

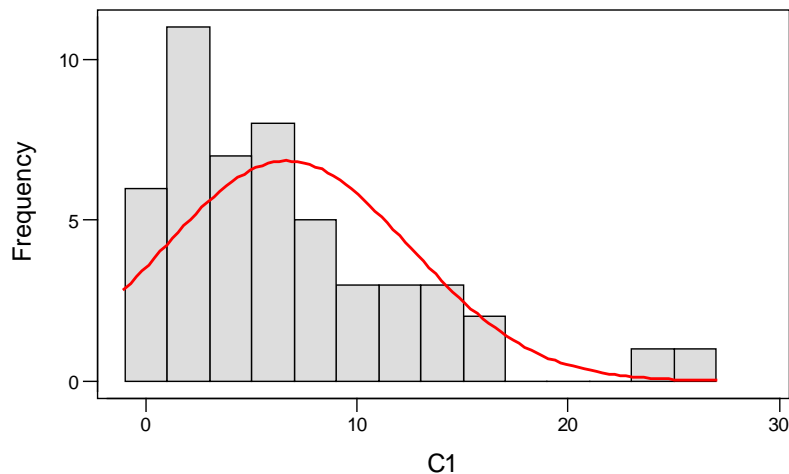
## Testing Assumptions: Normality and Equal Variances

So far we have been dealing with parametric hypothesis tests, mainly the different versions of the  $t$ -test. As such, our statistics have been based on comparing means in order to calculate some measure of significance based on a stated null hypothesis and confidence level. But is it always correct to compare means? No, of course not; parametric statistics, by definition, assume that the data we want to test are normally distributed, hence the use of the mean as the measure of central tendency. Sample data sets are often skewed to the right for various reasons, and if we cannot *normalize* the data we should not compare means (more on normalizing data sets later). In other words, in order to be consistent we need to formally test our assumptions of normality. Luckily this is very easy in MINITAB.

For example, say we have a sample of fifty ( $n = 50$ ) excavated units and we are interested in the artifact density per unit. Before we think about comparing the data set to another sample (for example) we need to see if the data is normal. To do this we run our descriptive statistics as usual and produce some graphics:

Descriptive Statistics						
Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
C1	50	6.702	5.679	6.099	5.825	0.824
Variable	Min	Max	Q1	Q3		
C1	0.039	25.681	2.374	9.886		

Histogram of C1, with Normal Curve

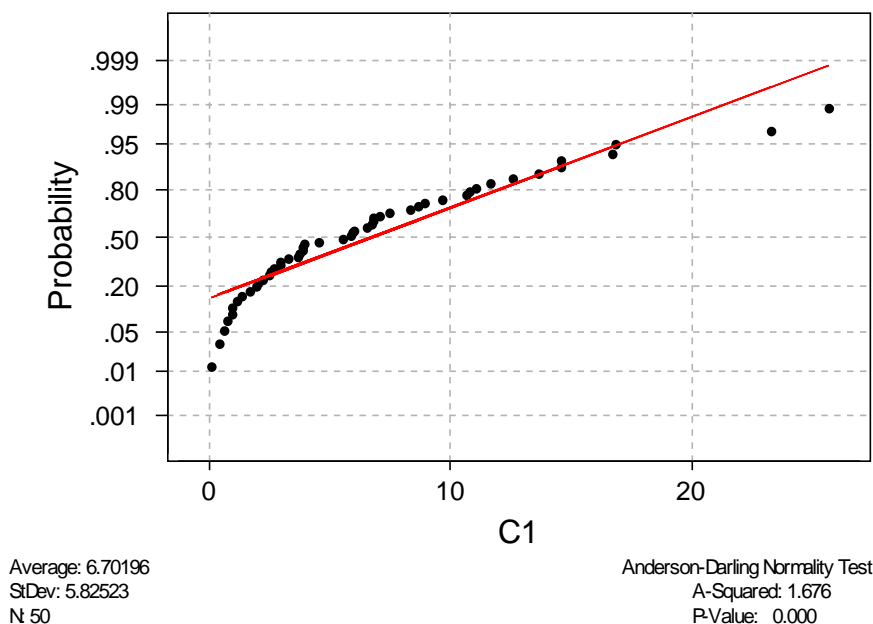


In this case we see that the data set is skewed to the right, and looks more like an exponential distribution than a normal distribution. To test formally for normality we use either an *Anderson-Darling* or a *Shapiro-Wilk* test. The way these tests work is by generating a normal

probability plot (sometimes called a *rankit* plot) based on what a normally distributed data set of a given sample size should look like. They then test the correlation between the predicted normal data with the actual data. This correlation coefficient has some critical value based on the degrees of freedom (or sample size) of the data set so that we can compare our coefficient to the critical value as in all the other tests. However, MINITAB gives us a  $p$  value with both tests, and so we can automatically compare this value to our stated alpha level without having to bother looking up values in a table.

