

Remedial Statistics

Introduction

These lectures are intended to provide a cookbook of easy approaches to data analysis for any students without any statistical background, or for anyone who has forgotten any statistics they used to know. The basic methods covered should be familiar to most of you and certainly to anyone who has done any of the quantitative courses in previous years. I will almost certainly offend anyone with a strongly mathematical background since the emphasis will be on the pragmatic rather than the elegant. I will also try to show you how you can design your projects so that you simplify the data analysis. It is often the case that a relative small change in the research protocol or research goal will make any statistics much simpler - and unfortunately it is often the simplest research procedures that can lead to horrific data.

Honestly, it is much easier to think of the statistics that you want to do first, and then design an appropriate experiment to generate the data you want than to produce some data and then try to analyse it afterwards. These lectures are organised as a set of recipes showing you what you do with data and what sorts of experiments lead to that sort of data. They do not provide a glossary of statistical terms or an encyclopaedia of statistical methods. For that sort of information you need to consult a textbook. There are loads of good statistics textbooks and you should buy yourself one you like the look of. For those of you who don't want to spend any money, there are even some web based books. The following pages will point you in the right directions:

```
http://statistics.com/  
http://www.stat.ucla.edu/  
http://www.stat.ufl.edu/vlib/statistics.html  
http://sportsci.org/resource/stats/index.html  
http://www.psychstat.smsu.edu/sbk00.htm  
http://www.stat.ucla.edu/textbook/  
http://davidmlane.com/hyperstat/index.html
```

When I get stuck, I refer to *Biostatistical Analysis* by Jerrold H. Zar and published by Prentice Hall. It's a good book for the algebraically minded, but probably not for the beginners. The *Statistics and Data Analysis* chapter in Martin and Bateson's *Measuring Behaviour* (Cambridge University Press) is extremely good although very short and *The OU Project Guide* by Chalmers and Parker (Open University) has a lot of good stuff in it.

What I'm not going to do is instruct you how to use the various statistics computer programs. There's too much personal preference involved. I quite like SPSS, but most of you will probably want to use the Data Analysis option under the Tools menu in *Excel* which is perfectly acceptable for most occasions.

NB. All example datasets have been artificially generated!

One-Sample or Descriptive Analysis

Firstly, I am going to start with so-called one-sample or descriptive statistics. These are useful when you are only taking measurements from a single sample. The problem with this sort of analysis is that it is only rarely appropriate which is a shame because it is certainly the easiest.

Here are two examples of when they are useful.

Firstly and most straightforwardly, when you simply want to a value for a population parameter. For example you've just discovered a new species of primate and you want and you want to know the **mean** crown-rump length. You would simply collect your specimens, measure their body mass and calculate the mean value.

Individual ID	Crown Rump Length (cm)
1	7.9
2	8.5
3	7.6
4	9.6
5	7.9
Mean	8.3

You would report this value in your paper describing the new species and you would state the number of individuals you measured. This is important because what you have measured is the mean of your sample of 5 individuals. What your readers want are much more interested in is the mean of all the individuals in the population, and the mean of your sample is a good estimate of the more general value. It is well known and generally true for all statistics that the larger the sample size, the better the estimate. In actual fact, I generated these numbers randomly and the mean should have been exactly 8 (and the standard deviation 0.75) so 8.3 isn't a bad estimate. The mean is a particularly well behaved statistic in that the mean of the sample is the best estimate of the mean of the population. Many statistics don't work this way. For example, the standard deviation (which is another parameter that you might want to calculate) of a sample underestimates the standard deviation of the population so you need to perform a slightly different calculation when you are estimating a population standard deviation from a sample than when you are calculating the population standard deviation directly.

Secondly, single sample statistics are useful when you are comparing a sample to a well known population and asking the question, "Is there any difference between the sample and the well-known population?" As a quick aside, this isn't the question we would actually ask. For classical statistics to work you have assert that two things are the same, "The sample is the same as the well-known population", and then you see whether you can show that this statement is unlikely to be true.

For example, you've been looking at a group of people who are working in the radiation industry and you want to know whether their job has affected their lifestyle. You could measure an assortment of life history parameters such as body weight and you could compare the value obtained with that in a standard reference such as *Regional Trends* (a regular governmental publication). This is where experimental design comes in. If you check what's available beforehand you can make sure that all the parameters you are going to measure are in the standard reference and your life is straightforward. If you pick non-standard measurements, then you have a great deal more work to do. The other thing to note is that there is a separate value in *Regional Trends* for men and women because this has a major effect on mean body weight, so you would also have to choose your sample to match this.

