

What does my data look like (Excel)?

We have many tools at our disposal to help describe a data set, such as means, medians, modes, percentages, and a variety of charts. It may also be appropriate to use inferential statistics, and a histogram can be used to determine whether a sample is parametric (which will influence the inferential statistical tests you are able to use). MS Excel can be used for most of these.

Descriptive statistics

Statistics such as means, mode etc can be obtained using the appropriate formula (eg. =average for the mean), or using the 'descriptive statistics' item from the list in the 'data analysis' section of the 'data' tab.

See: <http://www.excel-easy.com/examples/descriptive-statistics.html>

Charts

If you wish to depict your data in a chart or graph, highlight the data involved, and use: 'insert, chart' in Excel to select the chart you prefer

Histogram

A histogram is a quick, visual way to determine whether your data follow a parametric (ie. normal) or non-parametric distribution. The distribution (and shape) of your data determines the types of statistical analysis tests you are able to use.

See: <http://www.excel-easy.com/examples/histogram.html>

1. Find the standard deviation of your data using the formula =STDEV
2. Create a 'bin range' – this shows the points where the 'bars' in a histogram start and finish. Standard practice is to use a bin range from 3 standard deviations below the mean to three standard deviations above the mean
3. Under the "Data Analysis" tab, select "Histogram"
4. Your "Input Range" is your original data and your "Bin Range" is the bin range you have created. Don't forget to select an appropriate cell in which to place your histogram ("Output Range")
5. To produce a visual representation of your data as well as a histogram, select "Chart Output" before clicking "OK"

Inferential Statistical tests in Excel

There are a variety of inferential statistical tests that can be used to look for differences (including T-Tests, ANOVA Tests, and Chi-Square Tests). Remember that whenever you are running a test for difference, you must first establish your "null hypothesis" (there is no difference between these groups) and your "alternate hypothesis" (there is a difference between these groups). In all cases, the test will generate a test statistic (eg. T, F), and a **p value**. The p value is the key number to

note, and by comparing with your predetermined significance level (eg. 0.05), you can decide whether to accept or reject your null hypothesis. The smaller the p value, the more likely it is that you will reject the null hypothesis, and accept the alternate hypothesis.

You can also run inferential statistical tests to look for relationships (including regression and correlation analyses).

Don't forget that for all inferential statistical tests except Chi-Square Tests, your data must be parametric (ie. normally distributed) and scale (or interval/ratio). If your data is non-parametric, or if it is nominal/ordinal, you will need to look at using other types of tests to establish differences and/or relationships.

T-Tests

T-Tests determine whether there is a *significant* difference between sample means. T-Tests assume that the data involved follow a normal distribution (ie. parametric data), with equivalent variances (this can be tested using an F test), though t tests can be run with unequal variances also.

The key types of t-test are:

- “paired” t-test (the same population measured twice, generally for before/after situations)
- “independent” t-test (two different groups undergoing the same trial, called a “Two Sample” t-test by Excel).

<http://www.excel-easy.com/examples/t-test.html>

1. Under the “Data Analysis” tab choose your appropriate T-Test (paired, two-sample with equal variance, two-sample with unequal variance)
2. Your “Variable 1 Range” will be your first population and your “Variable 2 Range” will be your second population. Your “Hypothesized mean difference” should be entered as “0”. Don't forget to select an appropriate cell in which to place your T-Test output (“Output Range”)
3. What is your p-value? Is it less than the significance level you have set? If yes, you can reject your null hypothesis and accept the alternate hypothesis – there is a significant difference between the two groups.
4. Additionally: does your test statistic (t stat) lie between the negative t-Critical two-tail and the positive t-Critical two-tail? If yes, you can accept your null hypothesis – there is no significant difference between the two groups.

