Brief notes on statistics: Part 3
Null hypothesis significance tests and confidence intervals

Michael Wood (Michael.wood@port.ac.uk)

22 October 2012

Introduction and links to electronic versions of this document and the other parts at http://woodm.myweb.port.ac.uk/stats. The data in the tables, and the figures, are in the spreadsheet, http://woodm.myweb.port.ac.uk/stats/StatNotes.xls . For a rough, but still useful, understanding, you can ignore the numbered endnotes, which give extra detail.

Sometimes, particularly when conclusions are based on a small sample, the question arises of whether another sample might give a different result. **Can we be sure of the result, or could it just be the result of chance?** We'll look at two approaches to this question – null hypothesis (significance) tests, and, very briefly, confidence intervals.

This problem of ruling out chance as an explanation occurs **all over the place**. In October 2007, heart transplants were stopped at Papworth Hospital because 7 out of 20 patients had died instead of the normal 10% (Garfield, 2008 at http://tinyurl.com/5yyckp), but perhaps this was just chance[1]? Global temperatures seem to be rising, but temperature always have varied from year to year so can we be sure that what we are seeing now is more than this ordinary chance fluctuation? Schools are judged by league tables, speed cameras are justified by statistics showing falling levels of road accidents, but it is easy to forget to check whether the variations are bigger than can be accounted for by chance. (See the newspaper article at http://tinyurl.com/2wee89 for more on this issue.)

Both hypothesis testing, and the estimation of confidence intervals, are complex topics. My aim here is to explain the **basic ideas**, so that you should understand what the results mean and where they are important. You may also be able to use some of the methods we will look at in your own research. However, explaining all the relevant methods and their justification would take far too long, so you may need help in applying the methods to your own particular situation – in which case please ask.

Hypothesis testing

Testing hypotheses is one of the most widespread uses of statistical methods. Strictly, we should talk of ***null hypothesis* testing**, because the hypotheses tested are always ***null*** ones – what this

means should become clearer when you look at the examples. The tests are sometimes referred to as ***significance tests***.

The basic idea of statistical (null) hypothesis testing is very simple. However, the details of how to carry out a test – the mathematical formulae, statistical tables, and so on – tend to be complicated. And you will find that each different situation requires a different test, with different formulae and different tables. Furthermore, although the underlying idea *is* simple, many people find it confusing!

We'll start with an example where the underlying idea is pretty obvious – an **experiment on telepathy**.

An experiment on telepathy

Telepathy is communication which does not rely on any of the known senses: by some unknown method one person becomes aware of what another person is thinking, despite the absence of any opportunity to communicate. For this experiment, a pack of cards was shuffled and split so one of the cards was visible to a volunteer, Annette, who then concentrated hard on it. In another room, another volunteer, Brian, tried to see if he could determine what card Annette was thinking about. He got the card right! Conditions were checked, but there was definitely no way Brian could have known which card was chosen except by means of telepathy. There were only two viable explanations: either Brian was guessing and was lucky, or he was communicating telepathically with Annette.

The general idea of hypothesis testing is that you set up a ***null hypothesis*** – typically a hypothesis that nothing interesting is happening – and then check if the data you've got is consistent with this hypothesis.

In this case the null hypothesis is that Brian was guessing. The probability of Brian guessing the card correctly is about 2% (one chance in 52). Do you think this suggests that telepathy is the more likely explanation?

If you think that telepathy is *not* the more likely explanation, this is probably because you think telepathy is very unlikely or even impossible. Brian may be unlikely to get the card right by guessing, but if telepathy is impossible, this is the only viable explanation.

We could get stronger evidence by asking them to repeat the experiment. Let's imagine they guessed another card, under exactly the same circumstances, and got this one right too. Their score is now two out of two. What is the probability of this happening ***if the null hypothesis (guessing) is true***?

