Understanding Probability Distributions

By Jim Frost — 40 Comments

A probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume. In other words, the values of the variable vary based on the underlying probability distribution.

Suppose you draw a random sample and measure the heights of the subjects. As you measure heights, you can create a distribution of heights. This type of distribution is useful when you need to know which outcomes are most likely, the spread of potential values, and the likelihood of different results.

In this blog post, you'll learn about probability distributions for both discrete and continuous variables. I'll show you how they work and examples of how to use them.

General Properties of Probability Distributions

Probability distributions indicate the likelihood of an event or outcome. Statisticians use the following notation to describe probabilities:

p(x) = the likelihood that random variable takes a specific value of x.

The sum of all probabilities for all possible values must equal 1. Furthermore, the probability for a particular value or range of values must be between 0 and 1.

Probability distributions describe the dispersion of the values of a random variable. Consequently, the kind of variable determines the type of probability distribution. For a

single random variable, statisticians divide distributions into the following two types:

- Discrete probability distributions for discrete variables
- Probability density functions for continuous variables

0

You can use equations and tables of variable values and probabilities to represent a probability distribution. However, I prefer graphing them using probability distribution plots. As you'll see in the examples that follow, the differences between discrete and continuous probability distributions are immediately apparent. You'll see why I love these graphs!

Related post: Data Types and How to Use Them

Discrete Probability Distributions

Discrete probability functions are also known as probability mass functions and can assume a discrete number of values. For example, coin tosses and counts of events are discrete functions. These are discrete distributions because there are no in-between values. For example, you can have only heads or tails in a coin toss. Similarly, if you're counting the number of books that a library checks out per hour, you can count 21 or 22 books, but nothing in between.

For discrete probability distribution functions, each possible value has a non-zero likelihood. Furthermore, the probabilities for all possible values must sum to one. Because the total probability is 1, one of the values must occur for each opportunity.

For example, the likelihood of rolling a specific number on a die is 1/6. The total probability for all six values equals one. When you roll a die, you inevitably obtain one of the possible values.

If the discrete distribution has a finite number of values, you can display all the values with their corresponding probabilities in a table. For example, according to a study, the likelihood for the number of cars in a California household is the following:

Number of Cars	Probability
0	0.03
1	0.13
2	0.70
3	0.10
4+	0.04

0

Types of Discrete Distribution

There are a variety of discrete probability distributions that you can use to model different types of data. The correct discrete distribution depends on the properties of your data. For example, use the:

- Binomial distribution to model binary data, such as coin tosses.
- Poisson distribution to model count data, such as the count of library book checkouts per hour.
- Uniform distribution to model multiple events with the same probability, such as rolling a die.

To learn more in depth about several probability distributions that you can use with binary data, read my post Maximize the Value of Your Binary Data.

To learn how to determine whether a specific discrete distribution is appropriate for your data, read my post Goodness-of-Fit Tests for Discrete Distributions.

