# Chapter 1

## In the beginning...

<table>
<thead>
<tr>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 My favorite papers</td>
</tr>
<tr>
<td>1.1.1 Hurlbert, 1984. Pseudoreplication</td>
</tr>
<tr>
<td>1.1.2 Ruxton, 2006. The Welch t-test</td>
</tr>
<tr>
<td>1.1.3 Warton et al. 2011. The arcsine is asinine</td>
</tr>
<tr>
<td>1.1.4 Iannidia, 2005. Why most published research findings are false</td>
</tr>
<tr>
<td>1.2 Introduction</td>
</tr>
<tr>
<td>1.3 Effective note taking strategies</td>
</tr>
<tr>
<td>1.4 It’s all $\Gamma$ to me</td>
</tr>
<tr>
<td>1.5 Which computer package?</td>
</tr>
<tr>
<td>1.6 FAQ - Frequently Asked Question</td>
</tr>
<tr>
<td>1.6.1 Accessing journal articles from home</td>
</tr>
<tr>
<td>1.6.2 Downloading from the web</td>
</tr>
<tr>
<td>1.6.3 Printing 2 pages per physical page and on both sides of the paper</td>
</tr>
<tr>
<td>1.6.4 Is there an on-line textbook?</td>
</tr>
</tbody>
</table>

The suggested citation for this chapter of notes is:


In Course Notes for Beginning and Intermediate Statistics.


## 1.1 My favorite papers

Ironically, we will start with an overview of some of my favorite papers for users of statistic in ecology. Many of these paper will be explored in more detail in subsequent chapters.
CHAPTER 1. IN THE BEGINNING...

1.1.1 Hurlbert, 1984. Pseudoreplication


Hurlbert (1984) has become one of the most widely cited papers in the biological literature – it has been awarded the Citation Classic status. There is no more devastating review of a report, than a simple one-liner indicating that the researcher has fallen prey to pseudo-replication. This is a MUST READ paper before planning any field experiment.

1.1.2 Ruxton, 2006. The Welch t-test


You should ALWAYS ALWAYS use the two-sample unequal variance $t$-test.

1.1.3 Warton et al. 2011. The arcsine is asinine.


About the only sensible transformation in ecology is the logarithmic transformation. Many older texts still recommend arcsine or sqrt transformations. This is an old fashioned approach to dealing with the analysis of proportions or small counts. There are better statistical methods available (logistic and Poisson regression and ANOVA). These are covered in later chapters in these notes.

1.1.4 Iannidia, 2005. Why most published research findings are false.


This is scary stuff – it turns out that most published research finding are false positives and that is why so few of them can be replicated.

A nice non-technical summary of the issues is found in an article from the Economist on how science gets it wrong at: [http://www.economist.com/news/briefing/21588057-scientists-think-science-s]
1.2 Introduction

To many students, statistics is synonymous with sadistics and is not a subject that is “enjoyable”. Obviously, I think this view is mistaken and hope to present some of the interesting things that can be done with statistics.

Statistics is all about discovery - how to extract information in the face of uncertainty. In the past, learning about statistics was tedious because of the enormous amount of arithmetic that needed to be done. Now, we let the computer do the heavy lifting, but it now vitally important that you UNDERSTAND what a computer package is doing – after all, these computer packages don’t have a conceptual understanding of what the data are about. They will quite happily compute the average sex (where 0 codes males and 1 codes females) – only you can decide that this is a meaningless statistic to compute.

These note try to operate at a conceptual level. There are many example which show how a typical analysis might be performed using a statistical package. There is often no unique answer to a problem with several good alternatives always available, so don’t let my notes constrain your thinking.

Statistics is fun! Just ask my family:

1.3 Effective note taking strategies

This section is taken from: The Tomorrow’s Professor Listserv - an email list server on topics of general interest to higher education. It is available at Stanford University at [http://sll.stanford.edu/](http://sll.stanford.edu/).
It will soon become apparent, if it hasn’t already, that not all lectures are fascinating and stimulating, and that not all lecturers are born with a gift for public speaking.

However, the information and ideas that they are trying to impart are just as important, and any notes that you take in the lectures must be understandable to you, not only five minutes after the lecture has finished, but in several months’ time, when you come to revise from them. The question, then, is how to retain your concentration and produce a good set of notes.

