How to Identify the Distribution of Your Data

By

You're probably familiar with data that follow the normal distribution. The normal distribution is that nice, familiar bell-shaped curve. Unfortunately, not all data are normally distributed or as intuitive to understand. You can picture the symmetric normal distribution, but what about the Weibull or Gamma distributions? This uncertainty might leave you feeling unsettled. In this post, I show you how to identify the probability distribution of your data.

You might think of nonnormal data as abnormal. However, in some areas, you should actually expect nonnormal distributions. For instance, income data are typically right skewed. If a process has a natural limit, data tend to skew away from the limit. For example, purity can't be greater than 100%, which might cause the data to cluster near the upper limit and skew left towards lower values. On the other hand, drill holes can't be smaller than the drill bit. The sizes of the drill holes might be right-skewed away from the minimum possible size.

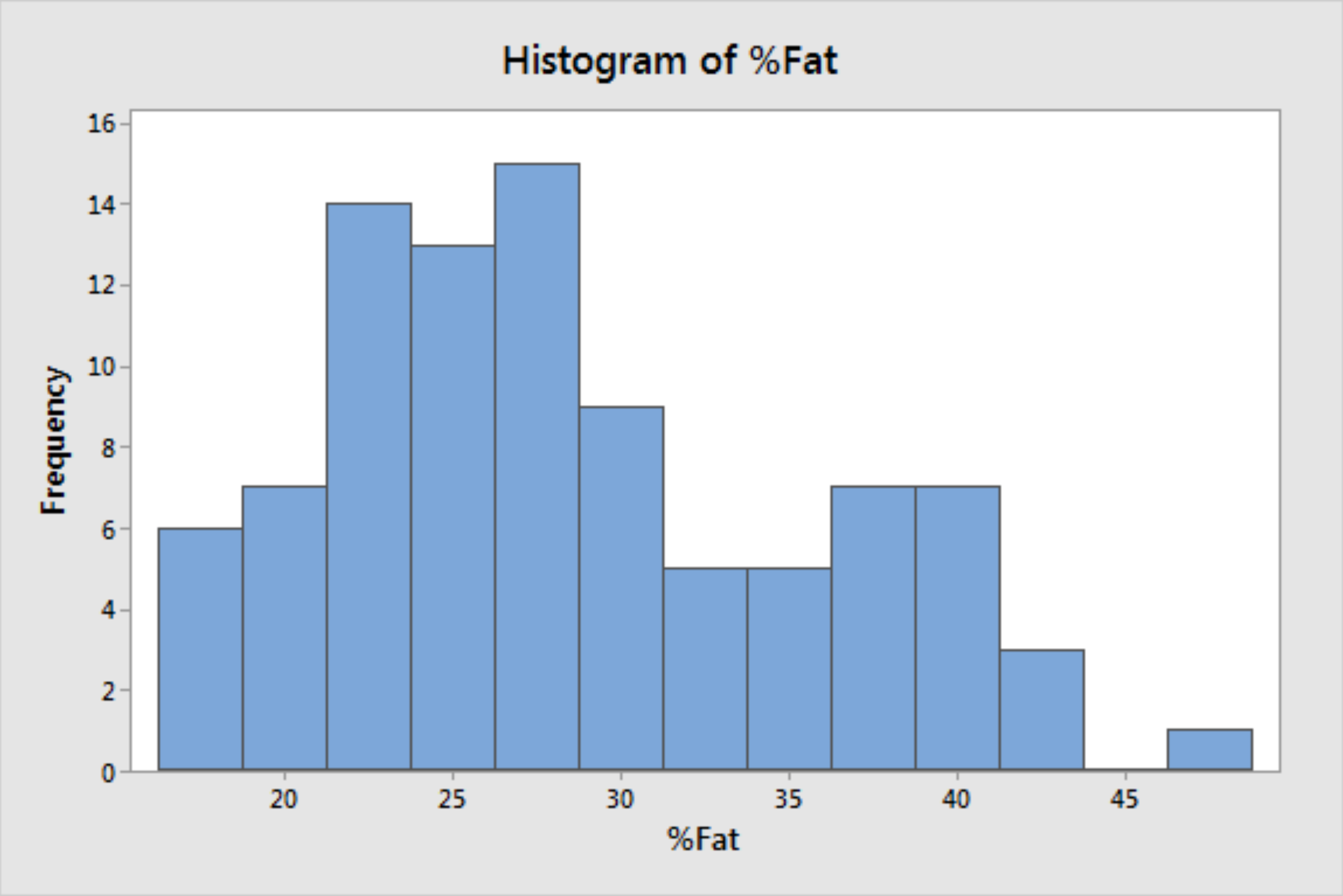
Data that follow any probability distribution can be valuable. However, many people don't feel as comfortable with nonnormal data. Let's shed light on how to identify the distribution of your data!