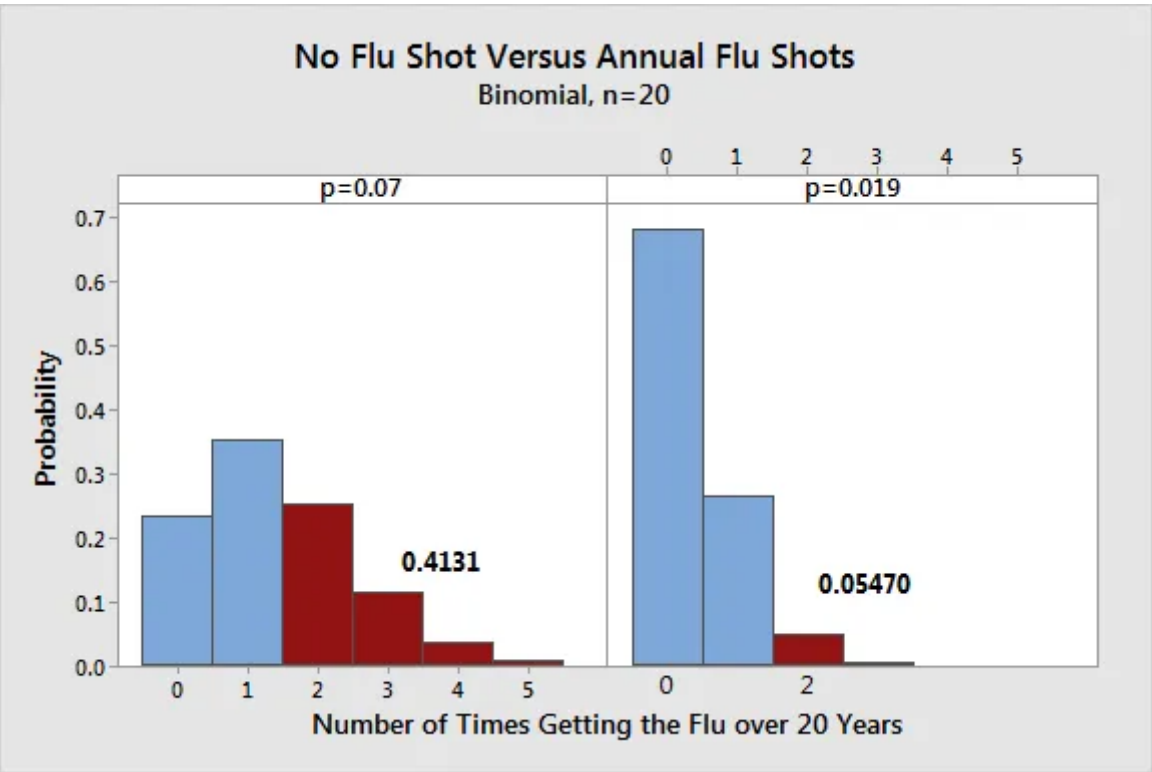
Example of How to Use Discrete Probability Distributions

All of the examples I include in this post will show you why I love to graph probability distributions. The case below comes from my blog post that presents a statistical analysis of flu shot effectiveness. I use the binomial distribution to answer the question—how many times can I expect to catch the flu over 20 years with and without annual vaccinations?

This example uses binary data because the two possible outcomes are either being infected by the flu or not being infected by the flu. Based on various studies, the long-term probability of a flu infection is 0.07 annually for the unvaccinated and 0.019 for the vaccinated. The graph plugs these probabilities into the binomial distribution to display the pattern of outcomes for both scenarios over twenty years. Each bar indicates the likelihood of catching the flu the specified number of times. Additionally, I've shaded

the bars red to represent the cumulative probability of at least two flu infections in 20 years. The left panel displays the expected outcomes with no vaccinations while th right panel shows the outcomes with annual vaccinations.

0



A significant difference jumps out at you—which demonstrates the power of probability distribution plots! The largest bar on the graph is the one in the right panel that represents zero cases of the flu in 20 years when you get flu shots. When you vaccinate annually, you have a 68% chance of not catching the flu within 20 years! Conversely, if you don't vaccinate, you have only a 23% of escaping the flu entirely.

In the left panel, the distribution spreads out much further than in the right panel. Without vaccinations, you have a 41% chance of getting the flu at least twice in 20 years compared to 5% with annual vaccinations. Some unlucky unvaccinated folks will get the flu four or five times in that time span!

Continuous Probability Distributions

Continuous probability functions are also known as probability density functions. You know that you have a continuous distribution if the variable can assume an infinite number of values between any two values. Continuous variables are often measurements on a scale, such as height, weight, and temperature.

Unlike discrete probability distributions where each particular value has a non-zero likelihood, specific values in continuous distributions have a zero probability. For example, the likelihood of measuring a temperature that is exactly 32 degrees is ze

0

Why? Consider that the temperature can be an infinite number of other temperatures that are infinitesimally higher or lower than 32. Statisticians say that an individual value has an infinitesimally small probability that is equivalent to zero.

