logical sense, however the distances between the various levels are not exactly known. Examples of ordinal variables include the age group of an individual (teenager, middle aged, old), the opinion on some matter (absolutely disagree, disagree, rather disagree, rather agree, agree, absolutely agree), the grade of a student in some exam (A, B, C, D, F), etc.

2 Frequency Tables

A frequency table is a table that lists items and uses tally marks to record and show the number of times they occur. Each entry of such a table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample. Frequency tables can be used for any dependent (e.g. answer to a poll question) or independent (e.g. age, gender, etc.) variable. It is often useful that percentages are included, taking into account the missing values. Below, frequency tables for three different variables are included.

Table 1: Do you agree with Newsweek's cover suggesting that Obama must go?

	Frequency	Percent
Yes. we need a new president No.	11	36.7
Obama is the most suitable one I do	9	30.0
not care	10	33.3
Total	30	100.0

Table 2: Gender

	Frequency	Percent
Male	15	50.0
Female	15	50.0
Total	30	100.0

Table 3: Age Groups

	Frequency	Percent
15-24	3	10.0
25-44	15	50.0
45-64	9	30.0
65+ years	3	10.0
Total	30	100.0

3 Crosstabs or Contingency Tables

Cross-tabulation is the process of creating a table from the multivariate frequency distribution of two statistical variables, tabulating the results of one variable against the other. Such tables are called contingency tables and give a basic picture of the interrelation of the two variables.

In contingency tables, independent variables (e.g. the gender) are usually displayed as rows and dependent variables (e.g. an answer to a poll question) as columns. It is often useful that percentages by row, by column, or total percentages are included. Some examples are given below.

3.1 Percentages by row

Table 4: Gender * Do you agree with Newsweek's cover suggesting that Obama must go? Crosstabulation

			Do you agree with Newsweek's cover suggesting that Obama must go?			
			Yes. we need a new president	No. Obama is the most suitable one	I do not care	Total
Gender	Male	Count	6	5	4	15
		% within Gender	40.0%	33.3%	26.7%	100.0%
	Female	Count	5	4	6	15
		% within Gender	33.3%	26.7%	40.0%	100.0%

Total	Count	11	9	10	30
	% within Gender	40.0%	33.3%	26.7%	100.0%

3.2 Percentages by column

Table 5: Gender * Do you agree with Newsweek's cover suggesting that Obama must go? Crosstabulation

			Do you agree with Newsweek's cover suggesting that Obama must go?			
			Yes. we need a new president	No. Obama is the most suitable one	I do not care	Total
Gender	Male	Count	6	5	4	15
		% within Do you agree with Newsweek's cover suggesting that Obama must go?	40.0%	33.3%	26.7%	100.0%
	Female	Count	5	4	6	15
		% within Do you agree with Newsweek's cover suggesting that Obama must go?	33.3%	26.7%	40.0%	100.0%
Total		Count	11	9	10	30
		% within Do you agree with Newsweek's cover suggesting that Obama must go?	40.0%	33.3%	26.7%	100.0%

3.3 Total percentages

Table 6: Gender * Do you agree with Newsweek's cover suggesting that Obama must go? Crosstabulation

			Do you agree with Newsweek's cover suggesting that Obama must go?			
			Yes. we need a new president	No. Obama is the most suitable one	I do not care	Total
Gender	Male	Count	6	5	4	15
		% of Total	40.0%	33.3%	26.7%	100.0%
	Female	Count	5	4	6	15
		% of Total	33.3%	26.7%	40.0%	100.0%
Total		Count	11	9	10	30
		% of Total	40.0%	33.3%	26.7%	100.0%

Contingency tables can analogously be defined for three or more variables, however for more than three variables they are hard to use and are usually avoided. An example of contingency table for three variables is given in Table 7 below.