ANOVA Test

ANOVA (Analysis of Variables) is (for our purposes, at least) a way of finding the same information you would get from a T-Test, but for more than two samples. So the test is generally used to determine whether there is a *significant* difference

between sample means. The test involves the computation of a F statistic. Options for ANOVA tests include:

- “one-way” ANOVA (>2 samples, one variable),
- a “two-way” ANOVA with replication (>2 samples, >2 variables),
- or a “two-way” ANOVA without replication (multiple individuals, multiple variables).

<http://www.excel-easy.com/examples/anova.html>

1. Under the “Data Analysis” tab, select the ANOVA test you wish to run (see above).
2. Your Data will be your “Input Range”. Remember to identify whether your populations are organised by row or column, and tick the “Labels” box. Don’t forget to select an appropriate cell in which to place your ANOVA output (“Output Range”)
3. If your F-value is greater than your F-critical value, you must reject your null hypothesis.
4. What is your p-value? What does this tell you?

Chi-square Test

A chi-square test determines the difference between what has been observed and what is expected. Chi-square tests are incredibly useful as they can be run using nominal and ordinal, as well as scale data. Excel doesn’t have an easy way to run a chi-square test for you, so the process is a bit more complicated than some of our other tests.

1. In order to determine the difference between observed and expected values, you must first *have* both the observed and expected values. Your observed values are the result of your data collection (e.g. experiment, observation, etc.) You must calculate your own set of expected values. In the worksheet example, we can calculate the expected value of violent crimes committed by under-18s by using the formula: (total crimes committed by under-18s * violent crimes committed by all ages)/total crimes committed by all ages. You could use the same formula for expected values of eye colours in a class for each gender, except you would now have: (total number of males * total number of brown-eyed people)/ total number of people.
2. Once you have your expected values, you can run the chi-square test using the Excel function =chisq.test. Excel will prompt you to enter your observed values (remember that this is only the values, and NOT the totals) followed by your expected values (again, this is only the values and NOT the totals).
3. This formula provides you with a p-value, which you can use to determine whether there is a *significant* difference between what has been observed and what is expected.

See: <https://www.youtube.com/watch?v=UPawNLQOv-8>

Regression Analysis

Regression analysis helps to describe the nature of the relationship between two continuous, scale (or interval/ratio) variables. One variable is the 'independent' variable, and the other variable depends on this, and it therefore called the 'dependent' variable. It is crucial that you select your variables in the right way for the question you are asking (as, for example, weight is likely to be a dependent variable on height in adults, but height is much less likely to be a dependent variable on weight).

<http://www.excel-easy.com/examples/regression.html>

1. Under the "Data Analysis" tab choose "Regression".
2. Your "Input Y Range" must be your dependent (or predictor) variable. Your "Input X Range" must be your independent (or explanatory) variable.
3. Remember to tick the boxes for both "Labels" and "Residuals". Don't forget to select an appropriate cell in which to place your regression output ("Output Range")
4. Your "R square" value gives you the degree to which your relationship can be explained, and the closer it is to 1, the better. For example, an R square of 0.785 means that 78.5% of the variance in your dependent variable Y can be explained by the independent variable X.
5. You must also examine the F and p- values to determine whether or not the relationship is statistically *significant*.

Correlation Analysis

Correlation relates to the sort of variables as in regression analysis, and correlation analysis tells us how strongly two variables are related. Correlation coefficients (the degree of correlation) can be negative (as one variable goes up, the other goes down) or positive (as one variable goes up, the other goes up). The closer your correlation coefficient is to 1 (or -1) the stronger the relationship. The closer your correlation coefficient is to 0, the weaker the relationship.

<http://www.excel-easy.com/examples/correlation.html>

1. Under the "Data Analysis" tab choose "Correlation"
2. Your "Input Range" should be your entire data set. Don't forget to identify whether each variable has its own Row or Column.
3. Remember to tick the box for "Labels". Don't forget to select an appropriate cell in which to place your correlation output ("Output Range")
4. Use the graph to determine the degree to which your variables are correlated with one another.