There are a few misconceptions on the part of students as to what can be expected of a lecture session. Firstly, that the responsibility for the success of the lecture is entirely the instructor’s, and that the student’s role is to sit and listen or to take verbatim notes. Secondly, that the purpose of a lecture is to impart information which will be needed for an exam question. And thirdly, that attending the lecture, and taking notes, is an individual, even competitive, activity.

This page aims to correct these ideas, and to help you develop successful note-taking strategies.

BEFORE THE LECTURE

- If you know the subject of the lecture, do some background reading beforehand. This way, you will go into the lecture with a better understanding, and find it easier to distinguish the points worth noting.
- Read through the notes of the previous lecture in the series just before the present one begins. This helps orient your thoughts to the subject in hand, especially if you have just come from a lecture on a completely different subject.

DURING THE LECTURE

- Think of a lecture as an active, learning process, rather than a passive, secretarial exercise. Writing verbatim what the lecturer says, or copying everything down from overheads, does not involve much thought, and subsequent reading of these notes often makes little sense.
- Pages of continuous prose are the least helpful for revision. Some things said in the lecture are obviously more important than others, and the notes you take should reflect this. Try to give them some structure, by using headings and sub-headings, by HIGHLIGHTING or underlining key ideas and realizing the links between them. Alternative noting forms to linear notes, such as flow diagrams or star charts, can be used, although these are often more helpful to revise your notes (see After the Lecture).
- In some situations, you may be directed in the amount of note-taking necessary. For example, the lecturer may start off by giving you some references on the subject he/she is to lecture on. A good strategy to adopt in this case would be to note down carefully the references, then just LISTEN to the lecture, making brief notes about the main points or specific examples. Taking notes from books is far easier than in lectures, as you can go at your own pace, stop and think about something, re-read a section etc. Use the lecture to try and understand the concepts being explained.
- Or, the lecturer may give hand-outs to accompany the lecture. In this case, you don’t need to make copious notes of your own. Again, listen to what is being said, and annotate the hand-out with any extra information. It gives you more time to think, and perhaps raise questions of your own.
- On the subject of questions, it is commonly believed by students that lecturers are not to be interrupted when they are in full flow. You may find that this isn’t always the case, and there is nothing wrong with asking individual lecturers if they mind taking questions during the lecture. It is best to establish this at the beginning of the course of lectures.
• However, there is also the problem of speaking out in front of your peers, perhaps asking something foolish, or not having the time to frame your question well. In this case, write down the question in the margin of your notes, to ask the lecturer later, or check with friends or in a textbook. It is far easier to recall the question you wanted to ask in this way, rather than rely on remembering after the lecture has finished (or even when you come to revise from your notes!)

AFTER THE LECTURE

• The best time to review your lecture notes is immediately following the lecture, although this is not always possible if, for example, you have to go straight to another one. However, the sooner you do it, with the lecture still fresh in your mind, the better chance you have to produce a good set of notes for revision.

• Revising your notes does not mean writing them out neatly!

• Try swapping notes with a friend, to check the accuracy/omissions of your own, and your understanding of the key points.

• If you feel that your notes are incomplete, or if you jotted down any questions during the lecture, follow this up by asking your tutor, or by reading round the subject.

• Transforming your lecture notes by using a different noting form can sometimes make them clearer, e.g. a flow diagram

• Highlight key points; produce summaries for revision purposes.

• Think how this topic relates to previous ones, or to other courses that you are studying, and begin to recognize themes and relationships.

• Meet with a few friends after lectures, to discuss the lecture topic and answer each others questions. Discussion with your peers often leads to a better understanding of a subject, which in turn makes it easier to remember. Your group could also establish a reading syndicate, whereby reading lists can be divided between members, who each take notes on their allotted texts and give copies to the rest of the group.

STORING YOUR NOTES

• A little time spent at this stage in organizing your notes will make life much easier when you come to revise from them some months later.

• Numbering pages, making a contents page, or using dividers in your file will all make your notes more accessible.

1.4 It’s all $\Gamma$reek to me

There are several common Greek letters and special symbols that are used in Statistics. In this section, we illustrate the common Greek letters and notation used.

