Chapter 02

Describing data with Numerical Measures

To describe quantitative variable, most of the statistical characteristics for ordinal variable description can be used (frequency, relative frequency cumulative frequency and cumulative relatife frequency

Apart from those, there are two additional ones:

- Measures of location: those indicate a typical distribution of the variable values.
- Measures of variability: those indicate a variability (variance) of the values around their typical position.

Definition: Numerical descriptive measures associated with a population of measurements are called parameters; those computed from sample measurements are called statistics.

① Measures of central tendency

These are three measures of central tendency that are commonly used to describe the average value of a set data. these are the "mode", the "mean" and the "median."

The mean

* Definition①: the mean is calculated by dividing the sum of the values by the number of values. It is defined by the following formula: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

where: x_i : are values of the variable.

n : size of the sample population (number of the values of the variable

- Sample mean $\bar{x}$.
- Population mean μ.

* Properties of the arithmetical mean

1/ $\sum_{i=1}^{N} (x_i - \bar{x}) = 0$

2/ $\forall a \in \mathbb{R}: \bar{x} = \frac{\sum x_i}{N} \Rightarrow \frac{\sum (a + x_i)}{N} = a + \bar{x}$.

3/ $\forall b \in \mathbb{R}: \bar{x} = \frac{\sum x_i}{N} \Rightarrow \frac{\sum (b x_i)}{N} = b \bar{x}$.

- Arithmetical mean is not always the Best way to calculate the mean of the sample population.

For example, if we work with a variable representing relative changes we use the so-called geometrical mean. To calculate mean when the variable has a form of a unit, harmonical mean is often used.

* Example ① the following data shows ages of musicians who performed at a concert. Calculate Mean.

22 - 82 - 27 - 43 - 19 - 47 - 41 - 34 - 34 - 42 - 35.

Solution: we use arithmetical mean:

$$\bar{x} = \frac{1}{N}\Sigma x_i = \frac{1}{11}(22+82+27+\dots+42+35) = 38{,}7 \text{ years.}$$

the musicians average age is 38,7 years.

Definition ②: * the median is the value in the middle of an ordered set of data.

Key Point: the sample median is obtained by first ordering the N observations from smallest to largest (with any repeated values included so that every sample observation appears in the ordered list). Then,

$$\tilde{x} = \begin{cases} \text{The single middle value if } N \text{ is odd} = \left(\frac{N+1}{2}\right)^{th} \text{ ordered value} \\ \text{The average of the two middle values if } N \text{ is even} = \\ \quad \text{average of } \left(\frac{N}{2}\right)^{th} \text{ and } \left(\frac{N}{2}+1\right)^{th} \text{ ordered values.} \end{cases}$$

Example ①: the median value: $N = 11 \Rightarrow N$ is odd

so $\left(\frac{N+1}{2}\right)^{th} = \frac{12}{2} = 6 \Rightarrow \tilde{x} = 35$

$13 - 22 - 27 - 34 - 34 - \underset{\tilde{x}}{\underset{\downarrow}{35}} - 41 - 42 - 43 - 47 - 82$

Definition ③: The mode is the most commonly occuring value.

- The mode and the modal class

the following table shows the scores on 25 rolls of a die, where 2 is the mode because it has the highest frequency.

Score on dice	1	2	3	4	5	6
Frequency	5	6	5	3	2	4

In a set of grouped data in which raw values cannot be seen, we can find the modal class, which is the class with the highest frequency.

* Other Measures of location: Quartiles, Percentiles

The median (population or sample) divides the data set into parts of equal size. To obtain finer measures of location, we could divide the data into more than two such parts. Roughly speaking, quartiles divide the data set into four equal parts of the data set, the second quartile being ~~ind~~ identical to the median, and the first quartile separating the lower quarter from the upper three-quarters. Similarly, a data set (sample or population) can be even more finely divided using percentiles; the 99th percentile separates the highest 1% from the bottom 99%, and so on. Unless the number of observations is a multiple of 100, care must be exercised in obtaining percentiles.

* Calculating sample Quartiles

- When the measurements are arranged in order of magnitude, the lower quartile Q_1, is the value of x in position $0{,}25(n+1)$, and the upper Quartile Q_3 is the value of x in position $0{,}75(n+1)$.

