Dr. Mohanad Faris Abdulhameed
Department of Public Health
College of Veterinary Medicine
University of Basrah
Second Class-Biostatistics
محاضرة رقم 2

Data collection and measure

- **In research field,** the question is how data can be obtained or generate good data? We can test two way of hypothesis: by observation naturally what is happening or manipulate some aspect of environment and observe the effects on variable interest us in the study.
- Therefore, data can be defined as the process of gathering and measure information on variables of interest using three ways: **censuses, sample surveys, and administrative data**.

1- Census (counts): a census can be referred to data collection of every unit in a group or population. If a researcher collected data about the heights of the students in the veterinary medicine school that would be all students from different stages regardless gender.

2- Sample survey: Example, collection the blood samples from the cattle and buffalo to identify presence of the blood parasites or collection milk samples from dairy cattle with mastitis to identify bacteria or may be fungi causes.

3- **Administrative data**: Administrative data are collected as a result of an organization's day-to-day operations. Examples include data on births, deaths, marriages, divorces car registrations, and crimes.

The data collection component of research is common to all fields of study including physical and social sciences, humanities, business. Data collection is a very demanding job which needs thorough planning, hard work, patience, perseverance and more to be able to complete the task successfully. Data collection starts with determining what kind of data required followed by the selection of a sample from a certain population. After that, you need to use a certain instrument to collect the data from the selected sample. All data need to be managed very well based on the date collected, sufficient details of a subject, provide computerised copy (in excel).

Table 1. shows numbers and frequency of persons with categorical age.

Age in years	No. of persons	Frequency
0-12	1124	98
12-24	1026	217
24-36	809	382
36-48	427	269
48-60	158	138
60-72	20	15
72-84	5	2
84-96	3	2
96-108	1	1
108-120	0	0

- **Classification of data**

The data are classified in to two types:

A- Primary data

Data that has been collected from first-hand-experience is known as primary data. Primary data has not been published yet and is more reliable, authentic and objective. Primary data has not been changed or altered by researchers; therefore, its likely to be valid than the secondary data.

Importance of Primary Data: in statistical surveys it is necessary to get information from primary sources and work on primary data. **For example**, the statistical records of number of farmers are registered by veterinary or agriculture authorities in the local area. A research can be conducted without secondary data but the research is based on only secondary data is least reliable and may have biases because secondary data has already been manipulated by human beings.

Sources of Primary Data: Sources for primary data are limited and at times it becomes difficult to obtain data from primary source because of either scarcity of population or lack of cooperation. Following are some of the sources of primary data:

1-Data originated from experiments: experiments require an artificial or natural setting in which to perform logical study to collect data. Experiments are more suitable for medicine, psychological studies, nutrition and for other scientific studies.

2- Survey: Survey is most commonly used method in social sciences, management, marketing and psychology to some extent. Surveys can be conducted in different methods. Or in veterinary medicine, data may collect from animals to present absence or presence of a disease.

Another examples vaccination status of important diseases may infect animals in the farms (livestock or poultry farm)

3- Questionnaire: It is the most commonly used method in survey. Questionnaires are a list of questions either open-ended or close-ended for which the respondents give answers. Questionnaire can be conducted via telephone, mail, live in a public area, or in an institute, through electronic mail or through fax and other methods.

4- Interview: Interview is a face-to-face conversation with the respondent.

5-Observations: Observation can be done while letting the observing person know that she/he is being observed or without letting him know. For example, if your colleague having a scowl facial expression probably, he/she is in a difficult situation (economic or social problem)

B- Secondary data

Data collected from a source that has already been published in any form is called as secondary data. The review of literature in any research is based on secondary data. It is collected by someone else for some other purpose (but being utilized by the investigator for another purpose). For examples, Census data being used to analyse the impact of education on career choice and earning. Common sources of secondary data for social science include censuses, organizational records and data collected through qualitative methodologies or qualitative research. Secondary data is essential, since it is impossible to conduct a new survey that can adequately capture past change and/or developments.

Sources of Secondary Data: The following are some ways of collecting secondary data –

- Books.
- Records.
- Biographies.
- Newspapers.
- Published censuses or other statistical data.
- Data archives.
- Internet articles.
- Research articles by other researchers (journals).
- Databases.

