**UNIT-IV**

**Data Processing: Editing, Coding, Tabulating**

After collecting data, the method of converting raw data into meaningful statement; includes data processing, data analysis, and data interpretation and presentation.

Data reduction or processing mainly involves various manipulations necessary for preparing the data for analysis. The process (of manipulation) could be manual or electronic. It involves editing, categorizing the open-ended questions, coding, computerization and preparation of tables and diagrams.

**Editing data:**

Information gathered during data collection may lack uniformity. Example: Data collected through questionnaire and schedules may have answers which may not be ticked at proper places, or some questions may be left unanswered. Sometimes information may be given in a form which needs reconstruction in a category designed for analysis, e.g., converting daily/monthly income in annual income and so on. The researcher has to take a decision as to how to edit it.

Editing also needs that data are relevant and appropriate and errors are modified. Occasionally, the investigator makes a mistake and records and impossible answer. “How much red chilies do you use in a month” The answer is written as “4 kilos”. Can a family of three members use four kilo chilies in a month? The correct answer could be “0.4 kilo”.

Care should be taken in editing (re-arranging) answers to open-ended questions. Example: Sometimes “don’t know” answer is edited as “no response”. This is wrong. “Don’t know” means that the respondent is not sure and is in a double mind about his reaction or considers the questions personal and does not want to answer it. “No response” means that the respondent is not familiar with the situation/object/event/individual about which he is asked.

**Coding of data:**

Coding is translating answers into numerical values or assigning numbers to the various categories of a variable to be used in data analysis. Coding is done by using a code book, code sheet, and a computer card. Coding is done on the basis of the instructions given in the codebook. The code book gives a numerical code for each variable.

Now-a-days, codes are assigned before going to the field while constructing the questionnaire/schedule. Pose data collection; pre-coded items are fed to the computer for processing and analysis. For open-ended questions, however, post-coding is necessary. In such cases, all answers to open-ended questions are placed in categories and each category is assigned a code.

Manual processing is employed when qualitative methods are used or when in quantitative studies, a small sample is used, or when the questionnaire/schedule has a large number of open-ended questions, or when accessibility to computers is difficult or inappropriate. However, coding is done in manual processing also.

Data classification/distribution:

Sarantakos (1998: 343) defines distribution of data as a form of classification of scores obtained for the various categories or a particular variable. There are four types of distributions:

1. Frequency distribution
2. Percentage distribution
3. Cumulative distribution
4. Statistical distributions

1. **Frequency distribution:**

In social science research, frequency distribution is very common. It presents the frequency of occurrences of certain categories. This distribution appears in two forms:

Ungrouped: Here, the scores are not collapsed into categories, e.g., distribution of ages of the students of a BJ (MC) class, each age value (e.g., 18, 19, 20, and so on) will be presented separately in the distribution.

Grouped: Here, the scores are collapsed into categories, so that 2 or 3 scores are presented together as a group. For example, in the above age distribution groups like 18-20, 21-22 etc., can be formed)

**2. Percentage distribution:**

It is also possible to give frequencies not in absolute numbers but in percentages. For instance instead of saying 200 respondents of total 2000 had a monthly income of less than Rs. 500, we can say 10% of the respondents have a monthly income of less than Rs. 500.

**3. Cumulative distribution:**

It tells how often the value of the random variable is less than or equal to a particular reference value.

**4. Statistical data distribution:**

In this type of data distribution, some measure of average is found out of a sample of respondents. Several kind of averages are available (mean, median, mode) and the researcher must decide which is most suitable to his purpose. Once the average has been calculated, the question arises: how representative a figure it is, i.e., how closely the answers are bunched around it. Are most of them very close to it or is there a wide range of variation?

**Tabulation of data:**

After editing, which ensures that the information on the schedule is accurate and categorized in a suitable form, the data are put together in some kinds of tables and may also undergo some other forms of statistical analysis.

Table can be prepared manually and/or by computers. For a small study of 100 to 200 persons, there may be little point in tabulating by computer since this necessitates putting the data on punched cards. But for a survey analysis involving a large number of respondents and requiring cross tabulation involving more than two variables, hand tabulation will be inappropriate and time consuming.

**Usefulness of tables:**

Tables are useful to the researchers and the readers in three ways:

1. The present an overall view of findings in a simpler way.
2. They identify trends.
3. They display relationships in a comparable way between parts of the findings.

By convention, the dependent variable is presented in the rows and the independent variable in the columns.

