Unit – 2 Measures of central tendency

Definition of Measures of Central Tendency

- A measure of central tendency is a measure that tells us where the middle of a bunch of data lies.
- The three most common measures of central tendency are the mean, the median, and the mode.

More about Measures of Central Tendency

- Mean: Mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers in a set of data. This is also known as average.
- Median: Median is the number present in the middle when the numbers in a set of data are arranged in ascending or descending order. If the number of numbers in a data set is even, then the median is the mean of the two middle numbers.
- Mode: Mode is the value that occurs most frequently in a set of data.

Examples of Measures of Central Tendency

- For the data 1, 2, 3, 4, 5, 5, 6, 7, 8 the measures of central tendency are

Mean = $\frac{1+2+3+4+5+5+6+7+8}{9} = \frac{41}{9} = 4.56$

Median = 5

Mode = 5

Solved Example on Measures of Central Tendency

Find the measures of central tendency for the data set 3, 7, 9, 4, 5, 4, 6, 7, and 9.

Choices:

A. Mean = 6, median = 6 and modes are 4, 7 and 9

B. Mean = 6, median = 6 and mode is 4

C. Mean = 6, median = 6 and modes are 4 and 9

D. Mean = 6, median = 9 and modes are 4, 7 and 9

Correct Answer: A

Solution:

Step 1: Mean, median and mode of a data set are the measures of central tendency.

Step 2: Mean of the data set = $\frac{sum\ of\ the\ data\ values}{number\ of\ the\ data\ values}$ [Formula.]

Step 3: $= \frac{3+7+9+4+5+4+6+7+9}{9}$ [Substitute the values.]

Step 4: $= \frac{54}{9} = 6$ [Add the data values in the numerator and divide.]

Step 5: The data set in the ascending order is 3, 4, 4, 5, 6, 7, 7, 9, and 9. So, Median of the set is 6. [Median is the middle data value of the ordered set.]

Step 6: Mode is/are the data value(s) that appear most often in the data set. So, the modes of the data set are 4, 7 and 9.

Step 7: So, the measures of central tendency of the given set of data are mean = 6, median = 6 and modes are 4, 7, and 9.

==========

Mean

Definition of Mean

- Mean of a set of numbers is the sum of the numbers divided by the number of items in the list. Mean of a set of n numbers a_1, a_2, a_3, ..., a_n is given by $\frac{a_1 + a_2 + a_3 + \ldots\ldots + a_n}{n}$.

More about Mean

- Mean can also be called as average or arithmetic mean.

Example of Mean

- In order to find the mean of 4, 5, 6, 3, and 7, first we have to add the numbers and then divide the sum by the number of items.
  4 + 5 + 6 + 3 + 7 = 25 i.e. the sum of the numbers is 25.
- Mean = $\frac{Sum\ of\ the\ values}{Number\ of\ items}$ = $\frac{4 + 5 + 6 + 3 + 7}{5}$ = $\frac{25}{5}$ = 5
- So, the mean of the data set 4, 5, 6, 3, and 7 is 5.

Solved Example on Mean

Find the mean weight of the data set shown.
5 lb, 48 lb, 31 lb, 31 lb, 41 lb, 20 lb, 19 lb, 5 lb

Choices:

A. 27 lb

B. 25 lb

C. 26 lb

D. 24 lb

Correct Answer: B

Solution:

Step 1: Mean weight = $\frac{\text{Sum of the weights}}{\text{Number of listed weights}}$

Step 2: Sum of the weights = 5 + 48 + 31 + 31 + 41 + 20 + 19 + 5

Step 3: = 200 lb [Add the weights.]

Step 4: Number of weights listed = 8

Step 5: Mean weight = $\frac{200}{8}$ = 25 lb [Substitute and simplify.]

Step 6: So, the mean weight of the data set is 25 lb.

==========

Median

Definition of Median

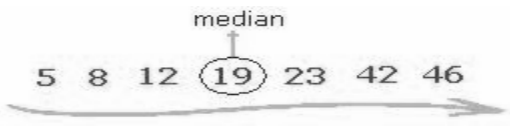
- Median is the middle data value of an ordered data set.

More about Median

- If there are two middle values, then the median is the mean of the two numbers.
- There will be two middle values when the number of values in the data set is even.

Examples of Median

- 12, 23, 8, 46, 5, 42, 19



The median in the above data set is 19.

 The median for the data set 2, 4, 7, 9, 3 is 4.
2, 3, 4, 7, and 9 is the ascending order of the data set 2, 4, 7, 9, 3. The middle number in the ordered data set is 4.

Let us find the median of a data set with even number of items in it, e.g. 33, 30, 42, 22, 18, and 31.

Arranging the above data set in ascending order, we find 18, 22, 30, 31, 33, and 42. The middle numbers from the above data set are 30 and 31. As there are two middle numbers we have to find the mean of those numbers.

$\frac{30+31}{2} = \frac{61}{2}$ = 30.5. So, 30.5 is the median (middle value) of the data set 33, 30, 42, 22, 18, and 31.

