Biostat|lec(6)|Descriptive Stat pt(3)

Measures of dispersion (Variability)
•A measure of variability is a summary statistic that represents the amount of dispersion in a dataset.
The 5 measures of dispersion are :
(1)Range
(2)Interquartile Range
(3)Variance
(4)SD
(5)C.V
•Why do we need to look at measures of dispersion ?
→ Helps us to get a better idea of how well the mean .
Meaning that ; If the spread of values in the data set is large, the mean is not as representative of the data as if the spread of data is small. This is bcz a **large spread indicates that there are probably large diff b/w individual scores**.
<Small Variance → More accurate ; less dispersion >
(1)The Range
*Formula : R=Highest V -Lowest V
* It is the simplest measure of variability & Easy to calculate .
* It provides a quick summary of a distribution's variability
* It also provides useful information about a distribution when there are extreme values.
*One problem w/ the range is that it is influenced by extreme values at either end ; If we have two extreme values one very low and the other very high then the range is useless .
*Value of range is only determined by two values → Not representative .

(2)SD
*Formula : Sq root of variance

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

*expressing by how much the members of a gp differ from the mean value for the gp;tells us about how the data is distributed about the mean value .
*The **smaller the sd**, the better is the mean as the summary of a typical score.
<<Empirical Rule → statistical rule which states that for a normal distribution, almost all data will fall within three standard deviations of the mean , 3 above the mean and 3 below the mean >>
***SD for a population ; you divide by N only** not (n-1) like in sample sd .
(3)Variance
*Formula :

$$\text{Sample Variance} = s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

* measure of dispersion of the data around the mean, mathematically the variance is the **average squared deviation from the mean.**
*(n-1) → indirect measure of variability , it's called the **degree of freedom**.

NOTES :
(1)The more the data are spread out, the greater the range, variance, and sd.
(2)The less the data are spread out, the smaller the range, variance, and sd.
(3)If the values are all the same (no variation), all these measures will be zero .

(4)C.V
*The c.v is a measure of comparing of two dispersions or more, mathematically the standard deviation is divided by the mean.
*It is relative variability around the mean. This can be used to compare two distributions directly to see which has more dispersion because it does not depend on units of the distribution.

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

*props :
1.Useful for comparing the variability of two or more variables (compare b/w different things).
2. It is independent of unit of measurement.
3. Measures the relative variation.
4. Always in percentage (%).

(5)IQR
*the diff b/w Q3 & Q1.
*It is not sensitive to extreme values
*One problem w/ IQR that it does not take into account the variability of the total data (only the central 50%). We are "throwing out" half of the data.

Relative Standing → provides info about the position of an individual score value within a distribution scores.
Types :
(1)Percentile Ranks
^ It is the percentage of scores in the distribution that fall **at or below a given value**.

$$\text{Rank} = \left(\frac{\text{Number of scores below or equal to desired score}}{\text{Total number of scores}} \right) \times 100$$

Rank ⟶ Percentile Rank

^ Percentiles are symbolized by the letter P, with a subscript indicating the percentage below the score value.
^The statement P40= 55 means that 40% of the values in the distribution fall below the score 55.
(2)Standard scores
^There are scores that are expressed in terms of their relative distance from the mean.
^It provides information not only about rank but also distance b/w scores.
<It often called Z-score → Is a standard score that indicates how many SDs from the mean a particular values lies. >

Formula for the z-score:

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

^It tells you how many SDs (s) an observation is from the mean. Thus, it is a way of quickly assessing how "unusual" an observation is.