Check that the following symbols and small equations.

• $\alpha$ - the Greek letter alpha
1.5 Which computer package?

Modern Statistics relies heavily upon computing – many would say that many modern statistical methods would be infeasible without modern software. Rather than spending time on tedious arithmetic, or on trying to reinvent the wheel, many people rely upon modern statistical packages.

Here are some of the most common packages in use today.

• **SAS** ([http://www.sas.com](http://www.sas.com)). Available in Windoze and Unix flavours. Modern Macintoshes with the Intel chip can run SAS under Windoze.

  One of the best packages around. SAS can handle nearly any type of data (dates, times, characters, numbers) with many possible analyses (over 100 different base analyses are currently available) and allows virtually arbitrary input formats and structures. SAS is extremely flexible and powerful but has a very steep learning curve. This is the premier statistical procedure – virtually all statistical analyses can be done with SAS. This is the package that I, as a Professional Statistician, use regularly in my job.

  Not only does SAS have modern statistical procedures, but is also a premiere database management system. It is designed for heavy duty computing.

  The SAS program includes a module SAS/INSIGHT that is virtually identical to **JMP** (see below).

• **SPSS** ([http://www.spss.com](http://www.spss.com)). This is a fairly powerful package (but not nearly as broad as SAS). It is very popular with Social Sciences researchers, but I personally, prefer SAS.

---

1 The Statistical Society of Canada has undertaken a program to accredit statisticians in Canada. Visit [http://www.ssc.ca](http://www.ssc.ca) for more details. Yours truly proudly bears the title P.Stat. 007.
• Minitab (http://www.minitab.com). This is one of the first user-friendly interactive statistical packages produced. Unfortunately, it is showing its age and relies on an awkward combination of graphical and text inputs. A fairly comprehensive set of routines, but most people will prefer JMP.

• JMP (http://www.jmp.com). JMP was originally developed by one of the two SAS developers (who were the 68 and 138 richest people in USA/NA in 2003). John Saul developed JMP. He did this originally for the MacIntosh platform and called it John’s MacIntosh Product ergo the name JMP.

JMP runs on Macintosh, Linux, and Windoze platforms.

JMP is easy to use and fairly powerful package. You should be able to do most things in this course in JMP.

JMP does not have the range of procedures as in SAS, nor can it deal with as complex data structures. However, my guestimation is that most people can do 80% of their statistical computing using JMP.

• SYSTAT (http://www.systat.com). This package has good graphical procedures, a fairly wide range of statistical procedures, but the package is showing its age. I find SYSTAT clumsy compared to using JMP and SAS and everytime I use it, I quickly get frustrated by its limitations and clumsy operations.

A review of SYSTAT is available in

American Statistician, 62, 177-178
http://dx.doi.org/10.1198/000313008x299339

• STATA (http://www.stata.com). I have never used STATA but a nice review of the package is found in


According to this review, Stata would be of interest to biostatisticians, mediacal/health outcomes, econometric, and social science research.

• R (http://www.r-project.org). R is based on the S-programing language. R can be freely downloaded from the WWW. It does not have the nice graphical interface.

This package is commonly used by statisticians when developing new statistical procedures. It is very flexible, but require a somewhat steep learning curve.

R is free, but not cheap!

• Excel. This is the standard spreadsheet program in the MSOffice Suite. Excel comes with some basic statistics but nothing too fancy. While EXCEL has its uses, you will find quickly that it can’t handly more complex analysis situations and gives wrong results without warning!

Except for very simple statistics, I RECOMMEND AGAINST the use of Excel to do statistical analyses.

People are wedded to Excel for often spurious reasons:

– "Its free." So is R and you get a much superior product.
– "It is easy to use". Yes, and easy to get WRONG answers without warnings.
– "It has good graphs". Excel has the largest selection of BAD graphs in the world. Hardly any of them are useful!

The following articles discuss some of the problems with Excel. They can also be accessed directly from the web by clicking on their respective links.

- Yet more problems with Excel are discussed in: *Practical Stat with Excel?* available at [http://www.practicalstats.com/xlsstats/excelstats.html](http://www.practicalstats.com/xlsstats/excelstats.html) and a copy of which is included in these notes.


