Date / /
Page No.

Understanding Variability in Statistics

What is variability

Variability refers to how spread out or dispersed the values in a dataset are. It provides insight into how much the number differ from each other. In simpler terms, variability measures the extent to which individual data points in a distribution deviate from the center (such as the mean or median)

Key Term

Variability, spread and dispersion all describe the same concept how spread the data is

Just like central dependency describes the "center" of the data variability describes how much individual data points differ from that center.

Range: A Simple Measure of variability

The range is one of the most basic and simplest measures of variability. It gives us an idea of how spread out the values in a dataset by calculating the difference between the highest and lowest scores

Limitation: The range only uses the extreme values (highest and lowest) which means it is very sensitive to outliers. If the dataset has extreme values, the range can give a misleading impression of how much the values are spread out.

Date / /
Page No.

Interquartile Range (IQR)

The interquartile range is a measure of variability that focuses on the middle 50% of the data. The IQR is particularly useful because it is not affected by the extreme values or outliers, making it a more robust measure of variability compared to the range

How to calculate the Interquartile Range

IQR = 75th percentile - 25th Percentile

Real world Example

Suppose you have the test scores of students in two different classes, Class A and Class B. You calculate IQR for both classes

- Class A IQR 4
- Class B IQR 8.

The class B IQR is higher, indicating that the middle 50% of the test scores in class B are more spread out compared to class A. class A has less variability in its central scores, while class B shows greater variability

Summary

- IQR focuses on the spread of the middle (50%) of the data, making it more robust to outliers than the range.
- The IQR is useful for understanding the concentration or dispersion of data without being influenced by extreme values.

Variance

Variance is a measure of the spread of a distribution of data. It quantifies how much the individual data points deviate from the mean (or center) of the distribution. A high variance means the data points are more spread out from the mean.

Formula for variance

1. Population Variance $\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$

σ^2 is the population variance
x represent inidividual data point
μ is the population mean
N is the total number of data point in the population

How to compute variance (Example)

- For Quiz I data : 9, 9, 9, 8, 8, 8, 8, 7, 7, 7, 7, 7, 6, 6, 6, 6, 6, 6, 5, 5

step I Mean (μ) : 7.0

step 2 compute the deviation from the mean ~~afor~~ each data point

(9-7)	(6-7)
(9-7)	(6-7)
(9-7)	(5-7)
(8-7)	(5-7)
(8-7)	
~~(8-8)~~	
(8-7)	
(7-7)	
(6-7)	
~~(6-7)~~	
(6-7)	
(6-7)	

Step 3 Square Each deviation
Squares $2^2 = 4$

$1^2 = 1$

$0^2 = 0$

$(-1)^2 = 1$

$(-2)^2 = 4$

Step 4 find the average of squared deviation, by adding all squared deviation

$= 30$

Step 5 Divide the total number of scores (N = 20)

$$\frac{\Sigma(x-\mu)^2}{N} = \frac{30}{20} = 1.5$$

∴ Variance for Quiz 1 is 1.5

For a sample, Lets assume we have a sample score of 1, 2, 4, 5

Sample Mean (M) = $(1+2+4+5)/4 = 3$

Step 1 compute the deviation from the mean

$(1-3) = -2$

$(2-3) = -1$

$(4-3) = 1$

$(5-3) = 2$

Step Squared each deviation

4, 1, 1, 4

Date / /
Page No

Step3 find the avg of squared deviation.
Add square deviation = 10

Step4 Divide by N-1 = 3 (since it is a sample)

$\frac{10}{3} \approx 3.33$

Therefore the sample variance is approximately 3.33

Key Concepts

- Variance measures how spread out the values in distribution are.
- The formula for variance for a population uses (N) as the denominator, while for a sample, the formula uses (N-1) to correct for bias and better estimate the population variance
- Sample variance is slightly larger than population variance to account for the fact that a sample might not represent the full diversity of a population

Standard Deviation

The standard Deviation is a measure of the spread or dispersion of a set of data points around the mean. It provides a way of quantifying how much the individual data points deviate from the mean of the distribution. The sd is the square root of the variance, as it has the same unit of measurement as the data points, which makes it easier to interpret Compared to the variance.

Date / /
Page No.

Formula for standard deviation

1 Population standard deviation $\sigma = \sqrt{\frac{\Sigma (X-\mu)^2}{N}}$

where σ is the population standard deviation
X represent individual data points
μ is the population Mean
N is the total number od data points in the popu

2 Sample standard deviation $s = \sqrt{\frac{\Sigma (X-M)^2}{N}}$

where s is the sample standard deviation
X represents individual data points
M is the sample mean
N is the number of data points in the sample

How to compute standard deviation (Example)
for Quiz I Data: 9, 9, 9, 8, 8, 8, 8, 7, 7, 7, 7, 7, 6, 6, 6, 6, 6, 6, 5, 5
Mean (x) = 7.0

step 1 Calculate the variance which is 1.5

Step 2 Take the square root of variance to get the standard deviation - $\sqrt{1.5} \simeq 1.257$

so the standard deviation for quiz I is approximately 1.257

Key Concepts

- standard deviation is a useful measure of variability because it is expressed in the same units as the data itself; making it easier to interpret then variance
- The larger the (std), the more spread out the data points are around the mean
- The smaller the (std), the more concentrated the data points are around the mean

Key Take away: (Mean, Median and Mode)

The mean, median and mode are all measures of central location, each answering the question "Where is the center of the data set?

Mean: Provides the Arithmetic average. It is the most common used measure but can be influenced by outliers or skewed distribution

Median: The middle value when the dat is ordered. It is less affected by outliers and is often the best choice for skewed distributions

Mode: The most frequent value in the data. It is especially used for categorical data and situations when you're interested in identifying the most common value.

Date / /
Page No.

Key Takeaway (The range, standard deviation, variance)

The range, standard deviation and variance all provide measures of the variability or spread of a data set. They help answer the question, " How variable are the data :

- Range gives a simple measure of how far apart the maximum and minimum values are
- Variance gives a measure of how much each data point differs from the mean, but in squared units.
- Standard deviation is the square root of the variance and provides a measure of spread in the original units of the data .