This probability comes to just under 0.04%, or about one in $2{,}700^2$. This is much less likely to happen by chance, which makes telepathy a more plausible explanation. Do you think this

stronger evidence suggests that volunteer Y really is telepathic? Again, there is no right answer here. If you think telepathy is completely impossible, then you will cling to the chance explanation, however unlikely it is to result in two correct guesses.

These probabilities are called ***p values*** or ***significance levels***. They tell us how likely the data is to have arisen from the null hypothesis. The ***lower*** the ***p value,*** the less plausible the null hypothesis is, and so the more likely is the alternative hypothesis – telepathy.

I made up this story of Annette and Brian, but i n the 1920s and 30s, the psychologist, J B Rhine, found a number of people who appeared to be telepathic. In one series of experiments, Hubert Pearce Junior, did a card guessing experiment 8,075 times, and got the card right on 3,049 occasions. There were five cards in the pack, so guesswork would have produced about 1615 hits. Rhine argues that Pearce's performance is so much better than guesswork that telepathy must be involved; others have taken the hypothesis that Pearce was cheating more seriously. Working out the *p* value for Rhine's experiment is more difficult. You either need to use more advanced probability or computer simulation (see Wood, 2003, p. 104 and Chapter 5).

The general procedure for any null hypothesis test

This is very simple, and comprises three steps:

Step 1: ***Formulate the null hypothesis****.* This is an imaginary world, in which the thing, which really interests you, does ***not*** happen. In the telepathy test, the null hypothesis is that volunteer Y is *not* telepathic and can only guess. More typically, the null hypothesis might be the hypothesis that there is *no* difference, on average, between two groups, or *no* relationship between two variables. It is usually a null, "nothingy" hypothesis. If the null hypothesis is true, any difference or relationship in the data must just be due to chance. An alternative name would be the ***chance hypothesis.***

Step 2: ***Estimate the p value****.* This stands for 'probability' - of results like those actually observed, or more extreme than those observed, ***if the null hypothesis is true***. It tells you how likely the results are to have occurred, ***if*** the null hypothesis is true. An alternative term for a ***p*** value is ***significance level***. In the telepathy test, the *p* value is 2% for the one card experiment, and 0.04% for the two card experiment.

Step 3: ***Draw your conclusions****.* The ***lower*** the ***p*** value, the ***less*** plausible the null hypothesis is, and so the ***more*** plausible it is that another hypothesis is true. In the telepathy experiment, the ***p*** value for the one card experiment was 2%, and for the two-card experiment it was 0.04%. The second experiment provides far stronger evidence against the null hypothesis, and so for the hypothesis of telepathy, than the first: this is indicated by the lower *p* value for the second experiment.

The important thing to remember is the null hypothesis. Whenever you see a ***p*** value, there *must be a null hypothesis which has been tested*. The null hypothesis represents a ***baseline***

assumption against which reality is tested. ***It is important to imagine this null hypothesis, and to see how it works***. Then you will be in a position to see how the ***p*** value is estimated, because it is always estimated on the assumption that the null hypothesis is true.

You may also have another hypothesis that you are trying to prove – e.g. that the volunteer *is* telepathic. This is sometimes called the **alternative hypothesis.**

This should all be clearer after we have looked at another example.

Sex differences in an examination

The table below shows the marks obtained by ten students in an exam.

Exam marks and sex of ten students

Mark	Sex
37	F
46	F
56	F
49	F
78	F
50	M
55	M
81	M
55	M
53	M

The average (mean) of the marks for the female (F) students is 53.2%, whereas for the males (M) it is 58.8%. The difference is 53.2%-58.8% or –5.6%. On average the males did 5.6% better.

However this is just a small group of students. With another group of students we may get a different answer. Can we be sure that males really do better overall?

Quick question 1. What do you think the answer is? Does this evidence prove the point, or could it be a fluke?

Let's now do a hypothesis test. The ***null hypothesis*** is that whether a student is male or female has no relationship to the mark they are likely to get. There are no systematic differences between males and females, and the average mark for ***all*** male students (who might have taken the exam) is the same as the average for ***all*** female students. The actual difference we have observed in our small sample of ten students is a 5.6% difference in the average marks of five males and five females. How likely is this to have occurred if this null hypothesis is true?