- When $0{,}25(n+1)$ and $0{,}75(n+1)$ are not integers, the quartile are found by interpolation, using the values in the two adjacent positions.

__Example__: Find the lower and upper quartiles for this set of measurements: 16 - 25 - 4 - 18 - 11 - 13 - 20 - 8 - 11 - 9

__solution__: Rank the $n = 10$ measurements from smallest to largest: 4 - 8 - 9 - 11 - 11 - 13 - 16 - 18 - 20 - 25

Calculate:

$Q_1 = 0{,}25(n+1) = 0{,}25(10+1) = 2{,}75$

$Q_3 = 0{,}75(n+1) = 0{,}75(10+1) = 8{,}25$

Since these positions are not integers, the lower quartile is taken to be the value 3/4 of the distance between the second and third ordered measurements, and the upper quartile is taken to be the value 1/4 of the distance between the eighth and ninth ordered measurements.

therefore: $Q_1 = 8 + 0{,}75(9-8) = 8{,}75$

$Q_3 = 18 + 0{,}25(20-18) = 18{,}5$.

- Because the median and the quartiles divide the data distribution into four parts, each containing approximately 25% of the measurements Q_1 and Q_3 are the upper and lower boundaries for the middle 50% of the distribution.

We can ~~measu~~ measure the range of this "middle 50%" of the distribution using a numerical measure called the "Interquartile range".

<u>Definition</u>: the interquartile range (IQR) for a set of measurements is the difference between the upper and lower quartiles, that is $IQR = Q_3 - Q_1$.

② Measures of variability:

Data sets may have the same center but look different because of the way the numbers spread out from the center.

- Measures of variability can help you create a mental picture of the spread of the data. We will present some of the more important ones. the siplest measure of variation is the "Range".

Definition: the Range R, of a set of n mesurements is defined as the difference between the largest and smallest measurements.

- We prefer, however, to overcome the difficulty caused by the signs of the deviations by working with their sum of squares. From the sum of squared deviations, a single measure called the variance is calculated

Definition: the variance of a population of N measurements is the average of the squares of the deviations of the measurements about their mean μ. the population variance is denoted by δ^2 :

$$\delta^2 = \frac{1}{N} \Sigma \left(x_i - \mu\right)^2$$

⇒ Most often, you will not have all Population measurements avaible but will need to calculated the variance of a sample of n measurements.

Definition: the variance of a sample of n measurement is the sum of the squared deviations of the measurement about their mean $\bar{x}$ divided by $(n-1)$. the sample variance is denoted by s^2 and is given by:

$$s^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2$$

Definition: the standard deviation of a set of measurements is equal to the positive square root of the variance:

Key point: the variance and std cannot be negative numbers.

Notation:

n: numbers of measurements in the sample.

N: " " " " " population.

S^2: sample variance.

δ^2 = population " .

$S = \sqrt{S^2}$: sample standard deviation.

$\delta = \sqrt{\delta^2}$: population " " .

Key point: if your are using your calculator, make sure to choose the correct key for the sample std.

Key Concepts and Formulas.

① Measures of center of a data distribution

a/ Arithmetic mean or average.

- Population μ.

- Sample of n measurements: $\bar{x} = \frac{\sum x_i}{n}$

b/ Median: position of the median $= 0{,}5(n+1)$

c/ Mode.

d/ the median may be prefered to the mean if the data are highly skewed.

② Measures of variability.

a/ Range: R = largest - smallest.

b/ Variance:

1/ Population of N measurements: $\delta^2 = \frac{\sum (x_i - \mu)^2}{N}$

2/ Sample of n measurements: $S^2 = \frac{\sum (x_i - \bar{x})^2}{(n-1)}$

c/ Standard deviation:

1/ Population $\delta = \sqrt{\delta^2}$

2/ Sample: $S = \sqrt{S^2}$

③ Measures of Relative standing.

a/ p^{th} percentile, $P\%$ of the measurements are smaller, and $(100-P)\%$ are larger.

b/ Lower quartile: Q_1: Position of $Q_1 = 0{,}25(n+1)$

c/ Upper quartile Q_3 Position of $Q_3 = 0{,}75(n+1)$

d/ Interquartile range: $IQR = Q_3 - Q_1$.

Rq: The five-number summary:

Min Q_1 Median Q_3 Max