Variable

In statistics, a **variable** has two defining characteristics:

* A variable is an attribute that describes a person, place, thing, or idea.
* The value of the variable can "vary" from one entity to another.

For example, suppose we let the variable *x* represent the color of a person's hair. The variable *x* could have the value of "blond" for one person, and "brunette" for another.

When the value of a variable is the outcome of a [statistical experiment](https://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment), that variable is a **random variable**.

If a [variable](https://stattrek.com/Help/Glossary.aspx?Target=Variable) can take on any value between two specified values, it is called a **continuous variable**; otherwise, it is called a **discrete variable**.

Some examples will clarify the difference between discrete and continuous variables.

* Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
* Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

Probability Distribution

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence. Consider a simple experiment in which we flip a coin two times. An outcome of the experiment might be the number of heads that we see in two coin flips.

## **Discrete Probability Distributions**

If a [random variable](https://stattrek.com/Help/Glossary.aspx?Target=Random_variable) is a discrete variable, its [probability distribution](https://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) is called a **discrete probability distribution**.

An example will make this clear. Suppose you flip a coin two times. This simple [statistical experiment](https://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment) can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable.

The probability distribution for this statistical experiment appears below.

|  |  |
| --- | --- |
| **Number of heads** | **Probability** |
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

The above table represents a *discrete* probability distribution because it relates each value of a discrete random variable with its probability of occurrence. On this website, we will cover the following discrete probability distributions.

* [Binomial probability distribution](https://stattrek.com/Lesson2/Binomial.aspx)
* [Hypergeometric probability distribution](https://stattrek.com/Lesson2/Hypergeometric.aspx)
* [Negative binomial distribution](https://stattrek.com/probability-distributions/negative-binomial.aspx)
* [Poisson probability distribution](https://stattrek.com/Lesson2/Poisson.aspx)

## **Continuous Probability Distributions**

If a [random variable](https://stattrek.com/Help/Glossary.aspx?Target=Random_variable) is a continuous variable, its [probability distribution](https://stattrek.com/Help/Glossary.aspx?Target=Probability_distribution) is called a **continuous probability distribution**.

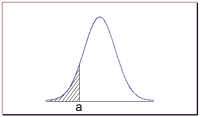
A continuous probability distribution differs from a discrete probability distribution in several ways.

* The probability that a continuous random variable will assume a particular value is zero.
* As a result, a continuous probability distribution cannot be expressed in tabular form.
* Instead, an equation or formula is used to describe a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a **probability density function**. Sometimes, it is referred to as a **density function**, a **PDF**, or a **pdf**. For a continuous probability distribution, the density function has the following properties:

* Since the continuous random variable is defined over a continuous range of values (called the **domain** of the variable), the graph of the density function will also be continuous over that range.
* The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
* The probability that a random variable assumes a value between *a* and *b* is equal to the area under the density function bounded by *a* and *b*.

For example, consider the probability density function shown in the graph below. Suppose we wanted to know the probability that the random variable *X* was less than or equal to *a*. The probability that *X* is less than or equal to *a* is equal to the area under the curve bounded by *a* and minus infinity - as indicated by the shaded area.



**Note:** The shaded area in the graph represents the probability that the random variable *X* is less than or equal to *a*. This is a [cumulative probability](https://stattrek.com/Help/Glossary.aspx?Target=Cumulative_probability). However, the probability that *X* is *exactly* equal to *a* would be zero. A continuous random variable can take on an infinite number of values. The probability that it will equal a specific value (such as *a*) is always zero.

On this website, we cover the following continuous probability distributions.

