Measures of Central Tendency:

In statistics, a central tendency (or measure of central tendency) is a central or typical value for a probability distribution. It may also be called a center or location of the distribution. Colloquially, measures of central tendency are often called averages. The term central tendency dates from the late 1920s.

The most common measures of central tendency are the arithmetic mean, the median and the mode. A central tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution. Occasionally authors use central tendency to denote "the tendency of quantitative data to cluster around some central value.

The central tendency of a distribution is typically contrasted with its dispersion or variability; dispersion and central tendency are the often characterized properties of distributions. Analysts may judge whether data has a strong or a weak central tendency based on its dispersion.

The following may be applied to one-dimensional data. Depending on the circumstances, it may be appropriate to transform the data before calculating a central tendency. Examples are squaring the values or taking logarithms. Whether a transformation is appropriate and what it should be, depend heavily on the data being analyzed:

1. The arithmetic mean:

(or mean or average) is the most commonly used and readily understood measure of central tendency in a data set. In statistics, the term average refers to any of the measures of central tendency. The arithmetic mean of a set of observed data is defined as being equal to the sum of the numerical values of each and every observation divided by the total number of observations.

Symbolically, if we have a data set consisting of the values $\{x_1, x_2, \cdots, x_n\}$ then the arithmetic mean $\bar{x}$ is defined by the formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\{x_1 + x_2 + \cdots + x_n\}}{n}$$

Example: find mean for these numbers $y_i = \{9\,,8\,,6\,,5\,,7\}$

Solve

$$\therefore \bar{y} = \frac{\sum y_i}{n} = \frac{9+8+6+5+7}{5} = \frac{35}{5} = 7$$

The arithmetic mean has several properties that make it useful, especially as a measure of central tendency. These include:

1. If numbers $\{x_1, x_2, \cdots, x_n\}$ have mean $(\bar{x})$, then $\sum(x_1 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0$. Since is the distance from a given number to the mean, one way to interpret this property is as saying that the numbers to the left of the mean are balanced by the numbers to the right of the mean. The mean is the only single number for which the residuals (deviations from the estimate) sum to zero. In same example above:

$$\sum(y_i - \bar{y}) = (9-7) + (8-7) + (6-7) + (5-7) + (7-7) = 0$$

2.If it is required to use a single number $\{x_1, x_2, \cdots, x_n\}$ as a "typical" value for a set of known numbers , then the arithmetic mean of the numbers does this best, in the sense of minimizing the sum of squared deviations from the typical value: the sum of $\sum(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$. In same example above:

$$\sum(y_i - \bar{y})^2 = (9-7)^2 + (8-7)^2 + (6-7)^2 + (5-7)^2 + (7-7)^2 = 10$$

2. The geometric mean:

In mathematics, the geometric mean is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the nth root of the product of n numbers, i.e., for a set of numbers $\{x_1, x_2, \cdots, x_n\}$ the geometric mean is defined as:

$$\bar{G} = \sqrt[n]{x_1 * x_2 * \cdots * x_n}$$

Example: find geometric mean for these numbers $y_i = \{9\ ,8\ ,6\ ,5\ ,7\}$

Solve

$$\bar{G} = \sqrt[5]{y_1 * y_2 * y_3 * y_4 * y_5} = \sqrt[5]{9 * 8 * 6 * 5 * 7} = \sqrt[5]{15120} = 6.85$$

A geometric mean is often used when comparing different items—finding a single "figure of merit" for these items—when each item has multiple properties that have different numeric ranges. For example, the geometric mean can give a meaningful "average" to compare two companies which are each rated at 0 to 5 for their environmental sustainability, and are rated at 0 to 100 for their financial viability. If an arithmetic mean were used instead of a geometric mean, the financial viability is given more weight because its numeric range is larger—so a small percentage change in the financial rating (e.g. going from 80 to 90) makes a much larger difference in the arithmetic mean than a large percentage change in environmental sustainability (e.g. going from 2 to 5). The use of a geometric mean "normalizes" the ranges being averaged, so that no range dominates the weighting, and a given percentage change in any of the properties has the same effect on the geometric mean. So, a 20% change in environmental sustainability from 4 to 4.8 has the same effect on the geometric mean as a 20% change in financial viability from 60 to 72.

3. Median:

The median is the value separating the higher half from the lower half of a data sample (a population or a probability distribution). For a data set, it may be thought of as the "middle" value. For example, in the data set$\{1, 3, 3, 6, 7, 8, 9\}$, the median is(6), the fourth largest, and also the fourth smallest, number in the sample. For a continuous probability distribution, the median is the value such that a number is equally likely to fall above or below it.

