◉Sheet ○Slides

Number

7

Done by:

Omayma Hassanin

corrected by:

Rana Njada

Doctor

Mahmoud Alhussami

Five numbers and Z score:

Five numbers→ Minimum value, Maximum value, Quartile 1 (Q_1), Quartile 3 (Q_3), Median/ Quartile 2 (Q_2)

(Through which quartiles are divided into four cuts); ex1: If we have values for the grades of six students where: The minimum value= 50; Q_1 = 60; Q_3 = 79; Maximum value= 100; Median (Q_2) =

Q_1- Min = 60-50=10

Max- Q_3 = 100-90= 10

The greater value would indicate the nature of skewness (positive or negative).

In the example above, the data is not skewed because both values are equal which indicates normal distribution.

Ex2: If→ Q_1 = 70; min = 50; Max= 100; Q_3 = 90, so→

Q_1- Min = 70 – 50 = 20

Max- Q_3 = 100 – 90= 10

Data in this example is skewed negatively.

So→

If the value of [Q_1- Min] is greater than [Max- Q_3], data is skewed negatively.

If the value of [Max- Q_3] is greater than [Q_1- Min], data is skewed positively.

If both values are equal, data is not skewed (normal distribution).

Data can be standardized which would result in standardized normal distribution.

Z score:

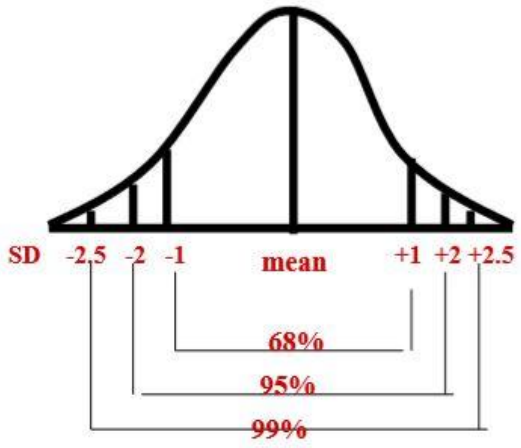
Z score is used to deal with standardized normal distribution.

The formula is as follows:

Z = [(Score of value (x) – mean of score of sample ($\bar{x}$)] ÷ standard deviation of the sample (S)

-In normal distribution, we deal with repetition. For example, a sample of the scores of a group of students might present repetition through the mode (many students scoring C for example), so that part of the data would be normally distributed. In other words, most students achieved about 70, so if the standard deviation is 5 points, then 68 percent of students have achieved C (refer to the diagram below for further understanding). Such an example is an application of:

-Frequency distribution→



Worldwide, repetition is not usually used. Instead, Standardized scores are used. In other words, every value or score is changed into a z-score.

So, for example if the score is 70 and the mean is 70 and by default the standard deviation is 5, by applying the formula mentioned earlier, the z-score would be equal to zero.

Therefore, the mean, mode, and median are always in standard deviation (Z score) are equal to zero. (?)

And 99.6 from the data is between 3 z-scores positive and 3 z-scores negative. In other words, z score value starts from negative 3 and ends with positive 3.

For project:

Z-score can be useful in understanding outliers.

Any value out of the z-score range (between -3 and +3) is considered an outlier.

For example, a researcher measured the pressures of a sample of students, and the mean of the pressures turned out to be 70. In this case, 70 (the mean) represents zero in the z-score range, so if a pressure value is below 70 (which is equivalent to 0 on the z-score scale), the researcher should calculate its z-score equivalent to determine whether such pressure value is an outlier or not. If the z-score equivalent of such value was below -3, it would be considered an outlier. The same scenario would be applied to any pressure values above 70.

In other words, we determine the mean of the values in a sample of data and because the mean is always equivalent to 0 on the z-score scale, we determine the z-score equivalents of any value above or below the mean in order to find outliers in the data.