How to Find Probabilities for Continuous Data

Probabilities for continuous distributions are measured over ranges of values rather than single points. A probability indicates the likelihood that a value will fall within an interval. This property is straightforward to demonstrate using a probability distribution plot—which we'll get to soon!

On a probability plot, the entire area under the distribution curve equals 1. This fact is equivalent to how the sum of all probabilities must equal one for discrete distributions. The proportion of the area under the curve that falls within a range of values along the X-axis represents the likelihood that a value will fall within that range. Finally, you can't have an area under the curve with only a single value, which explains why the probability equals zero for an individual value.

Characteristics of Continuous Probability Distributions

Just as there are different types of discrete distributions for different kinds of discrete data, there are different distributions for continuous data. Each probability distribution has parameters that define its shape. Most distributions have between 1-3 parameters. Specifying these parameters establishes the shape of the distribution and all of its probabilities entirely. These parameters represent essential properties of the distribution, such as the central tendency and the variability.

Related posts: Understanding Measures of Central Tendency and Understanding Measures of Variability

The most well-known continuous distribution is the normal distribution, which is also known as the Gaussian distribution or the "bell curve." This symmetric distribution fits a wide variety of phenomena, such as human height and IQ scores. It has two parameters —the mean and the standard deviation. The Weibull distribution and the lognormal distribution are other common continuous distributions. Both of these distributions can fit skewed data.

Distribution parameters are values that apply to entire populations. Unfortunately, population parameters are generally unknown because it's usually impossible to measure an entire population. However, you can use random samples to calculate estimates of these parameters.

