Statistics for SciFest

There is no point in doing a scientific experiment unless you have a good chance of being able to answer the question clearly. It is a waste of time to have an experiment that is under-powered: the sample being too small to get a clear Yes or No. If you want to test if a coin is biased you'll have to toss it more than twice; because two heads in a row is not unlikely. 20 heads in a row, and the coin has a case to answer – does it, for example, have two heads?

In science, especially biology and psychology, you are unlikely to get a clear Yes or No. There will be some fuzziness in your data. It is no longer true to say that only boys become doctors or engineers. But maybe there is a tendency in that direction? That is a scientific hypothesis and we could test it with some data. The data might be the results of a series of tests for 'thinking like an engineer' or 'knowledge about human physiology'. The tests administered to both boys and girls and their scores compared . . . statistically.

This short document offers a path through your data, pointing at the appropriate statistical tests to apply given different sorts of data. It might offer clues on a very important parallel issue – designing experiments so that the data can be analysed. Power analysis is a separate, more advanced, subject but you cannot go wrong by doubling the sample size that you first thought of. Running an experiment 'in triplicate' by repeating it three times, which is traditional in chemistry and physics, is likely to give you a fuzzy, statistically uncertain answer.

I hope this helps,

Andrew Lloyd
Dept Science & Health
Carlow IT

Rule 1 Statistics first. Design your experiments with an appropriate statistical test in mind. Poorly designed experiments may be impossible to analyse or have insufficient statistical power to effectively test your hypothesis.

Decision 1. Is my data continuous/numerical (height, concentration, speed)
or categorical/binnable (Y/N, M/F, black/white)

Note 1 you can bin numerical data into categories (such as high medium low) but you lose statistical power by doing so.

Numerical, continuous data.

Decision 2. Is my data distribution normal (bell-curved, Gaussian)
or not.

Note 2. Don't know? Plot the distribution using a histogram. If it looks more or less symmetrical with a hump in the middle, you can assume normal distribution.
Not enough data to plot a histogram? Or less than 10 (or better 20) cases. Then assume the data is not normal and use non-parametric statistics.

Decision 3. Do I have two groups or more than 2?

Decision 4. Is the data paired or independent. Paired data is when measurements are taken on the same individual, say before and after treatment. Unpaired data is when you've killed your zero time point to extract tissue to measure. It makes a difference in Figure 1.

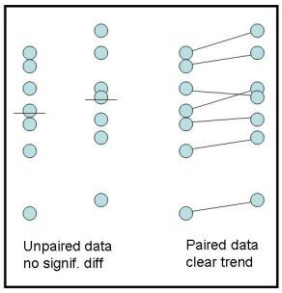

Figure 1

Tests to use:

Normal distribution; 2 groups; paired data	Paired t-test
Not normal distribution; 2 groups; paired data	Wilcoxon signed ranks test
Normal distribution; 2 groups; independent data	regular t-test
Not normal distribution; 2 groups; independent data	Wilcoxon rank sum test

Note 2. Mann-Whitney U test is the same as Wilcoxon test.

Normal distribution; >2 groups;	One way ANOVA
Not normal distribution; >2 groups;	Kruskal-Wallis test

Categorical data.

Decision 5. Two bins or more?

Tests to use:

2 groups; paired data	McNemar's test
2 groups; independent data	ChiSq test or Fisher's exact test
>2 groups	ChiSq test

Decision 6. do any of my bins have less than 5 members?

Tests to use: ChiSq with Yates correction
Or consider pooling small bins together to make up numbers.

Note 3. Degrees of freedom is what it says on the tin. Usually one less than the number of categories if the categories are mutually exclusive.
A 2 x 2 contingency table has 1 df (n-1)*(m-1) for the ChiSq test.

Here's one I did earlier which shows that sample size matters. Here is a 2 x 2 contingency table in which a marker has been measured in 59 controls and 13 subjects. It is human medical data, so subjects are hard to come by, hence the small sample size.

Abs	Pres		Tot	
35	24	|	59	Control (41% Present)
4	9	|	13	Subjects (69% Present)
39	33	|	72	

There appears to be an effect: the ratios are reversed in the Controls and Subjects. But when you apply statistical tests, you find that this "clear" difference is not significant:

ChiSq = 3.5 df = 1, p =0.06,
ChiSq with Yates correction = 2.4, df = 1, p = 0.12
Fisher's Exact Test p=0.07

A simple 2x2 ChiSq is one of the few useful statistical tests that you can do on the back of an envelope. If the numbers are represented by letters thus:

a	b	|	e
c	d	|	f
g	h	|	N

then ChiSq is (a*d – b*c)^2 * N / e*f*g*h Verify that ChiSq = 3.498 for the numbers above.

If you do this test in a statistical package you may well get a different answer because, by default, Yates continuity correction has been applied. This is a fudge to deal with the fact that for small numbers an increment of 1 is a large difference: 4 + 1 is a 25% increase in Absent markers in the Subjects for our dataset.

With Yates half is added or subtracted from the numbers while keeping row and column totals the same thus:

Abs	Pres		Tot	
34.5	24.5	|	59	Control
4.5	8.5	|	13	Subjects
39	33	|	72	

What to do? The obvious and desirable next step is to increase the number of subjects. You'd hope that the ratio will be the same in the new cases. Doubling the Subjects will give:

Abs	Pres		Tot	
35	24	|	59	Control
8	18	|	26	Subjects
43	42	|	85	

ChiSq = 5.9, df = 1, p = 0.02, ChiSq with Yates = 4.8, p 0.02, FisherExact p = 0.02. Bingo!

Correlation and regression

Having determined whether or not you have a statistically significant association between your variables, you will probably want to plot the results and maybe summarise the association with a correlation coefficient or use regression to predict one variable from the other.

Tests to use:

Normal distribution; correlation,	Pearson correlation - r
Not normal distribution; correlation	Spearman rank order correlation

Note 4. r^2 conveniently tells you the proportion of the variability in your dataset which is explained by the 2 way relationship. If you get r = 0.5, then r^2 = 0.25: 75% of the variability in your dataset is due to factors other than those you have measured!

Note 5. Regression, with its implication of causation rather than mere association, requires you to know which is the dependent and which the independent variable. The independent variable predicts the magnitude of the dependent variable, but not vice versa.

Note 6. Your stats package may report probabilities to 5 significant figures but it is unprofessional laziness to transcribe them thus for your report. 0.04 is a more accurate summary of the truth than 0.03945 and is easier to read and interpret

Many statistical tests are available on line: http://www.socscistatistics.com/tests/Default.aspx

Excel only really covers parametric (Normal, bell-curve, Gaussian) tests and these are not usually appropriate with small samples.