Shapes of distribution: (Unit 3- Shapes of distribution and Graphs)

Distribution of the data can be either normal (actual arithmetic mean and data; and if data needs transforming, we use the geometric mean) or standardized (z-score).

Characteristics of distribution:

1- Modality

2- Symmetry

3- Degree of Skewness

4- Kurtosis

Sometimes, even if the data is skewed, it might be considered within the normal range of the acceptance for the parametric techniques.

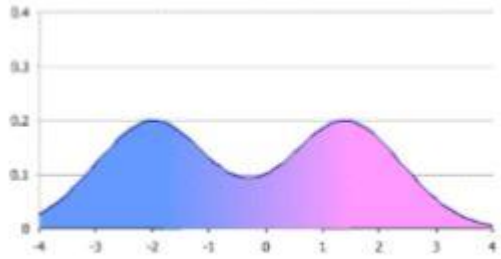
Therefore, we have to know how to calculate the index of skewness.

Modality:

The modality of a distribution concerns with how many peaks or high points there are.

Modality indicates how many bellies the curve has. In normal distribution, the curve should have one belly. A distribution with a single peak, one value a high frequency is a unimodal distribution.

A distribution with two (two modes) or more peaks called multimodal distribution.



*What is the importance of normal distribution? (Important question for the exam)

There are requirements for the parametric techniques in regards to the dependent variables.

The first requirement (which is an assumption) is that the dependent data must be normally distributed.

There are other assumptions for the parametric techniques which would be discussed later in the course.

Note: you have to calculate the skewness index in the project if you worked on a continuous data.

Symmetry and Skewness:

Shape can be described by degree of asymmetry (i.e., skewness):

mean > median positive or right-skewness

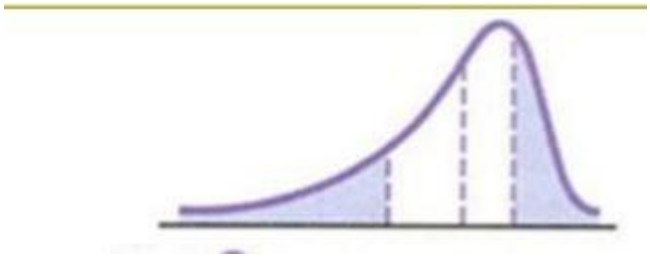
mean = median symmetric or zero-skewness

mean < median negative or left-skewness

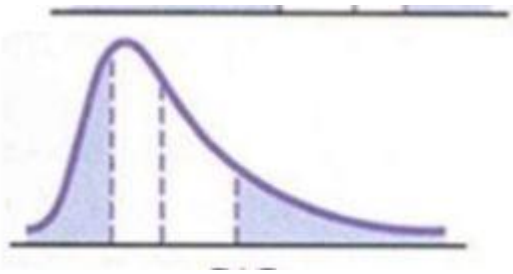
-Positive skewness can arise when the mean is increased by some unusually high values.
-Negative skewness can arise when the mean is decreased by some unusually low values.

Most common shapes of curves of frequency distribution:

Left (Negatively) skewed:

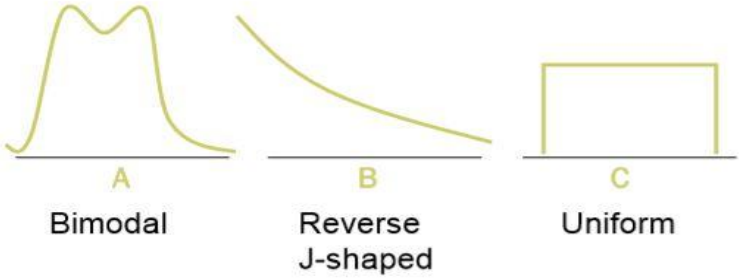


Right (Positively) skewed:



Symmetric (bell shaped):

Other kinds of curves present in literature:



When an extreme value is present in the data, we should measure the degree of skewness. Usually, the best skewness would be 0; however, it is impossible to encounter a degree of skewness equal to zero, unless the data is perfect (and usually data, especially when small in size, is not distributed logically/ perfectly).