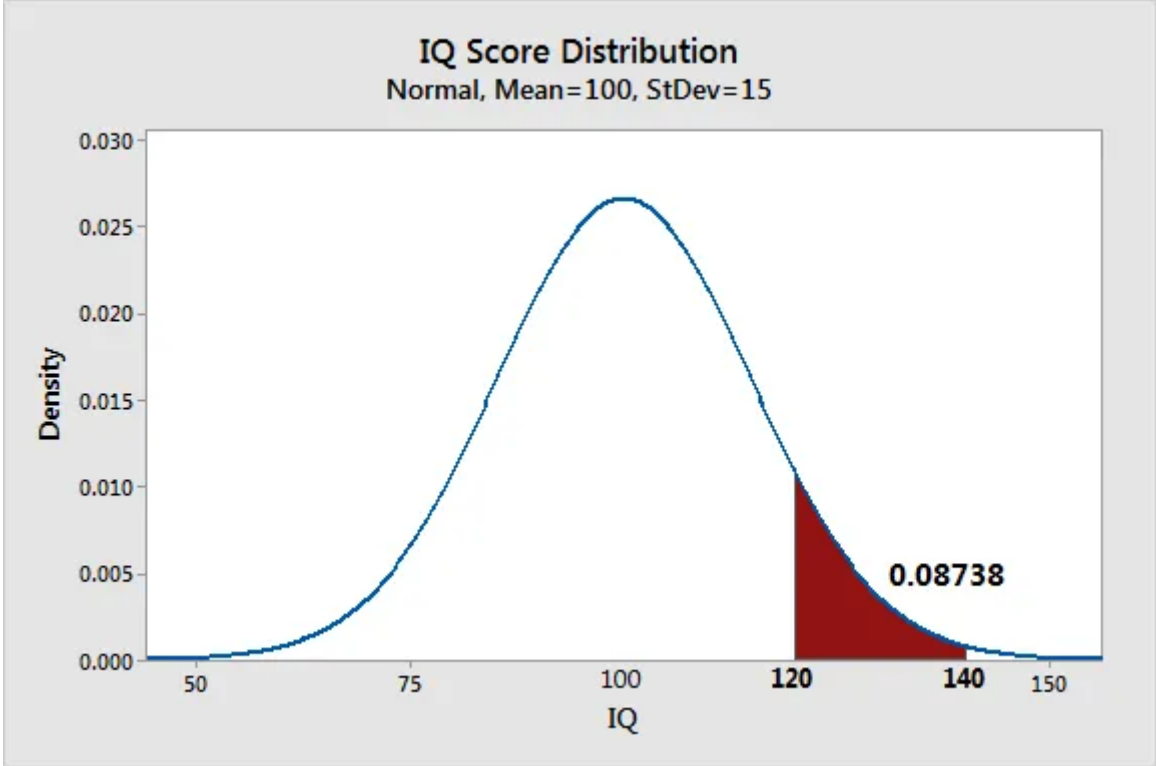
0

To learn how to determine which distribution provides the best fit to your sample data, read my post about How to Identify the Distribution of Your Data.

Example of Using the Normal Probability Distribution

Let's start off with the normal distribution to show how to use continuous probability distributions.

The distribution of IQ scores is defined as a normal distribution with a mean of 100 and a standard deviation of 15. We'll create the probability plot of this distribution. Additionally, let's determine the likelihood that an IQ score will be between 120-140.



Examine the properties of the probability plot above. We can see that it is a symmetric distribution where values occur most frequently around 100, which is the mean. The probabilities drops-off as you move away from the mean in both directions. The shaded area for the range of IQ scores between 120-140 contains 8.738% of the total area under the curve. Therefore, the likelihood that an IQ score falls within this range is 0.08738.

