

STATISTICS

Mean or Average

Mean in theory is defined as the sum of all the elements of a set divided by the number of elements. We can get a fairly good idea about the whole set of data by calculating its mean. Thus the formula for mean is:

$$\text{Mean } (\bar{x}) = \frac{\sum x}{n}$$

Median

Median is the middle value of a dataset. So, if a set consists of an odd number of values, then the middle value will be the median of the set. On the other hand, if the set consists of an even number of sets, then the median will be the average of the two middle values.

Thus, the median may be used to separate a set of data into two parts. To find the median of a set, we need to arrange the elements of the set in increasing order. Then find the middle value.

The diagram shows the word "Median" on the left. Two blue arrows branch out from it to the right. The top arrow points to the text "n is odd," followed by the formula $\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$. The bottom arrow points to the text "n is even," followed by the formula $\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2} + 1\right)^{th} \text{ observation}}{2}$.

Median

n is odd,
 $\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$

n is even,
 $\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2} + 1\right)^{th} \text{ observation}}{2}$

Mode

The mode in a dataset is the value that is most frequent in the dataset. The mode also summarizes the dataset with single information.

Variance

It measures the deviation of a set of data from their mean value. Variance measures variability from the average or mean. The variance of the particular dataset will always be positive.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

where

μ = mean

N = Number of scores.

©easycalculation.com

Standard Deviation

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. The Standard Deviation is a measure of how spread out numbers are

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{n}}$$

where,

σ = population standard deviation

\sum = sum of...

μ = population mean

n = number of scores in sample.

Covariance

Covariance is a measure of the joint variability of two random variables.

If the two variables increase and decrease simultaneously then the covariance value will be positive.

Conversely if one increases while the other decreases then the covariance will be negative.

$$\text{Cov}(X,Y) = \sum E((X-\mu)(Y-v)) / n-1$$

where

X and Y are variable.

μ is the mean of the random variable X

v is the mean of the random variable Y

n is the number of items in the data set

Calculate covariance for the following data set:

x: 2.1, 2.5, 3.6, 4.0 (mean = 3.1)

y: 8, 10, 12, 14 (mean = 11)

$$\begin{aligned}\text{Cov}(X,Y) &= \sum E((X-\mu)(Y-v)) / n-1 \\ &= (2.1-3.1)(8-11)+(2.5-3.1)(10-11)+(3.6-3.1)(12-11)+(4.0-3.1)(14-11) / (4-1) \\ &= (-1)(-3) + (-0.6)(-1) + (.5)(1) + (0.9)(3) / 3 \\ &= 3 + 0.6 + .5 + 2.7 / 3 \\ &= 6.8/3 \\ &= 2.267\end{aligned}$$

Analysis of Variance

Analysis of Variance also termed as ANOVA. It is procedure followed by statisticians to check the potential difference between scale-level dependent variable by a nominal-level variable having two or more categories. It was developed by Ronald Fisher in 1918 and it extends t-test and z-test which compares only nominal level variable to have just two categories.

Types of ANOVA

One-way ANOVA - One-way ANOVA have only one independent variable and refers to numbers in this variable. For example, to assess differences in IQ by country, you can have 1, 2, and more countries data to compare.

Two-way ANOVA - Two-way ANOVA uses two independent variables. For example, to access differences in IQ by country (variable 1) and gender (variable 2). Here you can examine the interaction between two independent variables

Binomial distribution

A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes. For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

Eg. For toss a coin whether their head or tail. ii. forgiving exam wheather pass or fail. Here we get only two outcomes, so it is Binomial Distribution. Binomial Distribution is a Discrete Distribution.

The binomial distribution formula is:

$$b(x; n, P) = nC_x \cdot P^x \cdot (1 - P)^{n - x}$$

or

$$P(X) = \frac{n!}{(n - X)! X!} \cdot (p)^X \cdot (q)^{n - X}$$

Where:

b = binomial probability

x = total number of “successes” (pass or fail, heads or tails etc.)