- An article by the Statistical Consulting Service at the University of Reading has a brief discussion of the pros and cons of using Excel or analyzing data at [http://www.rdg.ac.uk/ssc/software/excel/home.html](http://www.rdg.ac.uk/ssc/software/excel/home.html). Basically, the graphs presented by Excel are often inappropriate for data presentation and you quickly run into limitations of the analysis routines available.

- How to use the basic functions of Excel for Statistics. This page also has a link to discussion about regression in Excel. *Using Excel functions in Statistics* available at [http://physicsnt.clemson.edu/chrismo/tutorials/excel/stats.html](http://physicsnt.clemson.edu/chrismo/tutorials/excel/stats.html).


There are a number of “add ons” available for Excel that seem to be reasonable priced and extend the analyses available.

Nevertheless, the algorithms used in Excel to do the actual computations are flawed and can give INCORRECT results without any warning that something has gone wrong!

For this reason, I generally use Excel only for simple problems - for anything more complex than a simple mean, I reach for a package such a *JMP* or *SAS*.

**Friends don’t let friends do statistics in Excel!**
Practical Stats

Make sense of your data!

- Home
- Consulting
- Upcoming Classes
- Applied Environmental Stats
- Nondetects (NADA)
- Multivariate Relationships
- Newsletter
- Downloads
- Statistics With Excel?
- Which Test Do I Use?
- Who Else Has Taken These?
- Contact Us

Is Microsoft Excel an Adequate Statistics Package?
It depends on what you want to do, but for many tasks, the answer is ‘No’.

Excel is available to many people as part of Microsoft Office. It contains some statistical functions in its basic installation. It also comes with statistical routines in the Data Analysis Toolpak, an add-in found separately on the Office CD. You must install the Toolpak from the CD in order to get these routines on the Tools menu. Once installed, these routines are at the bottom of the Tools menu, in the "Data Analysis" command. People use Excel as their everyday statistics software because they have already purchased it. Excel’s limitations, and occasionally its errors, make this a problem. Below are some of the concerns with using Excel for statistics that are recorded in journals, on the web, and from personal experience.

Limitations of Excel

1. Many statistical methods are not available in Excel.
Excel's biggest problem. Commonly-used statistics and methods NOT available within Excel include:
   * Boxplots
   * p-values for the correlation coefficient
   * Spearman’s and Kendall’s rank correlation coefficients
   * 2-way ANOVA with unequal sample sizes (unbalanced data)
   * Multiple comparison tests (post-hoc tests following ANOVA)
* p-values for two-way ANOVA
* Levene’s test for equal variance
* Nonparametric tests, including rank-sum and Kruskal-Wallis
* Probability plots
* Scatterplot arrays or brushing
* Principal components or other multivariate methods
* GLM (generalized linear models)
* Survival analysis methods
* Regression diagnostics, such as Mallow’s Cp and PRESS (it does compute adjusted r-squared)
* Durbin-Watson test for serial correlation
* LOESS smooths

Excel's lack of functionality makes it difficult to use for more than computing summary statistics and simple univariate regression. Third-party add-ins to Excel attempt to compensate for these limitations, adding new functionality to the program (see "A Partial Solution", below).

2. Several Excel procedures are misleading.
Probability plots are a standard way of judging the adequacy of the normality assumption in regression. In statistics packages, residuals from the regression are easily, or in some cases automatically, plotted on a normal probability plot. Excel’s regression routine provides a Normal Probability Plot option. However, it produces a probability plot of the Y variable, not of the residuals, as would be expected.

Excel’s CONFIDENCE function computes z intervals using 1.96 for a 95% interval. This is valid only if the population variance is known, which is never true for experimental data. Confidence intervals computed using this function on sample data will be too small. A t-interval should be used instead.

Excel is inconsistent in the type of P-values it returns. For most functions of probabilities, Excel acts like a lookup table in a textbook, and returns one-sided p-values. But in the TINV function, Excel returns a 2-sided p-value. Look carefully at the documentation of any Excel function you use, to be certain you are getting what you want.

