TOPIC ONE: INTRODUCTION TO STATISTICS FOR BUSINESS RESEARCH

Why you need to use statistics in your research
It is obvious that society cannot be run effectively on the basis of hunches or trial and error, and that in business and economics much depends on the correct analysis of numerical information. Decisions based on data will provide better results than those based on intuition or gut feelings. What applies to this wider world applies to undertaking research into the wider world. Statistical thinking permeates all social interaction. Much of everyday life depends on making forecasts, and business cannot progress without being able to audit change or plan action. In your research, you may be looking at areas such as purchasing, production, capital investment, long-term development, quality control, human resource development, recruitment and selection, marketing, credit risk assessment or financial forecasts or others. And that is why the informed use of statistics is of direct importance to you while you are collecting your data and analysing them. If nothing else, your results and findings will be more accurate, more believable and, consequently more useful. Some of the reasons why you will be using statistics to analyse your data are the same reasons why you are doing the research. Ignoring the possibility that you are researching because the project or dissertation element of your qualification is compulsory, rather than because you very much want to find something out, you are likely to be researching because you want to:

- measure quantities;
- examine relationships;
- make predictions;
- test hypotheses;
- construct concepts and develop theories;
- explore issues;
- explain activities or attitudes;
- describe what is happening;
- present information;
- make comparisons to find similarities and differences; and
- draw conclusions about populations based only on sample results.

If you didn’t want to do at least one of these things, there would be no point to doing your research at all.

WHAT IS STATISTICS?
Put simply, statistics is a range of procedures for gathering, organising, analysing and presenting quantitative data. ‘Data’ is the term for facts that have been obtained and subsequently recorded, and, for statisticians, ‘data’ usually refers to quantitative data that are numbers. Essentially therefore, statistics is a scientific approach to analysing numerical data in order to enable us to maximise our interpretation, understanding and use. This means that statistics helps us turn data into information; that is, data that have been interpreted, understood and are useful to the recipient. Put formally, for your project, statistics is the systematic collection and analysis of numerical data, in order to investigate or discover relationships among phenomena so as to explain, predict and control their occurrence. The possibility of confusion comes from the fact that not only is statistics the techniques used on quantitative data, but the same word is also used to refer to the numerical results from statistical analysis.

TWO KINDS OF STATISTICS
Generally, there are two kinds of statistics:
• Descriptive Statistics
• Inferential Statistics

**Descriptive Statistics**
Descriptive statistics are used to describe the basic features of the data in a study. Historically, descriptive statistics began during Roman times when the empire undertook census of births, deaths, marriages and taxes. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. With descriptive statistics, you are simply describing what is or what the data show.

Descriptive statistics are used to present quantitative descriptions in a manageable form. In a research study, we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simplify large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simple summary. For instance, the profit of a firm may describes the general performance of the firm across a wide range of operations.

Descriptive statistics includes the construction of graphs, charts and tables and the calculation of various descriptive measures such as averages (mode, median and means) and measures of variation (e.g.: standard deviations). The purpose of descriptive statistics is to summarise, arrange and present a *set of data* in such a way that facilitates interpretation.

**Inferential Statistics**
Inferential statistics or statistical induction comprises the use of statistics to make inferences concerning some unknown aspect of a population. Inference is the act or process of deriving a conclusion based solely on what one already knows. In other words, you are trying to reach conclusions that extend beyond data obtained from your sample towards what the population might think. Among the widely used inferential statistical tools are t-test, analysis of variance, Pearson’s correlation, linear regression, multiple regression, etc.

**Descriptive or Inferential Statistics**
Descriptive statistics and inferential statistics are interrelated. You must always use techniques of descriptive statistics to organise and summarise the information obtained from a sample before carrying out an inferential analysis. Furthermore, the preliminary descriptive analysis of a sample often reveals features that lead you to the choice of the appropriate inferential method.

As you proceed through this course, you will obtain a more thorough understanding of the principles of descriptive and inferential statistics. You should establish the intent of your study. If the intent of your study is to examine and explore the data obtained for its own intrinsic interest only, the study is descriptive. However, if the information is obtained from a sample of a population and the intent of the study is to use that information to draw conclusions about the population, the study is inferential. Thus, a descriptive study may be performed on a sample as well as on a population. Only when an inference is made about the population, based on data obtained from the sample, does the study become inferential.

**Variables**
Before you can use a statistical tool to analyse data, you need to have data which have been collected. *What is data?* Data is defined as pieces of information which are processed or analysed to enable interpretation. Quantitative data consist of numbers while qualitative data consist of words and phrases. For example, the sales volumes of 50 firms in a year are referred to as data. To explain the sales performance of the 50 firms you need to process or analyse the sales volumes (or data) using a calculator or computer or manually. We collect and analyse data to explain a phenomenon. A phenomenon is explained based on the interaction between two or more variables. The following is an example of a phenomenon:
Educational Level and Attitude Influence Employee Performance

Note that there are THREE variables explaining the particular phenomenon, namely, Educational Level, Attitude and Employee Performance.

**What is a Variable?**

A variable is a construct that is deliberately and consciously invented or adopted for a special scientific purpose. For example, the variable "Educational Level" is a construct based on observation of presumably highly educated and less educated behaviours. Educational level can be specified by using level of schooling and measuring using some tests. Basically, a variable is something that “varies” and has a value. A variable is a symbol to which are assigned numerals of values. A variable can be either a **continuous variable** or **categorical variable**. In the case of the variable "gender" there are only two values, i.e. male and female, and it is called a categorical or **nominal variable**. Other examples of categorical variables are: graduate – non-graduate, low income – high income, citizen – noncitizen. They are also variables which have more than two values. For example, religion may have several values such as Islam, Christianity and Traditional.

Examples of continuous variables are salary, age, profit, size etc. When you use any statistical tool, you should be very clear on which variables have been identified as independent and which dependent variables are.

**Independent Variable**

An independent variable (IV) is the variable that is presumed to cause a change in the dependent variable (DV). The independent variables are the antecedents, while the dependent variable is the consequent. An independent variable can be manipulated. ‘Manipulated’ means the variable can be manoeuvred. Other names for the independent variable are treatment, factor and predictor variable.