**Problems in processing**

The problem concerning “Don’t know” (or DK) responses: While processing the data, the researcher often comes across some responses that are difficult to handle. One category of such responses may be ‘Don’t Know Response’ or simply DK response. When the DK response group is small, it is of little significance. But when it is relatively big, it becomes a matter of major concern in which case the question arises: Is the question which elicited DK response useless? The answer depends on two points viz., the respondent actually may not know the answer or the researcher may fail in obtaining the appropriate information. In the first case the concerned question is said to be alright and DK response is taken as legitimate DK response. But in the second case, DK response is more likely to be a failure of the questioning process.

How DK responses are to be dealt with by researchers? The best way is to design better type of questions. Good rapport of interviewers with respondents will result in minimising DK responses. But what about the DK responses that have already taken place? One way to tackle this issue is to estimate the allocation of DK answers from other data in the questionnaire. The other way is to keep DK responses as a separate category in tabulation where we can consider it as a separate reply category if DK responses happen to be legitimate, otherwise we should let the reader make his own decision. Yet another way is to assume that DK responses occur more or less randomly and as such we may distribute them among the other answers in the ratio in which the latter have occurred. Similar results will be achieved if all DK replies are excluded from tabulation and that too without inflating the actual number of other responses.

Use or percentages: Percentages are often used in data presentation for they simplify numbers, reducing all of them to a 0 to 100 range. Through the use of percentages, the data are reduced in the standard form with base equal to 100 which fact facilitates relative comparisons. While using percentages, the following rules should be kept in view by researchers:

1. Two or more percentages must not be averaged unless each is weighted by the group size from which it has been derived.
2. Use of too large percentages should be avoided, since a large percentage is difficult to understand and tends to confuse, defeating the very purpose for which percentages are used.
3. Percentages hide the base from which they have been computed. If this is not kept in view, the real differences may not be correctly read.
4. Percentage decreases can never exceed 100 per cent and as such for calculating the percentage of decrease, the higher figure should invariably be taken as the base.
5. Percentages should generally be worked out in the direction of the causal-factor in case of two-dimension tables and for this purpose we must select the more significant factor out of the two given factors as the causal factor.

It is an operation performed over data (raw facts ) to manipulate and convert the data into meaningful information.

Data processing involves 3 activities:

**1. Input**

In this process, the data that is collected is transformed into a form that the computer can understand.

It is the most important step because the output or the result depends completely on the data that is provided as the input.

The data input process involves the following set of activities:

* **Collection:**In this, we gather the raw data from various data sources and prepare it for the input process.
* **Encoding:**In this process, we convert the collected data into a form that it becomes easier to put in a data processing system.
* **Transmission:**In this stage, we send the input data to the various processors and carry it across various components.
* **Communication:**A set of activities which allows the sending of data from one data processing system to another.

**2. Process**

It is the process of transforming raw data into information by performing some actual data manipulation techniques.

The techniques that we use in the process stage are as follows:

* **Classification:**In this stage, we classify the data into different groups and subgroups so that it is easier to handle the data.
* **Storing:**In the storage technique, we store the data in an arranged order so that we can access the data quickly whenever we need it.
* **Calculation:**We apply this technique to the numeric data to calculate the required output from the raw figures.

**3. Output**

The information that we get as a result of the data processing is the output. We can present the output information in a visual form and can take further actions or make some decision on the basis of the output.

**Challenges in Data Processing**

Until now we understand the difference between the data and the information. Moreover, we came to know how data processing works.

But, it is not as simple as it appears. There are several challenges that come while processing the data. Let me put some light on the key challenges that appear while processing the data.

**1. Collection of Data**

The very first challenge in data processing comes in the collection or acquisition of the correct data for the input. We have the following data sources from which we can acquire data:

* Administrative data sources
* Mobile and website data
* Social media
* support calls
* Statistical surveys
* Census
* Purchasing data from third parties

There are many more, sometimes, the data collection agent walks door to door to collect the data that we need.

The challenge here is to collect the exact data to get the proper result. As the result directly depends on the input data. Hence, it is vital to collect the correct and exact data to get the desired result.

Choosing the right data collection technique can help to overcome this challenge. Below are the 4 different data collection techniques:

* **Observation:**Making direct observation is a quick and effective way to collect simple data with minimal intrusion.
* **Questionnaire:**Survey can be carried out to every corner of the globe and with this, the researcher can structure and precisely formulate the data collection plan.
* **Interview:**is the most suitable technique to interpret and understand the respondents.
* **Focus group session:**The presence of several relevant people simultaneously debating on the topic gives the researcher a chance to view both sides of the coin and build a balanced perspective.

