Brief Intro to Stats in Excel

Carlos and Rob

1. Excel: Getting your data into Excel and Making a Histogram
2. Excel: Working with Subsets of Data
3. Excel: Tables with PivotTables

1. Excel: Getting your data into Excel and Making a Histogram

See also:
http://faculty.chicagobooth.edu/EandS/ExcelExamples.html

There are different versions of Excel and, for better or worse, the menus are different.

Usually, it is not *too* hard to find what you want in an unfamiliar version.

For these notes I used Excel 2010, 2013 is very similar.

Before you start playing around with data, make sure you have the *Analysis Tookpack* active.

- Go to /File/options/Add-ins.
- You will see a list of "Active Application Add-ins" and a list of "Inactive Application Addins".
- You need the *Analysis Toolpack* to be active. If it is you are done.
- If it is not active, click **Go** at the bottom of the screen (next to Manage: Excel Add-ins)

Now you need to download the data file and copy it to a folder (directory) where you want to work.

I'm going to use the folder
C:\Documents and Settings\rem\My Documents\bstat-play.

I right clicked in "My Documents" and then did /New/Folder, and then entered the name bstat-play (which I just made up).

Next down-load the data.

For example, down-load the file Default.csv from the course web page (go to www.rob-mcculloch.org and click on "Data Sets").

On my Windows machine when I clicked on Default.csv, it downloaded the file to \Downloads.

Then I copied the file from Downloads to bstat-play.

(e.g. right-click/copy in Downloads and then right-click/paste in bstat-play).

Ok, we are finally ready to get into Excel.

Default.csv is a *comma separated file*: the fields are separated by commas.

Generally speaking, this format works well with Excel (and R).

On my machine, I just double-clicked on Default.csv and Excel launches with the Default data.

You may have to:

- /File/Open
- navigate to the directory where the data file is (/bstat-play in our example).
- Play with "Files of Type" to make sure .csv files show up.
- Double click on the file (Default.csv in our example).

On some versions of Excel you may have to step through a few more menus but it is usually pretty easy to get a csv file into Excel.

Two things I like to do to make life simpler are

(i) split the screen.
(ii) name the variables.

To split the screen go to top-right part of the spread-sheet (just to the right of the letters identifying the columns) and click and drag down on the little - (dash).

You can also go to /view/window/split.

Then adjust the top half so you can see the top of the data and the bottom half so you can see the bottom of the data.

If we name the variables we can refer to all the numbers for a variable by the name rather than using the cell range ("balance" rather than c2:c10001).

To name the variables:

- click on the top-left cell then shift-click on the bottom-right cell to select all the cells. Alternatively, click on a cell in the data and the do Ctrl-a.
- Got to /Formulas/Defined Names/Name Manager/Create from Selection.
- Click "top row", and then Ok.

Now click the down-arrow on the name box on the left side just above the data cells (just to the left of f_x). You should see the variable names there.

Just for fun:

- go to an empty cell and type "=average(balance)" and hit return.
- put "=average(c2:c10001)" in another cell (and hit return).
- put "=average(c2:c10001)" in another cell (and hit return).

To (*finally*) do the histogram of the variable balance:

- Go to /Data/Data Analysis/Histogram "OK".
- Put the cells with the balance values in the Input Range. You can just type in the range (e.g. c1:c10001 or $C$1:$C$10001) or you can click on the little box with the picture of spreadsheet on it and use it to select the first and last cells in the balance column.
- Then click the "Labels" and "Chart Output" boxes.
- Click OK.

A new sheet will be created in your workbook with the histogram chart and columns called Bin and Frequency. Frequency is the counts of observations in the intervals defined by Bin.

There are several things you can do to edit the look of the histogram.

