7.1: Using Spreadsheets for Statistics

Learning Objectives

- You can do most, maybe all of your statistics using a spreadsheet such as Excel. Here are some general tips.

Introduction

If you're like most biologists, you can do all of your statistics with spreadsheets such as Excel. You may spend months getting the most technologically sophisticated new biological techniques to work, but in the end you'll be able to analyze your data with a simple chi-squared test, t–test, one-way anova or linear regression. The graphing abilities of spreadsheets make it easy to inspect data for errors and outliers, look for non-linear relationships and non-normal distributions, and display your final results. Even if you're going to use something like SAS or SPSS or R, there will be many times when it's easier to enter your data into a spreadsheet first, inspect it for errors, sort and arrange it, then export it into a format suitable for your fancy-schmancy statistics package.

Some statisticians are contemptuous of Excel for statistics. One of their main complaints is that it can't do more sophisticated tests. While it is true that you can't do advanced statistics with Excel, that doesn't make it wrong to use it for simple statistics; that Excel can't do principal components analysis doesn't make its answer for a two-sample t–test incorrect. If you are in a field that requires complicated multivariate analyses, such as epidemiology or ecology, you will definitely have to use something more advanced than spreadsheets. But if you are doing well designed, simple laboratory experiments, you may be able to analyze all of your data with the kinds of tests you can do in spreadsheets.

The more serious complaint about Excel is that some of the procedures gave incorrect results (McCullough and Heiser 2008, Yalta 2008). Most of these problems were with procedures more advanced than those covered in this handbook, such as exponential smoothing, or were errors in how Excel analyzes very unusual data sets that you're unlikely to get from a real experiment. After years of complaining, Microsoft finally fixed many of the problems in Excel 2010 (Keeling and Pavur 2011). So for the statistical tests I describe in this handbook, I feel confident that you can use Excel and get accurate results.

A free alternative to Excel is Calc, part of the free, open-source OpenOffice.org package. Calc does almost everything that Excel does, with just enough exceptions to be annoying. Calc will open Excel files and can save files in Excel format. The OpenOffice.org package is available for Windows, Mac, and Linux. OpenOffice.org also includes a word processor (like Word) and presentation software (like PowerPoint).

Gnumeric sounds like a good, free, open-source spreadsheet program; while it is primarily used by Linux users, it can be made to work with Mac. I haven't used it, so I don't know how well my spreadsheets will work with it.

The instructions on this web page apply to both Excel and Calc, unless otherwise noted.

Basic spreadsheet tasks

I'm going to assume you know how to enter data into a spreadsheet, copy and paste, insert and delete rows and columns, and other simple tasks. If you're a complete beginner, you may want to look at tutorials on using Excel here or here. Here are a few other things that will be useful for handling data.

Separate text into columns

Excel

When you copy columns of data from a web page or text document, then paste them into an Excel spreadsheet, all the data will be in one column. To put the data into multiple columns, select the cells you want to convert, then choose "Text to columns..." from the Data menu. If you choose "Delimited," you can tell it that the columns are separated by spaces, commas, or some other character. Check the "Treat consecutive delimiters as one" box (in Excel) or the "Merge Delimiters" box (in Calc) if numbers may be separated by more than one space, more than one tab, etc. The data will be entered into the columns to the right of the original column, so make sure they're empty.

If you choose "Fixed width" instead of "Delimited", you can do things like tell it that the first 10 characters go in column 1, the next 7 characters go in column 2, and so on.

If you paste more text into the same Excel spreadsheet, it will automatically be separated into columns using the same delimiters. If you want to turn this off, select the column where you want to paste the data, choose "Text to columns..." from the Data menu, and choose "Delimited." Then unclick all the boxes for delimiters (spaces, commas, etc.) and click "Finish." Now paste your data into the column.

Series fill

You'll mainly use this for numbering a bunch of rows or columns. Numbering them will help you keep track of which row is which, and it will be especially useful if you want to sort the data, then put them back in their original order later. Put the first number of your series in a cell and select it, then choose "Fill: Series..." from the Edit menu. Choose "Rows" or "Columns" depending on whether you want the series to be in a row or a column, set the "Step value" (the amount the series goes up by; usually you'll use 1) and the "Stop value" (the last number in the series). So if you had a bunch of data in cells $B2$ through $E101$ and you wanted to number the rows, you'd put a 1 in cell $A2$, choose "Columns", set the "Step value" to 1 and the "Stop value" to 100, and the numbers 1 through 100 would be entered in cells $A2$ through $A101$.