Individual ID	Mass (kg)
1	66.7
2	81.0
3	65.8
4	63.5
5	52.5
Mean	65.9

You see from the literature that the mean weight of adult males is 62 kg. Does this mean that radiation causes weight gain?

There are two separate issues here. The first is statistical. From our sample, our estimate of the mean of the population is 65.9 kg which is indeed higher than the 62 kg we would expect. So what we want to know is, "How good is our estimate?" Should we expect an estimate that's based on 5 individuals to be off by a few kilograms? This is where confidence limits come in. By making certain assumptions about the population we can quantify how confident we are about the mean of the population the sample comes from (in our case the population is 'people who work in the radiation industry'). *Excel* will calculate what's known as the 95% confidence limit for us using the Descriptive Statistics command and gives a value of 12.6. This means that from this sample, we can be 95% certain (that's 1 in 20 times, the level most commonly chosen in biology) that the mean of the population that the sample was taken from is greater than $65.9 - 12.6$ and less than $65.9 + 12.6$ (in other words between 53.5 and 78.5). That means that we shouldn't be at all surprised that we didn't get 62 kg – it's well within the range. That means that from these data we cannot exclude the **null hypothesis** (the name for our original assertion in statistical jargon).

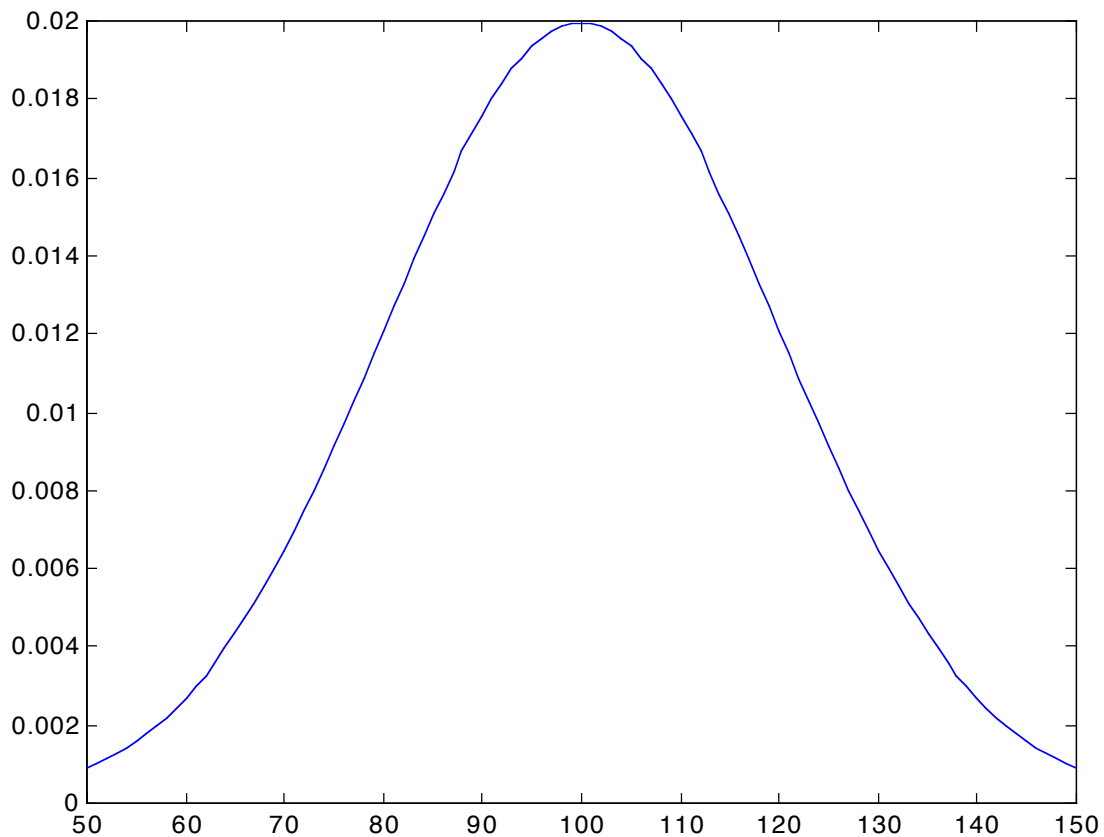
Individual ID	Mass (kg)
1	66.7
2	81
3	65.8
4	63.5
5	52.5