Solved Example on Median

The given data shows the number of burgers sold at a bakery in the last 14 weeks.
17, 13, 18, 17, 13, 16, 18, 19, 17, 13, 16, 18, 20, 19
Find the median number of burgers sold.

Choices:
A. 18.5

B. 17

C. 18

D. 17.5

Correct Answer: B

Solution:
Step 1: 13, 13, 13, 16, 16, 17, 17, 17, 18, 18, 18, 19, 19, 20 [Arrange the data in increasing order.]

Step 2: Number of observations, n = 14.

Step 3: n is an even number.

Step 4: Median is the mean of the 7th and 8th observations in the ordered list.

Step 5: Median = $\frac{17+17}{2}$ = 17

Step 6: So, the median number of burgers sold is 17.

==========

Mode

Definition of Mode

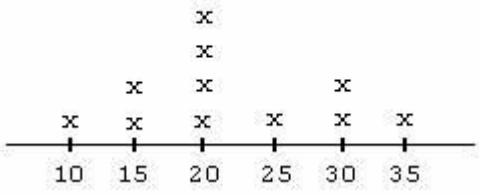
- Mode is a number that occurs most frequently in the data set.

More about Mode

- The data set with more than one mode is called Multimodal.

Examples of Mode

10, 20, 15, 20, 25, 30, 35, 20, 20, 30, 15



- In the given line plot, most number of cross (4) are shown against 20. So 20 is the mode of the given line plot.
- The mode of the set M, S, R, S, S, M, M, R, M, R is M, as M is occurred more frequently than S and R.
- 60, 55, 59, 56, 61, 62, 62, 62, 57, 61
  60 in the data set occur only once.
  55 in the data set occur only once.
  59 in the data set occur only once.
  56 in the data set occur only once.
  61 in the data set occur twice.
  62 in the data set occur thrice.
  57 in the data set occur only once.
  So, the mode for the above data set is 62 as it occurred most frequently.

Solved Example on Mode

The data shown below are the weights (in pounds) of different vegetables that Ashley bought.
16, 11, 14, 16, 7, 16, 14, 11, 16.
What is the mode of the data?
Choices:
A. 11
B. 7
C. 16
D. 14
Correct Answer: C
Solution:
Step 1: The number that occurs most frequently in a data set is called the mode.
Step 2: 16, 11, 14, 16, 7, 16, 14, 11, 16 [Original scores.]
Step 3: Since the number 16 appears four times, the mode of the data set is 16.

• Measures of variability

Definition of Measure of Variation

- Measure of variation is a measure that describes how spread out or scattered a set of data. It is also known as measures of dispersion or measures of spread.

Examples of Measure of Variation

- There are three measures of variation:
  The range, the variance, and the standard deviation.

Solved Example on Measure of Variation

The heights in cm of ten students are: 157, 152, 165, 151, 160, 156, 155, 162, 158, 163. Find the range of the data.

Choices:

A. 10
B. 13
C. 15
D. 14
Correct Answer: D

Solution:
Step 1: Maximum height = 165.
Step 2: Minimum height = 151.
Step 3: Range = 165 - 151 = 14. [Range = maximum height - minimum height.]

==========

Range

The range is the simplest measure of variation to find. It is simply the highest value minus the lowest value.

RANGE = MAXIMUM - MINIMUM

Since the range only uses the largest and smallest values, it is greatly affected by extreme values, that is - it is not resistant to change.

==========

Variance

Definition of Variance

- Variance is a statistical measure that tells us how measured data vary from the average value of the set of data.
- In other words, variance is the mean of the squares of the deviations from the arithmetic mean of a data set.

More about Variance

- Variance is the square of the standard deviation.
- The formula for variance

$$\sigma^2 = \left(\sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}} \right)^2$$

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}, \text{ where } x_1, x_2, x_3, \ldots, x_n \text{ are the values in the data set.}$$

Solved Example on Variance

Find the variance of the data set {1, 2, 3, 4, 10}.

Choices:

A. 10

B. 9

C. 8

D. 7

Correct answer: A

Solution:

Step 1: The mean of the data set {1, 2, 3, 4, 10} is $\bar{x} = \frac{1+2+3+4+10}{5} = 4$. [Use the formula for mean.]

Step 2: The standard deviation of the data set is

$$\sigma = \sqrt{\frac{1}{5}(1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (10-4)^2}$$

[Use the formula for mean.]

Step 3: $= \sqrt{\frac{50}{5}} = \sqrt{10}$

Step 4: The variance of the data set is $\sigma^2 = (\sqrt{10})^2 = 10$. [Substitute $\sigma = \sqrt{10}$.]

Standard Deviation

Standard deviation is a widely used measure of variability or diversity used in statistics and probability theory. It shows how much variation or "dispersion" there is from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values.

The standard deviation of a statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation.[1][2] A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.

In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times. The reported margin of error is typically about twice the standard deviation – the radius of a 95 percent confidence interval. In science, researchers commonly report the standard deviation of experimental data, and only effects that fall far outside the range of standard deviation are considered statistically significant – normal random error or variation in the measurements is in this way distinguished from causal variation. Standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the volatility of the investment.