P = probability of a success on an individual trial

n = number of trials

Normal distribution

A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. Height is one simple example of something that follows a normal distribution pattern: Most people are of average height and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

$$y = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma}$$

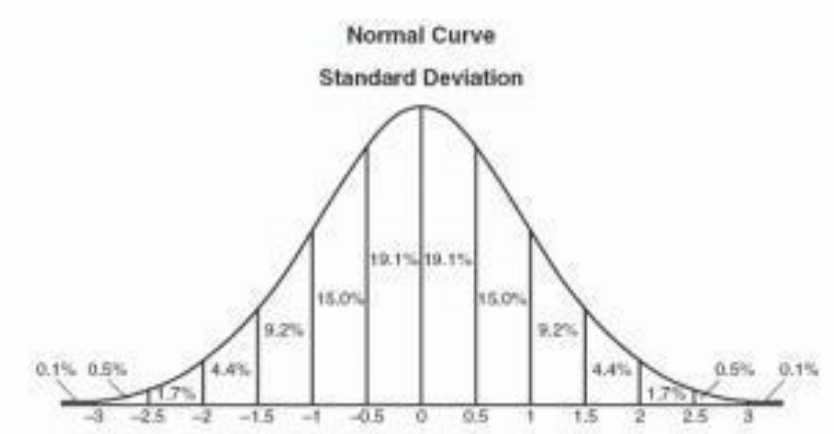
μ = Mean

σ = Standard Deviation

$\pi \approx 3.14159$

$e \approx 2.71828$

Here's an example of a normal distribution curve:



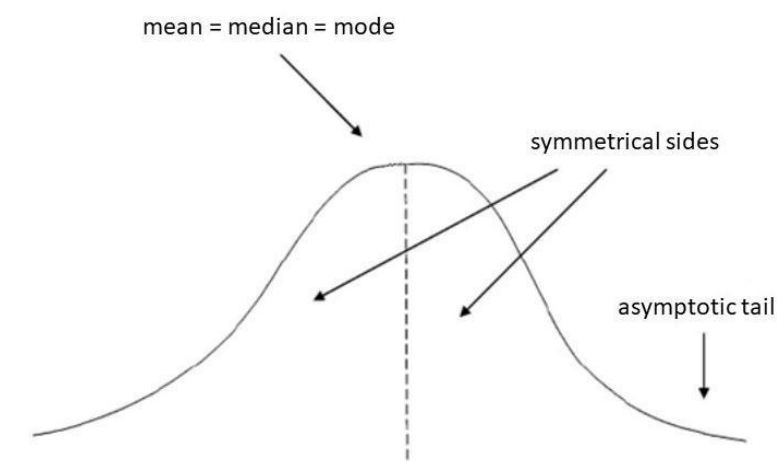
Normal distribution represents the behaviour of most of the situations in the universe (That is why it's called a "normal" distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the centre is a mirror image of the left side.

The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution tend to cluster around the mean, and the further a value is from the mean, the less likely it is to occur. The tails are asymptotic, which means that they approach but never quite meet the horizon (i.e. x-axis).

For a perfectly normal distribution the mean, median and mode will be the same value, visually represented by the peak of the curve.



The normal distribution is often called the bell curve because the graph of its probability density looks like a bell. It is also known as called Gaussian distribution, after the German mathematician Carl Gauss who first described it.

Poisson Distribution

The Poisson distribution is the discrete probability distribution of the number of events occurring in a given time period given the average number of times the event occurs over that time period.

In statistics, a Poisson distribution is a probability distribution that can be used to show how many times an event is likely to occur within a specified period of time. In other words, it is a count distribution. Poisson distributions are often used to understand independent events that occur at a constant rate within a given interval of time



The Poisson distribution is a discrete function, meaning that the variable can only take specific values in a (potentially infinite) list. Put differently, the variable cannot take all values in any continuous range. For the Poisson distribution (a discrete distribution), the variable can only take the values 0, 1, 2, 3, etc., with no fractions or decimals.