**Dependent Variable**

A dependent variable is a variable dependent on other variable(s). The dependent variable in the study “Educational Level and Attitude Influence Employee Performance” is the employee performance which cannot be manipulated by the researcher. Educational level and attitude are the independent variables presumed to cause change in employee performance. Other names for the dependent variable are outcome variable, results variable and criterion variable.

**OPERATIONAL DEFINITION OF VARIABLES**

As mentioned earlier, a variable is “deliberately” constructed for a specific purpose. Hence, a variable used in your study may be different from a variable used in another study even though they have the same name. For example, the variable “employee performance” could different thing to different people unless you further explain what you mean when you say “employee performance”. It can be computed based on employee’s performance in one-off specific task or cumulative response to task or appraisal score from superiors or punctuality, a combination of a number of measures.

An **operational definition** is a detailed specification of how one would go about measuring a given variable in the context of the study. This is done to facilitate measurement and to eliminate confusion. Thus, it is essential that you stipulate clearly how you have defined variables specific to your study. Operational definitions can range from very simple and straightforward to quite complex depending on the nature of the variable and the needs of the researcher.

**Levels of Measurement**

"The level of measurement of a variable in mathematics and statistics is a classification that was proposed in order to describe the nature of information contained within numbers assigned to objects and, therefore, within the variable.” (source: [Babylon online dictionary](https://www.babylon.com)).
Levels of measurement or scales of measure basically refer to different types of variables. This categorisation is important, because it fundamentally affects the meaning of the variables and what we can do with them statistically, as we will see. There are three basic levels of measurement (some authors distinguish 4, but for all practical purposes a distinction of three categories is sufficient), nominal, ordinal and continuous (interval/ratio).

- **Nominal** or categorical variables are measured at the lowest level. These are variables like place of birth, where any numbers we give to the values (e.g. 1 for Akan, 2 for Northerner and 3 for Voltarian) only serve to replace a name. The values cannot be placed in order. We can’t say ‘Voltarian is more than Northerner’, for example; so in this case we can’t say 3 is more than 2. Nominal variables just have categories, which can’t be ordered in any way. Any numbers given are merely a descriptor of that category (e.g. 1 = ‘Akan’).

- **Ordinal** variables do possess a natural ordering of categories. An example of an ordinal variable is a rating scale item such as 1-disagree totally, 2-somewhat disagree, 3-somewhat agree and 4-agree totally. These values can clearly be ordered, in that someone who agrees totally ‘agrees more’ than someone who somewhat agrees, and so on. This is different from the situation we described above with the variable ‘ethnic of birth’. Therefore, ordinal variables allow you to ‘order’ the values given. What you cannot do however is to ‘measure’ exactly the distance between the scale points. When you have a ruler, you know that the distance between 23 and 24 cm is exactly the same as the distance between 10 and 11 cm, i.e. 1 cm. This is not the case when we look at a rating scale. Is the distance between ‘agree totally’ and ‘somewhat agree’ the same as between ‘somewhat agree’ and ‘somewhat disagree’? In order to know this, we would have to find out how people thought about these categories, i.e. are these differences the same or different in respondents’ minds? And does this differ between different respondents? As we cannot know this, we cannot assume that the distance between each scale point is exactly the same like it is for a ruler. All these "agree – disagree" type variables can therefore be considered as ordinal.

- **Interval** variables are variables for which their central characteristic is that they can be measured along a continuum and they have a numerical value (for example, temperature measured in degrees Celsius or Fahrenheit). So the difference between 20°C and 30°C is the same as 30°C to 40°C. However, temperature measured in degrees Celsius or Fahrenheit is NOT a ratio variable.

Interval scales may be either numeric or semantic. Study the examples below.

---

**Figure 3.3 Examples of interval scales in numeric and semantic formats**

<table>
<thead>
<tr>
<th>Characteristics of Service provider</th>
<th>Circle the appropriate score on each line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable</td>
<td>5 4 3 2 1</td>
</tr>
<tr>
<td>Customer friendly</td>
<td>5 4 3 2 1</td>
</tr>
<tr>
<td>Understand needs of customers</td>
<td>5 4 3 2 1</td>
</tr>
<tr>
<td>Give Value for money</td>
<td>5 4 3 2 1</td>
</tr>
<tr>
<td>Attractively packaged</td>
<td>5 4 3 2 1</td>
</tr>
</tbody>
</table>

(a)
Please indicate your views on Balkan Olives by ticking the appropriate responses below:

<table>
<thead>
<tr>
<th>Characteristics of Service provider :</th>
<th>Excellent</th>
<th>Very Good</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customer friendly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understand needs of customers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Give Value for money</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attractively packaged</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Ratio** variables are interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. So, temperature measured in degrees Celsius or Fahrenheit is not a ratio variable because 0°C does not mean there is no temperature. However, temperature measured in Kelvin is a ratio variable as 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever. Other examples of ratio variables include height, mass, distance and many more. The name "ratio" reflects the fact that you can use the ratio of measurements. So, for example, a distance of ten metres is twice the distance of 5 metres.

**CATEGORICAL AND CONTINUOUS VARIABLES**

As mentioned earlier, a variable is a symbol to which numerals or values are assigned. A variable can be either a categorical variable or a continuous variable. Categorical variables are also known as discrete or qualitative variables. Categorical variables can be further categorized as either **nominal**, **ordinal** or **dichotomous** (**dichotomous** variables are nominal variables which have only two categories or levels. For example, if we were looking at gender, we would most probably categorize somebody as either "male" or "female". Also, "Yes" or "No" responses)

**Ambiguities in classifying a type of variable**

In some cases, the measurement scale for data is ordinal, but the variable is treated as continuous. For example, a Likert scale that contains five values - strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree - is ordinal. However, where a Likert scale contains seven or more value - strongly agree, moderately agree, agree, neither agree nor disagree, disagree, moderately disagree, and strongly disagree - the underlying scale is sometimes treated as continuous (although where you should do this is a cause of great dispute).

It is worth noting that how we categorise variables is somewhat of a choice. Whilst we categorised gender as a dichotomous variable (you are either male or female), social scientists may disagree with this, arguing that gender is a more complex variable involving more than two distinctions, but also including measurement levels like genderqueer, intersex and transgender. At the same time, some researchers would argue that a Likert scale, even with seven values, should never be treated as a continuous variable.