Descriptive Statistics

Mean	65.9
Standard Error	4.54961537
Median	65.8
Mode	#N/A
Standard Deviation	10.1732492
Sample Variance	103.495
Kurtosis	1.83361141
Skewness	0.40505312
Range	28.5
Minimum	52.5
Maximum	81
Sum	329.5
Count	5
Confidence Level(95.0%)	12.6317835

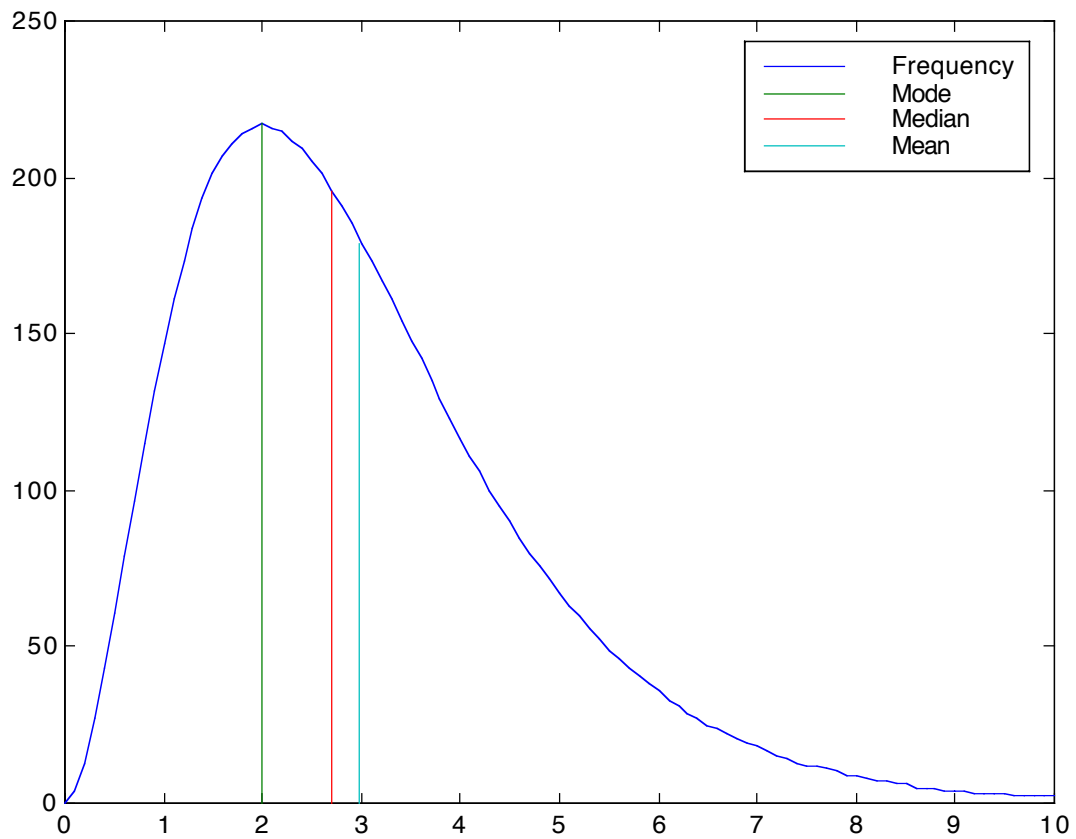
The other issue is equally important. It is to do with data interpretation. If our data had shown that there was a statistically significant weight gain in our population compared to the UK population on the whole, then what does that actually mean? Can we say that this is because of the job? Maybe heavier people choose to work in the radiation industry because it doesn't involve much physical exertion? This is where the biology comes in! Statistics only provides the evidence, and the biology is needed to do something sensible with it.

Also important, you may remember that I said that we have to make certain assumptions for the calculation of confidence limits. This is true. The confidence limits assume that the data approximately follows the **normal distribution**. The normal distribution is a relatively complex mathematical construct that does model the frequencies often seen in the real world. This graph shows a normal distribution with a mean of 100 and a standard deviation (a measure of the width of the hump) of 20.



If your data doesn't (even approximately) look like that then you need to use other approaches but, to be honest, you'll get away with it 9 times out of 10. If you are in any doubt, draw a graph of your data (you should probably do this anyway – at least in your lab book if not in your final write-up) and if most of the values cluster around the central area and it tails off at both low and high extremes then you'll be fine. And also remember that if your sample size is very low, you won't have enough data to check the shape of the distribution anyway.