We'll learn how to identify the probability distribution using body fat percentage data from middle school girls that I collected during an experiment. You can download the CSV data file: body_fat.

Related post: Understanding Probability Distributions and The Normal Distribution

Graph the Raw Data

Let's plot the raw data to see what it looks like.



The histogram gives us a good overview of the data. At a glance, we can see that these data clearly are not normally distributed. They are right skewed. The peak is around 27%, and the distribution extends further into the higher values than to the lower values.

These data are not normal, but which probability distribution do they follow? Fortunately, statistical software can help us!

Related post: Assessing Normality: Histograms vs. Normal Probability Plots

Using Distribution Tests to Identify the Probability Distribution that Your Data Follow

Distribution tests are hypothesis tests that determine whether your sample data were drawn from a population that follows a hypothesized probability distribution. Like any statistical hypothesis test, distribution tests have a null hypothesis and an alternative hypothesis.

- H_0: The sample data follow the hypothesized distribution.
- H_1: The sample data do not follow the hypothesized distribution.

For distribution tests, small p-values indicate that you can reject the null hypothesis and conclude that your data were not drawn from a population with the specified distribution. However, we want to identify the probability distribution that our data follow rather than the distributions they don’t follow! Consequently, distribution tests are a rare case where you look for high p-values to identify candidate distributions.

Before we test our data to identify the distribution, here are some measures you need to know:

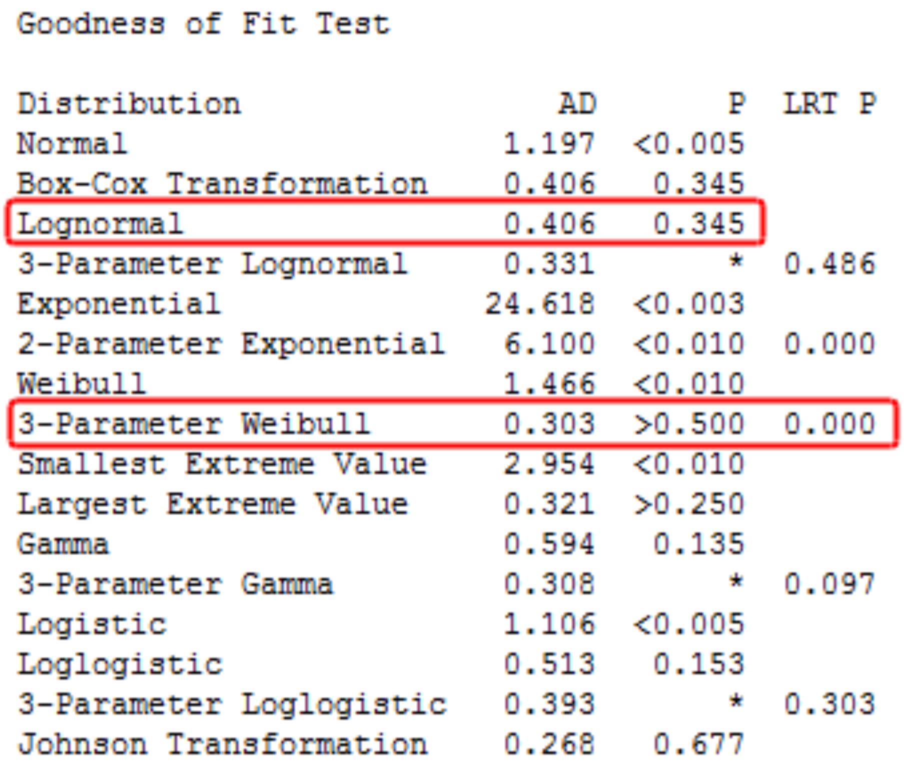
Anderson-Darling statistic (AD): There are different distribution tests. The test I’ll use for our data is the Anderson-Darling test. The Anderson-Darling statistic is the test statistic. It’s like the t-value for t-tests or the F-value for F-tests. Typically, you don’t interpret this statistic directly, but the software uses it to calculate the p-value for the test.