**Continuous rating scales**

For continuous rating scale respondents are asked to give a rating by placing a mark at the appropriate position on a continuous line. The scale can be written on card and shown to the respondent during the interview. Two versions of a continuous rating scale are depicted in figure 1.2.
Having been with your service provider for a year, how satisfied are you?

A. 

<table>
<thead>
<tr>
<th>Least Satisfied</th>
<th>very satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

B. 

<table>
<thead>
<tr>
<th>Least Satisfied</th>
<th>very satisfied</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When version B (Visual analogue scale) is used, the respondent's score is determined either by dividing the line into as many categories as desired and assigning the respondent a score based on the category into which his/her mark falls, or by measuring the distance, in millimetres or inches, from either end of the scale. **Whichever of these forms of the continuous scale is used, the results are normally analysed as interval scaled.**

**Why does it matter whether a variable is categorical, ordinal or interval?**

Statistical computations and analyses assume that the variables have specific levels of measurement. For example, it would not make sense to compute an average hair colour. An average of a categorical variable does not make much sense because there is no intrinsic ordering of the levels of the categories. Moreover, if you tried to compute the average of educational experience as defined in the ordinal section above, you would also obtain a nonsensical result. Because the spacing between the four levels of educational experience is very uneven, the meaning of this average would be very questionable. In short, an average requires a variable to be interval. Sometimes you have variables that are "in between" ordinal and interval, for example, a five-point likert scale with values "strongly agree", "agree", "neutral", "disagree" and "strongly disagree". If we cannot be sure that the intervals between each of these five values are the same, then we would not be able to say that this is an interval variable, but we would say that it is an ordinal variable. However, in order to be able to use statistics that assume the variable is interval, we will assume that the intervals are equally spaced.

**Continuous Variables**

If a variable can take on any value between its minimum value and its maximum value, it is called a **continuous variable**; otherwise, it is called a **discrete variable**. Some examples will clarify the difference between discrete and continuous variables.

- Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
- Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.
Can Likert Scale Data ever be Continuous?
by KAREN GRACE-MARTIN

A very common question is whether it is legitimate to use Likert scale data in parametric statistical procedures that require interval data, such as Linear Regression, ANOVA, and Factor Analysis. A typical Likert scale item has 5 to 11 points that indicate the degree of agreement with a statement, such as 1=Strongly Agree to 5=Strongly Disagree. It can be a 1 to 5 scale, 0 to 10, etc.

The issue is that despite being made up of numbers, a Likert scale item is in fact a set of ordered categories.

One camp maintains that as ordered categories, the intervals between the scale values are not equal. Any mean, correlation, or other numerical operation applied to them is invalid. Only nonparametric statistics should be used on Likert scale data (i.e. Jamieson, 2004).

The other group maintains that while technically the Likert scale item is ordered, using it in parametric tests IS valid in some situations. For example, Lubke & Muthen (2004) found that it is possible to find true parameter values in factor analysis with Likert scale data, if assumptions about skewness, number of categories, etc., were met. Likewise, Glass et al. (1972) found that F tests in ANOVA could return accurate p-values on Likert items under certain conditions.

Meanwhile, the debate rages on.

What is a researcher with integrity supposed to do? In the absence of a definitive answer, these are my recommendations:

1. Understand the difference between a Likert type item and a Likert Scale. A true Likert scale, as Likert defined it, is made up of many items, which all measure the same attitude. But many people use the term Likert Scale to refer to a single item. Confusion about what a Likert Scale is, no doubt, has contributed to the debate.

2. Proceed with caution. Research the consequences of using your procedure on Likert scale data from your study design. The fact that everyone uses it is not sufficient justification. There are some circumstances and procedures for which it is more egregious than others.

3. At the very least, insist that the item have at least 5 points (7 is better), that the underlying concept be continuous, and that there be some indication that the intervals between points are approximately equal. Make sure the other assumptions (normality & equal variance of residuals, etc.) be met.

4. When you can, run the nonparametric equivalent to your test. If you get the same results, you can be confident about your conclusions.

5. If you do choose to use Likert data in a parametric procedure, make sure you have strong results before making claims. Use a more stringent alpha level, like .01 or even .005, instead of .05. If you have p-values of .001 or .45, it’s pretty clear what the result is, even if parameter estimates are slightly biased. It’s when p-values are close to .05 that the effect of bending assumptions is unclear.

6. Consider the consequences of reporting inaccurate results. Will anyone ever read your paper? Will your research be published? Will it be used to shape public policy or affect practices? The answers to these questions can inform the seriousness of potential problems.

Common Error: Note
A Likert scale is never an individual item; it is always a set of several items, with specific format features, the responses to which are added or averaged to produce an overall score or measurement.
Independent and Dependent Variables
When you use any statistical tool you should be very clear which variables have been identified as independent variables and which variables are dependent variables.

A researcher conducted a study to find whether there was a difference in satisfaction between male and female teenagers. The two variables studied is Gender and satisfaction and both these can be measured. However, there is a difference between these variables. Gender is the Independent Variable while satisfaction is the Dependent Variable.

a) Independent Variable
- An independent variable (IV) (treatment variable, predictor variables, input variable) is the variable that is presumed can cause a change in the dependent variable. The independent variable is the antecedent while the dependent variable is the consequent.
- The independent variable (Gender) can be manipulated. ‘Manipulated’ means the variable can manoeuvred, and in this case it is divided into ‘male’ and ‘female’. Other examples of independent variables are ethnic background (Ga, Akan, Kusase), socioeconomic status (high, middle, low).

b) Dependent Variable
The dependent variable DV (outcome variable, results variable, output variable, criterion variable) in this study is satisfaction which cannot be manipulated by the researcher. Satisfaction is a score and other examples of dependent variables are attitude (score on an attitude scale), financial performance in history and so forth.

Operational Definition of Variables
As mentioned earlier a variable is “deliberately” constructed for a specific purpose. Hence, a variable used in your study may be different from a variable used in another study even though they have the same name. For example, the variable “performance” used in your study may be computed based on profitability while in another study it may be return on asset. Operational definition (Bridgman, 1927) means that variables used in the study must be defined as it is used in the context of the study. This is done to facilitate measurement and to eliminate confusion. Thus, it is essential that you stipulate clearly how you have defined variables specific to your study.