There are other univariate statistics that you will come across and they are generally used when the data is not normally distributed. The most useful of these is the **median**. This is a different measure of central tendency that is less affected by extreme values. It's the value that represents the figure where 50% of the population are less than this value and 50% are higher and is often a more meaningful figure than the mean when the distribution is not symmetrical (for example, you'll often hear politicians talking about the median annual wage since it's rather lower than the mean wage which is pushed up by the relatively few people earning vast sums of money). Another is the **mode**, which is the most common value. The following figure shows the relationship between the mean, median and the mode (known collectively as **measures of central tendency**) in a non-normal distribution. In a normal distribution, the value of the all three values is the same – the top of the hump.



Modes don't get used much, but medians are encountered. You can calculate the 95% limits, although not in *Excel*, but to do so you need to make assumptions about the distribution which may or may not be true. If you need to do this, you'll need to check in a statistics textbook.

Two-Sample Analysis

The vast majority of student projects should aim to use two sample statistics. This is the classic case where you are interested in the effect of an experimental protocol where you have a control group and a treatment group. It's also useful in observational research where your subjects fall into two categories. If at all possible, you should arrange your experiment to fit this sort of framework.

The standard experimental protocol would be something along these lines. I'm interested in whether cholesterol affects blood clotting times. I extract 200ml of blood from a volunteer and divide it into 20 sample of 10ml. I add 5mg of cholesterol to 10 of them and then I time how long the blood takes to clot. This gives me my two groups: controls and experimental. I can calculate the mean clotting time for each sample, and they are very likely to be different. What I want to know is whether this difference is due to chance effects or whether it is likely that there is a real effect of the treatment.

This is a classic example of when to use the *Student's T-Test* (often just referred to as the *T-test*). This test is for comparing the means of two samples and it tells you the likelihood that the two samples could have been taken from the same population, in other word that the treatment has had no effect.

Control	Treatment
15.3818	14.4399
15.3783	17.1774
12.9247	13.2351
13.6546	21.5496
13.3493	14.5908

t-Test: Two-Sample Assuming Unequal Variances

	Control	Treatment
Mean	14.13774	16.19856
Variance	1.353300583	11.0129073
Observations	5	5
Hypothesized Mean Difference	0	
df	5	
t Stat	-1.31040877	
P(T<=t) one-tail	0.123510751	
t Critical one-tail	2.015049176	
P(T<=t) two-tail	0.247021502	
t Critical two-tail	2.570577635	

From this example we have data from our two samples (there don't need to be equal numbers of measurements in each sample). The means are different (it would be very unlikely if the means were identical), but as you can see from the P(T<=t) part, the probability of the difference being due to chance is 0.25 (25%). In other words, we wouldn't be able to throw out the null hypothesis that the two means were actually the same based on these data. It's important to realise that just

because the test doesn't show a significant difference doesn't mean there isn't one, it just mean that one cannot be detected. In this case, the first column is from a population with a mean of 13 and a standard deviation of 2, and the second from a population with a mean of 15 and a standard deviation of 3. However, it's clear from this that a *T-Test* with 5 in each sample is insufficient to pick up this relatively small difference. A larger sample size will pick up this difference, so for example with 10 in each sample:

Control	Treatment
13.2279	17.574
15.1335	18.762
13.1186	10.2188
12.8087	10.6771
11.3353	16.7134
13.5888	13.8003
10.3276	17.07
14.4286	17.4469
16.2471	17.1357
11.6164	18.8707

t-Test: Two-Sample Assuming Unequal Variances

	Control	Treatment
Mean	13.18325	15.82689
Variance	3.230795034	9.964306597
Observations	10	10
Hypothesized Mean Difference	0	
df	14	
t Stat	-2.301420327	
P(T<=t) one-tail	0.018624907	
t Critical one-tail	1.76130925	
P(T<=t) two-tail	0.037249814	
t Critical two-tail	2.144788596	

This time the P value is 0.04 (4%) which is less than the magic 0.05 (5%) value that is commonly used in biology.

The sharp eyed will have noticed that *Excel* (and other statistics packages) quotes what are called **one-tailed** and **two-tailed** P values. You will almost always use two-tailed values. One-tailed values are for those who are clutching at straws in the vain hope of finding something significant to report and if you are ever that desperate you need to check in a textbook to see when the one-tailed value might conceivably be appropriate in your case.