Quick question 2. What do you think this p value is? (There is no easy way of calculating it exactly: all you can do is guess what you think it is roughly.) My answer is on the next page, so I'll leave a page break to give you the opportunity to have a guess without seeing the answer …

The probability, or *p* value, is about 53%. This was worked out using the randomization (or shuffle) test which is explained at http://woodm.myweb.port.ac.uk/stats/ShuffleTest.pdf or http://youtu.be/Uyub9cYimWw .

This is a simple, but powerful, computer simulation method which has the advantage that it is easy to see how it works and what the final probability means. I would suggest you look at this link, but don't worry about the details because in practice there are more convenient ways of getting the answer.

The standard method used for this type of situation is a *t* test or ANOVA (analysis of variance): they both give the same answer for the *p* value: 55% (which is, not surprisingly, similar to the 53% from the randomization test). The mathematical details of these are complicated and uninformative for anyone without a detailed understanding of the background theory. In practice most people use a computer package to work out the *p* values (see below for more details).

Quick question 3. What conclusions can you draw from the high p value (53% or 55%)? (This is an important question – check my answer if you are in any doubt.)

Other hypothesis tests

There are many other methods of testing null hypotheses – e.g. the *t* test, analysis of variance (ANOVA), the X^2 (chi squared) test, the sign test, and many others. Almost all rely on mathematical probability theory, rather than simulation[3]. However, the general procedure (Steps 1, 2 and 3 above) is exactly the same for *any* test, and *any* test *can* be carried out by a simulation method like the one just described (but in practice other approaches may be easier).

One type of test which is widely used in management research is **a test of the null hypothesis that two variables are not related and that the overall correlation is zero** (e.g. Tables 1 and 2 in Moutafi et al, 2007, at http://tinyurl.com/29cbgt). Significant results (with a low *p* value) then indicate a genuine relationship that goes beyond the level that chance could produce.

If you need to do a hypothesis test as part of your research, you may be able to use the shuffle (randomization) test. Alternatively, and probably much easier, you can use a statistical package like SPSS (see http://woodm.myweb.port.ac.uk/stats/StatNotes0.pdf) to do one of the other tests. Deciding which test to do may be a problem, but if you analyze your data using a package like SPSS the appropriate method is likely to be suggested. However, remember that the general procedure is *always* as above, and if you think in terms of the shuffle test you will have a fair idea of what the mathematics in the background is doing.

When you read research reporting these other tests, or use a package like SPSS to do them, you are likely to see **mysterious quantities** with names like ***t*** or ***F*** or $\mathbf{X^2}$ (chi square). These are part of

the mathematics used to calculate the *p* values. Unless you are interested in the mathematics, I would ignore them, and focus on the *p* values or "Sig" (significance) levels.

The interpretation of hypothesis tests

Many research papers (e.g. Tables 1 and 2 in Moutafi et al, 2007 at http://tinyurl.com/29cbgt give *p* values a **star rating** to indicate how *statistically significant* the results are. For example, one system is

*** means p<0.1%; ** means p<1%,; * means p<5%

The cut-off level of significance is often taken as 5%. If *p* is less than 5% it would be described as "significant at the 5% level", meaning that we have evidence that the null hypothesis is false (with the 5% describing the strength of the evidence). On the other hand, if *p* is more than 5%, this would be described as "not significant", meaning that evidence is not strong enough. Notice that *lower p* values get *more* stars because they indicate *stronger* evidence against the null hypothesis and so for the alternative hypothesis.

Customer service ratings from McGoldrick and Greenland (1992)

Let's see how this applies to some research on customer service in two different kinds of financial institution: banks and building societies (McGoldrick and Greenland, 1992). The data in the table below was obtained from a sample of customers who rated each institution on a scale ranging from 1 (very bad) to 9 (very good.). The above six dimensions are a selection from the 22 reported in the paper. NS means not significant - which in this table means that the *p* value is greater than 0.1. (This is a bit unusual: as mentioned above this level would normally be 5%.)