**2. Duplicacy of Data**

As the data is collected from different data sources, then many times it happens that there is duplicacy in data. The same entries and entities may present a number of times during the data encoding stage. This duplicate data is redundant and may produce an incorrect result.

Hence, we need to check the data for duplicacy and proactively remove the duplicate data.

Data **deduplication**is adapted to reduce the cost and free the storage space. Deduplication technology of data identifies the identical data blocks and eliminates the redundant data.
This technique significantly reduces the size of disk usage and also reduces the disk IO traffic. Hence, it enhances the processing performance and helps in achieving precise and high accuracy data.

**3. Inconsistency of data**

When we collect a huge amount of data and there is no guarantee that the data would be complete or all the fields that we need are filled correctly. Moreover, the data may be ambiguous.

As the input/raw data is heterogeneous in nature and is collected from autonomous data sources, the data may conflict with each other in three different levels:

* **Schema Level:**Different data sources have different data models and different schemas within the same data model.
* **Data representation level:**Data in different sources are represented in different structures, languages, and measurements.
* **Data value level:**Sometimes, the same data objects have factual discrepancies among various data sources. This occurs when we obtain two data objects from different sources and they are identified as versions of each other. But, the value corresponding to their attributes differ.

In this situation, we need to check for the completeness of the data. Also, we have to see the dependency and importance of the field (inconsistent field) to the desired result. furthermore, we need to proactively figure out bugs to ensure the consistency in the database.

**4. Variety of data**

The input data, as it is collected from different sources, can contain different forms. The rows and columns of a relational database don’t limit the data. The data varies from application to application and source to source. Much of these data is unstructured and cannot fit into a spreadsheet or a relational database.

There may be that the collected data is in text or tabular format. On the other hand, it may be a collection of photographs and videos and sometimes maybe just audio.

Sometimes to get the desired result, there is a need to process different forms of data altogether.

There are different techniques for resolving and managing data variety, some of them are as follows:

* **Indexing:**Different and incompatible data types can be related together with the indexing technique.
* **Data profiling:**This technique helps in identifying the abnormalities and interrelationship between the different data sources.
* **Metadata:**Meta description of data and its management helps in achieving contextual consistency in the data.
* **Universal format conversion:**In this technique, we can convert the collected data into a universally accepted format such as Extensible markup language (XML).

**5. Data Integration**

Data integration means to combine the data from various sources and present it in a unified view.

With the increased variety of data and different formats of data, the challenge to integrate the data enlarges.

The data integration consists of various challenges that are as follows:

* **Isolation:**Majority of the applications are developed and deployed in isolation which makes it difficult to integrate the data across various applications.
* **Technological Advancements:**With the advancement in the technology, the ways to store and retrieve data changes. The problem here occurs with the integration of newer data to the legacy data.
* **Data Problems:**The challenge in data integration rises when the data is incorrect, incomplete or is of the wrong format.

Then we have to figure the right approach to integrate the data so that the data remains consistent.

There are mainly three techniques for integrating data:

* **Consolidation:**It captures data from multiple sources and integrates it into a single persistent data store.
* **Federation:**It gives a single virtual view of multiple data sources. When it fires a query it returns data from the most appropriate data source.
* **Propagation:**Data propagation applications copies data from one source to another. Furthermore, it guarantees a two-way data exchange regardless of the type of data synchronization.

**6. Volume and Storage of data**

When processing big data, the volume of the data is considerably large. Big data consists of both structured and unstructured data. This includes the data available on the social networking sites, records of companies, data from surveillance sources, research and development data and much more. Here comes the challenge to store and manage this sheer volume of data. Also what amount of data is to present to the RAM so that the processing is faster and the resources utilization is smart.

Also, we need to back up the data to ensure the data protection from any sort of loss. The data loss could occur due to software or hardware loss, natural disaster or error made by humans.

Now, the data itself is huge in volume and we need to take the copy or backup of the data for safety. This increases the amount of stored data by up to 150% or even more.

Below are the possible approaches that we may use to store a large amount of data:

* **Object storage:**with this approach, it is easier to store very large sets of data.  It is a replacement of traditional tree-like file system.
* **Scale-out NAS:**is capable of scaling the capacity of the storage. It usually has its own distributed or clustered file system.
* **Distributed nodes:**most often the low-cost commodity implements this. It attaches directly to the computer server or even server memory.