The only variation in the experimental design that you need to be aware of is when you have a before and after approach. For example, if you have a wonder drug that reverses the reverses the effect of baldness, you would have a set of bald subjects, you would measure the scalp area covered with hair before treatment and the scalp area covered after treatment.

There would be nothing wrong with treating the before and after groups but to do so would be to throw away a lot of statistical power, since the variation in the subjects is diluting the effect of the treatment. For example, using the dataset shown, there is no significant difference using the normal *T-Test* ($P = 0.11$), but using the *Paired-Sample Student's T-Test* gives a highly significant result ($P = 0.003$) and you become a millionaire.

Before	After
28.3825	50.512
65.2176	111.5813
72.9641	114.4751
75.6966	140.2864
123.3384	166.9793

t-Test: Two-Sample Assuming Unequal Variances

	Before	After
Mean	73.11984	116.76682
Variance	1148.110676	1874.076641
Observations	5	5
Hypothesized Mean Difference	0	
df	8	
t Stat	-1.775327631	
P(T<=t) one-tail	0.056879167	
t Critical one-tail	1.85954832	
P(T<=t) two-tail	0.113758335	
t Critical two-tail	2.306005626	

t-Test: Paired Two Sample for Means

	Before	After
Mean	73.11984	116.76682
Variance	1148.110676	1874.076641
Observations	5	5
Pearson Correlation	0.952313338	
Hypothesized Mean Difference	0	
df	4	
t Stat	-6.458086058	
P(T<=t) one-tail	0.001480094	
t Critical one-tail	2.131846486	
P(T<=t) two-tail	0.002960187	
t Critical two-tail	2.776450856	

Again there are some assumptions that are made for these tests to be valid. The populations should be normally distributed but fortunately the *T-Test* is robust: the data has to be quite a way off being normal for it to make much difference. However, there are other equivalent tests that can be employed that rely on throwing away the actual values of the data and relying on ranks instead. The *Mann-Whitney Test*, and the *Wilcoxon Paired-Sample Test* are the equivalent of the *Student's T-Test* and the *Paired-Sample Student's T-Test*. Obviously, throwing out the numerical data and relying on the ranking of the data means that these tests are rather less powerful (requiring larger

sample sizes to produce significant results), however they generally give exactly the same result at the end of the day. You use them in exactly the same way as the *T-Test*, except that you'll need to use a more sophisticated statistics program.

Mann-Whitney Test from SPSS

GROUP	DATA		
1.00	28.38		
1.00	65.22		
1.00	72.96		
1.00	75.70		
1.00	123.34		
2.00	50.51		
2.00	111.58		
2.00	114.48		
2.00	140.29		
2.00	166.98		
- - - - - Mann-Whitney U - Wilcoxon Rank Sum W Test			
DATA			
by GROUP			
Mean Rank	Cases		
4.20	5 GROUP = 1.00		
6.80	5 GROUP = 2.00		
	--		
	10 Total		
		Exact	Corrected for ties
U	W	2-Tailed P	Z 2-Tailed P
6.0	21.0	.2222	-1.3578 .1745

Wilcoxon Paired-Sample Test from SPSS

BEFORE	AFTER	
28.38	50.51	
65.22	111.58	
72.96	114.48	
75.70	140.29	
123.34	166.98	
- - - - - Wilcoxon Matched-Pairs Signed-Ranks Test		
AFTER		
with BEFORE		
Mean Rank	Cases	
3.00	5	- Ranks (BEFORE LT AFTER)
.00	0	+ Ranks (BEFORE GT AFTER)
	0	Ties (BEFORE EQ AFTER)
	-	
	5	Total
Z =	-2.0226	2-Tailed P = .0431

There are several things to notice. Firstly that just as before the results are significant for the paired data and not for the unpaired data. Secondly, the level of significance is much lower (4% in the *Wilcoxon Matched Pairs* compared to 0.3% in the *Paired T-Test*. Thirdly, there are slightly different names that these tests are known by in different computer programs and text-books! This is a real nuisance and you'll just have to use your common sense. The *Mann-Whitney Test* is the same as the *Mann-Whitney-U Test* and the *Wilcoxon-Mann-Whitney-U Test*. There are no standard names. The ones used are the names of the authors of early papers on the techniques, and the letters in the names are the letters used to denote the statistic calculated in the paper. There's no standardisation on these either. Fourthly, look at the data format used. *Excel* tends to put data in multiple columns: one per sample. *SPSS* puts paired data into multiple columns, but non-paired data tends to be put in a single column with an index column that tells you which sample it belongs to. This latter approach seems unnecessarily complex for two variable data, but is very handy when you've got a more complex experimental design (which hopefully you never will).

