MarkQual®

Assessment and qualification system

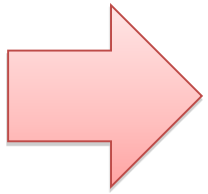
Statistical Glossary

Source : Murray R. Spiegel, STATISTICAL, McGraw-Hill.

SPECIAL ITEMS AND DEFINITIONS

Is the scientific study of methods to collect, organize, summarize and analyze data information and to draw valid conclusions rigorously and make reasonable decisions based on such analysis.

With this tool you can work in an objective, allows you to link cause and effect, present results with clarity and order. Is an essential input to interpret numerical information and then build curricula.



The statistic has three well-defined fields:

Descriptive

Inferential

Probability Theory.

Descriptive Statistics

Studies is made about the total number of individuals of a population in order to establish the main features of interest to the researcher.

Statistical Inference

It refers to studies that are done on a part of the population (sample) to obtain (infer) conclusions about the characteristics of interest throughout the population. It is a way of deduction at risk, with probability of error.

Probability Theory

It is a branch of mathematics of great importance in inferential studies, since the values obtained on the analysis of a sample are not exactly the same as for the population parameters. He studies the mathematical behavior of the random control of random phenomena.

Population

Entire group of individuals or objects that form the basis of interest to a statistical study. Is the set of all elements that match a particular feature that we want to measure and study.

Sample

A representative portion of a population. Is every subset of a population on which to conduct the study. The number of elements in the sample is called sample size.

Individual

In statistics it is considered an individual (object), each of the elements of the population.

Character

Each of the aspects or properties that can be studied in individuals of a population is called statistical or character. This allows classification of individuals.

The character can be quantitative if you can measure qualitative or if you can not measure but can be compared.

Data

Value or form that takes a variable for a given individual.

Estimators
Quantifiable characteristics that has a sample, and used to estimate population parameters.

Distributions
Organizational forms and tabular representation of data.

Statistical Experiment
Any process that generates a set of numerical data.

Sample Space
Set of all possible outcomes of a statistical experiment.

Variable
The set of values that can take a statistical character is called statistical variable. They are attributes that have or can be assigned to individuals in a population that differ from each other.

Classification of Variables

Qualitative: The defining qualities of individuals, usually can be subdivided into categories.

Example: Variable: Sex. Categories: M. F.

Indicator: Numeric values assigned to the categories of a qualitative variable.

Quantitative: When attributes that define them are quantifiable or measurable numerically. Quantitative variables can be discrete or continuous.

Discrete: When variables can only take certain values, (assume values one by one), that is can take a finite or infinite number of values.

Continuity: When can assume any value between two consecutive integers, that is can take all values in a range and as close as you want.

Class intervals

It is called class interval to each of the intervals that can be grouped wiht data in a statistical variable. They allow a more clear and concrete reality. By grouping the values of a statistical variable and classify at intervals, passing the variable to be considered continuous.

There are situations like the following that can occur:

- There is too much data for a single variable.
- Data may be few, but very scattered values.
- Interested in a particular classification of results.

In these cases, the pooling of data is a good technique for analyzing variables within a statistical study.

Class intervals - Method of Working Group I

1.Implement a data collection technique, such as: Stem and Leaf

2.To determine the range of information: $R = D_M - D_m$.

Where $D_M = \text{higher data}$ y $D_m = \text{lower data}$

3. Determining the Extent of intervals: Divide the range observed in two by the number of intervals in which you wish to group $A = \frac{R}{I}$

Where: A: Amplitude, R: Range and I: Number of intervals in which you want to group

4. If the amplitude does not give me an integer, you can adjust the following:

4.1. Adjust the amplitude obtained at the nearest integer.
4.2. With the above range and the number of intervals (I), found a new range (NR).
4.3.Distinguish: NR-VR, where VR is the old range. +
4.4. Original data adjusted according to the earlier dispute.

Class intervals - Working Method II

5. I set intervals, starting from the first date (or the first data set), and adding the range to cover the pre-defined number of intervals

6. The mark is calculated for each interval class $$Mc_i = \frac{L_{\text{inf}} + L_{\text{sup}}}{2}$$

where Mc_i = Class mark interval *i*; L_{inf} = Lower limit *i;*

L_{sup} = Upper limit of interval *i.*

Class Mark is a value that represents all the range or class. It is the midpoint between the extremes of each interval.

7. Identify the distribution of frequencies

Comments:

The number of intervals, can be defined previously at the discretion of the investigators, or by applying some techniques suggested to do so according to the type of study, a formula often used is that of Sturges: I = 1 + (3.3 log N), where N = total data.
In any case it is recommended that the number of intervals of not less than 5 nor more than 20.
Whenever you perform this grouping is a loss of information, taking into account the membership or not of each data to the interval but not its exact value.
It also fails the subsequent calculation of statistical parameters. The values that belong to the interval are represented by their class mark, and they may be higher or lower than this

Definitions

Absolute frequency(f_i)

Is the amount of times it appears and repeats the value data.