P-value: Distribution tests that have high p-values are suitable candidates for your data’s distribution. Unfortunately, it is not possible to calculate p-values for some distributions with three parameters.

LRT P: If you are considering a three-parameter distribution, assess the LRT P to determine whether the third parameter significantly improves the fit compared to the associated two-parameter distribution. An LRT P value that is less than your significance level indicates a significant improvement over the two-parameter distribution. If you see a higher value, consider staying with the two-parameter distribution.

Goodness of Fit Test Results for the Distribution Tests

I’m using Minitab, which can test 14 probability distributions and two transformations all at once. Let’s take a look at the output below. We’re looking for higher p-values in the Goodness-of-Fit Test table below.



```
Goodness of Fit Test

Distribution                    AD        P  LRT P
Normal                       1.197  <0.005
Box-Cox Transformation       0.406    0.345
Lognormal                    0.406    0.345
3-Parameter Lognormal        0.331        *  0.486
Exponential                 24.618  <0.003
2-Parameter Exponential      6.100  <0.010  0.000
Weibull                      1.466  <0.010
3-Parameter Weibull          0.303  >0.500  0.000
Smallest Extreme Value       2.954  <0.010
Largest Extreme Value        0.321  >0.250
Gamma                        0.594    0.135
3-Parameter Gamma            0.308        *  0.097
Logistic                     1.106  <0.005
Loglogistic                  0.513    0.153
3-Parameter Loglogistic      0.393        *  0.303
Johnson Transformation       0.268    0.677
```

As we expected, the Normal distribution does not fit the data. The p-value is less than 0.005, which indicates that we can reject the null hypothesis that these data follow the normal distribution.

The Box-Cox transformation and the Johnson transformation both have high p-values. If we need to transform our data to follow the normal distribution, the high p-values indicate that we can use these transformations successfully. However, we’ll disregard the transformations because we want to identify our probability distribution rather than transform it.

The highest p-value is for the three-parameter Weibull distribution (>0.500). For the three-parameter Weibull, the LRT P is significant (0.000), which means that the third parameter significantly improves the fit.

The lognormal distribution has the next highest p-value of 0.345.

Let’s consider the three-parameter Weibull distribution and lognormal distribution to be our top two candidates.