Numerical measures of data

Mean: it is obtained by adding together the observations in a data set and dividing by the number of observations in the set.

Median: it is the central value in the set of *n* observations which have been arranged in rank order

Mode: it is obtained from data that represent more frequent value in the dataset.

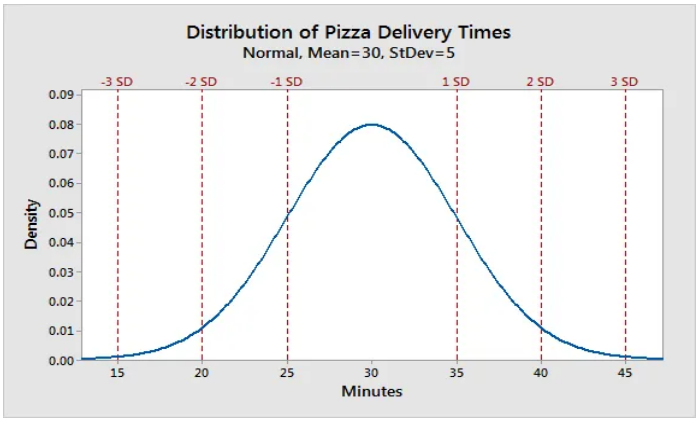
Range: is defined as the difference between the largest and smallest observations.

Normal Distribution | Examples, Formulas, & Uses

In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the centre.

Why do normal distributions matter?

All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.
Because normally distributed variables are so common, many statistical tests are designed for normally distributed populations.
Understanding the properties of normal distributions means you can use inferential statistics to compare different groups and make estimates about populations using samples.

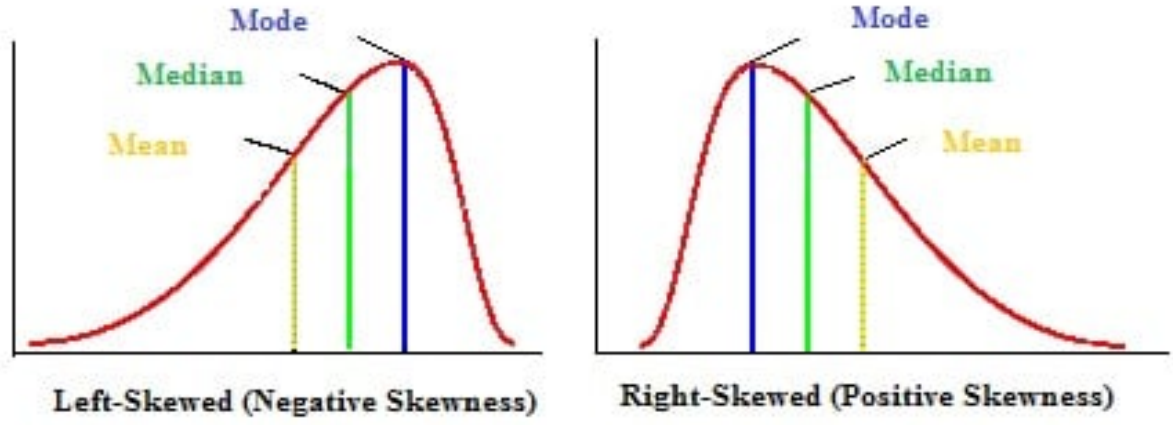


What are the properties of normal distributions?

Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry.



Standard deviation (SD)

Defined as a statistical measure of the amount of variation or dispersion in a set of data, representing how far individual data points are from the mean (average). A low SD indicates data points are close to the mean, while a high SD shows they are spread out.

Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{n}}$$

Calculate standard deviation (SD)

1. Determine the mean (the average of all the numbers) by adding up all the data pieces (xi) and dividing by the number of pieces of data (n).
2. Subtract the mean ($\bar{x}$) from each value.
3. Square each of those differences.
4. Determine the average of the squared numbers calculated in #3 to find the variance. (In sample sizes, subtract 1 from the total number of values when finding the average.)
5. Find the square root of the variance. That's the standard deviation!