Aspect of service	Banks' mean rating	Building Society's mean rating	Level of significance (p)
Sympathetic/understanding	6.046	6.389	0.000
Helpful /friendly staff	6.495	6.978	0.000
Not too pushy	6.397	6.644	0.003 (0.3%)
Time for decisions	6.734	6.865	0.028 (2.8%)
Confidentiality of details	7.834	7.778	NS
Branch manager available	5.928	6.097	0.090 (9%)

Quick question 4. What are the null hypotheses which are being tested here and on which these six p values are based?

Quick question 5. How many stars do each of the six results in the table get? For which aspect of service is the evidence for a difference between banks and building societies strongest?

Remember: the *lower* the *p* value, the *more* convincing the evidence is *against* the null hypothesis. Unfortunately, this is rather counter-intuitive and easily misinterpreted. Take care!

The main thing to remember is that the ***p*** value ***only*** tells you about the strength of the evidence against the null hypothesis[4]. With the exam marks the ***p*** value was large (53%) indicating that the data was consistent with the null hypothesis. However, this does ***not*** prove that there is no difference between males and females in their performance in this exam, just that there is not enough evidence to be certain. On other occasions, you may get a very low ***p*** value, suggesting very strong evidence for a difference, but the actual difference may be too small to be of interest. **You should always look at the size of any difference, as well as the *p* value.**

Quick question 6. Work out the size of the differences between each pair of means – e.g. for Sympathetic / understanding this difference is –0.343 (taking the second mean from the first) indicating that the mean rating for the banks is 0.343 less than the equivalent figure for the building societies. Do you think the differences between the banks and the building societies are large enough to be important? (If your answer to this is no, then there is little point in testing the null hypotheses formally, as the authors of this research article did.)

There is a strong argument that null hypothesis tests are **too prominent** in many fields of research (e.g. see Armstrong, 2006, at http://tinyurl.com/3a35fb, and Wood, 2003, for a more detailed discussion of some of the difficulties). Many articles state the null hypothesis and the alternative hypothesis formally (one of very many examples is Mathur et al, 2001 at http://tinyurl.com/yshfzx) with the aim of the research being to see which hypothesis fits the data better. In most cases, it is arguably better to state the main research aim as finding the **size** of the effect (or difference or correlation). The null hypothesis test is still useful as a check of whether the results can be explained by the chance hypothesis, but the main result would be the size of the effect. For example, the first row in the table above tells us that the building societies' results are on average 0.343 better than the banks', and that that this result is significant at the *** level.

An alternative approach to null hypothesis testing for assessing sampling error is the use of confidence intervals.

Confidence intervals

A typical political opinion poll uses a sample of 1000 electors to predict the result of an election involving millions. Obviously, the result of the poll can't be expected to be completely accurate, and so an error range is often given. For example, one poll predicted that the Conservatives would get 30% of the vote, and the error range given was +/- 3%. This means that the poll is predicting that the Conservatives will get somewhere between 27% and 33% of the vote in the election.

This error range is a *95% confidence interval*[5]. It means that we should be 95% confident that the truth is in the confidence interval. If the proportion voting Conservative turned out to be 28% or 31%, these would be in the confidence interval. If, on the other hand, only 26% voted Coservative, this would be outside the interval.

Quick question 7. What proportion of these predictions of election results would you expect to be outside the 95% confidence intervals?

Confidence intervals can be calculated using probability theory[6] to take account of sampling error – the fact that another sample would inevitably give a slightly different result. The method is based on the assumption that the sample is random, or a close approximation to a random sample.

Quick question 8. What confidence interval do you think the probability theory would predict if the sample only comprised 250 electors? (You probably won't be able to work it out exactly, but you should be able to guess whether the interval will be wider or narrower.)

Quick question 9. Confidence intervals take account of sampling error. Can you think of any other sources of error (which are not taken into account by these confidence intervals)?