Regression and Correlation

So far we have looked at situations where we want to differentiate between populations. Another thing that we often want to do is discover relationships between measurements taken from individuals within populations. For example, I might suspect that weight and height are related in humans so I would take a sample from a population and measure weight and height for each individual and then I would test to see whether there was any relationship. Regression and Correlation (specifically in our case *Simple Linear Regression* and *Simple Linear Correlation*) are two techniques to do that. They differ in that regression has what is known as the *independent variable* (often referred to as X) and the *dependent variable* (often Y) whereas correlation makes no assumption about dependency. Experimentally, this usually means that the independent variable is the one that is manipulated by the experimenter and the dependent variable is the one that is measured. In our weight/height relationship, neither variable is being directly manipulated and we have no real reason to believe that one of the variables determines the other (height and weight linkage is a relational chicken and egg syndrome – we don't really think that one 'came first' and caused the other) so we would use correlation, but if we were experimentally altering the amount of food we gave to a group of rats and measured their weight, then the amount of food would be the independent variable and the weight would be dependent.

Protein Intake (g)	Weight Gain (g)
0	2.6
1	8.2
2	11.3
3	13.1
4	17.3
5	24.3
6	24.2
7	33.3
8	36.7
9	36.6
10	42.2

SUMMARY OUTPUT

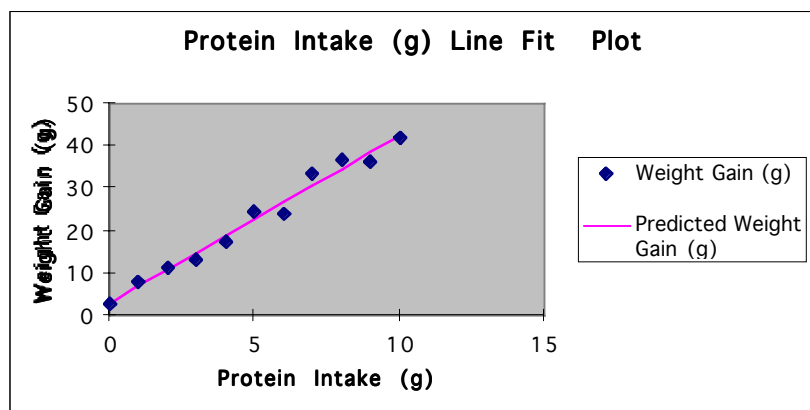
Regression Statistics	
Multiple R	0.991
R Square	0.982
Adjusted R Squ	0.980
Standard Error	1.882
Observations	11.000

ANOVA					
	df	SS	MS	F	Significance F
Regression	1.000	1721.075	1721.075	486.153	0.000
Residual	9.000	31.862	3.540		
Total	10.000	1752.937			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	2.925	1.061	2.756	0.022	0.524	5.325	0.524	5.325
Protein Intake (3.956	0.179	22.049	0.000	3.550	4.361	3.550	4.361

RESIDUAL OUTPUT

Observation	Actual Weight Gain	Residuals
1.000	2.925	-0.319
2.000	6.880	1.290
3.000	10.836	0.428
4.000	14.791	-1.698
5.000	18.747	-1.425
6.000	22.702	1.629
7.000	26.658	-2.507
8.000	30.613	2.730
9.000	34.569	2.138
10.000	38.524	-1.989
11.000	42.480	-0.277



The coefficients mean that the equation of the best-fit straight line for this experiment is:

$$\text{Weight Gain} = 2.925 + 3.956 \times \text{Protein Intake}$$

Since the data was generated artificially from the equation $y = 3 + 4x$, this is pretty good. The quality of the relationship is usually expressed by quoting r^2 value which is zero for no relationship and 1 for a perfect relationship. It is numerically equivalent to the square of the correlation coefficient which will be mentioned later. *Excel* tests the significance using *Analysis of Variance*, which is also perfectly acceptable.

As an aside, this sort of analysis is often useful to find these sorts of predictive equations and for calibrating items of equipment (e.g. calibrating a spectrophotometer to find out the actual concentration of a particular chemical in solution).

Lets go back to our height/weight correlation example.