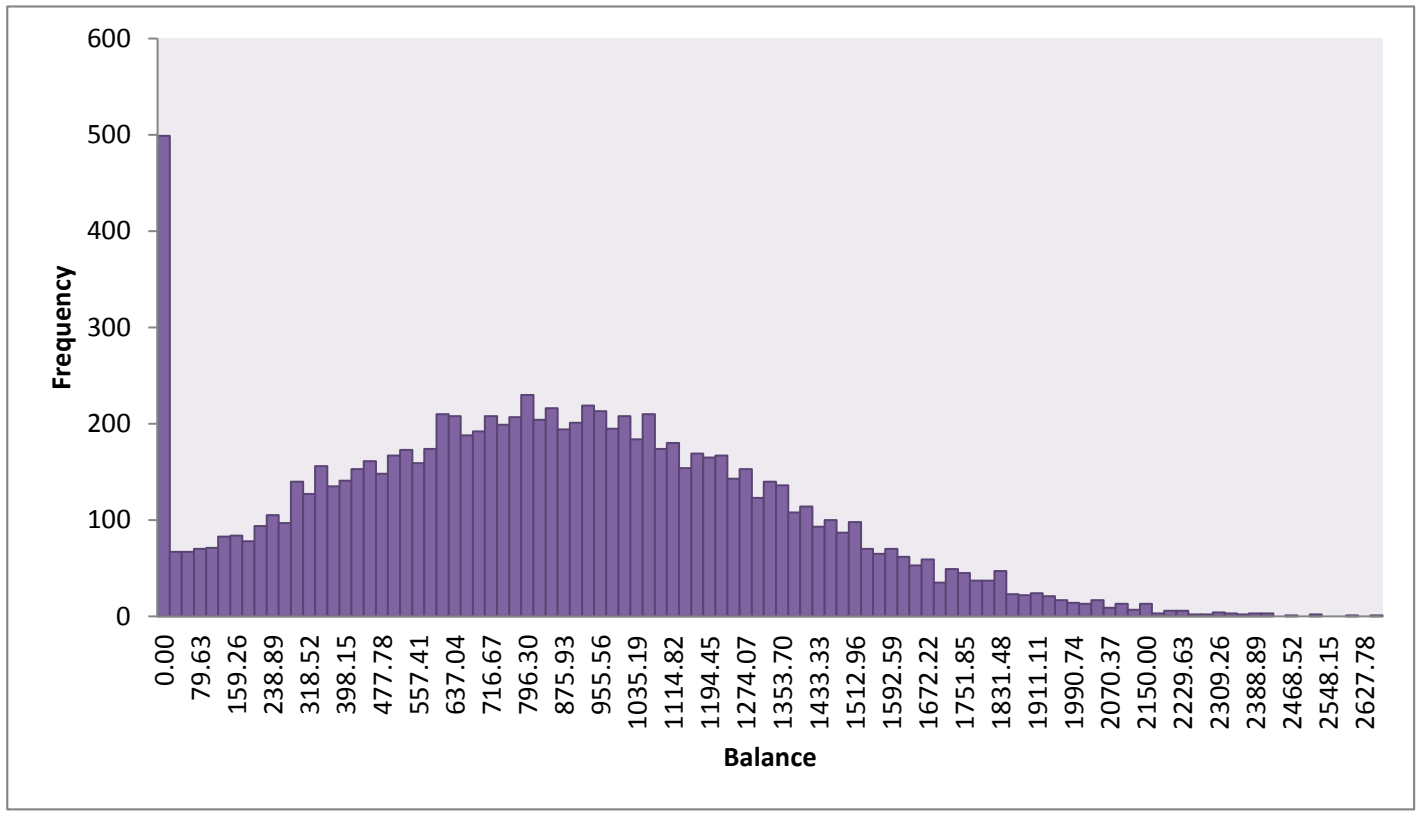
- You can click on any of the labels (e.g. Histogram) and edit (or delete) them. For example, you could change "Bin" to "Balance".
- You can select the Bin column, right-click/Format cells/Number and then change the number of decimal places.
- You can click on the chart (histogram) and then play around with "Chart Tools" *Design*, *Layout*, and *Format*.
- Under Design/Chart Layouts, I like to pick the one that puts the bars together, then under Design/Chart Styles, I pick one of the styles that highlights the bars.