Sampling
One of the most important decisions that any researcher makes is how to obtain the type of participants needed for the study. Who is in our study and how we sample them are critical aspects of business research. The sample that we draw for our study determines the generalizability of our findings. If we obtain a small sample of rare individuals, then our findings cannot apply to a broader population. We want to draw as unbiased a sample as we can from the population of interest.

Population: In most situations, it is impossible to study an entire population. We typically study a subset of people drawn from a larger population and use inferential statistics to make an inference from the sample back to the population. The validity of that inference depends on how representative the sample or subset is of the population from which it is drawn. Our goal, as researchers, is to obtain the most representative sample that we can. Some sampling strategies can get us pretty close to the population, others have problems that might result in a biased sample. As we learn about different sampling strategies, let's use the population drawn below as our starting point.
As we can see, this is a relatively diverse population. When we draw our sample, we want to have a good representation of all of the kinds of people in the population. Sampling procedures fall into two broad categories:

1. **probability sampling**
2. **nonprobability sampling**

Most probability sampling strategies have a random or chance component, though there are some important exceptions. It is this random component that give us confidence that our sample is a reasonably good representation of the population. This random component can be time-consuming and expensive, so there are other alternatives for selecting a study sample. We will discuss and give examples of each type of sample.

Probability sampling strategies typically use a random or chance process, although there are important exceptions to this rule. Random sampling is a strategy for selecting study participants in which each and every person has an equal and independent chance of being selected. What does it mean to be independent? The researchers select each person for the study separately.

Let's say you were asked to participate in an experiment, enjoyed it, and told your friends to contact the researcher to volunteer for the study. This would be an example of non-independent sampling. We assume that friends share similar values and by recruiting your friends to be in the study, the sample might represent you and your friends but not the whole population of interest. The "equal chance" and "independent" components of random sampling are what makes us confident that the sample has a reasonable chance of representing the population.

**Sampling Frame**

Where do we start? When we use probability sampling, we begin by defining our population. Once we have done this, we must have some sort of record or directory to use to select individual participants from the target population. The sampling frame is the population as it is defined and available through records. If our desired population is every one living in the Ghana, the census that is conducted every 10 years is our best record or directory of the population. Census data would be our sampling frame. If our target population is everyone on a University campus, the campus directory would be our best choice for a sampling frame. If you wanted to take a random sample from your campus, what options are available to you to use as a sampling frame and what problems might you encounter using or obtaining the sampling frame?

**Sample Size**

In order to be able to accurately project the results of a survey question from the sample to the entire population of the target market, the correct sample size must be used. The correct sample size can be calculated using this formula:

\[ n = \frac{Z^2 \times \sigma^2}{E^2} \]

where:

- \( n \) = Sample Size
- \( Z \) = Level of Significance (Expressed as a Z-Score)
- \( \sigma^2 \) = Population Standard Deviation (\( \sigma^2 \) = Population Variance)
- \( E \) = Acceptable Amount of Sampling Error

In the formula, \( n \) represents the sample size needed to achieve a certain level of confidence in the results of the survey. The value of \( Z \) depends on the desired level of confidence, which is typically 0.95 for a 95% confidence level. The value of \( \sigma^2 \) is an estimate of the variance of the population, and \( E \) is the acceptable margin of error. By plugging these values into the formula, we can calculate the required sample size for our study.
The Z-score in this equation can be looked up from a table that shows the probability of a sample error.

The population standard deviation is a number you will need to estimate. This is the most difficult part of determining the sample size. There are several ways to estimate the population standard deviation. The best ways are to use the results from a prior survey or to conduct a small pre-sample survey. A company will often have conducted a prior survey on a subject that relates closely to the issue you are researching. If this is the case, the standard deviation from this survey can be used as an estimate of the population standard deviation. If such a survey does not exist then the best thing to do is take a small sample survey and use the standard deviation from this survey as your estimate.

Another way to develop an estimate of population standard deviation is to use secondary data. It is possible that another firm conducted a similar survey. The results of this survey could be used to develop an estimate. If all else fails, you can always use judgment to estimate the standard deviation. This can be done by obtaining a potential range of answers from management that is in a position to make an educated guess regarding these figures.

A variant of equation 1.1 is sample size (n) for **infinite population** (population>50,000)

\[ n = \frac{Z^2 \times P(1-P)}{C^2} \]

where

- \(Z\) = Z-value (e.g., 1.96 for a 95 percent confidence level)
- \(P\) = Percentage of population picking a choice, expressed as decimal (can be determined from existing studies or pre-sample survey)
- \(C\) = Confidence interval, expressed as decimal (e.g., .04 = +/- 4 percentage points)

Adjustment to sample size of infinite population could be used to obtain the sample size \(n_0\) of **finite population** (population<50,000)

\[ n_0 = \left[ \frac{n}{1 + \frac{n-1}{N}} \right] \]

Where \(N\) is the size of the population.

There are also available sample size determination tables such as the Krejcie and Morgan (1970) table for determining sample size from a given population\(^1\) and the table developed by Bartlett, Kotrlik and Higgins (2001)\(^2\).

**Sampling Strategies/Techniques**

There are a number of probability sampling strategies that can be used that vary in their complexity. They are:

- Simple
- Systematic
- Stratified
- Cluster

**Simple random sampling** is the most straightforward of the random sampling strategies. We use this strategy when we believe that the population is relatively homogeneous for the

---

\(^1\) Available at [https://opa.uprrp.edu/InvInsDocs/KrejcieandMorgan.pdf](https://opa.uprrp.edu/InvInsDocs/KrejcieandMorgan.pdf)

characteristic of interest. For example, let's say you were surveying first-time parents about their attitudes toward mandatory seat belt laws. You might expect that their status as new parents might lead to similar concerns about safety.

**Simple Random Sampling, with Aid of a Spreadsheet Program**

Start with a list of all cases in the population, from which we want to draw a simple random sample of size \( n \).

General procedure: We have a list of names/identifiers for all cases in the population. We will reorder the list into a random order. We will then take the first \( n \) cases in the reordered list. These \( n \) cases will be a simple random sample from the population.