**7. Poor description and Meta Data**

One of the major sources of the input data is the data that are stored over the time in a relational database. But this data is not properly formatted and there is no meta description of the storage, structure and the relation of the data entities with each other.

The scenario even becomes worse when the amount of data is large and the database itself links to other databases. Without a proper documentation of the database, it is quite difficult to extract the correct input data from the databases.

* De-normalize the database for querying purpose.
* Use the stored procedure to allow complex data management task.
* Using NoSQL database for storing data

**8. Modification of Network data**

The data is distributed and simultaneously related to each other in a complex structure. The challenge here is to modify the structure of the data or add some data in this.

The internet is a network that consists of a variety of data, a lot of applications and websites generate data that are all of different forms and characteristics. **Schema** interconnects all of them.

A *schema* is the definition of the indexes, packages, table/rows and meta-data of a *database.*

It is difficult to transport data if a database doesn’t handle *Schema.*

Server Data Tools (SDT) includes a schema compare utility that we can use to compare two database definitions. The SDT can compare any combination of source and target *databases*.
Moreover, It also reports any discrepancies between schemas and detects mismatching data types and defaults of columns.

**9. Security**

Security plays the most important role in the data field. Hacking the data might result in a leak. Hence, it may cost highly to the data processing firm. The hacker might even change or delete the data that we have acquired and processed after a lot of struggle.

The reasons for the security breach in a database are mainly due to these reasons:

* Most of the data processing systems have a single level of protection.
* No *encryption*of Either the raw data or the result/ output data.
* Access of the data to unethical IT professional that risks in data loss.

To ensure the security of the data we should follow the below-mentioned practices:

* Do not connect to public networks.
* Keep personal information safe and secure with a strong password.
* Limit the access of humans to the data
* Encrypt and back up the data

**10. Cost**

Cost is the matter of consideration. When the amount of the data increases then the cost in each stage of the data processing increases gradually.

The cost of data processing depends on the following factors:

* The type of the processing data
* Turn around time to complete the processing of data and get the required result.
* The accuracy of the data.
* Workforce working on data processing.

The stakeholders or the management looking into the data processing must consider the budget and the expenses. Compressing the data reduces its size and thus the data occupy less disk space. With a proper planning of the costs and expenses, the firm could earn well with the data processing.

**Data Analysis: Editing, Coding, Tabular Representation of Data**

**Data Analysis** is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. In today’s business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation.

**EDITING**

EDITING is the process of checking and adjusting responses in the completed questionnaires for omissions, legibility, and consistency and readying them for coding and storage.

**Purpose of Editing**

Purpose of Editing For consistency between and among responses. For completeness in responses– to reduce effects of item non-response. To better utilize questions answered out of order. To facilitate the coding process.

**Basic Principles of Editing**

1. Checking of the no. of Schedules / Questionnaire)
2. Completeness (Completed in filling of questions)
3. Legibility.
4. To avoid Inconstancies in answers.
5. To Maintain Degree of Uniformity.
6. To Eliminate Irrelevant Responses.

**Types of Editing**

1. Field Editing

Preliminary editing by a field supervisor on the same day as the interview to catch technical omissions, check legibility of handwriting, and clarify responses that are logically or conceptually inconsistent.

1. Office Editing

Editing performed by a central office staff; often done more rigorously than field editing.

**CODING**

The process of identifying and classifying each answer with a numerical score or other character symbol. The numerical score or symbol is called a code, and serves as a rule for interpreting, classifying, and recording data.  Identifying responses with codes is necessary if data is to be processed by computer.

Coded data is often stored electronically in the form of a data matrix – a rectangular arrangement of the data into rows (representing cases) and columns (representing variables) The data matrix is organized into fields, records, and files:

Field: A collection of characters that represents a single type of data.

Record: A collection of related fields, i.e., fields related to the same case (or respondent).

File: A collection of related records, i.e. records related to the same sample.

**Tabular Representation of Data**

Presentation of data is of utter importance nowadays. After all everything that’s pleasing to our eyes never fails to grab our attention. Presentation of data refers to an exhibition or putting up data in an attractive and useful manner such that it can be easily interpreted.

**Tabular Representation**

A table facilitates representation of even large amounts of data in an attractive, easy to read and organized manner. The data is organized in rows and columns. This is one of the most widely used forms of presentation of data since data tables are easy to construct and read.