When $n = \{y_1 , y_2 , \cdots , y_n\}$ as observations and ascending arrange:

1. If (n) odd numbers then $\left(\bar{M}_e = \frac{n+1}{2}\right)$;

2. But at even numbers then$\left(\bar{M}_e = \frac{n}{2} + 1 , \bar{M}_e = \frac{n}{2}\right)$.

Example: find median for these numbers: $y_i = 80 , 82 , 76 , 87 , 84 \rightarrow n = 5$

Solve:

At first arranged numbers ascending as: $\{76 , 80 , 82 , 84 , 87\}$

Then

$$\therefore \bar{M}_e = \frac{n+1}{2} = \frac{5+1}{2} = \frac{6}{2} = 3 , \therefore \bar{M}_e = y_3 = 82$$

Example: find median for these numbers: $y_i = 5 , 4 , 8 , 7 , 3 , 12 , 9 , 2 \rightarrow n = 8$

Solve:

At first arranged numbers ascending as: $\{2 , 3 , 4 , 5 , 7, 8, 9, 12\}$

Then

$$\bar{M}_e = \frac{n}{2} + 1 , \frac{n}{2} \rightarrow \frac{8}{2} + 1 = \frac{8+2}{2} = \frac{10}{2} = 5 , \frac{8}{2} = 4$$

$$\therefore \bar{M}_e = \frac{y_4 + y_5}{2} = \frac{5+7}{2} = \frac{12}{2} = 6$$

4. Mode:

The mode of a set of data values is the value that appears most often. It is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

Example: find mode from these numbers$\{1, 2, 2, 3, 4, 7, 9\}$:

Solve:

Most frequent value in a data set =2

5. Weighted arithmetic mean:

The weighted arithmetic mean is similar to an ordinary arithmetic mean (the most common type of average), except that instead of each of the data points contributing equally to the final average, some data points contribute more than others. The notion of weighted mean plays a role in descriptive statistics and also occurs in a more general form in several other areas of mathematics.

$$\bar{y} = \frac{\sum y_i * w_i}{\sum w_i}$$

Example: find weighted arithmetic mean for this data:

Sites	S1	S2	S3	S4
y_i	30	35	40	25
w_i	80	75	60	90

Solve

Sites	y_i	w_i	$w_i * y_i$
S1	30	80	2400
S2	35	75	2625
S3	40	60	2400
S4	25	90	2250
$\sum$		305	9675

$$\bar{y} = \frac{\sum w_i * y_i}{w_i} = \frac{2400 + 2625 + 2400 + 2250}{305} = \frac{9675}{305} = 31.72$$

If all the weights are equal, then the weighted mean is the same as the arithmetic mean. While weighted means generally behave in a similar fashion to arithmetic means, they do have a few counterintuitive properties, as captured for instance in Simpson's paradox.

Measures of Dispersion or Variation Tendency:

A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.

Most measures of dispersion have the same units as the quantity being measured. In other words, if the measurements are in meters or seconds, so is the measure of dispersion. Examples of dispersion measures include:

1.Standard deviation

2.Interquartile range (IQR)

3.Range

4.Mean absolute difference (also known as Gini mean absolute difference)

5.Median absolute deviation (MAD)

6.Average absolute deviation (or simply called average deviation)

7.Distance standard deviation

These are frequently used (together with scale factors) as estimators of scale parameters, in which capacity they are called estimates of scale. Robust measures of scale are those unaffected by a small number of outliers, and include the IQR and MAD.

Other measures of dispersion are dimensionless. In other words, they have no units even if the variable itself has units. These include:

1.Coefficient of variation

2.Quartile coefficient of dispersion

3.Relative mean difference, equal to twice the Gini coefficient

There are other measures of dispersion:

1.Variance (the square of the standard deviation) – location-invariant but not linear in scale.

2.Variance-to-mean ratio – mostly used for count data when the term coefficient of dispersion is used and when this ratio is dimensionless, as count data are themselves dimensionless, not otherwise.

1. Range:

In statistics, the range of a set of data is the difference between the largest and smallest values.However, in descriptive statistics, this concept of range has a more complex meaning. The range is the size of the smallest interval which contains all the data and provides an indication of statistical dispersion. It is measured in the same units as the data. Since it only depends on two of the observations, it is most useful in representing the dispersion of small data sets:

$$a.\ y_i = 12\,,6\,,7\,,3\,,15\,,10\,,18\,,5 \rightarrow R = y_{max} - y_{min} = 18 - 3 = 15$$

$$b.\ y_i = 9\,,3\,,8\,,8\,,9\,,8\,,9\,,18 \rightarrow R = y_{max} - y_{min} = 18 - 3 = 15$$

2. Mean absolute difference:

The mean absolute difference (univariate) is a measure of statistical dispersion equal to the average absolute difference of two independent values drawn from a probability distribution. A related statistic is the relative mean absolute difference, which is the mean absolute difference divided by the arithmetic mean, and equal to twice the Gini coefficient. The mean absolute difference is also known as the absolute mean difference (not to be confused with the absolute value of the mean signed difference) and the Gini mean difference (GMD). The mean absolute difference is sometimes denoted by Δ or as MD.