Confidence intervals can also be used instead of hypothesis tests. Instead of giving *p* values, the authors could have given confidence intervals for the differences between the banks' and building societies' ratings. The first such difference is –0.343, and the 95% confidence interval might be –0.193 to –0.493[7]. (Not having the original data, this is just a guess, but it is consistent with what we are told.) This means that the true difference, taking *all* customers into account, is somewhere between about –0.2 and –0.5. Which, of course, means that we are confident it is not zero or positive; in other words the building societies have got a real, but slight, advantage.

Quick question 10. The width of this 95% confidence interval is 0.3. Do you think the 80% confidence interval would be wider or narrower?

Using confidence intervals instead of *p* values in this way has a number of advantages – they are, for example less likely to be misinterpreted. They are commonly used in some fields (e.g.

medicine), but very rarely in others (e.g. management). I think they deserve to be used more widely. I will show how they can be used in **regression** in Part 4 of these notes[8].

Sample sizes

It is possible to use the idea of hypothesis testing, or confidence intervals, to decide how large a sample needs to be. For example, the answer to *Quick Question 7* above is that the conclusion from the sample of 250 will have a wider confidence interval than the sample of 1000 – 24% to 36%, instead of 27% to 33%. We would now need to ask if this confidence interval indicates adequate reliability. Is it OK for the estimate to be correct to within 6% (of the best guess of 30%) with 95% confidence? If it is OK the sample of 250 is enough, otherwise we need a larger sample. If we want the answer to within 1%, the statistical calculations will tell us we need a sample of 9000.

Null hypothesis tests can be used in a similar way.

Summary

Both null hypothesis tests and confidence intervals provide a way of assessing how sure we can be of our results, bearing in mind the possibility of sampling error – the fact that one sample or set of results is likely to be different from another, despite the fact that they both come from the same source.

The starting point for a hypothesis test is always a *null hypothesis* – typically that there is *no* difference or *no* relation between variables or groups. You can think of the null hypothesis as the *chance* hypothesis – any differences or relations in the data are just due to chance. The alternative to the null hypothesis is sometimes called the *alternative* hypothesis. This is what you are trying to prove.

All tests result in a *p* value, which is the probability of obtaining results like the actual results, or more extreme, *if the null hypothesis is true*. The *p* value indicates the *plausibility* of the null hypothesis: *low p* values suggest that the null hypothesis is not plausible, and so there must be a *real difference or relation* between the variables or groups and the alternative hypothesis must be true. Often, 5% is taken as a cut-off level: *p* values which *are less than 5%* indicate *significant evidence against the null hypothesis* and *for the alternative hypothesis.*

Confidence intervals indicate the likely accuracy of an estimate of a statistic like a mean, a proportion or a correlation. For example, the (fictional) data in Wood (2004, copy at http://woodm.myweb.port.ac.uk/boot.pdf) indicates that we can be 95% confident that the correlation between age and time taken to run 10 km is between 0.22 and 0.57. This estimate takes account of the fact that the correlation is based on a limited sample of data.

Exercises

1 The table below is based on answers from 92 students to questions about their age, how many units of alcohol they had drunk the previous Saturday, Sunday and Monday, and the average number of cigarettes a day they smoked. The (Pearson) correlations indicate the relationship between the variables. (The data is at http://woodm.myweb.port.ac.uk/nms/drink.xls.)

		AGE	SATUNITS	SUNUNITS	MONUNITS	DAYCIGS
AGE	Pearson Correlation	1	-0.288	-0.130	-0.348	-0.097
	Sig. level (p-value)		0.005	0.218	0.001	0.359
	N	92	92	92	92	92
SATUNITS	Pearson Correlation	-0.288	1.000	0.591	0.729	0.554
	Sig. level (p-value)	0.005		0.000	0.000	0.000
	N	92	92	92	92	92
SUNUNITS	Pearson Correlation	-0.130	0.591	1.000	0.649	0.759
	Sig. level (p-value)	0.218	0.000		0.000	0.000
	N	92	92	92	92	92
MONUNITS	Pearson Correlation	-0.348	0.729	0.649	1.000	0.480
	Sig. level (p-value)	0.001	0.000	0.000		0.000
	N	92	92	92	92	92
DAYCIGS	Pearson Correlation	-0.097	0.554	0.759	0.480	1.000
	Sig. level (p-value)	0.359	0.000	0.000	0.000	
	N	92	92	92	92	92