Table 7: Gender * Do you agree with Newsweek's cover suggesting that Obama must go? * Age Group Crosstabulation

				Do you agree with Newsweek's cover suggesting that Obama must go?			
Age Group				Yes. we need a new president	No. Obama is the most suitable one	I do not care	Total
15-24	Gender	Male	Count	0	1	0	1
			% of Total	0.0%	33.3%	0.0%	33.3%
		Female	Count	1	0	1	2
			% of Total	33.3%	0.0%	33.3%	66.7%
	Total		Count	1	1	1	3
			% of Total	33.3%	33.3%	33.3%	100.0%
25-44	Gender	Male	Count	1	3	4	8
			% of Total	6.7%	20.0%	26.7%	53.3%
		Female	Count	2	2	3	7
			% of Total	13.3%	13.3%	20.0%	46.7%
	Total		Count	3	5	7	15
			% of Total	20.0%	33.3%	46.7%	100.0%
45-64	Gender	Male	Count	4	1	0	5
			% of Total	44.4%	11.1%	0.0%	55.6%
		Female	Count	1	2	1	4
			% of Total	11.1%	22.2%	11.1%	44.4%
	Total		Count	5	3	1	9
			% of Total	55.6%	33.3%	11.1%	100.0%
65+ years	Gender	Male	Count	1	0	0	1
			% of Total	33.3%	0.0%	0.0%	33.3%
		Female	Count	1	0	1	2
			% of Total	33.3%	0.0%	33.3%	66.7%
	Total		Count	2	0	1	3
			% of Total	66.7%	0.0%	33.3%	100.0%
Total	Gender	Male	Count	6	5	4	15
			% of Total	20.0%	16.7%	13.3%	50.0%
		Female	Count	5	4	6	15
			% of Total	16.7%	13.3%	20.0%	50.0%
	Total		Count	11	9	10	30
			% of Total	36.7%	30.0%	33.3%	100.0%

4 Pearson chi-square (χ^2) test

This test is used for contingency tables and tests whether two categorical variables are dependent or not. For example, one can check whether the answer to the question "Do you agree with Newsweek's cover suggesting that Obama must go?" depends on the gender or age of the person responding. This test makes use of a test Statistic proposed by Carl Pearson in 1900, which is a function of the squares of the deviations of the observed counts from their expected values, weighted by the reciprocals of their expected values.

5 Descriptives

Descriptive measures make sense for statistical analysis of quantitative data and include the following:

5.1 The Central Tendency

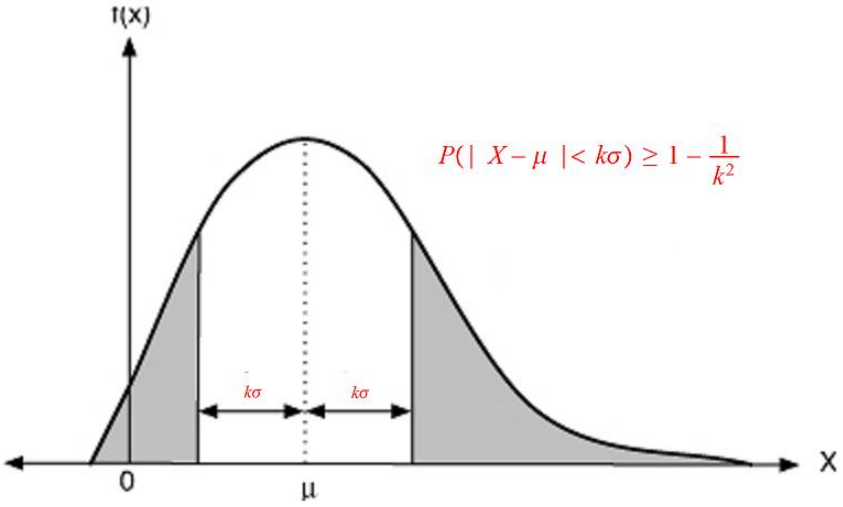
The central tendency includes Statistics which describe the location of the distribution of a quantitative variable. It includes the mean, the median, and the sum.