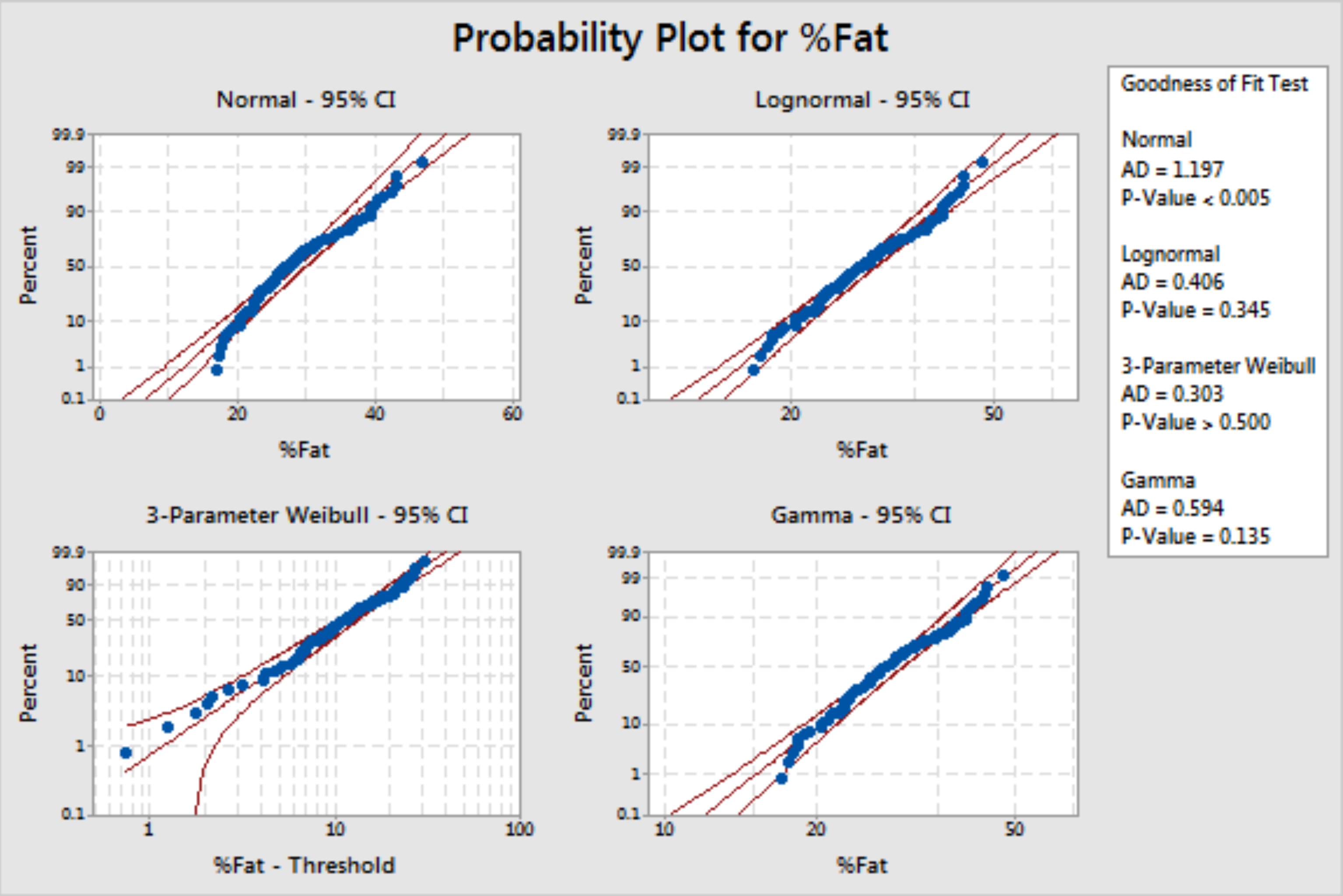
Using Probability Plots to Identify the Distribution of Your Data

Probability plots might be the best way to determine whether your data follow a particular distribution. If your data follow the straight line on the graph, the distribution fits your data. This process is very easy to do visually. Informally, this process is called the “fat pencil” test. If all the data points line up within the area of a fat pencil laid over the center straight line, you can conclude that your data follow the distribution.

These plots are especially useful in cases where the distribution tests are too powerful. Distribution tests are like other hypothesis tests. As the sample size increases, the statistical power of the test also increases. With very large sample sizes, the test can have so much power that trivial departures from the distribution produce statistically significant results. In these cases, your p-value will be less than the significance level even when your data follow the distribution.

The solution is to assess the probability plots to identify the distribution of your data. If the data points fall along the straight line, you can conclude the data follow that distribution even if the p-value is statistically significant.

The probability plots below include the normal distribution, our top two candidates, and the gamma distribution.



The data points for the normal distribution don't follow the center line. However, the data points do follow the line very closely for both the lognormal and the three-parameter Weibull distributions. The gamma distribution doesn't follow the center line quite as well as the other two, and its p-value is lower. Again, it appears like the choice comes down to our top two candidates from before. How do we choose?