What are the null hypotheses are being tested by the results in Table 7? Which results are significant at the 1% significance level? Explain what you can conclude about the relationships between

- the amount the students drink on Saturday and the amount they smoke
- their age and the amount they smoke.

2 An organization is worried about errors in invoices. A random sample of 20 invoices is studied in detail and 8 of them are found to have errors. This is considered unacceptable so they make some changes in the process, and then take a further sample of 20 invoices. This time there are only 4 errors. Does this prove that the process has been improved? How would you analyze this using a hypothesis test? What would the null hypothesis be?

Would it be helpful to use a larger sample of invoices?

3 Find some research articles which give *p* values as part of the results. Choose two or three *p* values and for each:

(a) Write down the null hypothesis

(b)Write down in plain English what the *p* value tells you about the results.

Additional exercises

4 Use the data at http://woodm.myweb.port.ac.uk/nms/drink.xls to compare the amount drunk by males and females, and check if the results are statistically significant.

In practice, a statistical package like SPSS is the best way to do this sort of analysis. SPSS is available on the network under Start - Academic applications, or from the library if you want to install it on your own computer. To use SPSS, you will need to download the file (http://woodm.myweb.port.ac.uk/nms/drink.xls), save it, and load it into SPSS (don't forget to tick the box telling SPSS to load an Excel file). Now use Analyze – Compare Means – Means to compare the amount that the males and females in this data set drink and smoke. Sex will be the independent variable here, and the variables giving details of drinking and smoking will be the dependent variables. You will need to tick Anova under Options to get significance levels. Make sure you understand the results including the *p* values (in the column headed "sig.").

Alternatively Excel has some statistical routines built in. Click on Data, then Data analysis and then Anova: single factor to compare means. (You will need to make sure the Analysis ToolPak is installed on your computer.)

Now do the same to compare the three courses (FB, FP, PP). (Compare means can compare more than two groups.)

You should always check data to make sure that there are no errors or anything which is not believable. Do you think any of the drink data might be inaccurate? If so, delete it, and redo your analysis. Is the result the same?

You should also be able to use SPSS (or Excel) to produce the table of correlations above. And there are lots of other things that SPSS can do with this data. There are a few more details towards the end of the last section of these notes on Practical aspects of using statistics (http://woodm.myweb.port.ac.uk/stats/StatNotes0.pdf).

5 Click the Exam sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls. You will see here that the assignments were marked by three markers, A, B and C. Work out the average (mean) of the marks given by each of these markers. Are there any differences, and are they statistically significant? What you say to students who are worried about the fairness of the marking process? (You should be able to use Compare means in SPSS as explained in the previous question.)

Answers to Quick questions and some exercises

1 Intuitively, I feel this data could be a fluke. The next sample might give a different result. However, your intuition might give a different answer. Intuition is unreliable in this context, which is why we need a formal way of testing hypotheses.

2 Answer in the text.

3 The probability of the results, or more extreme results, arising if there really is no difference between males and females is 53%. This high p value indicates that it is very possible that the null hypothesis is true. Note that we cannot say for sure that the null hypothesis is true: the data does, after all, indicate a 5.6% difference. There may or may not be a difference between male and female performance in the exam, but the data is consistent with the hypothesis that there is no difference.

4 There are six null hypotheses – one for each aspect of service in the table. The first states that there is no difference between the (overall) Building Society mean rating and the Bank mean rating for sympathetic understanding. The result of testing this null hypothesis is a very low p-value (0.000) which indicates that the null hypothesis is not plausible and so there must be a real difference between banks and building societies (although the size of the difference is small). The result is said to be very significant (statistically speaking, of course).