When only a sample of data from a population is available, the population standard deviation can be estimated by a modified quantity called the sample standard deviation, explained below.

Basic examples

Consider a population consisting of the following eight values:

$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9$$

These eight data points have the mean (average) of 5:

$$\frac{2+4+4+4+5+5+7+9}{8} = 5$$

To calculate the population standard deviation, first compute the difference of each data point from the mean, and square the result of each:

$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16 \end{array}$$

Next compute the average of these values, and take the square root:

$$\sqrt{\frac{(9+1+1+1+0+0+4+16)}{8}} = 2$$

This quantity is the **population standard deviation**; it is equal to the square root of the variance. The formula is valid only if the eight values we began with form the complete population. If they instead were a random sample, drawn from some larger, "parent" population, then we should have used 7 (which is n − 1) instead of 8 (which is n) in the denominator of the last formula, and then the quantity thus obtained would have been called the **sample standard deviation**. See the section Estimation below for more details.

A slightly more complicated real life example, the average height for adult men in the United States is about 70", with a standard deviation of around 3". This means that most men (about 68%, assuming a normal distribution) have a height within 3" of the mean (67"–73") — one standard deviation — and almost all men (about 95%) have a height within 6" of the mean (64"–76") — two standard deviations. If the standard deviation were zero, then all men would be exactly 70" tall. If the standard deviation were 20", then men would have much more variable heights, with a typical range of about 50"–90". Three standard deviations account for 99.7% of the sample population being studied, assuming the distribution is normal (bell-shaped).

Definition of population values

Let X be a random variable with mean value μ:

$$\mathrm{E}[X] = \mu.$$

Here the operator E denotes the average or expected value of X. Then the **standard deviation** of X is the quantity

$$\sigma = \sqrt{\mathrm{E}\left[(X-\mu)^2\right]}.$$

That is, the standard deviation σ (sigma) is the square root of the variance of X, i.e., it is the square root of the average value of $(X - \mu)^2$.

The standard deviation of a (univariate) probability distribution is the same as that of a random variable having that distribution. Not all random variables have a standard deviation, since these expected values need not exist. For example, the standard deviation of a random variable that follows a Cauchy distribution is undefined because its expected value μ is undefined.

Discrete random variable

In the case where X takes random values from a finite data set $x_1, x_2, \ldots, x_N$, with each value having the same probability, the standard deviation is

$$\sigma = \sqrt{\frac{1}{N}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2\right]}, \quad \text{where} \quad \mu = \frac{1}{N}(x_1 + \cdots + x_N),$$

or, using summation notation,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \quad \text{where} \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

If, instead of having equal probabilities, the values have different probabilities, let x_1 have probability p_1, x_2 have probability p_2, ..., x_N have probability p_N. In this case, the standard deviation will be

Continuous random variable

The standard deviation of a continuous real-valued random variable X with probability density function p(x) is

$$\sigma = \sqrt{\int_{\mathbf{X}} (x - \mu)^2 \, p(x)\, dx}, \quad \text{where} \quad \mu = \int_{\mathbf{X}} x \, p(x)\, dx,$$

and where the integrals are definite integrals taken for x ranging over the set of possible values of the random variable X.

In the case of a parametric family of distributions, the standard deviation can be expressed in terms of the parameters. For example, in the case of the log-normal distribution with parameters μ and σ^2, the standard deviation is $[(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)]^{1/2}$.

Estimation

One can find the standard deviation of an entire population in cases (such as standardized testing) where every member of a population is sampled. In cases where that cannot be done, the standard deviation σ is estimated by examining a random sample taken from the population. Some estimators are given below:

With standard deviation of the sample

An estimator for σ sometimes used is the **standard deviation of the sample**, denoted by s_N and defined as follows:

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

This estimator has a uniformly smaller mean squared error than the sample standard deviation (see below), and is the maximum-likelihood estimate when the population is normally distributed. But this estimator, when applied to a small or moderately sized sample, tends to be too low: it is a biased estimator.

The standard deviation of the sample is the same as the population standard deviation of a discrete random variable that can assume precisely the values from the data set, where the probability for each value is proportional to its multiplicity in the data set.

With sample standard deviation

The most common estimator for σ used is an adjusted version, the **sample standard deviation**, denoted by s and defined as follows:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

where $\{x_1, x_2, ..., x_N\}$ are the observed values of the sample items and $\overline{x}$ is the mean value of these observations. This correction (the use of $N-1$ instead of N) is known as Bessel's correction. The reason for this correction is that s^2 is an unbiased estimator for the variance σ^2 of the underlying population, if that variance exists and the sample values are drawn independently with replacement. However, s is not an unbiased estimator for the standard deviation σ; it tends to overestimate the population standard deviation.

The term standard deviation of the sample is used for the uncorrected estimator (using N) while the term sample standard deviation is used for the corrected estimator (using $N-1$). The denominator $N-1$ is the number of degrees of freedom in the vector of residuals, $(x_1 - \overline{x}, ..., x_n - \overline{x})$.

Difference between standard deviation population vs Sample:

When all available values are used, it is called a population; when only a subset of available values is used, it is called a sample.