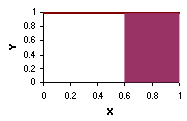
* [Normal probability distribution](https://stattrek.com/Lesson2/Normal.aspx)
* [Student's t distribution](https://stattrek.com/Lesson3/TDistribution.aspx)
* [Chi-square distribution](https://stattrek.com/Lesson3/ChiSquare.aspx)
* [F distribution](https://stattrek.com/Lesson3/FDistribution.aspx)

The probability distribution of a [continuous](https://stattrek.com/Help/Glossary.aspx?Target=Continuous%20variable) random variable is represented by an equation, called the **probability density function** (pdf). All probability density functions satisfy the following conditions:

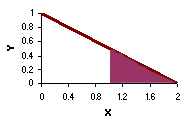
* The random variable Y is a function of X; that is, y = f(x).
* The value of y is greater than or equal to zero for all values of x.
* The total area under the curve of the function is equal to one.

The charts below show two continuous probability distributions. The first chart shows a probability density function described by the equation y = 1 over the range of 0 to 1 and y = 0 elsewhere. The second chart shows a probability density function described by the equation y = 1 - 0.5x over the range of 0 to 2 and y = 0 elsewhere. The area under the curve is equal to 1 for both charts.

**y = 1**



**y = 1 - 0.5x**



The probability that a continuous random variable falls in the interval between *a* and *b* is equal to the area under the pdf curve between *a* and *b*. For example, in the first chart above, the shaded area shows the probability that the random variable X will fall between 0.6 and 1.0. That probability is 0.40. And in the second chart, the shaded area shows the probability of falling between 1.0 and 2.0. That probability is 0.25.

**Note:** With a continuous distribution, there are an infinite number of values between any two data points. As a result, the probability that a continuous random variable will assume a particular value is always zero. For example, in both of the above charts, the probability that variable X will equal *exactly* 0.4 is zero.

# **Standard Normal Distribution**

The **standard normal distribution** is a special case of the [normal distribution](https://stattrek.com/Help/Glossary.aspx?Target=Normal%20distribution). It is the distribution that occurs when a [normal random variable](https://stattrek.com/Help/Glossary.aspx?Target=Normal%20random%20variable) has a mean of zero and a standard deviation of one.

## **Standard Score (aka, z-score)**

The normal random variable of a standard normal distribution is called a **standard score** or a **z-score**. Every normal random variable *X* can be transformed into a *z* score via the following equation:

*z* = (*X* - μ) / σ

where *X* is a normal random variable, μ is the mean of *X*, and σ is the standard deviation of *X*.

# **Chi-Square Distribution**

The distribution of the chi-square statistic is called the chi-square distribution. In this lesson, we learn to compute the chi-square statistic and find the probability associated with the statistic. And we'll work through some chi-square examples to illustrate key points.

## **The Chi-Square Statistic**

Suppose we conduct the following [statistical experiment](https://stattrek.com/Help/Glossary.aspx?Target=Statistical_experiment). We select a random sample of size *n* from a normal population, having a standard deviation equal to σ. We find that the standard deviation in our sample is equal to *s*. Given these data, we can define a [statistic](https://stattrek.com/Help/Glossary.aspx?Target=Statistic), called **chi-square**, using the following equation:

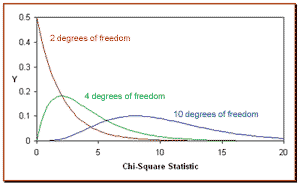
Χ2 = [ ( n - 1 ) \* s2 ] / σ2

The distribution of the chi-square statistic is called the chi-square distribution. The **chi-square distribution** is defined by the following [probability density function](https://stattrek.com/Help/Glossary.aspx?Target=Probability_density_function):

Y = Y0 \* ( Χ2 ) ( v/2 - 1 ) \* *e*-Χ2 / 2

where Y0 is a constant that depends on the number of degrees of freedom, Χ2 is the chi-square statistic, *v* = *n* - 1 is the number of [degrees of freedom](https://stattrek.com/Help/Glossary.aspx?Target=Degrees_of_freedom), and *e* is a constant equal to the base of the natural logarithm system (approximately 2.71828). Y0 is defined, so that the area under the chi-square curve is equal to one.

In the figure below, the red curve shows the distribution of chi-square values computed from all possible samples of size 3, where degrees of freedom is *n* - 1 = 3 - 1 = 2. Similarly, the green curve shows the distribution for samples of size 5 (degrees of freedom equal to 4); and the blue curve, for samples of size 11 (degrees of freedom equal to 10).