- The Mean is the arithmetic average i.e. the sum of values of the quantitative variable divided by the number of cases.
- The Median is the value above and below which half of the cases fall. It is also called the 50th percentile. If there is an even number of cases, the median equals the average of the two middle cases when they are sorted in ascending or descending order, while in an odd number of cases it equals the middle case, when the cases are sorted as above. 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).
- The Sum equals the sum of the values across all cases with non-missing values.

5.2 The Dispersion

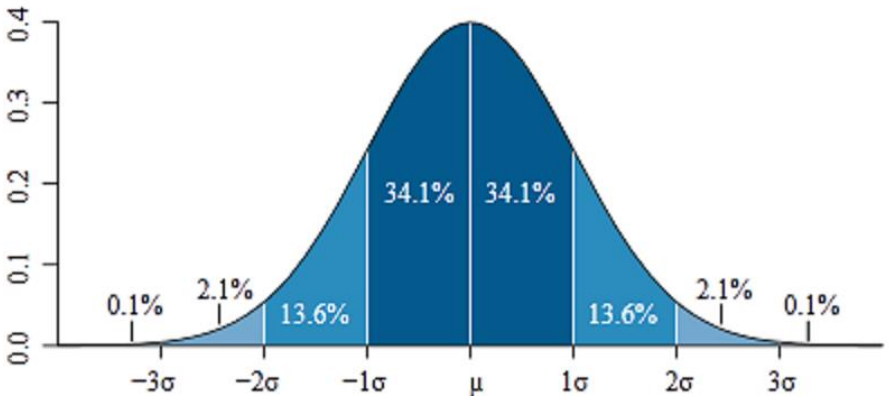
The dispersion includes Statistics which measure the amount of variation or spread in the data. They include the standard deviation, the variance, the range, the minimum, the maximum, and the standard error of the mean.

- The Standard deviation is a measure of dispersion around the mean. In any distribution, 93.75% of the cases fall within four standard deviation of the mean (Chebyshev's inequality). 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 years with a standard deviation of 10 years, in a normal distribution, 95% of the cases will have to be between 25 and 65 years of age.




Pafnuty Lvovich Chebyshev

Figure 1: Due to Chebychev's inequality, at least $(1-\frac{1}{k^2})100\%$ of the data lie in the white area bounded by the graph and the x-axis, for any $k = 2, 3, \ldots$



Carl Friedrich Gauß

Figure 2: Normal distribution with mean μ and standard deviation σ

- The Variance measures the amount of variation around the mean and is 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.
- The Range equals the difference between the largest and smallest values of a quantitative variable.
- The Minimum is the smallest value of a quantitative variable.
- The Maximum is the largest value of a quantitative variable.
- The Standard Error of the Mean measures how much the value of the mean may vary from sample to sample taken from the same distribution and can be used to roughly compare the observed mean to a hypothesized value.

5.3 The Distribution

The distribution includes the skewness and kurtosis which are Statistics describing the shape and symmetry of the distribution. These statistics are displayed with their standard errors.

- The Skewness is a measure of the asymmetry of a distribution. The normal distribution is symmetric and has zero skewness. Distributions with a significant positive skewness have a long right tail while distributions with a significant negative skewness have a long left tail. As a guideline, a skewness value more than twice its standard error is taken to indicate departure from symmetry. Examples of positively and negatively skewed distributions are shown in Figure 3.

