

Basic Statistics in Excel

Characterizing a sample

Open the file “mercury data.xls” from the class website. This contains the “data” we looked at in lecture.

Numeric characterization

The Excel functions for the arithmetic mean, the variance, and the standard deviation are AVERAGE, VAR, and STDEV, respectively.

A more thorough set of summary statistics are generated by “Descriptive Statistics” in the Data Analysis ToolPak. Select the input range and check “Summary Statistics”. You can also check “Confidence Level for Mean” and select a confidence level; this will give you the amount to add or subtract from the sample mean to generate the confidence interval for the mean based on the assumption of normality. Unfortunately, these are not dynamic formulas – they will not update if you put in new data.

You can get a dynamic calculation of the standard error of the mean by using the formula $=\text{STDEV}(A2:A31)/\text{SQRT}(30)$. More generically, you can use $=\text{STDEV}(A:A)/\text{SQRT}(\text{COUNT}(A:A))$ – the STDEV function ignores non-numeric values, and the COUNT function gives the number of cells with numbers in them.

To get a dynamic calculation of the normal-based confidence interval, we use the TINV function. This returns critical values of the t distribution; its arguments are the probability level (alpha) and the degrees of freedom. It gives the *two-sided* critical value – that is it returns a value t^* such that there is probability alpha that *either* $t < -t^*$ *or* $t > t^*$. Thus for a 95% confidence interval we would use an alpha of 0.05.

To get dynamic calculations of the most relevant quantities, set up your spreadsheet as shown in the picture below (*Excel tip: to toggle between showing values and formulas in your spreadsheet, use ctrl-` [backquote, the same key as tilde]*). Note that you can change the desired confidence levels by changing the value in cell D12; and that if you paste new data into column A all the calculations will be updated.

The screenshot shows an Excel spreadsheet titled "mercury data - work.xls". The data is organized as follows:

	A	B	C	D
1	Mercury concentration			
2	0.853511660795998		Sample mean	=AVERAGE(A:A)
3	0.391905706669006		Sample variance	=VAR(A:A)
4	0.143344302676524		Sample standard deviation	=STDEV(A:A)
5	0.198267856849027		Sample size	=COUNT(A:A)
6	0.266572366854502		Standard error of mean	=D4/SQRT(D5)
7	0.327306702032255		Minimum	=MIN(A:A)
8	0.834747834154153		Maximum	=MAX(A:A)
9	5.32261822012672		Range	=D8-D7
10	0.817037695737456		Confidence interval of mean	
11	0.157247166740851			
12	0.328456677061821		Confidence level (%)	95
13	3.79315352425584		Lower	=D2-D6*TINV(1-D12/100,D5-1)
14	0.513433214687504		Upper	=D2+D6*TINV(1-D12/100,D5-1)
15	0.502938252549826			
16	0.733454663222413			
17	0.279345254365329			
18	0.952473469852774			
19	0.742740502492076			
20	0.178309271400157			
21	0.469049645966509			

Graphical characterization

- Error bars:** In excel, set up your graph to plot the means as points. Double-click on the points to bring up the “Format data series” dialog, and select the “Y Error Bars” tab. Select “Both”, and “custom”. Then in the “+” and “-“ boxes select the cell that contains the values that describe the *length* of the error bars. For example, if your error bars represent +/- one standard error, put in the cell that contains the standard error of the mean.
- Box plot:** This can't be done in Exel. Well, I'm sure it could be, but it would be a lot of work!
- Histogram:** To use the “Histogram” tool in the analysis toolpak you need to first set up a column with the desired bin ranges. The tool then counts the number data points in each bin, and you can use the output to construct a histogram using the chart tool. Adjusting the bins is a lot of work!
- Dot plot:** I have created an excel spreadsheet that produces these (dotplot.xls, on the course webpage). Simply open this spreadsheet and paste your data according to the instructions at the top of the sheet.
- Empirical cumulative distribution function (ECDF):** First copy your data to a new sheet (in case the order of the values has meaning), and then sort it in ascending order. Assuming these values start in cell A1, enter “1” in cell B1, and =B1/(COUNT(A:A)+1) in cell C1. Fill columns B and C down to the same

length as your data. Then make a chart with column A as your X values and column C as your Y values.

Testing normality

Shapiro-Wilk test

The easiest way to do this is to download a set of Excel workbooks called the “Excel toolkit” from <http://esc.syrres.com/pracenter/productdownloads/exceltoolkit/exceltoolkitnodown.html>. One of these is set up so that all you do is paste your data and it calculates the test for you.

QQ Plot

Go back to the sheet where you were calculating the empirical CDF of the data. In cell D1, enter =NORMINV(C1,0,1). Fill this down. This formula calculates the values of z that are the same quantiles of the standard normal as the data are quantiles of the sample distribution. Then make a plot with column A as the Y data and column D as the X data. Note that if you plot column D against column C then you will get the CDF of the standard normal distribution.

If you want to ask whether the data are a sample from a lognormal distribution, this is equivalent to asking whether the log of the data come from a normal distribution. So log transform the data and perform the same tests as above on the transformed data.

Calculating Confidence Intervals

The formulas to calculate the confidence interval of the mean under the assumption that the population is normally distributed has already been described above under “Characterizing a sample”. We can also calculate confidence intervals for the variance: using the same setup that we had for Characterizing a sample, the lower end of the interval is

$$=(D5-1)*D3/CHIINV((1-D12/100)/2,D5-1)$$

and the upper end of the interval is

$$=(D5-1)*D3/CHIINV(1-(1-D12/100)/2,D5-1)$$

The complicated bit involving the confidence level (D12) is because we have to explicitly specify the two tails as $\alpha/2$ and $1-\alpha/2$.