Things to remember:

σ is the population standard deviation which is usually unknown.

s is the sample standard deviation which is an estimate of the unknown population standard deviation.

The sum of squares is divided by 1 less than the sample size to account for the error in estimation from the sample standard deviation (called the degrees of freedom).[3]

Standard error

The **standard error** is the standard deviation of the sampling distribution of a statistic.[1] The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate.

For example, the sample mean is the usual estimator of a population mean. However, different samples drawn from that same population would in general have different values of the sample mean. The **standard error of the mean** (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population. Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analyzed at the time.

A way for remembering the term standard error is that, as long as the estimator is unbiased, the standard deviation of the error (the difference between the estimate and the true value) is the same as the standard deviation of the estimates themselves; this is true since the standard deviation of the difference between the random variable and its expected value is equal to the standard deviation of a random variable itself.

In practical applications, the true value of the standard deviation (of the error) is usually unknown. As a result, the term standard error is often used to refer to an estimate of this unknown quantity. In such cases it is important to be clear about what has been done and to attempt to take proper account of the fact that the standard error is only an estimate

Range

The **range** is the length of the smallest interval which contains all the data. It is calculated by subtracting the smallest observation (sample minimum) from the greatest (sample maximum) and provides an indication of statistical dispersion.

Absolute deviation

It is measured in the same units as the data. Since it only depends on two of the observations, it is a poor and weak measure of dispersion except when the sample size is large.

For a population, the range is greater than or equal to twice the standard deviation, with equality only for the coin toss (Bernoulli distribution with p = ½).

The range, in the sense of the difference between the highest and lowest scores, is also called the **crude range**. When a new scale for measurement is developed, then a potential maximum or minimum will emanate from this scale. This is called the **potential (crude) range**. Of course this range should not be chosen too small, in order to avoid a ceiling effect. When the measurement is obtained, the resulting smallest or greatest observation, will provide the **observed (crude) range**.

The midrange point, i.e. the point halfway between the two extremes, is an indicator of the central tendency of the data. Again it is not particularly robust for small samples.

In statistics, the **absolute deviation** of an element of a data set is the absolute difference between that element and a given point. Typically the point from which the deviation is measured is a measure of central tendency, most often the median or sometimes the mean of the data set.

$$D_i = | x_i - m(X) |$$

where

D_i is the absolute deviation,
x_i is the data element
and m(X) is the chosen measure of central tendency of the data set—sometimes the mean ($\overline{x}$), but most often the median.

Measures of dispersion

Several measures of statistical dispersion are defined in terms of the absolute deviation.

Average absolute deviation

The **average absolute deviation** or simply **average deviation** of a data set is the average of the absolute deviations and is a summary statistic of statistical dispersion or variability. It is also called the **mean absolute deviation,** but this is easily confused with the median absolute deviation.

The average absolute deviation of a set $\{x_1, x_2, ..., x_n\}$ is

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - m(X)|.$$

The choice of measure of central tendency, m(X), has a marked effect on the value of the average deviation. For example, for the data set {2, 2, 3, 4, 14}:

Measure of central tendency m(X)	**Average absolute deviation**
Mean = 5	$\frac{\lvert 2-5\rvert + \lvert 2-5\rvert + \lvert 3-5\rvert + \lvert 4-5\rvert + \lvert 14-5\rvert}{5} = 3.6$
Median = 3	$\frac{\lvert 2-3\rvert + \lvert 2-3\rvert + \lvert 3-3\rvert + \lvert 4-3\rvert + \lvert 14-3\rvert}{5} = 2.8$
Mode = 2	$\frac{\lvert 2-2\rvert + \lvert 2-2\rvert + \lvert 3-2\rvert + \lvert 4-2\rvert + \lvert 14-2\rvert}{5} = 3.0$

The average absolute deviation from the median is less than or equal to the average absolute deviation from the mean. In fact, the average absolute deviation from the median is always less than or equal to the average absolute deviation from any other fixed number.

The average absolute deviation from the mean is less than or equal to the standard deviation; one way of proving this relies on Jensen's inequality.

For the normal or "Gaussian" distribution, the ratio of mean absolute deviation to standard deviation is $\sqrt{2/\pi}=0.79788456...$. Thus if X is a normally distributed random variable with expected value 0 then

$$\frac{E|X|}{\sqrt{E(X^2)}} = \sqrt{\frac{2}{\pi}}.$$

In other words, for a Gaussian, mean absolute deviation is about 0.8 times the standard deviation.

Mean absolute deviation

The mean absolute deviation (MAD), also referred to as the mean deviation, is the mean of the absolute deviations of a set of data about the data's mean. In other words, it is the average distance of the data set from its mean during certain number of time periods.

The equation for MAD is as follows:

MAD = $1/n \sum(|e_i|)$, where $e_i = F_i - D_i$

This method forecast accuracy is very closely related to the mean squared error (MSE) method which is just the average squared error of the forecasts. Although these methods are very closely related MAD is more commonly used because it does not require squaring.

The equation for MSE is as follows:

MSE = $1/n \Sigma(e_i^2)$, where $e_i = F_i - D_i$

Median absolute deviation (MAD)

The median absolute deviation is the median of the absolute deviation from the median. It is a robust estimator of dispersion.