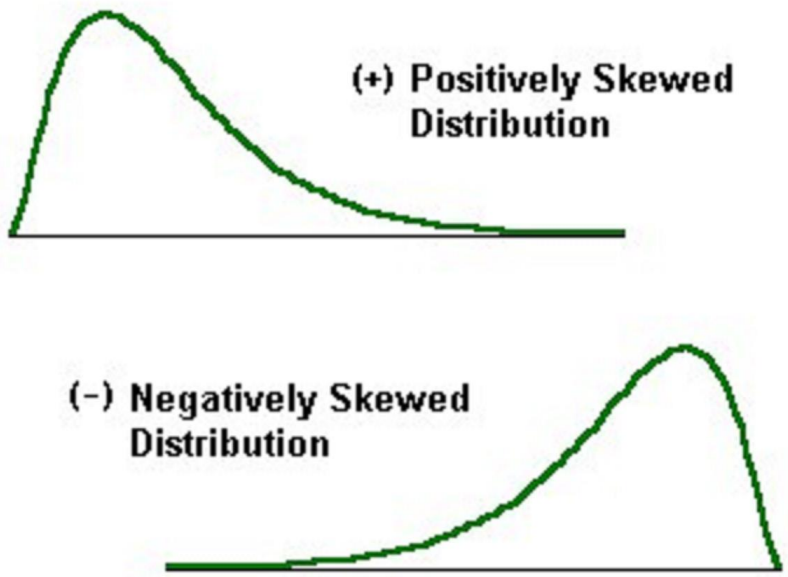


Figure 3: Positive and negative skewness examples.

- The Kurtosis is a measure of the extent to which observations cluster around a central point. Positive kurtosis indicates that, relative to a normal distribution of the same mean and variance, the observations are more clustered about the center of the distribution and have thinner tails towards the extreme values of the distribution. At these points, the tails of leptokurtic distributions are thicker relative to a normal distribution. Negative kurtosis indicates that, relative to a normal distribution of the same mean and variance, the observations cluster less and have thicker tails towards the extreme values of the distribution. At these point, the tails of platykurtic distributions are thinner relative to a normal distribution. Examples of distributions with different kurtosis are shown in Figure 4.

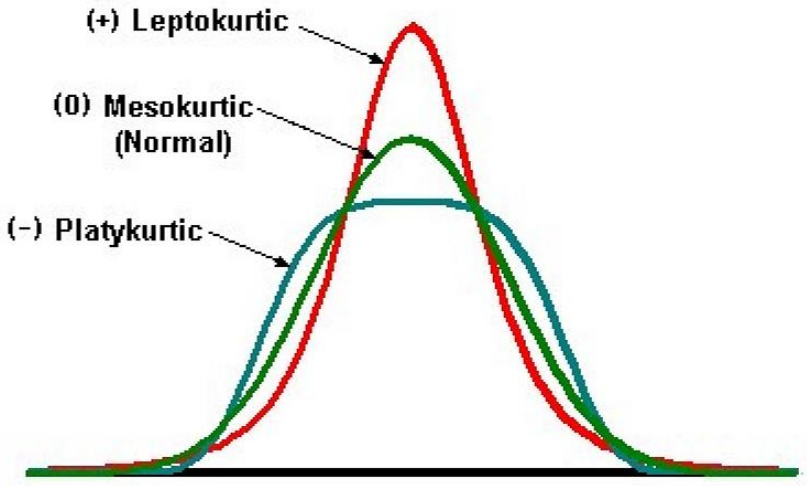


Figure 4: Kurtosis examples.

An example of descriptive statistics is shown in Table 8.

Table 8: Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Household income in thousands	5000	1064.00	9.00	1073.00	275203.00	55.0406	.78552	55.54475	3085.219	5.125	.035	56.259	.069
Valid N (listwise)	5000												

5.4 Tables

As shown in Section 3, frequency tables record the distribution of the values of a categorical variable, while crosstabs record the correlation of at least two categorical variables. A different, pictorial way, to study correlations of variables is by using graphs.

5.5 Graphs

Graphs enable us to understand the relationship between variables, and interpret the behavior of objects under consideration in a simple, pictorial way, easily understood by almost everyone.

In the following table we present a list of graphs suitable for the study of variables or combinations of variables of specific type.

Type of Graph	Type of Variable	Diagram
One Variable	Nominal	Bar Chart Pie Chart
	Numerical	Histogram Box Plot
Two Variables	Two Nominal	Clustered Bar Charts
	One Nominal & One Numerical	Box Plot Error Bars
	Two Numerical	Scatter Plot Line Chart Area Chart
At least three Variables	Numerical	Scatter Plot Matrix

Figure 5: Diagrams suitable for certain variable types.