*uniform curve is not normally distributed.

Degree of Skewness:

A skewness index can readily be calculated most statistical computer program in conjunction with frequency distributions

The index has a value of 0 for perfectly symmetric distribution.

A positive value if there is a positive skew, and negative value if there is a negative skew.

A skewness index that is more than twice the value of its standard error can be interpreted as a departure from symmetry.

Two types of measurement for skewness:

1- Pearson's skewness coefficient (index): It is non-algebraic and easily calculated. Also, it is useful for quick estimates of symmetry. It is normal between -1 and +1 because the numbers from which we subtract are large. (Arithmetic mean)

It is defined as: skewness = mean-median/SD; (SD: standard deviation)

2- Fisher's measure of skewness: It is based on deviations from the mean to the third power. Normal: from -0.2 to +0.2→ because the measurement uses third power, the numbers would be very small. (Geometric mean) used in computer programs.

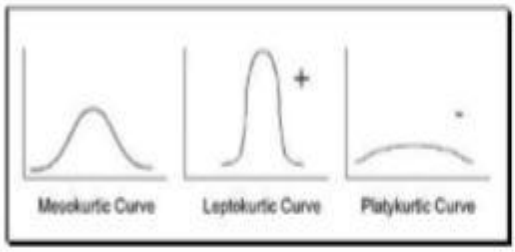
Kurtosis:

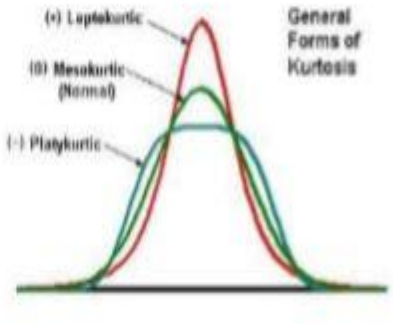
The distribution's kurtosis is concerned with how pointed or flat its peak.

Sometimes, the data is presented as clusters, so the curve would be sharp and thin. There is an index which would limit a curve from being too narrow or too flat.

If the peak is very sharp/ thin, the distribution is called leptokurtic.

If the peak is flat, the distribution is called platykurtic.





The value of the standard deviation determines if the curve is flat or narrow.

The larger the standard deviation, the flatter is the curve. The smaller the standard deviation, the thinner is the curve.

However, an extremely thin curve is not favorable. A normal (mesokurtic) distribution is the best (index of kurtosis = 0).

There is a statistical index of kurtosis that can be computed when computer programs are instructed to produce a frequency distribution

For kurtosis index, a value of zero indicates a shape that is neither flat nor pointed.

Positive values on the kurtosis statistics indicate greater peakedness, and negative values indicate greater flatness.

*Skewness is used more frequently than kurtosis. Because if skewness is normal the kurtosis will be normal too.

Fishers' measure of Kurtosis:

Fisher's measure is based on deviation from the mean to the fourth power.

A z-score is calculated by dividing the measure of kurtosis by the standard error for kurtosis.

The value ranges from -0.2 to 0.2 .

Graphical Methods:

1- Frequency Distribution

2- Histogram

3- Frequency Polygon

4- Cumulative Frequency Graph

5- Pie Chart

Presenting Data:

Tables:

Condenses data into a form that can make them easier to understand.

Shows many details in summary fashion.

But, since table shows only numbers, it may not be readily understood without comparing it to other values.

Principles of Table Constructing:

Don't try to do too much in a table. Use white space effectively to make table layout pleasing to the eye, the font size should be 12 and not bold. Make sure tables & test refer to each other. Use some aspect of the table to order & group rows & columns. If appropriate, frame table with summary statistics in rows & columns to provide a standard of comparison. Round numbers in table to one or two decimal places to make them easily understood. When creating tables for publication in a manuscript, double-space them unless contraindicated by journal.