Related Post: Using the Normal Distribution

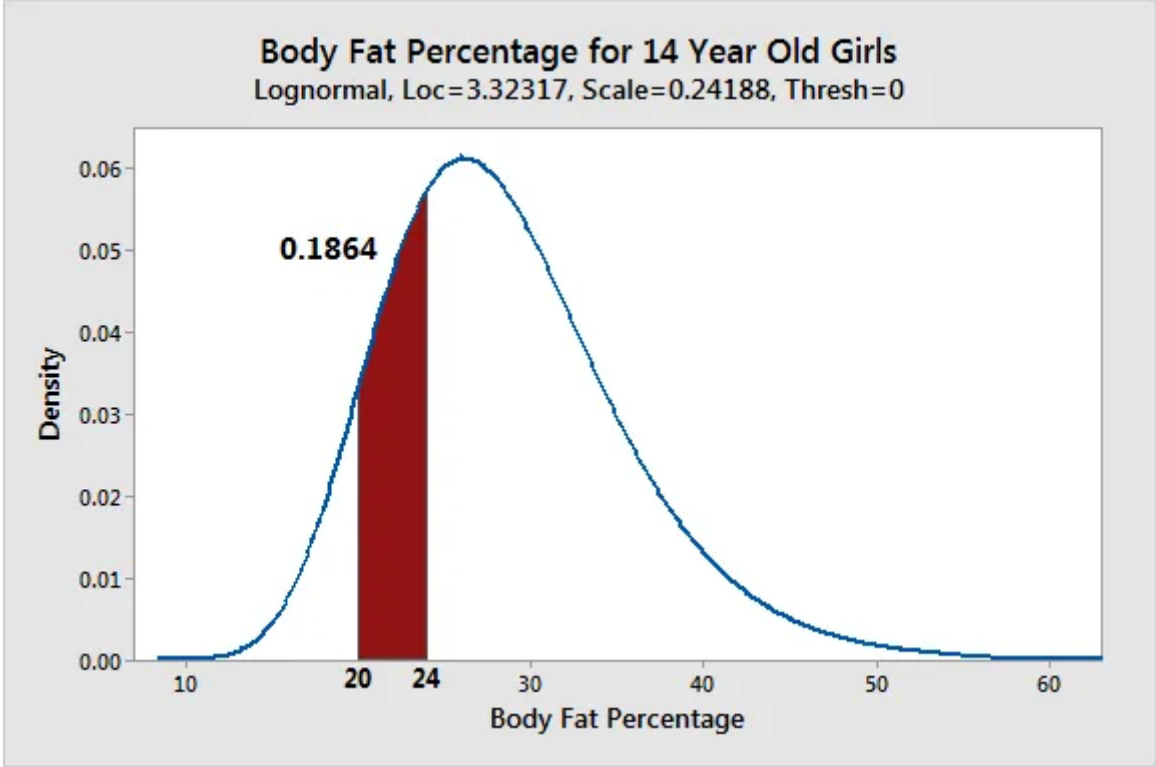
0

Example of Using the Lognormal Probability Distribu

As I mentioned, I really like probability distribution plots because they make distribution properties crystal clear. In the example above, we used the normal distribution. Because that distribution is so well-known, you might have guessed the general appearance of the chart. Now, let's look at a less intuitive example.

Suppose you are told that the body fat percentages for teenage girls follow a lognormal distribution with a location of 3.32317 and a scale of 0.24188. Furthermore, you're asked to determine the probability that body fat percentage values will fall between 20-24%. Huh? It's probably not clear what the shape of this distribution is, which values are most common, and how often values fall within that range!

Most statistical software allow you to plot probability distributions and answer all of these questions at once.



The graph displays both the shape of the distribution and how our range of interest fits within it. We can see that it is a right-skewed distribution and the most common values fall near 26%. Furthermore, our range of interest falls below the curve's peak and contains 18.64% of the occurrences.

As you can see, these graphs are an effective way to report complex distribution information to a lay audience.

0

This distribution provides the best fit for data that I collected for a study. Learn ho identified the distribution of these data.

Hypothesis Testing Uses Special Probability Distributions

Statistical hypothesis testing uses particular types of probability distributions to determine whether the results are statistically significant. Specifically, they use sampling distributions and the distributions of test statistics.