**Components of Data Tables**

* Table Number: Each table should have a specific table number for ease of access and locating. This number can be readily mentioned anywhere which serves as a reference and leads us directly to the data mentioned in that particular table.
* Title: A table must contain a title that clearly tells the readers about the data it contains, time period of study, place of study and the nature of classification of data.
* Headnotes: A headnote further aids in the purpose of a title and displays more information about the table. Generally, headnotes present the units of data in brackets at the end of a table title.
* Stubs: These are titles of the rows in a table. Thus a stub display information about the data contained in a particular row.
* Caption: A caption is the title of a column in the data table. In fact, it is a counterpart if a stub and indicates the information contained in a column.
* Body or field: The body of a table is the content of a table in its entirety. Each item in a body is known as a ‘cell’.
* Footnotes: Footnotes are rarely used. In effect, they supplement the title of a table if required.
* Source: When using data obtained from a secondary source, this source has to be mentioned below the footnote.

**Construction of Data Tables**

There are many ways for construction of a good table. However, some basic ideas are:

* The title should be in accordance with the objective of study: The title of a table should provide a quick insight into the table.
* Comparison: If there might arise a need to compare any two rows or columns then these might be kept close to each other.
* Alternative location of stubs: If the rows in a data table are lengthy, then the stubs can be placed on the right-hand side of the table.
* Headings: Headings should be written in a singular form. For example, ‘good’ must be used instead of ‘goods’.
* Footnote: A footnote should be given only if needed.
* Size of columns: Size of columns must be uniform and symmetrical.
* Use of abbreviations: Headings and sub-headings should be free of abbreviations.
* Units: There should be a clear specification of units above the columns.

**The Advantages of Tabular Representation**

* Ease of representation: A large amount of data can be easily confined in a data table. Evidently, it is the simplest form of data presentation.
* Ease of analysis: Data tables are frequently used for statistical analysis like calculation of central tendency, dispersion etc.
* Helps in comparison: In a data table, the rows and columns which are required to be compared can be placed next to each other. To point out, this facilitates comparison as it becomes easy to compare each value.
* Economical: Construction of a data table is fairly easy and presents the data in a manner which is really easy on the eyes of a reader. Moreover, it saves time as well as space.

**Hypothesis: Framing Null Hypothesis and Alternative Hypothesis**

**Hypothesis**

A hypothesis (plural: hypotheses), in a scientific context, is a testable statement about the relationship between two or more variables or a proposed explanation for some observed phenomenon. In a scientific experiment or study, the hypothesis is a brief summation of the researcher’s prediction of the study’s findings, which may be supported or not by the outcome. Hypothesis testing is the core of the scientific method.

The researcher’s prediction is usually referred to as the alternative hypothesis, and any other outcome as the null hypothesis — basically, the opposite outcome to what is predicted. (However, the terms are reversed if the researchers are predicting no difference or change, hypothesizing, for example, that the incidence of one variable will not increase or decrease in tandem with the other.) The null hypothesis satisfies the requirement for falsifiability: the capacity for a proposition to be proven false, which some schools of thought consider essential to the scientific method. According to others, however, testability is adequate, on the grounds that if there is sufficient support for a hypothesis it is not necessary to be able to conceive of a contrary outcome.

**Framing Null Hypothesis**

The null hypothesis is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups. Testing (accepting, approving, rejecting, or disproving) the null hypothesis—and thus concluding that there are or are not grounds for believing that there is a relationship between two phenomena (e.g. that a potential treatment has a measurable effect)—is a central task in the modern practice of science; the field of statistics gives precise criteria for rejecting a null hypothesis.

**A null hypothesis is a precise statement about a population that we try to reject with sample data.**

We don’t usually believe our null hypothesis (or H0) to be true. However, we need some exact statement as a starting point for statistical significance testing.



**Null Hypothesis Examples**

Often -but not always- the null hypothesis states there is no association or difference between variables or subpopulations. Like so, some typical null hypotheses are:

* The correlation between frustration and aggression is zero (correlation-analysis);
* The average income for men is similar to that for women (independent samples t-test);
* Nationality is (perfectly) unrelated to music preference (chi-square independence test);
* The average population income was equal over 2012 through 2016 (repeated measures ANOVA).

