Assumption of Normality

One of the assumptions of many statistical procedures (including the t-test) is that the population from which you are sampling is normally distributed. The t-test is said to be rather ‘robust’ in terms of this assumption, which means that reality can deviate from this assumption a fair amount without seriously affecting the validity of the analysis.

This is particularly true when the size of the samples is large (thanks to the Central Limit Theorem). Some deviations from normality can pose a problem for the t-test, specifically those that involve getting extreme scores more frequently than you would if the distribution were normal.

Statistical Software Packages provides two statistical tests for deviation from normality, the ***'Kolomogorov-Smirnov'*** family of tests and the ***'Shapiro-Wilk'*** test.

The 'Kolomogorov-Smirnov' test can be used to test if two data sets are distributed according to the same distribution. It can also be used to test if one data set comes from a specified distribution, such as the normal distribution. (As such, the normal distribution must be specified as an argument to the function.)

For the purposes of this module, we will only use a special case of the 'Kolomogorov-Smirnov' test, known as the **‘*Anderson-Darling'*** test of normality.

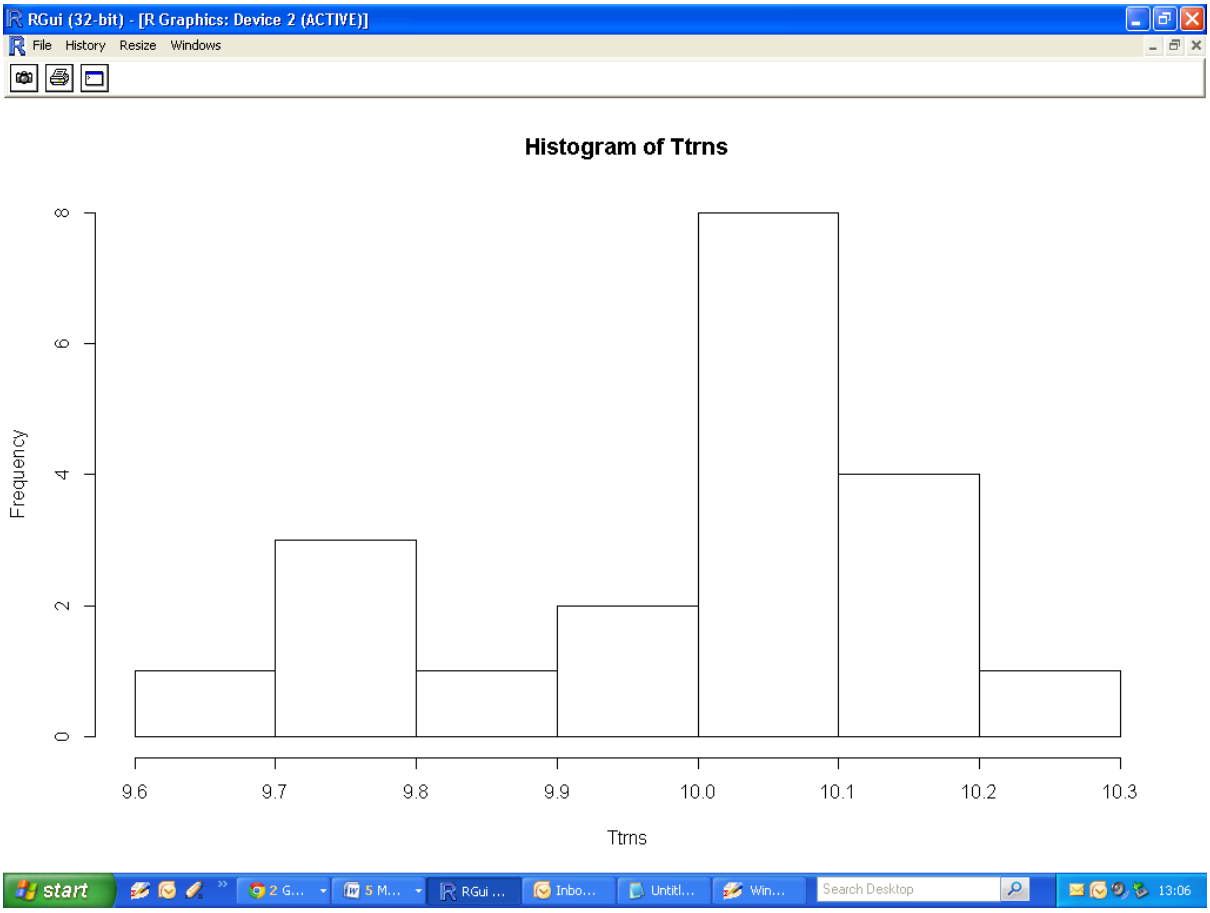
The ‘Anderson-Darling’ test can not be implemented directly in ***R***. Using the test requires the installation of the *nortest* package. We will look at packages in greater detail later in the semester.

The null hypothesis of both the `Anderson-Darling’ and `Shapiro-Wilk’ tests is that the population is normally distributed, and the alternative hypothesis is that the data is not normally distributed.

Let us use both tests to assess whether the titration data set (the combined scores from all four students as one data set) is normally distributed.

Judging by this histogram – do you think the data set is normally distributed?

(Remark : it is skewed to the right)



Using the ‘Shapiro-Wilk’ Test

```
> Ttrns = c(X.A, X.B, X.C, X.D)
> shapiro.test(Ttrns)

        Shapiro-Wilk normality test

data:  Ttrns
W = 0.9188, p-value = 0.09394
```

Using the ‘Anderson-Darling’ Test

```
> library(nortest)
> ad.test(Ttrns)

        Anderson-Darling normality test

data:  Ttrns
A = 0.6961, p-value = 0.0583
```

In both cases we fail to reject the null hypothesis that the data set is normally distributed.

However, the p-values were still quite low in both cases.

Limitations of Tests

There are some important limitations to the usefulness of these tests.

If you reject H_0 you can conclude that the population is not normally distributed, but if you don't reject H_0 then you only conclude that you failed to show the population is not normally distributed. In other words, you can prove the population is not normally distributed but you can't prove it is normally distributed.

Rejecting H_0 means that the population is not normally distributed, but it doesn't tell you whether it is because it is a fat-tailed distribution, a thin-tailed distribution, a skewed distribution, or something else.

The tests are influenced by power. If you have a small sample the test may not have enough power to detect non-normality in the population.

Graphical Methods

The quantile-quantile (Q-Q) plot is an excellent way to see whether the data deviate from normal (the plot can be set up to see if the data deviate from other distributions as well but here we are only interested in the normal distribution).

The process used for creating a QQ plot involves determining what proportion of the 'observed' scores fall below any one score, then the "z-score" that would fit that proportion if the data were normally distributed is calculated, and finally that "z- score" that would cut off that proportion (the 'expected normal value') is translated back into the original metric to see what raw score that would be.

A scatter plot is then created that shows the relationship between the actual 'observed' values and what those values would be 'expected' to be if the data were normally distributed.

If the data is normally distributed then the circles on the resulting plot (each circle representing a score) will form a straight line.

A trend line can be added to the plot to assist in determining whether or not this relationship is linear.

```
> qqnorm(Ttrns)
> qqline(Ttrns)
```

The Q-Q plot will look like this. How well do the covariates follow the trendline? Compare your conclusion to the relatively low p-values of the formal tests.