Here is the Anderson-Darling output for our data set:

### Normal Probability Plot



We are primarily concerned with the  $p$  value in the bottom right corner of the graph, which in our case is  $p = 0.000$ . The null hypothesis (as usual) states that there is no difference between our data and the generated normal data, so that we would reject the null hypothesis as the  $p$  value is less than any stated alpha level we might want to choose; the data is highly non-normal and we should not use parametric statistics on the raw data of excavated units. The straight line on the graph is the null hypothesis of normality, so that we want our data to be as close to that line as possible in order to assume normality. The  $p$  value tells us whether our data are significantly different from this line or not. The Shapiro-Wilk test produces the same graph using a slightly different test statistic, but is equally as valid.

In MINITAB there are two ways of conducting a normality test. The normal probability plot is generated by the following procedure:

```
>STATS
```

**>BASIC STATISTICS**

**>NORMALITY TEST**

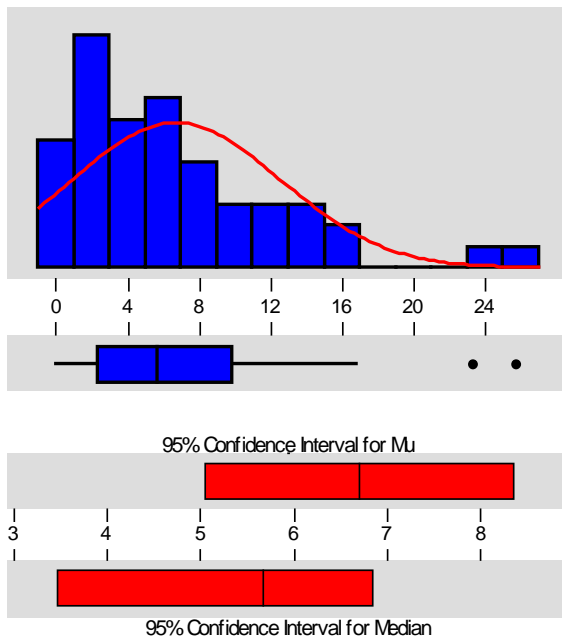
**>Put your data column in the VARIABLE BOX (leave the reference box empty)**

**>Choose ANDERSON-DARLING or RYAN-JOINER (same as Shapiro-Wilk)**

**>OK**

The other way is to choose the GRAPHICAL SUMMARY output option under the GRAPHICS for the DESCRIPTIVE STATISTICS: This output includes an Anderson-Darling test for normality at the top on the left.

### Descriptive Statistics



Variable: C1

Anderson-Darling Normality Test	
A-Squared:	1.676
P-Value:	0.000
Mean	6.70196
StDev	5.82523
Variance	33.9333
Skewness	1.27891
Kurtosis	1.40201
N	50
Minimum	0.0386
1st Quartile	2.3741
Median	5.6790
3rd Quartile	9.8855
Maximum	25.6808
95% Confidence Interval for Mu	
5.0464	8.3575
95% Confidence Interval for Sigma	
4.8660	7.2590
95% Confidence Interval for Median	
3.4860	6.8595

## Equal Variances: The $F$ -test

---

The different options of the  $t$ -test revolve around the assumption of equal variances or unequal variances. We have learned that we can usually eye-ball the data and make our assumption, but there is a formal way of going about testing for equal variances; the  $F$ -test. The  $F$ -test is not only used for  $t$ -tests, but for any occasion when you are interested comparing the variation in two data sets. As usual, the test calculates an  $F_{STAT}$  that is compared to a  $F_{CRIT}$  in a statistical table, which can then be turned into a  $p$  value. The  $F$ -test is very easy.

$$F_{STAT} = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$$

Of course, what is going on here is that if the sample variances are equal, the ratio of their differences should be around 1. The test calculates whether the sample variances are close enough to 1, given their respective degrees of freedom.

For example, say we had two samples:  $n_1 = 25$ ,  $s_1 = 13.2$ , and  $n_2 = 36$ ,  $s_2 = 15.3$ . Remember the ratio is the *variance* not the standard deviation, so

$$F_{STAT} = \frac{s_2^2}{s_1^2} = \frac{15.3^2}{13.2^2} = \frac{234.09}{174.24} = 1.34$$

The degrees of freedom are  $v_2 = 36 - 1 = 35$  and  $v_1 = 25 - 1 = 24$  for the larger and smaller variances respectively. In an  $F$  table we would look for the column  $v$  for the larger sample variance ( $v_2 = 35$ ) along the top of the table, and the row relating to the smaller variance ( $v_1 = 24$ ). In our case, we are not given all the exact degrees of freedom so we assume our critical value is less than the next highest value give, which would be  $F_{CRIT} = 1.79$ . As our  $F_{STAT} < F_{CRIT}$  we can assume the sample variances are equal. Notice that we cannot calculate a  $p$  value from the table.

MINITAB does not do  $F$ -tests, but EXCEL does.

The formula is `=FTEST(array1, array2)`, so `=FTEST(Xi:Xj, Yi:Yj)`, and EXCEL will return a  $p$  value, which you can then compare to an alpha level of your choosing.