Tables of standard distributions such as the normal and t distributions return p-values for tests, or are used to confidence intervals. With Excel, the user must be careful about what is being returned. To compute a 95% t confidence interval around the mean, for example, the standard method is to look up the t-statistic in a textbook by entering the table at a value of alpha/2, or 0.025. This t-statistic is multiplied by the standard error to produce the length of the t-interval on each side of the mean. Half of the error (alpha/2) falls on each side of the mean. In Excel the TINV function is entered using the value of alpha, not alpha/2, to return the same number.

For a one-sided t interval at alpha=0.05, standard practice would be to look up the t-statistic in a textbook for alpha=0.05. In Excel, the TINV function must be called using a value of 2*alpha, or 0.10, to get the value for alpha=0.05. This nonstandard entry point has led several reviewers to state that Excel’s distribution functions are incorrect. If not incorrect, they are certainly nonstandard. Make sure you read the help menu descriptions carefully to know what each function produces.

3. Distributions are not computed with precision.
NEW In reference (1), the authors show that all problems found in Excel 97 are still there in Excel 2000 and
XP. They say that "Microsoft attempted to fix errors in the standard normal random number generator and the inverse normal function, and in the former case actually made the problem worse." From this, you can assume that the problems listed below are still there in the current versions of the software.

Statistical distributions used by Excel do not agree with better algorithms for those distributions at the third digit and beyond. So they are approximately correct, but not as exact as would be desired by an exacting statistician. This may not be harmful for hypothesis tests unless the third digit is of concern (a p-value of 0.056 versus 0.057). It is of most concern when constructing intervals (multiplying a std dev of 35 times 1.96 give 68.6; times 1.97 gives 69.0) As summarized in reference 2:

"...the statistical distributions of Excel already have been assessed by Knusel (1998), to which we refer the interested reader. He found numerous defects in the various algorithms used to compute several distributions, including the Normal, Chi-square, F and t, and summarized his results concisely: So one has to warn statisticians against using Excel functions for scientific purposes. The performance of Excel in this area can be judged unsatisfactory."

4. Routines for handling missing data were incorrect.
This was the largest error in Excel, but a 'band-aid' has been added in Office 2000. In earlier versions of Excel, computations and tests were flat out wrong when some of the data cells contained missing values, even for simple summary statistics. See (3), (5), and page 4 of (6). Error messages are now displayed in Excel 2000 when there are missing values, and no result is given. Although this is still inferior to computing correct results it is somewhat of an improvement.

In reference to pre-2000,
"Excel does not calculate the paired t-test correctly when some observations have one of the measurements but not the other." E. Goldwater, ref. (5)

5. Regression routines are incorrect for multicollinear data.
This affects multiple regression. A good statistics package will report errors due to correlations among the X variables. The Variance Inflation Factor (VIF) is one measure of collinearity. Excel does not compute collinearity measures, does not warn the user when collinearity is present, and reports parameter estimates that may be nonsensical. See (6) for an example on data from an experiment. Are multicollinear data of ‘practical’ problems? I think so -- I find many examples of collinearity in environmental data sets.

Excel also requires the X variables to be in contiguous columns in order to input them to the procedure. This can be done with cut and paste, but is certainly annoying if many multiple regression models are to be built.

6. Ranks of tied data are computed incorrectly.
When ranking data, standard practice is to assign tied ranks to tied observations. The value of these ranks should equal the median of the ranks that the observations would have had, if they had not been tied. For example, three observations tied at a value of 14 would have had the ranks of 7, 8 and 9 had they not been tied. Each of the three values should be assigned the rank of 8, the median of 7, 8 and 9.

Excel assigns the lowest of the three ranks to all three observations, giving each a rank of 7. This would result in problems if Excel computed rank-based tests. Perhaps it is fortunate none are available.

7. Many of Excel's charts violate standards of good graphics.
Use of perspective and glitz (donut charts?) violate basic principles of graphics. Excel's charts are more suitable to USA Today than to scientific reports. This bothers some people more than others.

"Good graphs should...[a list of traits]...However, Excel meets virtually none of these criteria. The vast majority of chart types produced by Excel should never be used!" -- Jon Cryer, ref (3).

"Microsoft Excel is an example of a package that does not allow enough user control to consistently make readable and concise graphs from tables."