5 The first two are three star results, Not too pushy is two star, Time for decisions is one star, and the other two are no star. These last two are not significant at the 5% significance level. The evidence for a difference is strongest for the first two results.

6 The biggest difference is the second which 0.483. Remembering that this is a 1 to 9 scale, this seems too small to be interesting to me, but you may, of course, disagree.

7 5%.

8 As the sample is smaller, the degree of accuracy will be reduced so the confidence interval will be wider. In fact statistical theory, or simulation, shows that the interval will be +/- 6% or 24% to 36%.

9 There are several other sources of error. People may not tell the truth about their intentions to opinion pollsters, or they may change their mind between the opinion poll and the election. Remember that confidence intervals, and hypothesis tests, *only* deal with sampling error.

10 You would always expect an 80% confidence interval to be narrower than a 95% one.

Exercise 1: The significance levels, or p-values, are based on the null hypotheses that all the correlations are actually zero and there are no relationships between the data. Low p-values mean that this hypothesis is not plausible, so the null hypothesis should be rejected and we should conclude that the correlation is genuine. *All* the correlations *except* those between AGE and SUNUNITS and AGE and DAYCIGS are significant at the 1% significance level.

The correlation between SATUNITS and DAYCIGS is 0.554. This suggests a reasonably strong tendency for people who drink a lot on Saturday night to be among the heavier smokers. The p-value for this is 0.000, which is very low indicating a significant result – this means that this is *not* likely to be a chance effect, so there is (almost certainly) a positive correlation between these two variables.

On the other hand, the correlation between AGE and DAYCIGS is -0.097. This suggests a slight tendency for older people to smoke less – but the high value of the significance level (0.359) suggests that the null hypothesis is plausible and this could well be a chance effect.

Exercise 2: The data obviously does not prove that the process has improved. The apparent improvement might be just due to chance. To analyze this using a null hypothesis test we would set up the null hypothesis that the process is unchanged, and then work out the *p* value which tells us how probable it is that the data could be a result of chance. Using a larger sample of invoices would help: the result would be more likely to be significant.

The *p* value comes to about 0.3 or 30%. This is not significant (*p* is not less than 5%), so the result could be due to chance and there is no strong evidence that the process has improved.

If you want to work this out for yourself (not compulsory!), you can either use the randomization test as described at http://woodm.myweb.port.ac.uk/stats/ShuffleTest.pdf, or SPSS. In either case, use the code 1 for an invoice that contains errors, and 0 for an invoice that is error free, and the two groups you need to compare might be called Before (the change) and After. You will have 20 Before rows and 20 After rows making 40 rows of data in all. The randomization test should be self-explanatory. Using SPSS, click on Analyze – Descriptive statistics – Crosstabs. It doesn't matter which variable is Row and which is Column, but you will need to click on Statistics and tick Chi square. The best answers are using the continuity correction or the Fisher exact test.

References

Armstrong, J.S. (2006). *Significance tests harm progress in forecasting*. Wharton School, University of Pennsylvania. Downloaded on 24 February 2008 from http://marketing.wharton.upenn.edu/ideas/pdf/Armstrong/StatSigIJF36.pdf .

Ayres, I. (2007). *Super crunchers: how anything can be predicted.* London: John Murray.

Garfield, S. (2008). What went wrong? *The Observer magazine, 6 April.*

McGoldrick, P. M., & Greenland, S. J. (1992). Competition between banks and building societies. *British Journal of Management*, 3, 169-172.

Mathur, I., Singh, M., & Gleason, K. C. (2001). The evidence from Canadian firms on multinational diversification and performance. *The Quarterly Review of Economics and Finance, 41*, 561-578.

Moutafi, J, Furnham, A, & Crump, J. (2007). Is managerial level related to personality? *British Journal of Management, 18*, 272-280.

Wood, M. *Making sense of statistics: a non-mathematical approach*. Basingstoke: Palgrave, 2003.