Poisson Distribution is a Discrete Distribution.

It is use to find probability in between some time period/ time interval.

Here we consider some time interval. The events are dependent on interval.

Poisson Distribution Formula


$$P(x) = \frac{e^{-\lambda} * \lambda^x}{x!}$$
 

Where:

e is Euler's number (e = 2.71828...)

x is the number of occurrences

x! is the factorial of x

λ is equal to the expected value of x when that is also equal to its variance

A Poisson distribution can be used to analyze the probability of various events regarding how many customers go through the drive-through. It can allow one to calculate the probability of a lull in activity (when there are 0 customers coming to the drive-through) as well as the probability of a flurry of activity (when there are 5 or more customers coming to the drive-through). This information can, in turn, help a manager plan for these events with staffing and scheduling.

Example

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

We know the following:

$\mu = 2$; since 2 homes are sold per day, on average.

$x = 3$; since we want to find the likelihood that 3 homes will be sold tomorrow.

$e = 2.71828$; since e is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$$P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$$

$$P(3; 2) = (2.71828^{-2}) (2^3) / 3!$$

$$P(3; 2) = (0.13534) (8) / 6$$

$$P(3; 2) = 0.180$$

Thus, the probability of selling 3 homes tomorrow is 0.180

Exponential Distribution

The exponential distribution is one of the widely used continuous distributions. It is often used to model the time elapsed between events. To predict the amount of waiting time until the next event (i.e., success, failure, arrival, etc.)

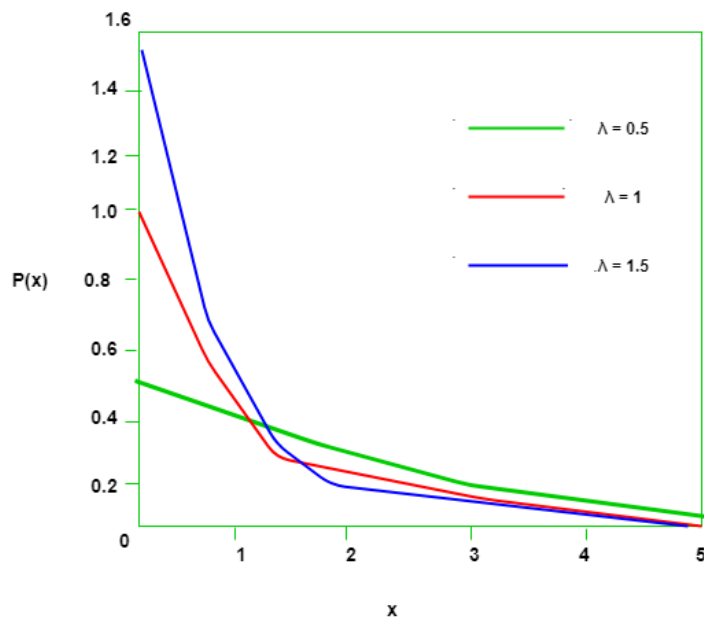
The exponential distribution is often concerned with the amount of time until some specific event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution.

A continuous random variable X is said to have an exponential distribution with parameter $\lambda > 0$, shown as $X \sim \text{Exponential}(\lambda)$ if its PDF is given by

$$f(x) = \{\lambda e^{-\lambda x}, x > 0$$

0, otherwise

The parameter λ is called rate parameter and its effects on the density function are illustrated below



For example, we want to predict the following:

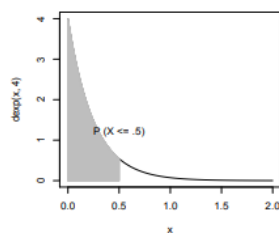
The amount of time until the customer finishes browsing and actually purchases something in your store (success).