For example: Take the values 2, 1, 3, 2 and 4.

1. Determine the mean (average):

2 + 1 +3 + 2 + 4 = 12
12 ÷ 5 = 2.4 (mean)

2. Subtract the mean from each value:

2 - 2.4 = -0.4
1 - 2.4 = -1.4
3 - 2.4 = 0.6
2 - 2.4 = -0.4
4 - 2.4 = 1.6

3. Square each of those differences:

-0.4 x -0.4 = 0.16
-1.4 x -1.4 = 1.96
0.6 x 0.6 = 0.36
-0.4 x -0.4 = 0.16
1.6 x 1.6 = 2.56

4. Determine the average of those squared numbers to get the variance.

0.16 + 1.96 + 0.36 + 0.16 + 2.56 = 5.2
5.2 ÷ 5 = 1.04 (variance)

5. Find the square root of the variance.

Square root of 1.04 = 1.01

The standard deviation of the values 2, 1, 3, 2 and 4 is 1.01.

Percentile and Quartile

A percentile (or a centile) is a measure used in statistics indicating the value *below which* a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found. The term percentile and the related term *percentile rank* are often used in the reporting of scores from norm-referenced tests. For example, if a score is at the 86th percentile, where 86 is the percentile rank, it is equal to the value below which 86% of the observations may be found. In contrast, if it is in the 86th percentile, the score is at or below the value of which 86% of the observations may be found.
The 25th percentile is also known as the first quartile (Q1), the 50th percentile as the median or second quartile (Q2), and the 75th percentile as the third quartile (Q3). In general, percentiles and quartiles are specific types of quantiles.

Percentage is different from percentile which include the characterisation of interest divided by the total number of observations
For example: if you have three positive samples from the total of 30, that means 3/30*100
Which give equal to 10%.

1-Percentile

- Percentiles are used to understand and interpret data. They indicate the values below which a certain percentage of the data in a data set is found.

- Percentiles can be calculated using the formula n = (P/100) x N, where P = percentile, N = number of values in a data set (sorted from smallest to largest), and n = ordinal rank of a given value.
- Percentiles are frequently used to understand test scores and biometric measurements.

Percentile Formula

Percentiles for the values in a given data set can be calculated using the formula:
n = (P/100) x N

where N = number of values in the data set, P = percentile, and n = ordinal rank of a given value (with the values in the data set sorted from smallest to largest). For example, take a class of 20 students that earned the following scores on their most recent test: 75, 77, 78, 78, 80, 81, 81, 82, 83, 84, 84, 84, 85, 87, 87, 88, 88, 88, 89, 90. These scores can be represented as a data set with 20 values: {75, 77, 78, 78, 80, 81, 81, 82, 83, 84, 84, 84, 85, 87, 87, 88, 88, 88, 89, 90}. The data should be arranged in the ascending order.

Note: We should organise the data base on descending level

We can find the score that marks the 20th percentile by plugging in known values into the formula and solving for *n*:

n = (20/100) x 20

n = 4

Interpretation the result: The fourth value in the data set is the score 78. This means that 78 marks the 20th percentile; of the students in the class, 20 percent earned a score of 78 or lower.

Quartile

Quartiles often are **used in sales and survey data to divide populations into groups**. For example, you can use QUARTILE to find the top 25 percent of incomes in a population.

Mean divided the data point into four parts or quarter

- The first quartile (Q_1) is defined as the middle number between the smallest number (minimum) and the median of the data set. It is also known as the *lower* or *25th empirical* quartile, as 25% of the data is below this point.
- The second quartile (Q_2) is the median of a data set; thus 50% of the data lies below this point.
- The third quartile (Q_3) is the middle value between the median and the highest value (maximum) of the data set. It is known as the *upper* or *75th empirical* quartile, as 75% of the data lies below this point

1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:

1,3,3,4,5,6,6,7,8,8

In this case Quartile 2 is half way between 5 and 6:

$$Q2 = (5+6)/2 = \mathbf{5.5}$$

And the result is:

- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**
-

Use quartile in an excel sheet (25%,50%,75%)