Task 4 {Basmala}

Measures of spread :

> gives an idea of how students differ.

“***it’s concerned with how far are points from one another.”***

▼ ***MEASURES OF SPREAD:***

- *Range*
- *Interquartile range*
- *Standard deviation*
- *Variance*

- -

<< Range & *Interquartile range*>>

HISTOGRAM:

“***Most common visual for quantitative data”.***

How it works:

▼ The histogram creator chooses how the `binning` occurs .

-`Binning`: the process of making a category from which the certain elements lies between certain limits.

Ex: the values 1,2,2,4 lies in a bin called from 1-4.

▼ The number of values determine the hight of each histogram bin.

5 number summary:

> one of the most common ways to measure the spread.

“Gives values for calculating the range and interquartile range”.

▼ **Consists of:**

- Minimum.
- First Quartile (Q1).
- Second Quartile “median”(Q2).
- Third Quartile(Q3).
- Maximum.

Side Note:

- First we order the values → which makes it easier to detect the minimum, maximum and the median(Q2)
- Second Quartile (median)⟶ “50% of the data or 2/4 fall bellow this value”.
- First & Third Quartile ⟶ “Are considered the medians of the data on either sides of Q2”.
- First Quartile ⟶ “25% of the data fall bellow this value”.
- Third Quartile ⟶“75% of the data fall bellow this value”.

`THE RANGE = MAX - MIN`

`Interquartile range = Q3 - Q1`

Box plot:

“the values of five number summary marked”

“Useful for quickly comparing the spread of two data sets across some key metrics like quartiles, maximum and minimum.”

- -

<< Variance & *Standard deviation*>>

“the most common way to measure the spread with only one value”

Standard deviation

Also called “root mean square error”

→ On average ,how much each point varies from the mean of the points in a dataset.

→ gives a measure of variation, or spread within this dataset.

→ used to compare spreads of different groups.

→ If there had been more variation between points, the standard deviation would have been even larger.

→ if there had been less variation the standard deviation would have been smaller.

→ is often deemed as a more useful measurement of spread as it shares the units of the original data set, while the variance shares units of original data set `squared` which doesn’t make sense.

$s.n$:

when data concerns money or economy having higher Standard deviation is associated with higher risk

for comparison to be fair : all data should be in the same unit

_________ ______- _______ ___

Variance

> Standard deviation = sqrt(Variance)

“Average square different of each observation from the mean (xi - x bar)^2 /number of elements”

Shape :

> Give a more complete picture.

“histograms are used to determine shape associated with data”.

“shape of distribution can tell us a lot about the measures and spread”.

Shapes of histogram:

1. Left skewed :

- has shorter bins on the left and taller ones on the right.
- mean < median.

- Ex on Left-skewed distribution:

- GPA
- Age of death
- Asset price changes

2. Right skewed :

- has taller bins on the left and shorter ones on the right.
- mean > median

- Ex on Right-skewed distribution:

- Amount of drug left in blood
- Wealth distribution
- Athletic abilities

3. Symmetric distribution:

- the right side mirrors the left side.
- ex: normal distribution (bell curve)(Gaussian distribution).
- mean = median = mode.

- Ex on Bell-shaped distribution:

- Heights

- Weights
- Scores
- Precipitation
- Mean of a distribution
- Errors in manufacturing process

Outliers :

> Data points that fall very far from the rest of the values in our dataset.

- standard deviation & mean are not great measures in this case.
- The median is a better measure of the center.
- outliers greatly increase the mean& standard deviation.
- Reporting the five maximum summary is better than the mean and standard deviation when outliers exist.

Bell-shaped data:

- You can find every little detail about the data by finding the mean and standard deviation.

Skewed data:

- Five-number summary is the best for this case.

— — — — — — — — — — — — — —

Descriptive statistics :

> Describing the data we’ve collected

“used regularly by scientists to briefly summarize the key features of a dataset or population”.

Scientists typically use descriptive statistics to:

1. Concisely summarize the characteristics of a population or dataset.
2. Determine the distribution of measurement errors or experimental uncertainty

Inferential statistics :

> Drawing conclusions about a population based on data collected from a sample of individuals from that population.

Population → entire group of interest.

Sample → subset from our population.

Statistics → any numeric summary calculated from the sample.

Parameter→ any numeric summary from the population.