**“Null” Does Not Mean “Zero”**

A common misunderstanding is that “null” implies “zero”. This is often but not always the case. For example, a null hypothesis may also state that

**The correlation between frustration and aggresion is 0.5.**

No zero involved here and -although somewhat unusual- perfectly valid.

The “null” in “null hypothesis” derives from “nullify”: the null hypothesis is the statement that we’re trying to refute, regardless whether it does (not) specify a zero effect.

**Null Hypothesis – Limitations**

Thus far, we only concluded that the population correlation is probably not zero. That’s the only conclusion from our null hypothesis approach and it’s not really that interesting.

What we really want to know is the population correlation. Our sample correlation of 0.25 seems a reasonable estimate. We call such a single number a point estimate.

Now, a new sample may come up with a different correlation. An interesting question is how much our sample correlations would fluctuate over samples if we’d draw many of them. The figure below shows precisely that, assuming our sample size of N = 100 and our (point) estimate of 0.25 for the population correlation.



**Framing Alternative Hypothesis**

An alternative hypothesis is one in which a difference (or an effect) between two or more variables is anticipated by the researchers; that is, the observed pattern of the data is not due to a chance occurrence. This follows from the tenets of science, in which empirical evidence must be found to refute the null hypothesis before one can claim support for an alternative hypothesis (i.e. there is in fact a reliable difference or effect in whatever is being studied). The concept of the alternative hypothesis is a central part of formal hypothesis testing.

An alternative hypothesis states that there is statistical significance between two variables. In the earlier example, the two variables are Mentos and Diet Coke. The alternative hypothesis is the hypothesis that the researcher is trying to prove. In the Mentos and Diet Coke experiment, Arnold was trying to prove that the Diet Coke would explode if he put Mentos in the bottle. Therefore, he proved his alternative hypothesis was correct.

The alternative hypothesis is generally denoted as H1. It makes a statement that suggests or advises a potential result or an outcome that an investigator or the researcher may expect. It has been categorized into two categories: directional alternative hypothesis and non-directional alternative hypothesis.

**Key Differences between Null and Alternative Hypothesis**

The important points of differences between null and alternative hypothesis are explained as under:-

1. A null hypothesis is a statement, in which there is no relationship between two variables. An alternative hypothesis is a statement; that is simply the inverse of the null hypothesis, i.e. there is some statistical significance between two measured phenomenon.
2. A null hypothesis is what, the researcher tries to disprove whereas an alternative hypothesis is what the researcher wants to prove.
3. A null hypothesis represents, no observed effect whereas an alternative hypothesis reflects, some observed effect.
4. If the null hypothesis is accepted, no changes will be made in the opinions or actions. Conversely, if the alternative hypothesis is accepted, it will result in the changes in the opinions or actions.
5. As null hypothesis refers to population parameter, the testing is indirect and implicit. On the other hand, the alternative hypothesis indicates sample statistic, wherein, the testing is direct and explicit.
6. A null hypothesis is labelled as H0 (H-zero) while an alternative hypothesis is represented by H1 (H-one).
7. The mathematical formulation of a null hypothesis is an equal sign but for an alternative hypothesis is not equal to sign.
8. In null hypothesis, the observations are the outcome of chance whereas, in the case of the alternative hypothesis, the observations are an outcome of real effect.

**Conclusion**

There are two outcomes of a statistical test, i.e. first, a null hypothesis is rejected and alternative hypothesis is accepted, second, null hypothesis is accepted, on the basis of the evidence. In simple terms, a null hypothesis is just opposite of alternative hypothesis.

**Chi – Square Test**

**Chi – Square Test**

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

1. Achi-square goodness of fit test determines if a sample data matches a population. For more details on this type, see: Goodness of Fit Test.
2. A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
	* A very small chi square test statisticmeans that your observed data fits your expected data extremely well. In other words, there is a relationship.
	* A very large chi square test statistic means that the data does not fit very well. In other words, there isn’t a relationship.



The subscript “c” are the degrees of freedom. “O” is your observed value and E is your expected value. It’s very rare that you’ll want to actually *use* this formula to find a critical chi-square value by hand. The summation symbol means that you’ll have to perform a calculation for every single data item in your data set. As you can probably imagine, the calculations can get very, very, lengthy and tedious. Instead, you’ll probably want to use technology:

* Chi Square Test in SPSS.
* Chi Square P-Value in Excel.

A chi-square statistic is one way to show a relationship between two categorical variables. In statistics, there are two types of variables: numerical (countable) variables and non-numerical (categorical) variables. The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.