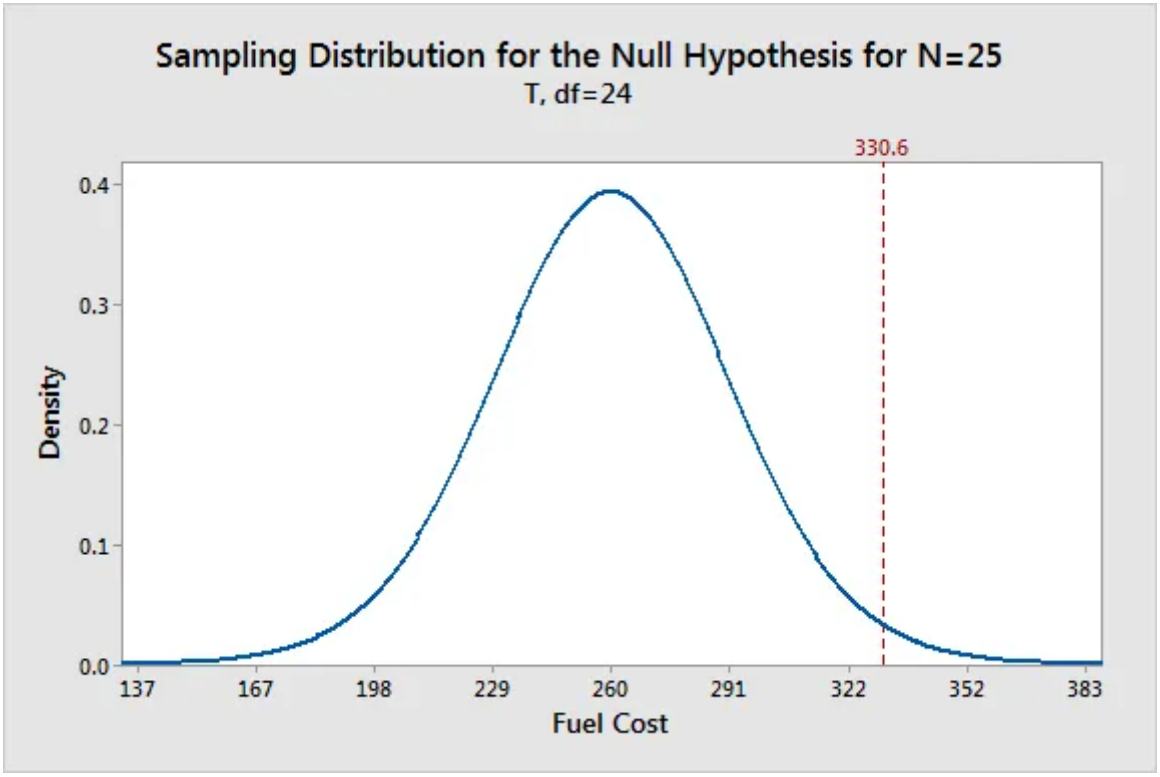
Sampling distributions

A vital concept in inferential statistics is that the particular random sample that you draw for a study is just one of a large number of possible samples that you could have pulled from your population of interest. Understanding this broader context of all possible samples and how your study's sample fits within it provides valuable information.

Suppose we draw a substantial number of random samples of the same size from the same population and calculate the sample mean for each sample. During this process, we'd observe a broad spectrum of sample means, and we can graph their distribution.

This type of distribution is called a sampling distribution. Sampling distributions allow you to determine the likelihood of obtaining different sample values, which makes them crucial for performing hypothesis tests.

The graph below displays the sampling distribution for energy costs. It shows which sample means are more and less likely to occur when the population mean is 260. It also displays the specific sample mean that a study obtains (330.6). The graph indicates that our observed sample mean isn't the most likely value, but it's not wholly implausible either. Hypothesis tests use this type of information to determine whether the results are statistically significant.



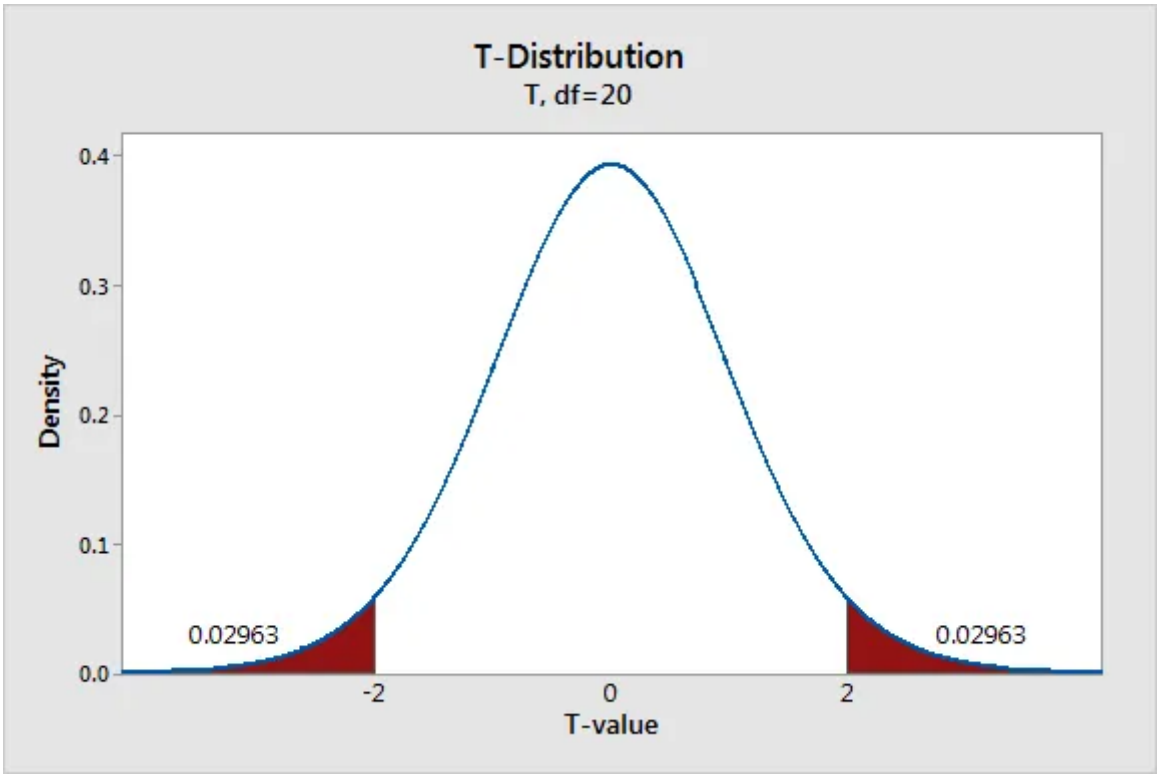
0

To learn more about sampling distributions, read my post about How Hypothesis Tests Work.