General spreadsheet procedure: We will place the list of names/identifiers for all cases in the population in a column. In the column to the right we will enter a column of random numbers. We now have a list of names, each name having an associated random number. We will now reorder (sort) the list of names and associated random numbers into the order of the random numbers. Since the names are now in a random order, we will take the first \( n \) names as a simple random sample of \( n \) from the population.

Specific spreadsheet steps to implement the procedure:

- Enter the name/identifier for each case in the population in one column. Let each case correspond to a row.
- Use the random number command to generate a column of random numbers, one number beside each of the names. In Excel: Tools, Data Analysis, Random Number Generation, choose uniform random numbers. (Note: If you do not see Data Analysis under the Tools menu, then under Tools/Add-Ins check Analysis TooPak.)
- Sort the names (and their associated random number) according to the random numbers. In Excel: Data, Sort.
- Choose cases from the first case on down to reach the desired sample size.

**Systematic sampling**

Systematic sampling yields a probability sample but it is not a random sampling strategy (it is one of our exceptions). Systematic sampling strategies take every \( n \)th person from the sampling frame. For example, you choose a random start page and take every 45th name in the directory until you have the desired sample size. Its major advantage is that it is much less cumbersome to use than the procedures outlined for simple random sampling.

**Systematic Sampling Using a Random Number Table for a Random Start**

Number the cases in your population sequentially. Divide the number of cases in your population \( (N) \) by the desired sample size \( (n) \) to determine the skip factor \((k)\), i.e. \( k = \frac{N}{n} \). Use a table of random numbers, by pointing to a psuedo-random starting place on the table without looking, to identify a case number between 1 and \( k \) that will be your starting point; if \( k \) is a 1-digit number you will sample a 1-digit number, if \( k \) is a 2-digit number you will sample a 2-digit number, etc. Your systematic sample will consist of the starting case that you identified, say case \( r \), case \( r+k \), case \( r+2k \), etc.

**Stratified Random Sampling**

It is common for people who are conducting statistical analysis to want to obtain information about key subsections of a population. Stratified random sampling is a technique which involves dividing a population or sampling frame into several, non-overlapping ‘strata’ (subgroups) according to a particular characteristic which reflects the variables of interest.
Once the population or sampling frame is divided appropriately, simple random samples would then be selected from within each stratum. This is more preferable than simple random sampling when selecting samples for strata. Examples of usual stratification characteristics are age-group, gender, income bracket and ethnic origin.

For instance, one may be particularly interested in survey responses by the ages of the respondents. If simple random sampling was used to select the sample, it would be possible that one/some groups may be over-represented and one/some may be under-represented. To ensure that adequate numbers of people from each age group are included in the sample, it would be necessary to conduct stratified sampling, with each age-group forming one stratum.

**Proportionate Allocation**

Normally sample sizes are proportionate to the size of the stratum which means that each stratum has the same sampling fraction. The proportionate allocation method of stratification is used to ensure that sample sizes for strata are of their expected size in relation to the population.

**Disproportionate Allocation**

With disproportionate stratification, the sampling fraction will vary from one stratum to another. This method is often used in cases when there is one (or more) minority group(s) within the population which are likely to be particularly under-represented or omitted from a simple random sample, unless specific attention is paid to them. It therefore gives larger than proportionate sample sizes for one or more strata to ensure that separate analyses by sub-group will be possible.

**Cluster sampling**

Cluster sampling involves splitting the population of interest into clusters. These could be geographical areas (e.g. towns, postcode sectors or local authorities) or they could be natural clusters (e.g. industries, schools or hospitals). After the population is divided, several clusters are chosen at random to form the sampling frame. Ideally, the chosen clusters should be dissimilar from one another to ensure that the sample is as representative of the population as possible. Clusters provide a more localised way of conducting the survey and, whilst some may be in different geographical locations, there will be less widespread dispersion and it would be possible to assign one interviewer to each cluster. Two forms of cluster sampling are described below:

**One-stage Cluster Sampling**

This involves splitting the population into suitable clusters, then randomly selecting (via simple random sampling) a proportion of those to be included for further analysis. All units contained within the sample clusters would then be chosen to participate in the survey. For example, the Ghana Government wishes to find out information about the diets of primary pupils under school feeding programme. Clearly we must create a sample, as it would be expensive and time consuming to survey all primary pupils in Ghana. We may decide that each school in Ghana represents one cluster and then select a random sample of 30 schools. In one-stage cluster sampling, we would now visit each of the 30 schools (clusters) and interview all of the primary pupils in each.

**Two-stage cluster sampling**

This involves splitting the population into suitable clusters and, once more, selecting a proportion of those to be included for further analysis. However, in this method, the units within each sample cluster would be subject to a further round of simple random sampling so that only a proportion of units in each cluster would actually be surveyed.
What is Descriptive Statistics?
Descriptive statistics are used to summarise the data collected and its presentation in a way that is clearly understood. For example, a researcher administered an instrument to measure self-esteem among 500 customers. How might these measurements be summarised and presented? There are two basic methods: numerical and graphical.

- Using the numerical approach the researcher might compute the mean, median, mode, standard deviation and variance. Numerical approaches are more precise and objective.
- Using the graphical approach the researcher might create a frequency table, bar-chart, a line graph or a box plot. These graphical methods display detailed information about the distribution of the scores. Graphical methods are better suited for identifying patterns in the data. Since the numerical and graphical approaches compliment each other, it is wise to use both.

Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics you are simply describing what is or what the data shows based on the sample. With inferential statistics, you are trying to reach conclusions based on the sample that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make "inferences" from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simply large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. For instance, consider the Grade Point Average (GPA). This single number describes the general performance of a student across a potentially wide range of course experiences. The single number describes a large number of discrete events such as the grade obtained for each subject taken.

However, every time you try to describe a large set of observations with a single indicator you run the risk of distorting the original data or losing important detail. The GPA doesn't tell you whether the student was in difficult courses or easy ones, or whether they were courses in his/her major field or in other disciplines. Even with these limitations, descriptive statistics provide a powerful summary that may enable comparisons across people or other units.

This section contains 3 main topics related to available techniques to describe and interpret statistic data, using numerical measures of centre, location, and variation.