Height (m)	Weight (kg)
1.5	50.3
1.53	56.2
1.73	96.7
1.42	51.8
1.5	61.7
1.49	59.8
1.72	102.3
1.54	71.6
1.41	48.3
1.42	47.5

	Height (m)	Weight (kg)
Height (m)	1	0.96063393
Weight (kg)	0.96063393	1

Correlation coefficients (also known as *Pearson's Correlation Coefficient*) can have values from -1 to $+1$. Minus values indicate that as one variable goes up, the other goes down, and positive values mean that the two variables go up together. A value of zero means that there is no association between the variables. Interestingly, *Excel* doesn't bother to tell us whether the correlation coefficient is significant – you'll have to look it up in a book of statistical tables or use a proper statistics program!

It is **ALWAYS** worth producing a scatter plot of measurements you are going to investigate with either regression or correlation. The methods above are for data that is linearly related (falls approximately on a straight line). A plot will tell you if this is a good idea. Any sign of a curved relationship and you'll need to think about transforming your data.

It is also worth thinking about how you intend to interpret your correlated/regressed data. Correlation in particular only shows that data is associated. It tells you absolutely nothing about any causal relationship there might (or might not) be. Commonly, two associated variables are associated because they both depend on a third factor, for example. In our case it is likely that a large weight does not cause a large height; the converse is probably not true either. However, it is a very likely that there is are heritable and environmental factors that affect both similarly.

Things to avoid

OK, we've covered the items the statistics that you should aim to use. What I now want to talk about are the situations that lead to ugly, difficult statistics and how you might avoid them. And what to do if you really feel that you have no choice!

Multiple Samples

This is probably the commonest level of increased complexity. Instead of having just two samples – A & B; treatment and controls – you have 3 or more samples. You cannot use a *t-test* in this situation. There are two approaches. Firstly, you may be able to reformulate your experimental design so that instead of comparing A with B and C simultaneously, you actually compare A with B, B with C, and C with A. This is OK if the you don't share data between the experiments – so the price you pay for simpler statistics is extra experiments (instead of one experiment comparing A, B and C, you have to do three experiments AB, BC and CA). For 3 cases this usually isn't too much extra work and if the experiments are reasonably straightforward, it's might be worth doing. The number of experiments you have to do to compare all combinations goes up very rapidly! If you insist on having multiple samples, then you need to use a technique called *Analysis of Variance* or *ANOVA* for short. There is even a non-parametric version called *Kruskal-Wallis Test*. You can do *ANOVA* in *Excel*, and it is pretty straightforward.

Consider an experiment where you have three strains of rats and you want to know whether they have different mean tail lengths:

A	B	C
87.0231	135.7275	126.2658
50.0325	135.6749	144.5158
103.76	98.871	118.2337
108.6303	109.8188	173.6637
65.6059		127.2721
		132.2786

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
A	5	415.0518	83.01036	623.371362
B	4	480.0922	120.02305	347.715365
C	6	822.2297	137.038283	397.005215

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	8127.84769	2	4063.92385	8.83196488	0.00438243	3.88529031
Within Groups	5521.65762	12	460.138135			
Total	13649.5053	14				

In this example, the P-value is clearly less than 0.05, which means that one or more of A, B and C is significantly different from the others. In this case (and you can't tell this from this analysis), A, B are samples from the same population and C is from a different population. To find out that it's only C that's different you would need to repeat the experiment looking at AB, BC and CA. However, if the difference is large, you might be able to get away with calculating the 95% confidence limits of the mean (see the single-sample section) and seeing the degree of overlap in each case. You may need to increase the sample sizes to get this to work though as *ANOVA* is quite a powerful technique.

Two-Factor Analysis of Variance

The next thing to avoid are experimental designs where you look simultaneously at the effects of two (or, heaven forbid, more than two) treatments. Examples are where you have controls and you are looking at the effect of two separate drugs. You can usually avoid this by looking at the drugs separately! The only excuses for looking at them together are if you suspect that the effect of one drug has some sort of cross-effect on the effect of the other drug (for example, if one drug inhibits the effect of the other drug), or it might be cheaper in terms of number of repeats to look at both simultaneously (for example, you may need to do this if you are killing animals). In actual fact, if you only have two-factors, it's not really too bad. Again, you can use *Excel* for this, but it's not very flexible and doesn't have all the options so you may need to brave a proper statistics package again.