$$MD = \frac{\sum|y_i - \bar{y}|}{n}$$

Example: find Median absolute deviation for these numbers $y_i = \{9\ ,8\ ,6\ ,5\ ,7\}$

Solve

$$\bar{y} = \frac{\sum y_i}{n} = \frac{35}{5} = 7$$

y_i	$y_i - \bar{y}$	$\lvert y_i - \bar{y}\rvert$
9	2	2
8	1	1
6	- 1	1
5	- 2	2
7	0	0
$\sum y_i = 35$	$\sum (y_i - \bar{y}) = 0$	$\sum \lvert y_i - \bar{y}\rvert = 6$

$$\therefore\ MAD\ = \frac{\sum |y_i - \bar{y}|}{n} = \frac{6}{5} = 1.20$$

3. Standard deviation

In statistics, the standard deviation (**SD**, also represented by the lower case Greek letter sigma **σ** or the Latin letter **s**) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

$$S = \sqrt{S^2} = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum y_i^2 - \frac{\sum (y_i)^2}{n}}{n-1}}$$

Where S^2 is called **Variance**

Example: find Standard deviation for these numbers $y_i = \{9\ ,8\ ,6\ ,5\ ,7\}$

Solve

$$\bar{y} = \frac{\sum y_i}{n} = \frac{35}{5} = 7$$

y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(y_i)^2$
9	2	4	81
8	1	1	64
6	- 1	1	36
5	- 2	4	25
7	0	0	49
$\sum y_i = 35$	$\sum (y_i - \bar{y}) = 0$	$\sum (y_i - \bar{y})^2 = 10$	$\sum (y_i)^2 = 255$

$$S = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.58$$

$$SS = \sum y_i^2 - \frac{\sum (y_i)^2}{n} = \sum 255 - \frac{\sum (35)^2}{5} = \frac{1275 - 1225}{5} = \frac{50}{5} = 10$$

$$\therefore S = \sqrt{\frac{\sum y_i^2 - \frac{\sum (y_i)^2}{n}}{n-1}} = \sqrt{\frac{10}{4}} = \sqrt{2.5} = 1.58$$

4. Coefficient of variation:

The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean multiplier with 100:

$$C.V. = \frac{S}{\bar{y}} * 100$$

The CV is widely used in analytical chemistry to express the precision and repeatability of an assay. It is also commonly used in fields such as engineering or physics when doing quality assurance studies and ANOVA gauge R&R. In addition, CV is utilized by economists and investors in economic models and in determining the volatility of a security.

Application: $\{(T_i) = (23), (38), (20), ((35)), (18), (13), (40), (30), (28), (33), (15), (25)\}$, then find the fallowing :

a) $\bar{T}_i$ **b)** $M.D._{T_i}$ **c)** $S^2_{T_i}$ **d)** $C.V.\%_{T_i}$

Solve

Rank	T_i	$T_i - \bar{T}$	$\lvert T_i - \bar{T}\rvert$	$(T_i - \bar{T})^2$	T_i^2
1	23	-3.50	3.50	12.25	529
2	38	11.50	11.50	132.25	1444
3	20	-6.50	6.50	42.25	400
4	35	8.50	8.50	72.25	1225
5	18	-8.50	8.50	72.25	324
6	13	-13.50	13.50	182.25	169
7	40	13.50	13.50	182.25	1600
8	30	3.50	3.50	12.25	900
9	28	1.50	1.50	2.25	784
10	33	6.50	6.50	42.25	1089
11	15	-11.50	11.50	132.25	225
12	25	-1.50	1.50	2.25	625
Σ	318	0.00	90	887	9314

a) $\bar{T}_i = \frac{\sum_{i=1}^{12} T_i}{n} = \frac{318}{12} = 26.50$

b) $M.D._{T_i} = \frac{\sum|T_i - \bar{T}|}{n} = \frac{90}{12} = 7.50$

c) $S^2_{T_i} = \frac{\sum(T - \bar{T})^2}{n-1} = \frac{887}{11} = 80.64$

or

$$S^2_T = \frac{\sum T_i^2 - \frac{\sum(T_i)^2}{n}}{n-1} = \frac{9314 - \frac{(318)^2}{12}}{11} = \frac{\frac{111768 - 101124}{12}}{11} = \frac{\frac{10644}{12}}{11} = \frac{887}{11} = 80.64$$

d) $C.V.\%_{T_i} = \frac{S_{T_i}}{\bar{T}} * 100$

$$S_{T_i} = \sqrt{S^2_{T_i}} = \sqrt{80.64} = 8.98$$

$$C.V.\%_{T_i} = \frac{S_{T_i}}{\bar{T}} * 100 = \frac{8.98}{26.50} * 100 = 33.89\%$$