1. Measures of Center and Location
2. Measures of Variation
3. Describing and comparing measures
Let me summarize what you will find in this chapter: Remember when you learned about nominal and ordinal data? Now we are going to use these concepts to understand what type of measurement is suitable to describe data.

Our first concept is the difference between a parameter and a statistic. When you are measuring data from the entire population, you are calculating a parameter, whereas when measuring data from a sample, you are calculating a statistic (Lind, 2005). It is important to keep these 2 concepts in mind all the time, because you will see them repeatedly through the course.

Types of Measurements
In statistics there are basically 2 types of measurements: a) measures of location and b) measures of variation

This module addresses seven measures of location and six measures of variation. At the end of the chapter you will learn how to integrate these measures in five indicators to reach conclusions about the data. Let’s start summarizing them in the following tables:

<table>
<thead>
<tr>
<th>Measures of location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure of Central Tendency</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Mode</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measures of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range</td>
</tr>
<tr>
<td>Interquartile Range</td>
</tr>
<tr>
<td>Population Variance</td>
</tr>
<tr>
<td>Sample Variance</td>
</tr>
<tr>
<td>Population Deviation</td>
</tr>
<tr>
<td>Sample Deviation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean and standard deviation combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient of Variation for Population</td>
</tr>
<tr>
<td>Coefficient of Variation for Sample</td>
</tr>
<tr>
<td>Empirical rule</td>
</tr>
</tbody>
</table>

MEASURES OF CENTRAL TENDENCY

What is central tendency? Generally, it is a measure of the "average". Let us look at an example to illustrate central tendency. You are in a class of 8 students and your teacher gave a mathematics test consisting of 10 questions. The teacher returns you test paper and you got a score of 5/10. Are you happy with the score of 5 over 10? Are you disappointed? Like most students, you ask your classmates, "What did you get?" You might ask the teacher, "How did the class do?"

In other words, the additional information you are asking is how you score compares to other students' score. If your score of 5 is the highest in the class you will be ecstatic! On the other hand, if your score of 5 is among the lowest score in the class, you will be pretty disappointed. Examine the table to the left which presents 3 possible sets of results:
<table>
<thead>
<tr>
<th>Student A</th>
<th>RESULTS A</th>
<th>RESULTS B</th>
<th>RESULTS C</th>
<th>RESULTS D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>YOURS</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

- Look at **Result A**, all your classmates obtained a score of 5. In other words, your score is *at the exact centre of the distribution*. "You did as well as everyone else or everyone else did just as well as you".
- Look at **Result B**, which will make you happy. All other seven students scored lower than you. In other words, you are *above the centre of the distribution*.
- Look at Result C, which may be depressing. All other seven students score higher than you. In other words, you are *below the centre of the distribution*.
- Look at Result D, which make you both happy and depressed. Three students scored higher than you, three students scored lower than you and one student score similar to you. In other words, you are *somewhere in the middle of the distribution*.

Central tendency is a loosely defined concept that has to do with measuring the location of the middle or the centre of the distribution. There are several measures of central tendency and the three most common are the mean, median and mode.

**A) MEAN**
The mean is the most widely used measure of central tendency in social science research. The formula for computation of the mean is show below:

\[
\bar{X} = \frac{\sum X}{N}
\]

**Computing the Mean**
The mean or X (pronounced as X bar) is the figure obtained when the sum of all the items in the group is divided by the number of items (N). Say for example you have the score of 10 (i.e. N) students on a science test. The sum (\( \sum X \)) of all the ten scores

\[
\sum X = 23 + 22 + 26 + 21 + 30 + 24 + 20 + 27 + 25 + 32 = 250
\]

\[
\bar{X} = \frac{250}{10} = 25
\]

**Mean Influenced by Extreme Score**
In the computation of the mean, every item counts. As a result, extreme values at either end of the group or series of scores severely affects the value of the mean. The mean could be "pulled towards" as a result of the extreme scores which may give a distorted picture of the groups or series of scores or data. However, in general, the mean is a good measure of central tendency for roughly symmetric distributions but can be misleading in *skewed* distributions (see example below) since it can be greatly influenced by extreme scores.

**Example:**
If we were to change the first score from 23 to 75, see what happens to the mean

\[ 75 + 22 + 26 + 21 + 30 + 24 + 20 + 27 + 25 + 32 = 302 \]

The extreme score has changed the mean to 30.2 which may give the impression that students performed better in the test when in fact only one student scored high.

**NOTE:** Keep in mind this characteristic when interpreting the means obtained from a set of data.

**B) MEDIAN**

The Median is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in ascending order, and then locate the score in the center of the sample. For example, if we order the following 7 scores as shown below, we would get:

\[ 12, 18, 22, 25, 30, 37, 40 \]

Score 25 is the median because it represents the halfway point or midpoint for the distribution of scores.

Look at these set of 8 scores. What is the median score?

\[ 15, 15, 15, 18, 20, 21, 25, 36 \]

When there is an even number of scores, the median is the average of the two middle scores (i.e. you interpolate to determine the mean). The fourth score (18) and the fifth score (20) represent the halfway point. Thus, the median is \( 18 + 20 \) divide by 2 = 38/2 = 19.

**B) MODE**

The Mode is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown below, and then count each one.

\[ 15, 15, 15, 20, 20, 21, 25, 36 \]

The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the mode. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

**SHOULD YOU USE THE MEAN OR THE MEDIAN?**

The mean and median are two common measures of central tendencies of a typical score in a sample discussed. Which of these two should you use when describing your data? It depends on your data. In other words, you should ask yourself whether the measure of central tendency you have selected gives a good indication of the typical score in your sample. If you suspect that the measure of central tendency selected does not give a good indication of the typical score, then you most probably have chosen the wrong one.

The MEAN is the most frequently used measure of central tendency and it should be used if you are satisfied that it gives a good indication of the typical score in your sample. However, there is a problem with the mean. Since it uses all the scores in a distribution, it is sensitive to extreme score,
Example:

The Mean for these set of nine scores:

\[
20 + 22 + 25 + 26 + 30 + 31 + 33 + 40 + 42 \quad \text{is} \quad 29.89
\]

If we were to change the last score from 42 to 70, see what happens to the mean

The Mean:

\[
20 + 22 + 25 + 26 + 30 + 31 + 33 + 40 + 70 \quad \text{is} \quad 33.00
\]

Obviously, this mean is not a good indication of the typical score in this set of data. The extreme score has changed the mean from 29.89 to 33.00. If these were test scores, it may give the impression that students performed better in the latter test when in fact only one student scored high.