Growth Rate	Control	Tetracycline	Penicillin
pH 7	25.6744	51.709	87.8637
	13.3442	66.0015	87.8375
	31.2533	50.8892	69.4355
	32.8768	48.5653	74.9094
	18.5353	37.5148	72.6196
pH 8	41.9092	54.4162	67.1994
	41.8916	29.9573	80.8869
	29.6237	60.7149	61.1753
	33.2729	74.3534	102.7478
	36.6743	39.6234	67.9541

Anova: Two-Factor With Replication

SUMMARY	Control	Tetracycline	Penicillin	Total
<i>pH 7</i>				
Count	5	5	5	15
Sum	121.684	254.6798	392.6657	769.0295
Average	24.3368	50.93596	78.53314	51.2686333
Variance	69.2633508	103.329459	76.1247414	595.629173
<i>pH 8</i>				
Count	5	5	5	15
Sum	183.3717	259.0652	379.9635	822.4004
Average	36.67434	51.81304	75.9927	54.8266933
Variance	28.9762199	305.087019	275.321282	455.034819
<i>Total</i>				
Count	10	10	10	
Sum	305.0557	513.745	772.6292	
Average	30.50557	51.3745	77.26292	
Variance	85.9439462	181.732121	157.990965	

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	94.9484322	1	94.9484322	0.66389607	0.42319989	4.25967528
Columns	10973.241	2	5486.62051	38.3634117	3.3481E-08	3.40283179
Interaction	303.646563	2	151.823281	1.06157498	0.36159996	3.40283179
Within	3432.40829	24	143.017012			
Total	14804.2443	29				

Chi-Squared Test

Chi-squared tests are used for so called *goodness of fit* testing. There are 2 main occasions where this can be used, but the chances are you won't come across them. Firstly, when you have a theoretic prediction about expected frequencies and you want to see whether the observed frequencies are a good match. The classic example here is when you have a genetic trait and you want to see whether the frequencies of the phenotypes fit the predicted Mendelian ratios. Secondly, when you want to see whether two samples could have come from the same population where you are using

frequencies of some characteristic as your defining criteria – for example you might be interested in frequencies of different hair colours among Scottish and English populations. Excel can be persuaded to do simple Chi-squared, but you have to do quite a lot of work yourself. It's probably easier to use another program.

	Actual	Expected	=CHITEST(B2:B3,C2:C2)
Green	50	48.75	0.72030033
Yellow	15	16.25	
Sum	65		
	Actual	Expected	=CHITEST(B2:B3,C2:C2)
Green	56	48.75	0.0378262
Yellow	9	16.25	
Sum	65		

Multivariate Analysis

Don't go there! The worst and ugliest statistics are needed for non-experimental data. This comes up when you collect data by survey or by observation without any clear hypothesis to test. Often referred to as fishing trips, you simply collect as much data as you can with the hope that something interesting will turn up. You should not do this! Or at least if you do, you should be aware of the difficulties of trying to do anything rigorous with your data. With this sort of data you will almost certainly need to use the big class of multivariate statistical methods and these are not for the faint of heart. With less rigour, you may be able to get away with selecting pairs of data from within your dataset that you can then analyse using the basic *T-Test* and *correlation* style methods. If you do this, however, the significance levels will not be what the computer program spits back at you so interpret with caution.

The Bottom Line

You should be able to define the statistics you need right at the beginning. No one should ever come to their supervisor and say, "I've collected some data, now what do I do with it?" Deciding on the analysis up-front generally makes it a trivial exercise so discuss it with your supervisor **BEFORE** you collect the data. If you are not happy with statistics, design the project/experiment so that you don't need to do anything more complicated than a *T-Test*. It can almost always be done. If you are keen, then read a statistics book and find an example in the book that matches what you are trying to do and just plug in your numbers. Again if you are keen it is probably worth getting to grips with a proper statistics program such as *Minitab*, *SPSS*, *Statview* or whatever. However, *Excel* does a creditable job nowadays and is probably all you will need. And remember, the right graph can be far more impressive than lots of numbers!

Exercises

Split up into groups of 4 or 5 and work out an outline design for the experiment, including the sample size, the data you would collect and how you would analyse the data. Consider what alternatives there are and how they would affect the data analysis and any difficulties there might be in interpretation. Be specific about the hypotheses you are testing.

1. Design an experiment to investigate the effect of alcohol consumption on reaction time.
2. Design an experiment to investigate the effect of growth hormone on birth weight in mice.
3. Design an experiment to investigate whether Scottish students do better at Edinburgh than English ones.
4. Design an experiment to investigate the effect of age at death on skeletal mass.
5. Design an experiment to investigate whether eye colour is linked to height.