The amount of time you need to wait until the bus arrives (arrival)

Example

If jobs arrive every 15 seconds on average, $\lambda = 4$ per minute, what is the probability of waiting less than or equal to 30 seconds, i.e .5 min?

$$P(T \leq .5).$$



$$\begin{aligned}
 P(T \leq .5) &= \int_0^{.5} 4e^{-4t} dt \\
 &= [-e^{-4t}]_{t=0}^{.5} \\
 &= 1 - e^{-2} \\
 &= 0.86
 \end{aligned}$$

Activate Windows
Go to PC settings to activate Windows.

Chi Square Distribution

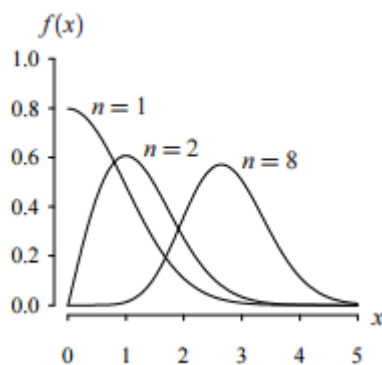
If X_1, X_2, \dots, X_m are m independent random variables having the standard normal distribution, then the following quantity follows a Chi-Squared distribution with m degrees of freedom. Its mean is m , and its variance is $2m$.

$$V = X_1^2 + X_2^2 + \dots + X_m^2 \sim \chi_{(m)}^2$$

This distribution describes the square root of a variable distributed according to a chi-square distribution.; with $df = n > 0$ degrees of freedom has a probability density function of:

$$f(x) = 2^{-(n/2+1)} x^{(n/2-1)} e^{-(x/2)} / \Gamma(n/2)$$

For values where x is positive.



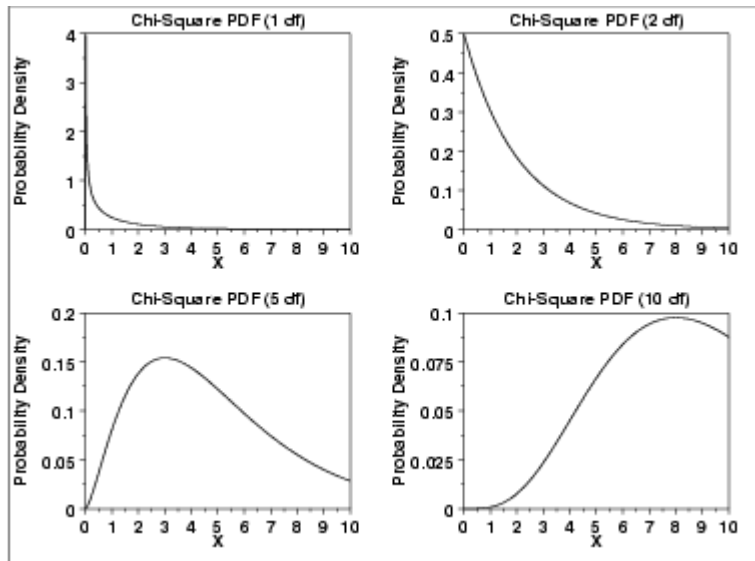
The Chi-square distribution results when v independent variables with standard normal distributions are squared and summed.

Chi-squared distribution is widely used by statisticians to compute the following:

To check the relationships between categorical variables

To conduct a The chi-square test (a goodness of fit test).

The following is the plot of the chi-square probability density function for 4 different values of the shape parameter.



Let's say you have a random sample taken from a normal distribution. The chi square distribution is the distribution of the sum of these random samples squared. The degrees of freedom (k) are equal to the number of samples being summed. For example, if you have taken 10 samples from the normal distribution, then $df = 10$. The degrees of freedom in a chi square distribution is also its mean. In this example, the mean of this particular distribution will be 10. Chi square distributions are always right skewed. However, the greater the degrees of freedom, the more the chi square distribution looks like a normal distribution