Sorting

To sort a bunch of data, select the cells and choose "Sort" from the Data menu. If the first row of your data set has column headers identifying what is in each column, click on "My list has headers." You can sort by multiple columns; for example, you could sort data on a bunch of chickens by "Breed" in column A, "Sex" in column C, and "Weight" in column B, and it would sort the data by breeds, then within each breed have all the females first and then all the males, and within each breed/sex combination the chickens would be listed from smallest to largest.

If you've entered a bunch of data, it's a good idea to sort each column of numbers and look at the smallest and largest values. This may help you spot numbers with misplaced decimal points and other egregious typing errors, as they'll be much larger or much smaller than the correct numbers.

Graphing

See the web page on graphing with Excel. Drawing some quick graphs is another good way to check your data for weirdness. For example, if you've entered the height and leg length of a bunch of people, draw a quick graph with height on the X axis and leg length on the Y axis. The two variables should be pretty tightly correlated, so if you see some outlier who's 2.10 meters tall and has a leg that's only 0.65 meters long, you know to double-check the data for that person.

Absolute and relative cell references

In the formula "$=B1+C1$", $B1$ and $C1$ are relative cell references. If this formula is in cell $D1$, "$B1$" means "that cell that is two cells to the left." When you copy cell $D1$ into cell $D2$, the formula becomes "$=B2+C2$"; when you copy it into cell $G1$, it would become "$=E1+F1$". This is a great thing about spreadsheets; for example, if you have long columns of numbers in columns A and B and you want to know the sum of each pair, you don't need to type "$=B1+C1$" into cell $D1$, then type "$=B2+C2$" into cell $D2$, then type "$=B3+C3$" into cell $D3$, and so on; you just type "$=B1+C1$" once into cell $D1$, then copy and paste it into all the cells in column D at once.

Sometimes you don't want the cell references to change when you copy a cell; in that case, you should use absolute cell references, indicated with a dollar sign. A dollar sign before the letter means the column won't change when you copy and paste into a different cell. If you enter "$=\$B1+C1$" into cell $D1$, then copy it into cell $E1$, it will change to "$=\$B1+D1$"; the $C1$ will change to $D1$ because you've copied it one column over, but the $B1$ won't change because it has a dollar sign in front of it. A dollar sign before the number means the row won't change; if you enter "$=B\$1+C1$" into cell $D1$ and then copy it to cell $D2$, it will change to "$=B\$1+C2$". And a dollar sign before both the column and the row means that nothing will change; if you enter "$=\$B\$1+C1$" into cell $D2$ and then copy it into cell $E2$, it will change to "$=\$B\$1+D2$". So if you had 100 numbers in column B, you could enter "$=B1-\text{AVERAGE}(B\$1:B\$100)$" in cell $C1$, copy it into cells $C2$ through $C100$, and each value in column B would have the average of the 100 numbers subtracted from it.

Paste Special

When a cell has a formula in it (such as "$=B1*C1+D1^2$"), you see the numerical result of the formula (such as "7.15") in the spreadsheet. If you copy and paste that cell, the formula will be pasted into the new cell; unless the formula only has absolute cell references, it will show a different numerical result. Even if you use only absolute cell references, the result of the formula will change every time you change the values in $B1$, $C1$ *or* $D1$. When you want to copy and paste the number that results from a

function in **Excel**, choose "Paste Special" from the Edit menu and then click the button that says "Values." The number (7.15, in this example) will be pasted into the cell.

In **Calc**, choose "Paste Special" from the Edit menu, uncheck the boxes labeled "Paste All" and "Formulas," and check the box labeled "Numbers."

Change number format

The default format in Excel and Calc displays 9 digits to the right of the decimal point, if the column is wide enough. For example, the P value corresponding to a chi-square of 4.50 with 1 degree of freedom, found with "=CHIDIST(4.50, 1)", will be displayed as 0.033894854 This number of digits is almost always ridiculous. To change the number of decimal places that are displayed in a cell, choose "Cells..." from the Format menu, then choose the "Number" tab. Under "Category," choose "Number" and tell it how many decimal places you want to display. For the P value above, you'd probably just need three digits, 0.034. Note that this only changes the way the number is displayed; all of the digits are still in the cell, they're just invisible.