The chi-square distribution has the following properties:

* The mean of the distribution is equal to the number of degrees of freedom: μ = *v*.
* The variance is equal to two times the number of degrees of freedom: σ2 = 2 \* *v*
* When the degrees of freedom are greater than or equal to 2, the maximum value for Y occurs when Χ2 = *v* - 2.
* As the degrees of freedom increase, the chi-square curve approaches a normal distribution.

## **Test Your Understanding**

**Problem 1**

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

**Solution**

We know the following:

* The standard deviation of the population is 4 minutes.
* The standard deviation of the sample is 6 minutes.
* The number of sample observations is 7.

To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

Χ2 = [ ( n - 1 ) \* s2 ] / σ2   
Χ2 = [ ( 7 - 1 ) \* 62 ] / 42 = 13.5

where Χ2 is the chi-square statistic, *n* is the sample size, *s* is the standard deviation of the sample, and σ is the standard deviation of the population.

**Problem 2**  
  
Let's revisit the problem presented above. The manufacturing department ran a quality control test, using 7 randomly selected batteries. In their test, the standard deviation was 6 minutes, which equated to a chi-square statistic of 13.5.

Suppose they repeated the test with a new random sample of 7 batteries. What is the probability that the standard deviation in the new test would be greater than 6 minutes?

**Solution**

We know the following:

* The sample size *n* is equal to 7.
* The degrees of freedom are equal to *n* - 1 = 7 - 1 = 6.
* The chi-square statistic is equal to 13.5 (see Example 1 above).

Given the degrees of freedom, we can determine the cumulative probability that the chi-square statistic will fall between 0 and any positive value. To find the cumulative probability that a chi-square statistic falls between 0 and 13.5, we enter the degrees of freedom (6) and the chi-square statistic (13.5) into the [Chi-Square Distribution Calculator](https://stattrek.com/Tables/ChiSquare.aspx). The calculator displays the cumulative probability: 0.96.

This tells us that the probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This means (by the [subtraction rule](https://stattrek.com/Help/Glossary.aspx?Target=Subtraction_rule)) that the probability that the standard deviation would be *greater than* 6 minutes is 1 - 0.96 or 0.04.

Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans but not All-Stars, and 10% are veteran All-Stars.

Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

* **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  + Null hypothesis: The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.
  + Alternative hypothesis: At least one of the proportions in the null hypothesis is false.
* **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a [chi-square goodness of fit test](https://stattrek.com/Help/Glossary.aspx?Target=Chi-square%20goodness%20of%20fit%20test) of the null hypothesis.
* **Analyze sample data**. Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the [degrees of freedom](https://stattrek.com/Help/Glossary.aspx?Target=Degrees%20of%20freedom), we determine the [P-value](https://stattrek.com/Help/Glossary.aspx?Target=P-value).

DF = k - 1 = 3 - 1 = 2 (Ei) = n \* pi  
(E1) = 100 \* 0.30 = 30  
(E2) = 100 \* 0.60 = 60  
(E3) = 100 \* 0.10 = 10  
Χ2 = Σ [ (Oi - Ei)2 / Ei ]   
Χ2 = [ (50 - 30)2 / 30 ] + [ (45 - 60)2 / 60 ] + [ (5 - 10)2 / 10 ]  
Χ2 = (400 / 30) + (225 / 60) + (25 / 10) = 13.33 + 3.75 + 2.50 = 19.58

where DF is the degrees of freedom, k is the number of levels of the categorical variable, n is the number of observations in the sample, Ei is the expected frequency count for level i, Oi is the observed frequency count for level i, and Χ2 is the chi-square test statistic.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.58.

We use the [Chi-Square Distribution Calculator](https://stattrek.com/Tables/ChiSquare.aspx) to find P(Χ2 > 19.58) = 0.0001.

* **Interpret results**. Since the P-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis.

**Note:** If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the variable under study was categorical, and each level of the categorical variable had an expected frequency count of at least 5.