NOTE: Keep in mind this characteristic when interpreting the means obtained from a set of data.

If you find that you have extreme score and you are unable to use the mean, then you should use the MEDIAN. The median is not sensitive to extreme scores. If you examine the above example, the median is 30 in both distributions. The reason is simply that the median score does not depend on the actual scores themselves beyond putting them in ascending order. So the last score in a distribution could be 80, 150 or 5000 and the median still would not change. It is this insensitivity to extreme scores that make the median useful when you cannot use the mean.

**Skewness and Symmetry**

Now, let’s take a look at other 2 important concepts – Skewed and Symmetric distributions. Data sets are symmetric when their values are evenly spread around the center, and to confirm this, median and mean must be equal. Take a look at the following curve:

![Graph showing skewness and symmetry](image)

When this doesn’t happen, data might be left-skewed distributed \(\rightarrow\) the mean is smaller than (to the left of) the median. Or right-skewed distributed \(\rightarrow\) the mean is larger than (to the right of) the median.

Here is another concept to remember: The mean can be highly affected by extreme values. If one of the observations has very low or very high values, it affects the mean making it lower or higher, respectively.
**Note on skewness**

1) Skewness characterizes the degree of asymmetry of a distribution around its mean.
2) Positive skewness indicates a distribution with an asymmetric tail extending towards more positive values (right skewed)
3) Negative skewness indicates a distribution with an asymmetric tail extending towards more negative values (left skewed)
4) Most often, median is used as a measure of central tendency when data sets are skewed.
5) Normal distributions will have a skewness value of approximately zero.
6) Typically, the skewness value will range from negative 3 to positive 3.
7) **As a rule of thumb, if skewness is more than +/- 1 (more accurately, if the absolute value of skewness is more than twice the standard error of skewness [ses]), consider using median rather than (or along with) mean. But, this is a rule of thumb. Use of particular statistics is at the judgment of a researcher.**
8) Example – sometimes, the existence of super rich people like Bill Gates makes the distribution highly positively skewed, which makes median a better choice for describing the data than mean.
Measures of variation
If all the data are not the same value you have got VARIATION, isn’t it an easy concept? Sometimes 2 data sets may have the same mean, but variation (or behavior) of their observations is different making one set more stable than other.

When measuring variation you may use any of the following 6 measures:

Range
Difference between maximum and minimum value in a data set:
\[ R = \text{Maximum value} - \text{Minimum value} \]

It’s useful when we want to have an idea of what is the general composition of the data, and how apart our maximum and minimum is.

Interquartile Range
Difference between 3rd and 1st quartile. It’s not affected by extreme values, more efficient than Range.
\[ \text{IQR} = \text{Third Quartile} - \text{First Quartile} \]

This range measures the information grouped within 25% and 75% of the data set, leaving out the data above and below the limits. It’s more accurate than the range, but still presents an important weakness: None of these two formulae use all the data for computations. To overcome this difficulty the following measures were created:

Population Variance
This is one of the most common measures of variation in Statistics. Many of the concepts that we will learn in the future regarding probabilities and hypothesis tests, rely on the accurate computation of the variance.

Variance is the average of the squared variations from the mean. As the formula suggests below, it’s necessary to compute the difference between each value and the mean, then square that difference and finally add up all this variations and divide them by the total number of observations.
\[
\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}
\]

Population Standard Deviation
Square root of Variance, explains how spread out a distribution is, and it’s very useful to make comparisons between data sets with the same mean. If distributions have the same mean, the one with the largest standard deviation has the greatest relative spread.
\[
\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}
\]

Sample Variance
The formula is similar to population variance, but notice that the denominator is n-1. The source of the information is a sample and not the entire population. Notice also that the variables are written in lower case and the mean is expressed by x bar and not m.
\[
s^2 = \frac{\sum x^2 - (\sum x)^2}{n - 1}
\]
Combining measurement tools

In this section we will combine the measurements of location and the measurement of variation and use it for applications in business.

**Coefficient of Variation (CV)**

This indicator combines the standard deviation and the mean in a very useful measure that provides information about variation of data sets, when their means are different. There is a coefficient of variation for population and for samples; their only difference is the type of standard deviation used.

**The Empirical Rule**

Combines information about (m) and (s), to explain approximately how much information in your data set is contained within a specific range. This is a very useful indicator for decision makers, because it identifies the outliers or extreme elements of our data. Refer to the table below. The table says that in a normal distribution of values, 68% of the observations will be contained in a range of one standard deviation from the mean. 95% of the observations are within 2 standard deviations from the mean, and virtually all the data values should be within 3 standard deviations from the mean.

### Table 4

<table>
<thead>
<tr>
<th>μ ± 1σ</th>
<th>Contains approx. 68% of the values</th>
</tr>
</thead>
<tbody>
<tr>
<td>μ ± 2σ</td>
<td>Contains approx. 95% of the values</td>
</tr>
<tr>
<td>μ ± 3σ</td>
<td>Contains virtually all of the values</td>
</tr>
</tbody>
</table>

**Sample Standard Deviation**

Square root of Sample variance:

\[
s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{n-1}}
\]

\[
\bar{x} = \text{sample mean}
\]
Note: Frequency distribution must be bell-shaped and symmetric to apply this rule.

Using Statistical Package to compute Descriptive Statistics

I have good news for you, using MSExcel/SPSS makes it easy to compute the descriptive statistics, as easy as 1-2-3.

MS Excel

Now that you have learned the operational part of computing measures of descriptive statistics, we will take a quick look at a tool provided by Excel to help us in this process. It’s the Data Analysis option. We can request a Summary Statistics report using the following commands:

a. Open Excel  
b. Click on TOOLS menu  
c. Click on DATA ANALYSIS  
d. Click on DESCRIPTIVE STATISTICS  
e. Follow the prompts and select the range with the data you want to input  
f. Click on SUMMARY STATISTICS  
g. Click OK

A table with all the information about: Mean, Median, Mode, Standard Deviation, Sample Variance, Range, Minimum, Maximum, Sum, Count will appear. You will be able to compare samples, populations and make a more informed decision about the variation or stability of the data set.