The disadvantage of setting the "Number" format to a fixed number of digits is that very small numbers will be rounded to 0. Thus if you set the format to three digits to the right of the decimal, "=CHIDIST(24.50,1)" will display as "0.000" when it's really 0.00000074 The default format ("General" format) automatically uses scientific notation for very small or large numbers, and will display $7.4309837243E-007$, which means 7.43×10^{-7}; that's better than just rounding to 0, but still has way too many digits. If you see a 0 in a spreadsheet where you expect a non-zero number (such as a P value), change the format to back to General.

For P values and other results in the spreadsheets linked to this handbook, I created a user-defined format that uses 6 digits right of the decimal point for larger numbers, and scientific notation for smaller numbers. I did this by choosing "Cells" from the Format menu and pasting the following into the box labeled "Format code":

$$[>0.00001]0.######;[<-0.00001]0.######;0.00E-00 \tag{7.1.1}$$

This will display 0 as $0.00E00$, but otherwise it works pretty well.

If a column is too narrow to display a number in the specified format, digits to the right of the decimal point will be rounded. If there are too many digits to the left of the decimal point to display them all, the cell will contain "###". Make sure your columns are wide enough to display all your numbers.

Useful spreadsheet functions

There are hundreds of functions in Excel and Calc; here are the ones that I find most useful for statistics and general data handling. Note that where the argument (the part in parentheses) of a function is "Y", it means a single number or a single cell in the spreadsheet. Where the argument says "Ys", it means more than one number or cell. See AVERAGE(Ys) for an example.

All of the examples here are given in Excel format. Calc uses a semicolon instead of a comma to separate multiple parameters; for example, Excel would use "=ROUND(A1, 2)" to return the value in cell $A1$ rounded to 2 decimal places, while Calc would use "=ROUND(A1; 2)". If you import an Excel file into Calc or export a Calc file to Excel format, Calc automatically converts between commas and semicolons. However, if you type a formula into Calc with a comma instead of a semicolon, Calc acts like it has no idea what you're talking about; all it says is "#NAME?".

I've typed the function names in all capital letters to make them stand out, but you can use lower case letters.

Math functions

ABS(Y) Returns the absolute value of a number.

EXP(Y) Returns e to the y^{th} power. This is the inverse of LN, meaning that "=EXP(LN(Y))" equals Y.

LN(Y) Returns the natural logarithm (logarithm to the base e) of Y.

LOG10(Y) Returns the base-10 logarithm of Y. The inverse of LOG is raising 10 to the Y^{th} power, meaning "=10^(LOG10(Y))" returns Y.

RAND() Returns a pseudorandom number, equal to or greater than zero and less than one. You must use empty parentheses so the spreadsheet knows that RAND is a function. For a pseudorandom number in some other range, just multiply; thus "=RAND()*79" would give you a number greater than or equal to 0 and less than 79. The value will change every time you enter something in any cell. One use of random numbers is for randomly assigning individuals to different treatments; you could enter "=RAND()" next to

each individual, Copy and Paste Special the random numbers, Sort the individuals based on the column of random numbers, then assign the first 10 individuals to the placebo, the next 10 individuals to $10mg$ of the trial drug, etc.

A "pseudorandom" number is generated by a mathematical function; if you started with the same starting number (the "seed"), you'd get the same series of numbers. Excel's pseudorandom number generator bases its seed on the time given by the computer's internal clock, so you won't get the same seed twice. There are problems with Excel's pseudorandom number generator that make it inappropriate for serious Monte Carlo simulations, but the numbers it produces are random enough for anything you're likely to do as an experimental biologist.

ROUND(Y,digits) Returns Y rounded to the specified number of digits. For example, if cell $A1$ contains the number 37.38, "=ROUND(A1, 1)" returns 37.4, "=ROUND(A1, 0)" returns 37, and "=ROUND(A1, -1)" returns 40. Numbers ending in 5 are rounded up (away from zero), so "=ROUND(37.35,1)" returns 37.4 and "=ROUND(-37.35)" returns -37.4.

SQRT(Y) Returns the square root of Y.

SUM(Ys) Returns the sum of a set of numbers.

Logical functions

AND(logical_test1, logical_test2,...) Returns TRUE if logical_test1, logical_test2... are all true, otherwise returns FALSE. As an example, let's say that cells $A1$, $B1$ and $C1$ all contain numbers, and you want to know whether they're all greater than 100. One way to find out would be with the statement "=AND(A1>100, B1>100, C1>100)", which would return TRUE if all three were greater than 100 and FALSE if any one were not greater than 100.