For the example {2, 2, 3, 4, 14}: 3 is the median, so the absolute deviations from the median are {1, 1, 0, 1, 11} (reordered as {0, 1, 1, 1, 11}) with a median of 1, in this case unaffected by the value of the outlier 14, so the median absolute deviation (also called MAD) is 1.

Maximum absolute deviation

The **maximum absolute deviation** about a point is the maximum of the absolute deviations of a sample from that point. It is realized by the sample maximum or sample minimum and cannot be less than half the range.

==========

Coefficient of variation

In probability theory and statistics, the **coefficient of variation (CV)** is a normalized measure of dispersion of a probability distribution. It is also known as **unitized risk** or the **variation coefficient.**

Definition

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ:

$$c_v = \frac{\sigma}{|\mu|}$$

which is the inverse of the signal-to-noise ratio. The CV is defined only for non-zero mean and the absolute value is taken for the mean to ensure it is always positive. It is sometimes expressed as a percent, in which case the CV is multiplied by 100%.

The coefficient of variation should be computed only for data measured on a ratio scale. To demonstrate this using an example of a group of temperatures is analyzed, the standard deviation does not depend on whether the Kelvin or Celsius scale is used since an object that changes its temperature by 1 K also changes its temperature by 1° C. However the mean temperature of the data set would differ in each scale by an amount of 273 and thus the coefficient of variation would differ. So the coefficient of variation may not have any meaning for data on an interval scale.

Comparison to standard deviation

Advantages

The coefficient of variation is useful because the standard deviation of data must always be understood in the context of the mean of the data. The coefficient of variation is a dimensionless number. So for comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.

Disadvantages

- When the mean value is close to zero, the coefficient of variation will approach infinity and is hence sensitive to small changes in the mean.
- Unlike the standard deviation, it cannot be used to construct confidence intervals for the mean.

==========

Correlation and dependence

In statistics, **dependence** refers to any statistical relationship between two random variables or two sets of data. **Correlation** refers to any of a broad class of statistical relationships involving dependence.

Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship.

Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several **correlation coefficients**, often denoted ρ or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function

of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation — that is, more sensitive to nonlinear relationships

Karl Pearson's product-moment coefficient

The most familiar measure of dependence between two quantities is the Pearson product-moment correlation coefficient, or "Pearson's correlation." It is obtained by dividing the covariance of the two variables by the product of their standard deviations. Karl Pearson developed the coefficient from a similar but slightly different idea by Francis Galton.

The population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where E is the expected value operator, cov means covariance, and, corr a widely used alternative notation for Pearson's correlation.

The Pearson correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy–Schwarz inequality that the correlation cannot exceed 1 in absolute value. The correlation coefficient is symmetric: corr(X,Y) = corr(Y,X).

The Pearson correlation is +1 in the case of a perfect positive (increasing) linear relationship (correlation), −1 in the case of a perfect decreasing (negative) linear relationship (**anticorrelation**), and some value between −1 and 1 in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship (closer to uncorrelated). The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables.

If the variables are independent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. For example, suppose the random variable X is symmetrically distributed about zero, and $Y = X^2$. Then Y is completely determined by X, so that X and Y are perfectly dependent, but their correlation is zero; they are uncorrelated. However, in the special case when X and Y are jointly normal, uncorrelatedness is equivalent to independence.

If we have a series of n measurements of X and Y written as x_i and y_i where i = 1, 2, ..., n, then the sample correlation coefficient can be used to estimate the population Pearson correlation r between X and Y. The sample correlation coefficient is written

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

where x and y are the sample means of X and Y, and s_x and s_y are the sample standard deviations of X and Y.

This can also be written as:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

If x and y are measurements that contain measurement error, as commonly happens in biological systems, the realistic limits on the correlation coefficient are not -1 to +1 but a smaller range

---

Linear regression

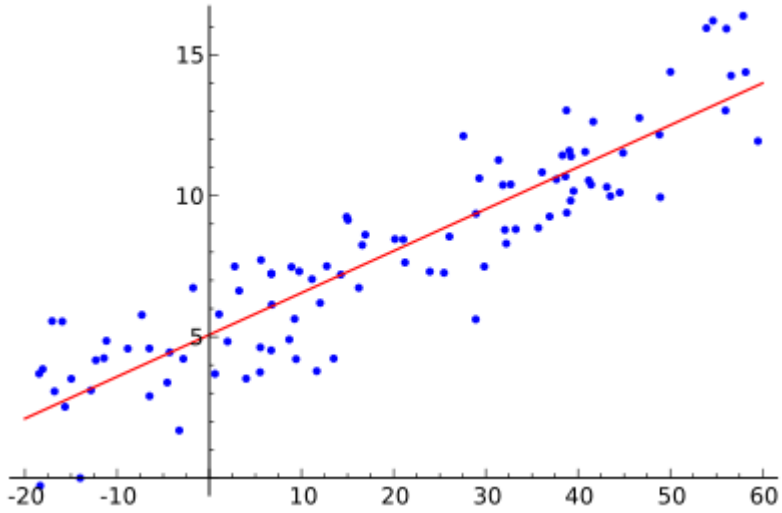
In statistics, **linear regression** is an approach to modeling the relationship between a scalar variable y and one or more variables denoted X. In linear regression, data are modeled using linear functions, and unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X. Less commonly, linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications of linear regression fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.
- Given a variable y and a number of variables X_1, ..., X_p that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j, to assess which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain redundant information about y.

Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the “lack of fit” in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression. Conversely, the least squares approach can be used to fit models that are not linear models. Thus, while the terms “least squares” and linear model are closely linked, they are not synonymous.



Introduction to linear regression

Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of regressors x_i is linear. This relationship is modeled through a so-called "disturbance term" ε_i — an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $'$ denotes the transpose, so that $x_i'\beta$ is the inner product between vectors x_i and β.

Often these n equations are stacked together and written in vector form as

$$y = X\beta + \varepsilon,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Some remarks on terminology and general use:

- y_i is called the regressand, endogenous variable, response variable, measured variable, or dependent variable (see dependent and independent variables.) The decision as to which variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables. Alternatively, there may be an operational reason to model one of the variables in terms of the others, in which case there need be no presumption of causality.

- x_i are called regressors, exogenous variables, explanatory variables, covariates, input variables, predictor variables, or independent variables (see dependent and independent variables, but not to be confused with independent random variables). The matrix X is sometimes called the design matrix.
    - Usually a constant is included as one of the regressors. For example we can take $x_{i1} = 1$ for $i = 1, ..., n$. The corresponding element of β is called the intercept. Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero.
    - Sometimes one of the regressors can be a non-linear function of another regressor or of the data, as in polynomial regression and segmented regression. The model remains linear as long as it is linear in the parameter vector β.
    - The regressors x_i may be viewed either as random variables, which we simply observe, or they can be considered as predetermined fixed values which we can choose. Both interpretations may be appropriate in different cases, and they generally lead to the same estimation procedures; however different approaches to asymptotic analysis are used in these two situations.
- β is a p-dimensional parameter vector. Its elements are also called effects, or regression coefficients. Statistical estimation and inference in linear regression focuses on β.
- ε_i is called the error term, disturbance term, or noise. This variable captures all other factors which influence the dependent variable y_i other than the regressors x_i. The relationship between the error term and the regressors, for example whether they are correlated, is a crucial step in formulating a linear regression model, as it will determine the method to use for estimation.

Example. Consider a situation where a small ball is being tossed up in the air and then we measure its heights of ascent h_i at various moments in time t_i. Physics tells us that, ignoring the drag, the relationship can be modeled as

$$h_i = \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_i,$$

where β_1 determines the initial velocity of the ball, β_2 is proportional to the standard gravity, and ε_i is due to measurement errors. Linear regression can be used to estimate the values of β_1 and β_2 from the measured data. This model is non-linear in the time variable, but it is linear in the parameters β_1 and β_2; if we take regressors $x_i = (x_{i1}, x_{i2}) = (t_i, t_i^2)$, the model takes on the standard form

$$h_i = x'_i\beta + \varepsilon_i.$$

Assumptions

Two key assumptions are common to all estimation methods used in linear regression analysis:

- The design matrix X must have full column rank p. For this property to hold, we must have $n > p$, where n is the sample size (this is a necessary but not a sufficient condition). If this condition fails this is called the multicollinearity in the regressors. In this case the parameter vector β will be not identifiable — at most we will be able to narrow down its value to some linear subspace of $\mathbf{R}^p$. Methods for fitting linear models with multicollinearity have been developed,[1][2][3][4] but require additional assumptions such as "effect sparsity" — that a large fraction of the effects are exactly zero.

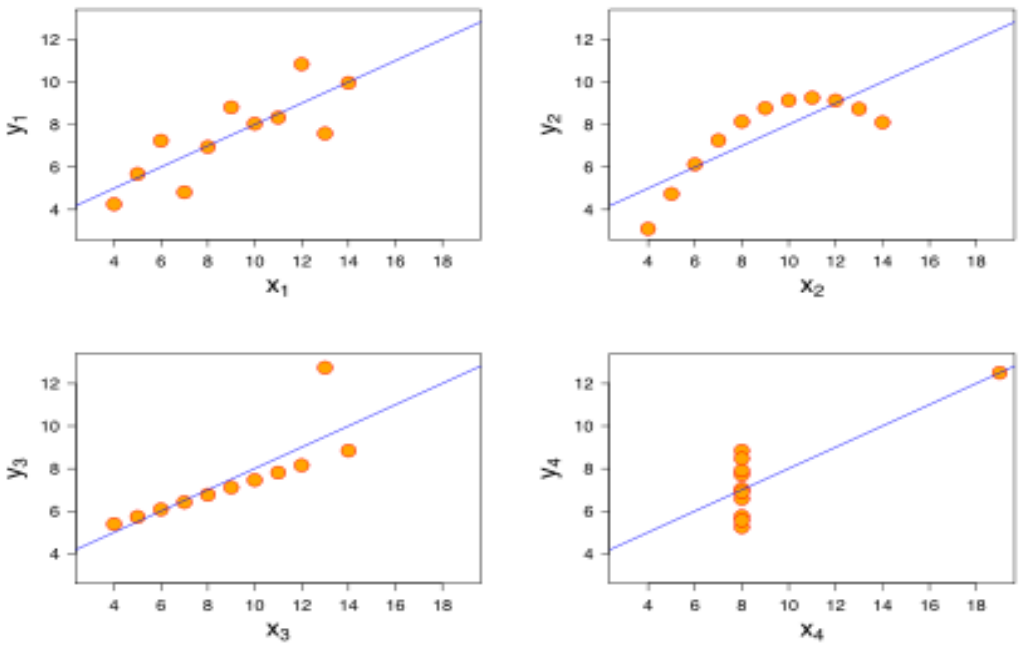
A simpler statement of this is that there must be enough data available compared to the number of parameters to be estimated. If there is too little data, then you end up with a system of equations with no unique solution. See partial least squares regression.