Distributions for test statistics

Each type of hypothesis test uses a test statistic. For example, t-tests use t-values, ANOVA uses F-values, and Chi-square tests use chi-square values. Hypothesis tests use the probability distributions of these test statistics to calculate p-values. That's right, p-values come from these distributions!

For instance, a t-test takes all of the sample data and boils it down to a single t-value, and then the t-distribution calculates the p-value. The probability distribution plot below represents a two-tailed t-test that produces a t-value of 2. The plot of the t-distribution indicates that each of the two shaded regions that corresponds to t-values of +2 and -2 (that's the two-tailed aspect of the test) has a likelihood of 0.02963—for a total of 0.05926. That's the p-value for this test!



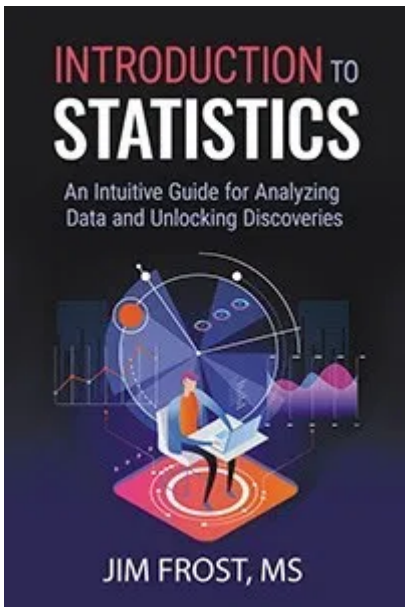
0

To learn more about how this works for different hypothesis tests, read my posts about:

- How t-Tests Work
- How the F-test Works in One-Way ANOVA
- Degrees of Freedom (There's a section about probability distributions.)

I hope you can see how crucial probability distributions are in statistics and why I think graphing them is a powerful way to convey results!

If you're learning about statistics and like the approach I use in my blog, check out my Introduction to Statistics eBook!



0

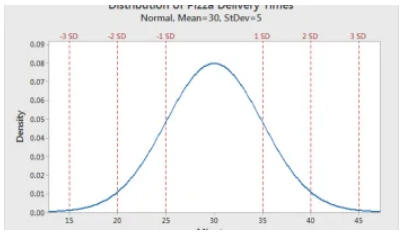
Learn more $9.00 USD

VISA AMEX DISCOVER PayPal

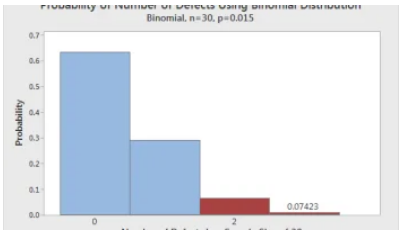
Share this:

Share 642 Share Tweet Salvar 2

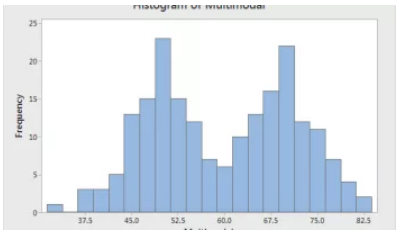
Related



Normal Distribution in Statistics
In "Basics"



Goodness-of-Fit Tests for Discrete Distributions
In "Hypothesis Testing"



Using Histograms to Understand Your Data
In "Basics"

Filed Under: Basics Tagged With: conceptual, data types, distributions, graphs, interpreting results, probability

Comments