Cumulative absolute frequency (F_A)

Cumulative absolute frequency is called a value to the sum of the absolute frequency values less than or equal to the measured value. $F_A = \sum f_i$

Frequency distribution

Tabular representation of data for a variable, including:

Relative Frequency (f_r)

Often called a value relative to the ratio of absolute frequency and the total number of data involved in the experiment.

Frecuencia Relativa Acumulada (F_R)

Cumulative relative frequency is called a value to the sum of all frequencies relative values less than or equal to the consideration. Can also be calculated as: $F_R = \frac{F_A}{N}$

Data Presentation: Tables, Charts and Diagrams

There are many different types and styles of tools used to represent the data for a statistical variable. It is essential to be clear, easy to understand and interpret, must conform strictly to the reality they represent.

There are two types of graphs to represent distributions Grouped by intervals:

Histogram: Vertical bar chart attached to the same extent and focus on the class mark. Interval associated with each area of a rectangle proportional to the frequency for that interval. The heights of these rectangles are the ratios of absolute frequencies and the lengths of the intervals that correspond. A equal class intervals, the heights are directly proportional to the frequencies.

Frequency Range: Graph of strokes or lines, locked, which is constructed by joining the upper middle end points of a histogram, namely the points corresponding to the frequencies of each value.

Block or Line Chart

They are used for one or more discrete quantitative variables.

Bar graph (vertical and horizontal)

Is a graph associated with each variable value bar (vertical or horizontal), proportional to the frequency it deserves. It is suitable for qualitative variables alone or in comparison.

Within the representation of bars, the bars are in percentage per component, very useful when dealing with populations with very different sizes.

Cake or pie slices

For a single qualitative or quantitative variable. Are useful to represent the different parts of a whole, the various components of a character. Each event is represented by a circular sector with an amplitude proportional to its frequency

Pictograms

Representations usually bars that rely on the facilities that offer computer graphics. Along the graph of the data you can see an image on or referred to objects that are measured.

Cartograms

Representations of data on a map.

Statistical parameters

Numerical characteristics which has a population quantifiable. Are obtained through a process of calculation from measured data. They are numbers that describe the behavior and general characteristics of a set of statistical data. They are grouped into two categories: centralization and dispersal

Measures of Central Tendency

Numerical values can be obtained from the distribution of a quantitative variable, and the results are found by the same distribution center, they are:

The mode (M_o), Median(M_{ed}), and the Media(X).

Mode (M_o)

Is the data that is most often complete within a distribution. If there are two mode in a distribution is a Bimodal distribution if more than two modes then we say a multimodal distribution. In the case of a continuous variable this value is meaningless.

Median (M_{ed})

Se denomina mediana al valor central de los datos cuando éstos se han organizado ordenadamente de menor a mayor. Es un valor que divide a la distribución en dos partes iguales, cada una de las cuales contiene el 50% de los datos por debajo y el otro 50% por encima.

Media (X)

Is defined as the sum of all values (data) that assumes a variable, divided by the total number of data.

Formula for Calculating Measures of Central Tendency

1 For non-group distributions

*** Mode** : Just look at the distribution, and identify the data that occurs most often

*** Median** : Depending on whether the total data is odd or even:

* If the total data is odd, the median is that data to take the place $\frac{n+1}{2}$
* If the total data is even, the Media is the average of the data to Occupy the positions $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)$ namely values that occupy the central positions.

*** Media:** The definition leads to the following formula, which is most used in the calculation of the mean:

$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$; if the data are grouped can be calculated as: $\overline{X} = \frac{\sum_{i=1}^{n} x_i f_i}{n}$

where x_i : Data iésimo; f : absolute frequency of data iésimo; n : total information data.

2. For clumped distributions (With intervals of equal amplitude)

Uses the following formulas:

Mode $$M_o = L_{\inf} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) A$$

Median $$M_{ed} = L_{\inf} + \left(\frac{\frac{n}{2} - \sum f_{ant}}{f_{abs}} \right) A$$

Media

* Long method $$\bar{X} = \frac{\sum M_{ci} f_i}{n}$$

* Short method $$\bar{X} = M_{ca} + \left(\frac{\sum \mu . f_i}{n} \right) A$$

$L_{\inf}$	: Lower limit of the class
A	: Interval width
M_{ci}	: Interval-class brand i.
Δ_1	: Difference between the frequency absolute modal interval and the absolute frequency of the previous interval
Δ_2	: Absolute frequency difference between the modal interval and the absolute frequency of the next interval
M_{ca}	: Arbitrary class mark
μ	: Number of times the size of the intervals, from arbitrary class brand.