An Additional Consideration for Three-Parameter Distributions

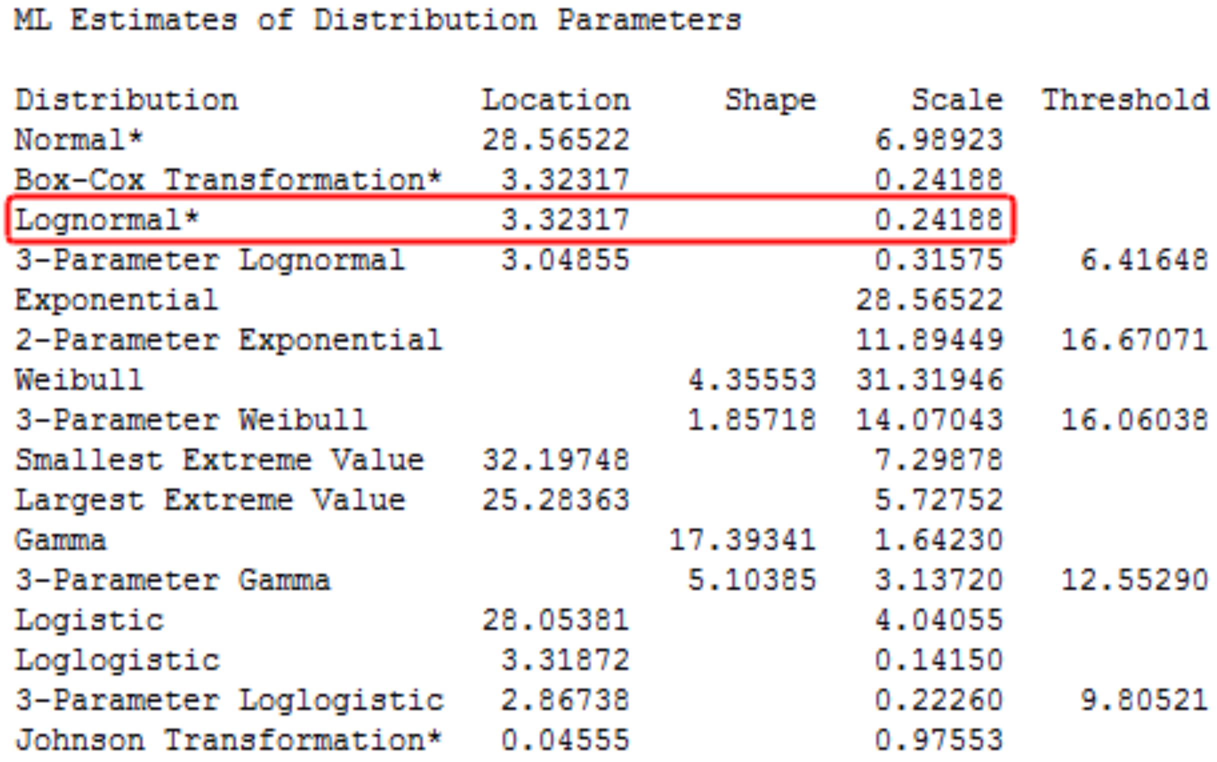
Three-parameter distributions have a threshold parameter. The threshold parameter is also known as the location parameter. This parameter shifts the entire distribution left and right along the x-axis. The threshold/location parameter defines the smallest possible value in the distribution. You should use a three-parameter distribution only if the location truly is the lowest possible value. In other words, use subject-area knowledge to help you choose.

The threshold parameter for our data is 16.06038 (shown in the table below).

This cutoff point is based on (but not equal to) the smallest value in our sample. However, in the full population of middle school girls, it is unlikely that there is a strict cutoff at this value. Instead, lower values are possible even though they are less likely. Consequently, I'll pick the lognormal distribution.

Parameter Values for Our Distribution

We've identified our distribution as the lognormal distribution. Now, we need to find the parameter values for it. Population parameters are the values that define the shape and location of the distribution. We just need to look in the distribution parameters table below!