A partial solution:
Some of these difficulties (parts of 1,2, 6 and 7) can be overcome by using a good set of add-in routines. One of the best is StatPlus, which comes with an excellent textbook, "Data Analysis with Microsoft Excel". With StatPlus, Excel becomes an adequate statistical tool, though still not in the areas of multiple regression and ANOVA for more than one factor. Without this add-in Excel is inadequate for anything beyond basic summary statistics and simple regression.
Data Analysis with Microsoft Excel by Berk and Carey published by Duxbury (2000).

Opinion: Get this book if you're going to use Excel for statistics.
(I have no connection with the authors of StatPlus and get no benefit from this recommendation. I'm just a satisfied user.)

Some advice from others:
"If you need to perform analysis of variance, avoid using Excel, unless you are dealing with extremely simple problems."
- Statistical Services Centre, Univ. of Reading, U.K. (at A, below)

"Enterprises should advise their scientists and professional statisticians not to use Microsoft Excel for substantive statistical analysis. Instead, enterprises should look for professional statistical analysis software certified to pass the (NIST) Statistical Reference Datasets tests to their users' required level of accuracy."
- The Gartner Group

References:
1) On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP

(2) On the accuracy of statistical procedures in Microsoft Excel '97

(3) Problems with using Microsoft Excel for statistics [pdf Download]

(4) Use of Excel for statistical analysis
Neil Cox, (2000), AgResearch Ruakura
(5) Using Excel for statistical data analysis  
Eva Goldwater, (1999), Univ. of Massachusetts Office of Information Technology  
[pdf download]

(6) Statistical analysis using Microsoft Excel [pdf download]  
Jeffrey Simonoff, (2002)  
[pdf download]

(7) Spreadsheet addiction  
Patrick Burns

(8) On the Accuracy of Statistical Distributions in Microsoft Excel 97  
Leo Knuesel  
[pdf download]

(9) Statistical flaws in Excel  
Hans Pottel  
[pdf download]

Guides to Excel on the web:
(A) A Beginner's Guide to Excel - Univ. of Reading, UK

(B) An Intermediate Guide to Excel - Univ. of Reading, UK

Note: All opinions other than those cited as coming from others are my own.

- Home >
- Statistics With Excel? >

© 2007 Practical Stats Email Us
1.6 FAQ - Frequently Asked Question

1.6.1 Accessing journal articles from home

I make reference to several journal articles in these notes. These are often available in e-journals so the link should take you there directly if you are authorized to access this journal. For example, if you try and access these articles from a computer with an SFU IP address, you should likely be granted permission without problems.

However, if you are trying to access these from home, you must go through the SFU library site and access the e-journal via the catalogue. You will then be prompted to enter your SFU ACS userid and password to grant you access to this journal.

1.6.2 Downloading from the web

When ever I try and download an Excel file, it seems to be corrupted and can’t be opened.

Through out the notes, reference is made to spreadsheets or SAS programs available at my web site.

The SAS programs and listings are simple text files and should transfer to your computer without much problem.

If you trying to download an Excel spreadsheet, be sure to specify that the file should be transfered as a source document rather than as text. If you transfer the sheets as text, you will find that the data are corrupted.

1.6.3 Printing 2 pages per physical page and on both sides of the paper

The notes look as if I could print 2 per page. Is this possible, and can I print on both sides of the paper?

Yes, it is possible to print two logical pages per physical page - the text is a bit small, but still readable.

On a Macintosh System with a recent OS, when you select Print, it presents the standard print options menu. Under the popdown menu is a Layout option. Select 2 logical pages per physical page. This will work with ALL applications that use the standard print dialogue.

I’m not familiar enough with Windoze machines to offer any advice.

To print on both sides of the paper, you need a printer capable of duplex printing, i.e. on both sides of the paper. I believe that most printers in the public areas of campus are capable of this. You will have to consult your own printer manual if you are printing at home. Otherwise, you have to print first the odd pages, then take the paper, reverse it, and print the even pages - a recipe for disaster.

1.6.4 Is there an on-line textbook?

Are there any online textbooks in statistics?
CHAPTER 1. IN THE BEGINNING...

Yes, there are several - it is easiest to search the web using google. Beware that some of the advice on the web may be less than perfect.