- The regressors x_i are assumed to be error-free, that is they are not contaminated with measurement errors. Although not realistic in many settings, dropping this assumption leads to significantly more difficult errors-in-variables models.

Beyond these two assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods:

- Some estimation methods are based on a lack of correlation, among the n observations $(y_i,\, x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n$. Statistical independence of the observations is not needed, although it can be exploited if it is known to hold.
- The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent.
- The variances of the error terms may be equal across the n units (termed homoscedasticity) or not (termed heteroscedasticity). Some linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity is present.
- The arrangement, or probability distribution of the predictor variables x has a major influence on the precision of estimates of β. Sampling and design of experiments are highly-developed subfields of statistics that provide guidance for collecting data in such a way to achieve a precise estimate of β.

Interpretation



The sets in the Anscombe's quartet have the same linear regression line but are themselves very different.

A fitted linear regression model can be used to identify the relationship between a single predictor variable x_j and the response variable y when all the other predictor variables in the model are "held fixed". Specifically, the interpretation of β_j is the expected change in y for a one-unit change in x_j when the other covariates are held fixed. This is sometimes called the unique effect of x_j on y. In contrast, the marginal effect of x_j on y can be assessed using a correlation coefficient or simple linear regression model relating x_j to y.

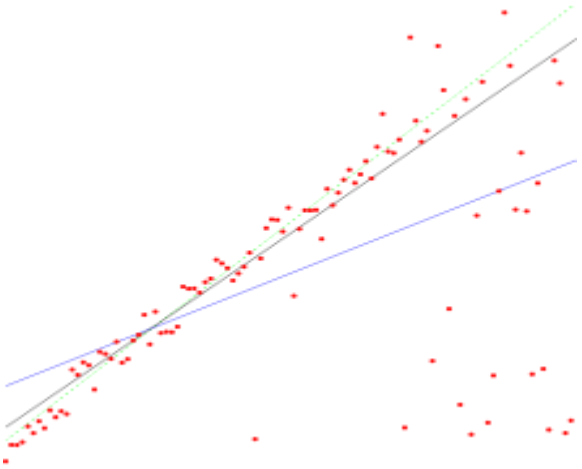
Care must be taken when interpreting regression results, as some of the regressors may not allow for marginal changes (such as dummy variables, or the intercept term), while others cannot be held fixed (recall the example from the introduction: it would be impossible to "hold t_i fixed" and at the same time change the value of t_i^2).

It is possible that the unique effect can be nearly zero even when the marginal effect is large. This may imply that some other covariate captures all the information in x_j, so that once that variable is in the model, there is no contribution of x_j to the variation in y. Conversely, the unique effect of x_j can be large while its marginal effect is nearly zero. This would happen if the other covariates explained a great deal of the variation of y, but they mainly explain variation in a way that is complementary to what is captured by x_j. In this case, including the other variables in the model reduces the part of the variability of y that is unrelated to x_j, thereby strengthening the apparent relationship with x_j.

The meaning of the expression "held fixed" may depend on how the values of the predictor variables arise. If the experimenter directly sets the values of the predictor variables according to a study design, the comparisons of interest may literally correspond to comparisons among units whose predictor variables have been "held fixed" by the experimenter. Alternatively, the expression "held fixed" can refer to a selection that takes place in the context of data analysis. In this case, we "hold a variable fixed" by restricting our attention to the subsets of the data that happen to have a common value for the given predictor variable. This is the only interpretation of "held fixed" that can be used in an observational study.

The notion of a "unique effect" is appealing when studying a complex system where multiple interrelated components influence the response variable. In some cases, it can literally be interpreted as the causal effect of an intervention that is linked to the value of a predictor variable. However, it has been argued that in many cases multiple regression analysis fails to clarify the relationships between the predictor variables and the response variable when the predictors are correlated with each other and are not assigned following a study design.[5]

Estimation methods



Comparison of the Theil–Sen estimator (black) and simple linear regression (blue) for a set of points with outliers.

Numerous procedures have been developed for parameter estimation and inference in linear regression. These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency.

Some of the more common estimation techniques for linear regression are summarized below.