<table>
<thead>
<tr>
<th>Revenue in dollars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Mode</td>
</tr>
<tr>
<td>Standard Deviation</td>
</tr>
<tr>
<td>Sample Variance</td>
</tr>
<tr>
<td>Kurtosis</td>
</tr>
<tr>
<td>Skewness</td>
</tr>
<tr>
<td>Range</td>
</tr>
<tr>
<td>Minimum</td>
</tr>
<tr>
<td>Maximum</td>
</tr>
<tr>
<td>Sum</td>
</tr>
</tbody>
</table>
SPSS
Method 1:
Analyze → Descriptive statistics → Descriptives
Send over the variables you want to examine
Click on the “Options” box
You can choose to calculate the mean, SD, variance, range, minimum and maximum score, standard error of the mean, as well as the kurtosis and skewness of the distribution
Hit “Continue” then “Ok”
An output file will be produced listing the information

<table>
<thead>
<tr>
<th></th>
<th>Adult illiteracy</th>
<th>level of intolerance for corruption</th>
<th>average rate of inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>180</td>
<td>180</td>
<td>180</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>27.712</td>
<td>4.3806</td>
<td>9.395</td>
</tr>
<tr>
<td>Median</td>
<td>16.000</td>
<td>3.4000</td>
<td>7.000</td>
</tr>
<tr>
<td>Mode</td>
<td>2.0</td>
<td>1.80a</td>
<td>4.5</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>26.8967</td>
<td>2.59234</td>
<td>8.3270</td>
</tr>
<tr>
<td>Variance</td>
<td>723.434</td>
<td>6.720</td>
<td>69.340</td>
</tr>
<tr>
<td>Skewness</td>
<td>.613</td>
<td>.597</td>
<td>2.619</td>
</tr>
<tr>
<td>Std. Error of Skewness</td>
<td>.181</td>
<td>.181</td>
<td>.181</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.052</td>
<td>-1.139</td>
<td>8.276</td>
</tr>
<tr>
<td>Std. Error of Kurtosis</td>
<td>.360</td>
<td>.360</td>
<td>.360</td>
</tr>
<tr>
<td>Range</td>
<td>84.5</td>
<td>7.90</td>
<td>41.9</td>
</tr>
<tr>
<td>Minimum</td>
<td>.5</td>
<td>1.30</td>
<td>2.5</td>
</tr>
<tr>
<td>Maximum</td>
<td>85.0</td>
<td>9.20</td>
<td>44.4</td>
</tr>
<tr>
<td>Percentiles 25</td>
<td>3.400</td>
<td>2.2000</td>
<td>4.100</td>
</tr>
<tr>
<td>50</td>
<td>16.000</td>
<td>3.4000</td>
<td>7.000</td>
</tr>
<tr>
<td>75</td>
<td>48.000</td>
<td>6.9000</td>
<td>12.000</td>
</tr>
</tbody>
</table>

a. Multiple modes exist. The smallest value is shown

Method 2:
Analyze → Descriptive statistics → Explore.
Send over the (dependent) variables you want to examine into the “Dependent List”. If you have a grouping (independent) variable, you can send it over to “Factor List”
Click on the “Statistics” box and check which summary measures you want.
Click on the “Plot” box and check which type of graphs you want.
Click “Continue” then “Ok.”
## Descriptives

<table>
<thead>
<tr>
<th></th>
<th>Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adult illiteracy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>27.712</td>
<td>2.0048</td>
</tr>
<tr>
<td>95% Confidence</td>
<td>Lower Bound</td>
<td>23.756</td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>Upper Bound</td>
<td>31.668</td>
</tr>
<tr>
<td>Median</td>
<td>16.000</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>723.434</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>26.8967</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>.5</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>85.0</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>84.5</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>44.6</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>.613</td>
<td>.181</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.052</td>
<td>.360</td>
</tr>
<tr>
<td><strong>Level of intolerance for corruption</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.3806</td>
<td>.19322</td>
</tr>
<tr>
<td>95% Confidence</td>
<td>Lower Bound</td>
<td>3.9993</td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>Upper Bound</td>
<td>4.7618</td>
</tr>
<tr>
<td>Median</td>
<td>3.4000</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>6.720</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>2.59234</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>1.30</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>9.20</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td>7.90</td>
<td></td>
</tr>
<tr>
<td>Interquartile Range</td>
<td>4.70</td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>.597</td>
<td>.181</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-1.139</td>
<td>.360</td>
</tr>
<tr>
<td><strong>Average rate of inflation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>9.395</td>
<td>.6207</td>
</tr>
<tr>
<td>95% Confidence</td>
<td>Lower Bound</td>
<td>8.170</td>
</tr>
<tr>
<td>5% Trimmed Mean</td>
<td>Upper Bound</td>
<td>10.620</td>
</tr>
<tr>
<td>Median</td>
<td>7.000</td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>69.340</td>
<td></td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>8.3270</td>
<td></td>
</tr>
<tr>
<td>Minimum</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Maximum</td>
<td>44.4</td>
<td></td>
</tr>
</tbody>
</table>
Levels of Measurement and Numerical Measures
After learning different numerical measures of central tendency and dispersion, we should probably know that some measures are appropriate for quantitative variables, but not for qualitative ones. For instance, it is senseless to calculate the mean or the standard deviation for a nominal variable, like sex or region.
However, those measures that can be applied to qualitative variables are also applicable to quantitative variables, though it may not be useful. Certainly mode is appropriate for qualitative variables, but we can also use mode to describe an interval or a ratio variable. The following table gives us some hints about appropriate measures for different levels of measurement, when we do our univariate analysis:

<table>
<thead>
<tr>
<th>Measurement Level</th>
<th>Statistical measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Mode, Frequencies, Percentage Frequencies</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Median, Range</td>
</tr>
<tr>
<td>Interval / ratio</td>
<td>Mean, Standard Deviation</td>
</tr>
</tbody>
</table>