Measures of dispersion

Are numerical values that give us information about the scattered or clustered, where the data for a quantitative variable in a statistical study. Possible to obtain an image of remoteness of the data on measures of centralization. There may be data sets with equal measures of central but very different aspect, as the dispersion of data.

The most commonly used measures of dispersion are:

***** Ranks:

Common or comprehensiveness: It is defined as the difference between the book ends on a quantitative variable, thus: Dato Major-Minor

Intercuartil: Q3-Q1; proporciona información sobre el 50% central de la variable.

Percentil: P90-P10; se utiliza cuando se quieren excluir algunos datos extremos de información; recoge información sobre el 90% central de los datos.

***La Varianza:** $\left(\sigma^2\right)$

***Las desviaciaciones:** La Desviación Media y la Desviación Típica o Standard (S.D.)

Coefficient of variation

Is a value that gives comprehensive information on the degree of dispersion of the statistical measure used, when the measure used is the arithmetic mean is defined as:

$$cv = \frac{SD}{\bar{x}}$$

It is very useful to statistically compare two populations or two different samples. It shows the relative variation of each population. Who owns the coefficient of variation greater the more heterogeneous.

Position measurements: The Quantile

This is the name of certain values within an information, that can divide into equal parts. We have seen that the median is the value that divides the dataset into two equal parts, well to the medians of each of these two equal parts that have been, called quartiles.
The quantiles used are

The quantiles*(Q):*
Are used to divide the information into four (4) equal parts, each of which contains 25% of the data. 4 quartiles are noted as: Q_1 - Q_2- Q_3 - *and* - Q_4.

The Deciles(D):
Are used to divide information into ten (10) equal parts, each of which contains 10% of the data. 10 Deciles are noted as: D_1- D_2 - - - D_{10}

The Percentiles(P):
Are used to divide information into one hundred equal parts, each of which contains 1% of the data. Are noted as one hundred percentiles: P_1 - P_2- P_3- - -P_{100}

The Variance $\left(\sigma^2\right)$

Provides comprehensive information on how the data vary in how it plays a major role in inferential statistics when it comes to estimates, since the variance analysis of quantitative information can deduce many results on the general behavior of the parameters of a population. Is defined as the average of the squared mean deviations. You can discover the variation between two samples of the same or different populations.

The variance is defined as :

I) $Var(x) = \sigma^2 = \dfrac{\sum (x_i - \bar{x})^2 f_i}{n}$; for ungrouped data.

II) $Var(x) = \sigma^2 = \dfrac{\sum (M_{ci} - \bar{x})^2 f_i}{n}$; for grouped data.

III) As an alternative formula for the variance, which is commonly used is:

$$Var(x) = \frac{\sum x^2 f_i}{n} - \left(\frac{\sum x_i f_i}{n}\right)^2 = \bar{x^2} - \left(\bar{x}\right)^2$$

The Deviations

A deviation is defined as the difference between the value of data and some statistical measure, the most common deviations are taken as the arithmetic mean, but you can take deviations from the mode, the median, or one of the quantiles,

the Diversion Media

Is defined as the average - the arithmetic mean - of the absolute deviations of a variable, taken in absolute value with respect to the arithmetic mean of the variable in formulas will be:

$$DM = \frac{\sum \left| x_i - \bar{x} \right|}{n}$$

$$DM = \frac{\sum \left| x_i - \bar{x} \right| f_i}{n}$$

Standard Deviation

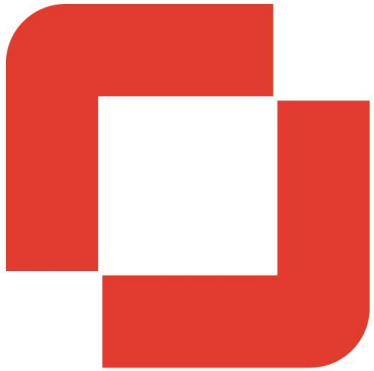
Is defined as the square root of the variance in the formula is:

$$SD = \sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f_i}{n}}$$, for ungrouped data.

$$SD = \sigma = \sqrt{\frac{\sum (Mc_i - \bar{x})^2 f_i}{n}}$$, for grouped data.

Note:

The medium is the average value, can be thought of as the physically "center of gravity" of the data set. You can imagine how fair value would be obtained by sharing all of its elements.
The standard deviation is a measure of what has been done equitable distribution. There is less balanced when there is greater deviation



MarkQual®

Assessment and qualification system