- **Ordinary least squares** (OLS) is the simplest and thus most common estimator. It is conceptually simple and computationally straightforward. OLS estimates are commonly used to analyze both experimental and observational data. The OLS method minimizes the sum of squared residuals, and leads to a closed-form expression for the estimated value of the unknown parameter β:

  $$\hat{\beta} = (X'X)^{-1}X'y = \left(\tfrac{1}{n}\sum x_i x_i'\right)^{-1}\left(\tfrac{1}{n}\sum x_i y_i\right)$$

  The estimator is unbiased and consistent if the errors have finite variance and are uncorrelated with the regressors[6]

  $$\mathrm{E}[\, x_i \varepsilon_i \,] = 0.$$

  It is also efficient under the assumption that the errors have finite variance and are homoscedastic, meaning that $E[\varepsilon_i^2|x_i]$ does not depend on i. The condition that the errors are uncorrelated with the regressors will generally be satisfied in an experiment, but in the case of observational data, it is difficult to exclude the possibility of an omitted covariate z that is related to both the observed covariates and the response variable. The existence of such a covariate will generally lead to a correlation between the regressors and the response variable, and hence to an inconsistent estimator of β. The condition of homoscedasticity can fail with either experimental or observational data. If the goal is either inference or predictive modeling, the performance of OLS estimates can be poor if multicollinearity is present, unless the sample size is large.
  In simple linear regression, where there is only one regressor (with a constant), the OLS coefficient estimates have a simple form that is closely related to the correlation coefficient between the covariate and the response.

==========

ANOVA (Analysis of variance)

The Analysis Of Variance, popularly known as the ANOVA test, can be used in cases where there are more than two groups.

When we have only two samples we can use the t-test to compare the means of the samples but it might become unreliable in case of more than two samples. If we only compare two means, then the t-test (independent samples) will give the same results as the ANOVA.

It is used to compare the means of more than two samples. This can be understood better with the help of an example.

ONE WAY ANOVA

EXAMPLE: Suppose we want to test the effect of five different exercises. For this, we recruit 20 men and assign one type of exercise to 4 men (5 groups). Their weights are recorded after a few weeks.

We may find out whether the effect of these exercises on them is significantly different or not and this may be done by comparing the weights of the 5 groups of 4 men each.

The example above is a case of one-way balanced ANOVA.

It has been termed as one-way as there is only one category whose effect has been studied and balanced as the same number of men has been assigned on each exercise. Thus the basic idea is to test whether the samples are all alike or not.

WHY NOT MULTIPLE T-TESTS?

As mentioned above, the t-test can only be used to test differences between two means. When there are more than two means, it is possible to compare each mean with each other mean using many t-tests.

But conducting such multiple t-tests can lead to severe complications and in such circumstances we use ANOVA. Thus, this technique is used whenever an alternative procedure is needed for testing hypotheses concerning means when there are several populations.

ONE WAY AND TWO WAY ANOVA

Now some questions may arise as to what are the means we are talking about and why variances are analyzed in order to derive conclusions about means. The whole procedure can be made clear with the help of an experiment.

Let us study the effect of fertilizers on yield of wheat. We apply five fertilizers, each of different quality, on four plots of land each of wheat. The yield from each plot of land is

recorded and the difference in yield among the plots is observed. Here, fertilizer is a factor and the different qualities of fertilizers are called levels.

This is a case of one-way or one-factor ANOVA since there is only one factor, fertilizer. We may also be interested to study the effect of fertility of the plots of land. In such a case we would have two factors, fertilizer and fertility. This would be a case of two-way or two-factor ANOVA. Similarly, a third factor may be incorporated to have a case of three-way or three-factor ANOVA.

CHANCE CAUSE AND ASSIGNABLE CAUSE

In the above experiment the yields obtained from the plots may be different and we may be tempted to conclude that the differences exist due to the differences in quality of the fertilizers.

But this difference may also be the result of certain other factors which are attributed to chance and which are beyond human control. This factor is termed as “error”. Thus, the differences or variations that exist within a plot of land may be attributed to error.

Thus, estimates of the amount of variation due to assignable causes (or variance between the samples) as well as due to chance causes (or variance within the samples) are obtained separately and compared using an F-test and conclusions are drawn using the value of F.

ASSUMPTIONS

There are four basic assumptions used in ANOVA.

- the expected values of the errors are zero
- the variances of all errors are equal to each other
- the errors are independent
- they are normally distributed

In statistics, **analysis of variance** (**ANOVA**) is a collection of statistical models, and their associated procedures, in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups. Doing multiple two-sample t-tests would result in an increased chance of committing a type I error. For this reason, ANOVAs are useful in comparing two, three or more means.

Models

There are three classes of models used in the analysis of variance, and these are outlined here.

Fixed-effects models (Model 1)

The fixed-effects model of analysis of variance applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see if the

response variable values change. This allows the experimenter to estimate the ranges of response variable values that the treatment would generate in the population as a whole.

Random-effects models (Model 2)

Random effects models are used when the treatments are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments differ from ANOVA model 1.

Mixed-effects models (Model 3)

A mixed-effects model contains experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.