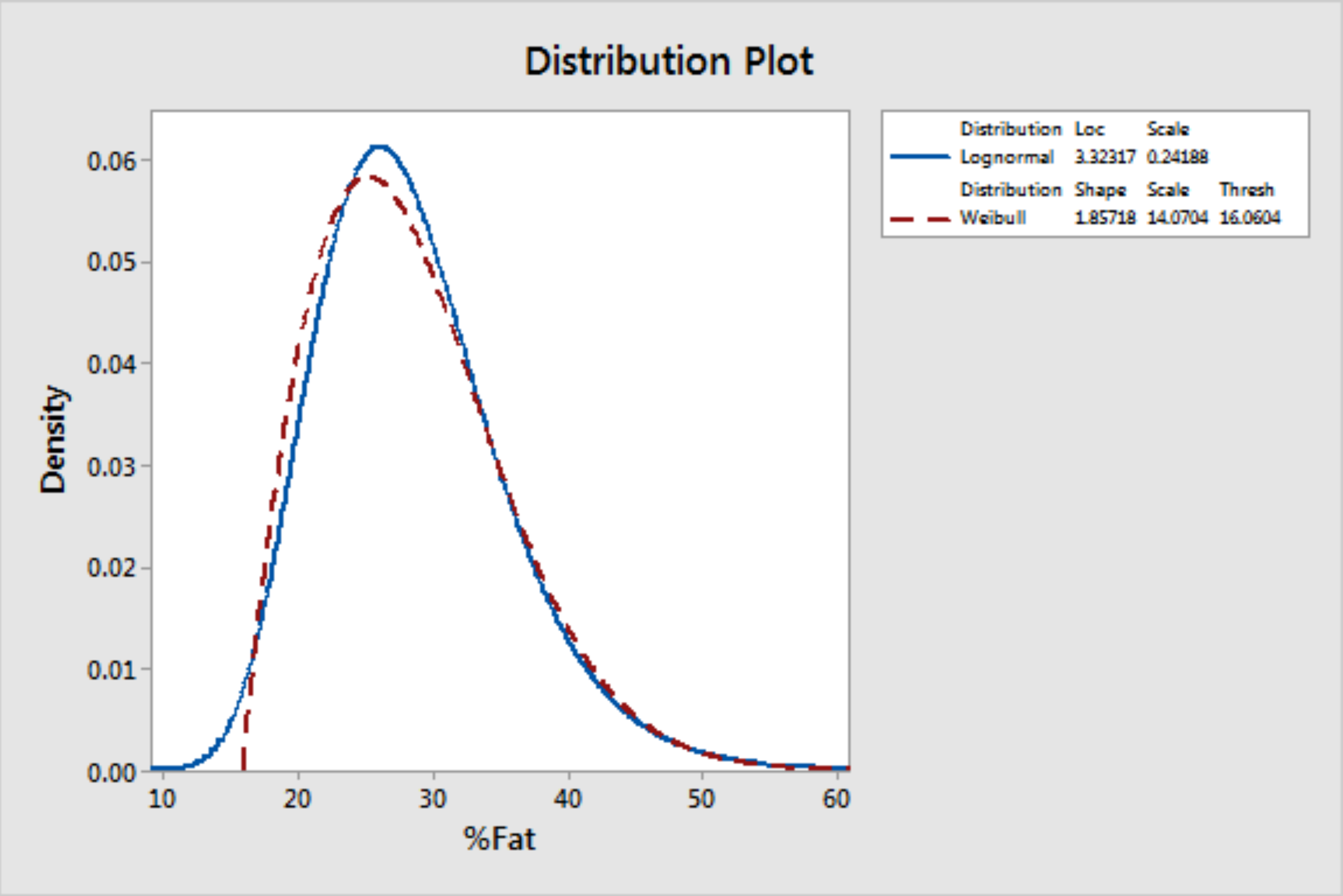


```
ML Estimates of Distribution Parameters

Distribution               Location     Shape     Scale  Threshold
Normal*                    28.56522             6.98923
Box-Cox Transformation*     3.32317             0.24188
Lognormal*                  3.32317             0.24188
3-Parameter Lognormal       3.04855             0.31575    6.41648
Exponential                                    28.56522
2-Parameter Exponential                        11.89449   16.67071
Weibull                               4.35553  31.31946
3-Parameter Weibull                   1.85718  14.07043   16.06038
Smallest Extreme Value     32.19748             7.29878
Largest Extreme Value      25.28363             5.72752
Gamma                                17.39341   1.64230
3-Parameter Gamma                     5.10385   3.13720   12.55290
Logistic                   28.05381             4.04055
Loglogistic                 3.31872             0.14150
3-Parameter Loglogistic     2.86738             0.22260    9.80521
Johnson Transformation*     0.04555             0.97553
```