When designing tables, a standardized style should be used.

The latest, frequently used style is APA 6th edition.

When designing the table (according to APA), you name the table (for example, Table 1), then you insert a line and then insert the data and place a second line beneath the inserted data.

Example:

Table 1

Error Rates of Older and Younger Groups

Level of difficulty	Mean error rate		Standard deviation		Sample size	
	Younger	Older	Younger	Older	Younger	Older
Low	.05	.14	.08	.15	12	18
Moderate	.05	.17	.07	.15	15	12
High	.11	.26	.10	.21	16	14

Note. From "Generations," by L.G. Elias and C.C. Bent, 2002, Journal of Geriatric Care, 5, p. 22.

The same procedure applies for the AMA style.

The difference between AMA and APA is that AMA depends on numbers in referencing (when citing, the first reference would be numbered '1', the next reference '2', and so on. And when reference number one for example is repeated elsewhere, you use '1' again. Then at the reference list*, you would mention the full name of reference '1', '2', and so on.)

As for APA style, it depends on the last name of the author; for example: (Cushman, 2008). If two people, ex.: (Cushman, 2008; Maloney and O'Dea, 2000). If more than 3, ex.: (Wainwright et al., 2000)

[Useful links for referencing and designing tables in your projects:

http://www.cite.auckland.ac.nz/2.html

https://www.library.auckland.ac.nz/subject-guides/edu/docs/APAbooklet.pdf

https://owl.english.purdue.edu/owl/resource/560/19/]

*Bibliography and reference list differ.

Reference list→ any reference mentioned in the body of the manuscript must be listed in the reference list.

Bibliography→ sometimes the researchers don't mention in the body of the manuscript some references they used. They would mention the references they used in the bibliography at the end of the body of the manuscript.

Frequency Distributions:

A useful way to present data when you have a large data set is the formation of a frequency table or frequency distribution.

Frequency – the number of observations that fall within a certain range of the data.

Frequency Table:

Age	Number of Deaths
<1	564
1-4	86
5-14	127
15-24	490
25-34	66
35-44	806
45-54	1,425
55-64	3,511
65-74	6,932
75-84	10,101
85+	9825
Total	34,524

Data Intervals	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
10-19	5	5		
20-29	18	23		
30-39	10	33		
40-49	13	46		
50-59	4	50		
60-69	4	54		
70-79	2	56		
Total				

Number of intervals:

There is no clear-cut rule on the number of intervals or classes that should be used.

Too many intervals – the data may not be summarized enough for a clear visualization of how they are distributed.

Too few intervals – the data may be over summarized and some of the details of the distribution may be lost.

Presenting Data: (Please refer to the slides from page 24 to 40)

Data can be presented in bars, charts, and pie charts.

The type of representation of data depends on the level of data.

For example, we cannot use the histogram for nominal data. For nominal data, pie charts are used because nominal data can be a percentage and pie charts can demonstrate percentages.

With ordinal data as well, histogram cannot be used. Instead, bar charts are used. In bar charts, the bars are separated, so they are used to represent separated data.

However, in continuous data, a histogram should be used.

Missing Data:

Surveys contain errors. Most participants do not fill out the questionnaire or complete it. Usually, some data is missing.

As a researcher, there are techniques to cover for or solve the problem of missing data.

One technique is to delete the items not filled, and if the uncompleted items in a questionnaire is more than 25%, then you have to omit the participant, and this depends on how many participants you collected the data from.

Another technique is to complete the questionnaire. Completing the questionnaire is through the mean, so the uncompleted value is replaced with a mean, unless the

researcher is aware of the participant's conditions and can answer instead of her/ him (intelligent guess).

As for outliers, as mentioned before, for z-score, we consider 3 standard deviations on both negative and positive directions (values between -3 and +3). Any value less than -3 or more than +3 would be considered an outlier and be deleted, and the resulting mean would be called 'trimmed mean'.