IF(logical_test, A, B) Returns A if the logical test is true, B if it is false. As an example, let's say you have 1000 rows of data in columns A through E, with a unique ID number in column A, and you want to check for duplicates. Sort the data by column A, so if there are any duplicate ID numbers, they'll be adjacent. Then in cell $F1$, enter "=IF(A1=A2, "duplicate","ok"). This will enter the word "duplicate" if the number in $A1$ equals the number in $A2$; otherwise, it will enter the word "ok". Then copy this into cells $F2$ through $F999$. Now you can quickly scan through the rows and see where the duplicates are.

ISNUMBER(Y) Returns TRUE if Y is a number, otherwise returns FALSE. This can be useful for identifying cells with missing values. If you want to check the values in cells $A1$ to $A1000$ for missing data, you could enter "=IF(ISNUMBER(A1), "OK", "MISSING")" into cell $B1$, copy it into cells $B2$ to $B1000$, and then every cell in $A1$ that didn't contain a number would have "MISSING" next to it in column B.

OR(logical_test1, logical_test2,...) Returns TRUE if one or more of logical_test1, logical_test2... are true, otherwise returns FALSE. As an example, let's say that cells $A1$, $B1$ and $C1$ all contain numbers, and you want to know whether any is greater than 100. One way to find out would be with the statement "=OR(A1>100, B1>100, C1>100)", which would return TRUE if one or more were greater than 100 and FALSE if all three were not greater than 100.

Statistical functions

AVERAGE(Ys) Returns the arithmetic mean of a set of numbers. For example, "=AVERAGE(B1..B17)" would give the mean of the numbers in cells $B1..B17$, and "=AVERAGE(7, A1, B1..C17)" would give the mean of 7, the number in cell $A1$, and the numbers in the cells $B1..C17$. Note that Excel only counts those cells that have numbers in them; you could enter "=AVERAGE(A1:A100)", put numbers in cells $A1$ to $A9$, and Excel would correctly compute the arithmetic mean of those 9 numbers. This is true for other functions that operate on a range of cells.

BINOMDIST(S, K, P, cumulative_probability) Returns the binomial probability of getting S "successes" in K trials, under the null hypothesis that the probability of a success is P. The argument "cumulative_probability" should be TRUE if you want the cumulative probability of getting S or fewer successes, while it should be FALSE if you want the probability of getting exactly S successes. (**Calc** uses 1 and 0 instead of TRUE and FALSE.) This has been renamed "BINOM.DIST" in newer versions of Excel, but you can still use "BINOMDIST".

CHIDIST(Y, df) Returns the probability associated with a variable, Y, that is chi-square distributed with df degrees of freedom. If you use SAS or some other program and it gives the result as "Chi-sq=78.34, 1 d.f., P<0.0001", you can use the CHIDIST function to figure out just how small your P value is; in this case, "=CHIDIST(78.34, 1)" yields 8.67×10^{-19}. This has been renamed CHISQ.DIST.RT in newer versions of Excel, but you can still use CHIDIST.

CONFIDENCE(alpha, standard-deviation, sample-size) Returns the confidence interval of a mean, *assuming you know the population standard deviation*. Because you don't know the population standard deviation, **you should never use this function**;

instead, see the web page on confidence intervals for instructions on how to calculate the confidence interval correctly.

COUNT(Ys) Counts the number of cells in a range that contain numbers; if you've entered data into cells $A1$ through $A9$, $A11$, and $A17$, "=COUNT(A1:A100)" will yield 11.

COUNTIF(Ys, criterion) Counts the number of cells in a range that meet the given criterion.
"=COUNTIF(D1:E1100,50)" would count the number of cells in the range $D1:E100$ that were equal to 50;
"=COUNTIF(D1:E1100,">50")" would count the number of cells that had numbers greater than 50 (note the quotation marks around ">50");
"=COUNTIF(D1:E1100,F3)" would count the number of cells that had the same contents as cell $F3$;
"=COUNTIF(D1:E1100,"Bob")" would count the number of cells that contained just the word "Bob". You can use wildcards; "?" stands for exactly one character, so "Bo?" would count "Bob" or "Boo" but not "Bobble", while "Bo*" would count "Bob", "Boo", "Bobble" or "Bodacious".