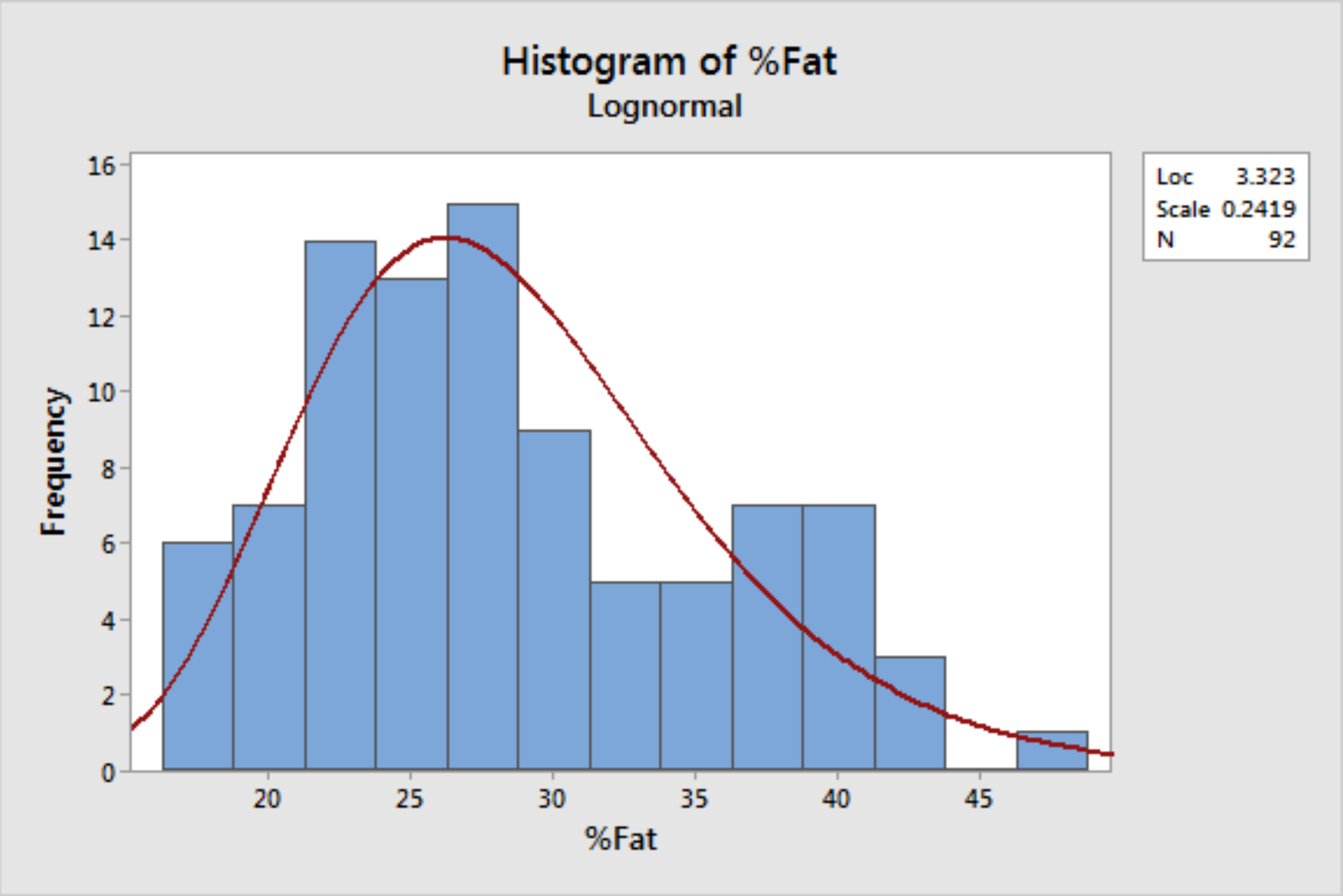
Our body fat percentage data for middle school girls follows a lognormal distribution with a location of 3.32317 and a scale of 0.24188.

Below, I created a probability distribution plot of our two top candidates using the parameter estimates. You can see how the three-parameter Weibull distribution stops abruptly at the threshold/location value. However, the lognormal distribution continues to lower values.



Identifying the probability distribution that your data follow can be critical for analyses that are very sensitive to the distribution, such as capability analysis. In a future blog post, I’ll show you what else you can do by simply knowing the distribution of your data. This post is all continuous data and continuous probability distributions. If you have discrete data, read my post about Goodness-of-Fit Tests for Discrete Distributions.

Finally, I’ll close this post with a graph that compares the raw data to the fitted distribution that we identified.



Note: I wrote a different version of this post that appeared elsewhere. I've completely rewritten and updated it for my blog site.

Related Posts on Statistics by Jim