There are a few variations on the chi-square statistic. Which one you use depends upon how you collected the data and which hypothesis is being tested. However, all of the variations use the same idea, which is that you are comparing your expected values with the values you actually collect. One of the most common forms can be used for contingency tables:



Where O is the observed value, E is the expected value and “i” is the “ith” position in the contingency table.

A low value for chi-square means there is a high correlation between your two sets of data. In theory, if your observed and expected values were equal (“no difference”) then chi-square would be zero — an event that is unlikely to happen in real life. Deciding whether a chi-square test statistic is large enough to indicate a statistically significant difference isn’t as easy it seems. It would be nice if we could say a chi-square test statistic >10 means a difference, but unfortunately that isn’t the case.

You could take your calculated chi-square value and compare it to a critical value from a chi-square table. If the chi-square value is more than the critical value, then there is a significant difference.

You could also use a p-value. First state the null hypothesis and the alternate hypothesis. Then generate a chi-square curve for your results along with a p-value (See: Calculate a chi-square p-value Excel). Small p-values (under 5%) usually indicate that a difference is significant (or “small enough”).

Tip: The Chi-square statistic can only be used on numbers. They can’t be used for percentages, proportions, means or similar statistical value. For example, if you have 10 percent of 200 people, you would need to convert that to a number (20) before you can run a test statistic.

**T-Test (Mean, Proportion)**

**The t test** is one type of inferential statistics. It is used to determine whether there is a significant difference between the means of two groups. With all inferential statistics, we assume the dependent variable fits a normal distribution. When we assume a normal distribution exists, we can identify the probability of a particular outcome. We specify the level of probability (alpha level, level of significance, p) we are willing to accept before we collect data (p < .05 is a common value that is used). After we collect data we calculate a test statistic with a formula. We compare our test statistic with a critical value found on a table to see if our results fall within the acceptable level of probability.

When the difference between two population averages is being investigated, a t test is used. In other words, a t test is used when we wish to compare two means (the scores must be measured on an interval or ratio measurement scale). We would use a t test if we wished to compare the reading achievement of boys and girls. With a t test, we have one independent variable and one dependent variable. The independent variable (gender in this case) can only have two levels (male and female). The dependent variable would be reading achievement. If the independent had more than two levels, then we would use a one-way analysis of variance (ANOVA).

The test statistic that a t test produces is a t-value. Conceptually, t-values are an extension of z-scores. In a way, the t-value represents how many standard units the means of the two groups are apart.

With a t test, the researcher wants to state with some degree of confidence that the obtained difference between the means of the sample groups is too great to be a chance event and that some difference also exists in the population from which the sample was drawn. In other words, the difference that we might find between the boys’ and girls’ reading achievement in our sample might have occurred by chance, or it might exist in the population. If our t test produces a t-value that results in a probability of .01, we say that the likelihood of getting the difference we found by chance would be 1 in a 100 times. We could say that it is unlikely that our results occurred by chance and the difference we found in the sample probably exists in the populations from which it was drawn.

**ASSUMPTIONS UNDERLYING THE T TEST**

* The samples have been randomly drawn from their respective populations
* The scores in the population are normally distributed
* The scores in the populations have the same variance (s1=s2) Note: We use a different calculation for the standard error if they are not.

**THREE TYPES OF T TESTS**



1. Pair-difference t test (a.k.a. t-test for dependent groups, correlated t test) df= n (number of pairs) -1

This is concerned with the difference between the average scores of a single sample of individuals who are assessed at two different times (such as before treatment and after treatment). It can also compare average scores of samples of individuals who are paired in some way (such as siblings, mothers, daughters, persons who are matched in terms of a particular characteristics).

1. t test for Independent Samples (with two options)

This is concerned with the difference between the averages of two populations. Basically, the procedure compares the averages of two samples that were selected independently of each other, and asks whether those sample averages differ enough to believe that the populations from which they were selected also have different averages. An example would be comparing math achievement scores of an experimental group with a control group.

* Equal Variance (Pooled-variance t-test) df=n (total of both groups) -2 Note: Used when both samples have the same number of subject or when s1=s2 (Levene or F-max tests have p > .05).
* Unequal Variance (Separate-variance t test) df dependents on a formula, but a rough estimate is one less than the smallest group Note: Used when the samples have different numbers of subjects and they have different variances —  s1<>s2 (Levene or F-max tests have p < .05).