Wood, M. (2004). Statistical inference using bootstrap confidence intervals. *Significance, Volume 1* (4), 180-182 (copy at http://woodm.myweb.port.ac.uk/boot.pdf).

Endnotes – these include more detail which can be ignored for a rough understanding

[1] The *p* value is about 0.2% (using the binomial distribution). This is the probability of 7 or more out of 20 dying if the overall death rate is 10%. The underlying concept (although not the method of calculation) should be clear when you have read to the end of these notes. A few weeks after the operations were stopped, an enquiry made 12 recommendations and agreed that transplants could be resumed, but had "no specific explanation as to why so many recent operations ended in failure" (Garfield, 2008, p. 31 at http://tinyurl.com/5yyckp).

[2] Imagine them doing the whole experiment lots of times. Brian will guess the first card correctly on about 2% of occasions. He will also get the second right on 2% of this 2% of occasions, which is 0.02x0.02 or 0.004. The exact answer is (1/52)x(1/52) which is 1/2704 or 0.00037. (This is the multiplication rule of probability.)

[3] If you have used any of these hypothesis tests before, you will probably have seen graphs very like Figures 1 and 2. The only difference is that the graphs for most tests are produced by the mathematics behind the probability theory, rather than by simulation.

[4] It may be tempting to read too much into *p* values. The ***p*** value in the one card telepathy experiment was 2% (see http://www.palgrave.com/skills4study/subjectareas/statistics/telepathy.asp). Can we assume this means that the probability of the volunteer guessing is 2%, so the probability of her being telepathic is 98%? Obviously we can't. You probably decided that the evidence was not strong enough to convince you that volunteer Y was telepathic, so it certainly doesn't make sense to

say that the probability of her being telepathic is 98%! This probability is the probability of volunteer Y getting the card wrong if she's guessing – not a very interesting probability. Similarly the 2.5% ***p*** value for the data in Table 4 doesn't mean that the probability of the null hypothesis is 2.5% so the probability that there is a difference between males and females is 97.5%. The ***p*** value gives an indication of how plausible the null hypothesis is, but it is ***not*** the probability of the null hypothesis being true. If you want this, you need to use Bayesian statistics (see Wood, 2003 for a very brief introduction).

[5] Most people are uncertain about some of the things they think they know. If the knowledge is numerical, this uncertainty can be expressed as a confidence interval. Ayres (2007, p. 113) suggests a good exercise to test the accuracy of self-assessed confidence intervals. The exercise comprises 10 questions with a definite answer (the first is "What was martin Luther King Junior's age at death?) with the instruction to assess one's knowledge with a confidence interval. For example, my 80% confidence interval for this first question was 25 to 50 years – the correct answer, 39 years, was within my interval. We would expect about 80% of such confidence intervals to include the correct answer. In practice, according Ayres, this figure is usually substantially less. because there is a tendency for people to overestimate their degree of confidence.

[6] Like hypothesis tests, you can use simulation methods for working out confidence intervals. The method used is known as bootstrapping – see, for example, Wood (2003, chapter 7) or Wood (2004 – copy at http://woodm.myweb.port.ac.uk/boot.pdf).

[7] Confidence intervals such as these are often worked out using a statistic called the *standard error*. This is the standard deviation which would be expected for the quantity you are measuring. In the example, this is a mean, and standard error of the mean would be 0.075. Using the normal distribution, the 95% confidence interval would then extend from about 2 standard errors (0.15) above the mean to 2 standard errors below the mean. (With small samples this is not quite right because theory then tells us we have to use the *t* distribution: with large samples this becomes closer and closer to the normal distribution.) Sometimes you are just told the standard error – which you can interpret by thinking of the plus or minus one standard error interval as being a 68% confidence interval.

[8] For one approach to deriving confidence intervals for regression slopes, and also a direct illustration of what conclusions might be derived from different samples from the same source, see http://woodm.myweb.port.ac.uk/BRLS.xls.