DEVSQ(Ys) Returns the sum of squares of deviations of data points from the mean. This is what statisticians refer to as the "sum of squares." I use this in setting up spreadsheets to do anova, but you'll probably never need this.

FDIST(Y, df1, df2) Returns the probability value associated with a variable, Y, that is F-distributed with $df1$ degrees of freedom in the numerator and $df2$ degrees of freedom in the denominator. If you use SAS or some other program and it gives the result as "F=78.34, 1, 19 d.f., P<0.0001", you can use the FDIST function to figure out just how small your P value is; in this case, "=FDIST(78.34, 1, 19)" yields 3.62×10^{-8}. Newer versions of Excel call this function F.DIST.RT, but you can still use FDIST.

MEDIAN(Ys) Returns the median of a set of numbers. If the sample size is even, this returns the mean of the two middle numbers.

MIN(Ys) Returns the minimum of a set of numbers. Useful for finding the range, which is MAX(Ys)-MIN(Ys).

MAX(Ys) Returns the maximum of a set of numbers.

NORMINV(probability, mean, standard_deviation) Returns the inverse of the normal distribution for a given mean and standard deviation. This is useful for creating a set of random numbers that are normally distributed, which you can use for simulations and teaching demonstrations; if you paste "=NORMINV(RAND(),5,1.5)" into a range of cells, you'll get a set of random numbers that are normally distributed with a mean of 5 and a standard deviation of 1.5.

RANK.AVG(X, Ys, type) Returns the rank of X in the set of Ys. If *type* is set to 0, the largest number has a rank of 1; if *type* is set to 1, the smallest number has a rank of 1. For example, if cells $A1:A8$ contain the numbers 10, 12, 14, 14, 16, 17, 20, 21 "=RANK(A2, A$1:A$8, 0)" returns 7 (the number 12 is the 7^{th} largest in that list), and "=RANK(A2, A$1:A$8, 1)" returns 2 (it's the 2nd smallest).

The function "RANK.AVG" gives average ranks to ties; for the above set of numbers, "=RANK.AVG(A3, A$1:A$8, 0)" would return 5.5, because the two values of 14 are tied for fifth largest. Older versions of Excel and Calc don't have RANK.AVG; they have RANK, which handled ties incorrectly for statistical purposes. If you're using Calc or an older version of Excel, this formula shows how to get ranks with ties handled correctly:

$$= \text{AVERAGE}(\text{RANK}(A1,\ A\$1:A\$8,\ 0),\ 1+\text{COUNT}(A\$1:A\$8)-\text{RANK}(A\$1,\ A\$1:A\$8,\ 1)) \quad (7.1.2)$$

STDEV(Ys) Returns an estimate of the standard deviation based on a population sample. This is the function you should use for standard deviation.

STDEVP(Ys) Returns the standard deviation of values from an entire population, not just a sample. **You should never use this function**.

SUM(Ys) Returns the sum of the Ys.

SUMSQ(Ys) Returns the sum of the squared values. Note that statisticians use "sum of squares" as a shorthand term for the sum of the squared deviations from the mean. SUMSQ does not give you the sum of squares in this statistical sense; for the statistical sum of squares, use DEVSQ. You will probably never use SUMSQ.

TDIST(Y, df, tails) Returns the probability value associated with a variable, Y, that is t-distributed with df degrees of freedom and *tails* equal to one or two (you'll almost always want the two-tailed test). If you use SAS or some other program and it gives the result as "t=78.34, 19 d.f., P<0.0001", you can use the TDIST function to figure out just how small your P value is; in this case, "=TDIST(78.34, 19, 2)" yields 2.55×10^{-25}. Newer versions of Excel have renamed this function T.DIST.2T, but you can still use TDIST.

VAR(Ys) Returns an estimate of the variance based on a population sample. This is the function you should use for variance.

VARP(Ys) Returns the variance of values from an entire population, not just a sample. **You should never use this function**.

References

Keeling, K.B., and R.J. Pavur. 2011. Statistical accuracy of spreadsheet software. American Statistician 65: 265-273.

McCullough, B.D., and D.A. Heiser. 2008. On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis 52: 4570-4578.

Yalta, A.T. 2008. The accuracy of statistical distributions in Microsoft Excel 2007. Computational Statistics and Data Analysis 52: 4579-4586.