**HOW DO I DECIDE WHICH TYPE OF T TEST TO USE?**



**F- Test, Z – Test**



**F- TEST**

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. Exact “F-tests” mainly arise when the models have been fitted to the data using least squares. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s

**Assumptions of F- Test**

Several assumptions are made for the test. Your population must be approximately normally distributed (i.e. fit the shape of a bell curve) in order to use the test. Plus, the samples must be independent events. In addition, you’ll want to bear in mind a few important points:-

* The larger variance should always go in the numerator (the top number) to force the test into a right-tailed test. Right-tailed tests are easier to calculate.
* For two-tailed tests, divide alpha by 2 before finding the right critical value.
* If you are given standard deviations, they must be squared to get the variances.
* If your degrees of freedom aren’t listed in the F Table, use the larger critical value. This helps to avoid the possibility of Type I errors.

**Common examples**

Common examples of the use of F-tests include the study of the following cases:

* The hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal. This is perhaps the best-known F-test, and plays an important role in the analysis of variance (ANOVA).
* The hypothesis that a proposed regression model fits the data well. See Lack-of-fit sum of squares.
* The hypothesis that a data set in a regression analysis follows the simpler of two proposed linear models that are nested within each other.

**F Test to compare two variances by hand: Steps**

**Warning:** F tests can get really tedious to calculate by hand, especially if you have to calculate the variances. You’re much better off using technology (like Excel — see below).

These are the general steps to follow. Scroll down for a specific example (watch the video underneath the steps).

**Step 1:** If you are given standard deviations, go to Step 2. If you are given variances to compare, go to Step 3.

**Step 2:** Square both standard deviations to get the variances. For example, if σ1 = 9.6 and σ2 = 10.9, then the variances (s1 and s2) would be 9.62 = 92.16 and 10.92 = 118.81.

**Step 3:** Take the largest variance, and divide it by the smallest variance to get the f-value. For example, if your two variances were s1 = 2.5 and s2 = 9.4, divide 9.4 / 2.5 = 3.76.

Why? Placing the largest variance on top will force the F-test into a right tailed test, which is much easier to calculate than a left-tailed test.

**Step 4**: Find your degrees of freedom. Degrees of freedom is your sample size minus 1. As you have two samples (variance 1 and variance 2), you’ll have two degrees of freedom: one for the numerator and one for the denominator.

**Step 5:** Look at the f-value you calculated in Step 3 in the f-table. Note that there are several tables, so you’ll need to locate the right table for your alpha level. Unsure how to read an f-table? Read What is an f-table?.

**Step 6:** Compare your calculated value (Step 3) with the table f-value in Step 5. If the f-table value is smaller than the calculated value, you can reject the null hypothesis.

**Z-TEST**

A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. For each significance level, the Z-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more convenient than the Student’s t-test which has separate critical values for each sample size. Therefore, many statistical tests can be conveniently performed as approximate Z-tests if the sample size is large or the population variance is known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large (n < 30), the Student’s t-test may be more appropriate.

A one-sample location test, two-sample location test, paired difference test and maximum likelihood estimate are examples of tests that can be conducted as z-tests. Z-tests are closely related to t-tests, but t-tests are best performed when an experiment has a small sample size. Also, t-tests assume the standard deviation is unknown, while z-tests assume it is known. If the standard deviation of the population is unknown, the assumption of the sample variance equaling the population variance is made.

**One-Sample Z-Test Example**

For example, assume an investor wishes to test whether the average daily return of a stock is greater than 1%. A simple random sample of 50 returns is calculated and has an average of 2%. Assume the standard deviation of the returns is 2.50%. Therefore, the null hypothesis is when the average, or mean, is equal to 3%. Conversely, the alternative hypothesis is whether the mean return is greater than 3%. Assume an alpha of 0.05% is selected with a two-tailed test. Consequently, there is 0.025% of the samples in each tail, and the alpha has a critical value of 1.96 or -1.96. If the value of z is greater than 1.96 or less than -1.96, the null hypothesis is rejected.

The value for z is calculated by subtracting the value of the average daily return selected for the test, or 1% in this case, from the observed average of the samples. Next, divide the resulting value by the standard deviation divided by the square root of the number of observed values. Therefore, the test statistic is calculated to be 2.83, or (0.02 – 0.01) / (0.025 / (50)^(1/2)). The investor rejects the null hypothesis since z is greater than 1.96, and concludes that the average daily return is greater than 1%.

A