In Excel 2013, there are tabs "Design" and "Format" *and* "plus", "pen", and "funnel" icons to the top right of the chart (after you click it). All of these can be used to modify the appearance of the histogram.

Once you have a chart you like you can use copy/paste to pop it into another document (e.g. Powerpoint or Word).

You an also do /File/Save&Send/Create PDF/XPS Document.

For example, I got the below.

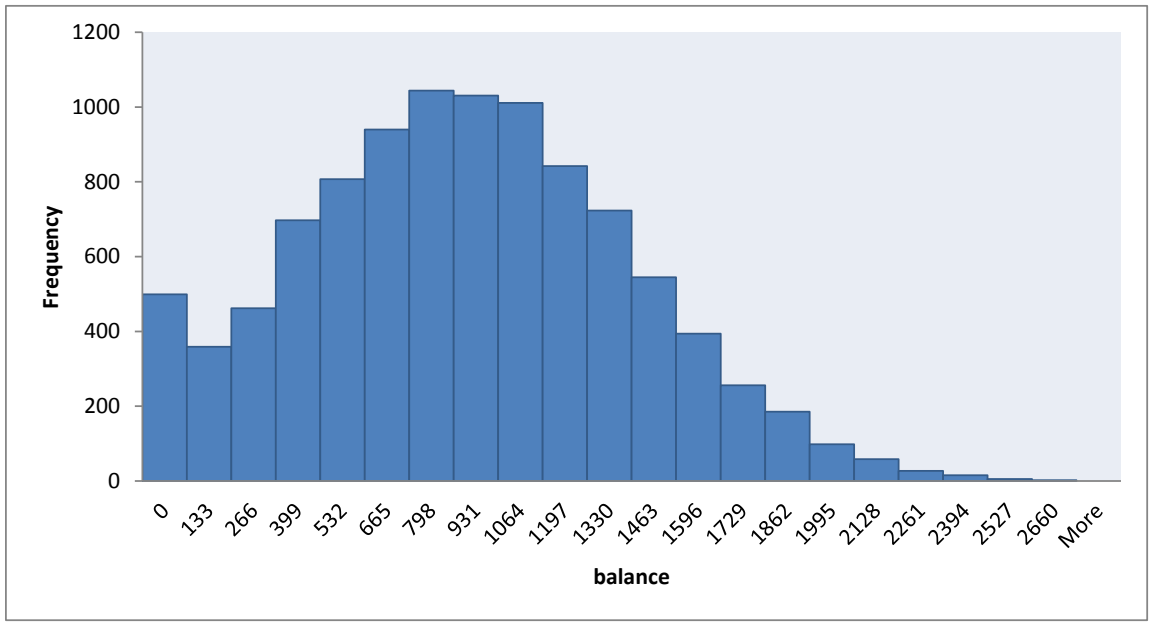


To have more control over the histogram, you have to set the bins yourself.
To create a series with cells with bin cutoffs in them:

- Put the minimum in a cell. (e.g. =min(balance) in F3).
- Put the maximum in a cell. (e.g. =max(balance) in F5).
- Put the number of cutoffs you want in a cell. (e.g. put 20 in F8).
- Compute the size of each cell using (max-min)/(# of cutoffs). (e.g. =(F5-F3)/F8 in F11).
- Put the value of minumum in a cell, this will be our initial cutoff (e.g. put 0 in H2).
- Click the cell with the initial cutoff in it.
- Use /Home/Editing/Fill/Series/ click columns, put the size of each cell in "step size" and put the maximum value in "Stop Value". "Fill" is indicated by a big blue down arrow right below Σ in the Editing section.
- Optionally, put a lable (e.g. "bins") above the cutoffs.

Now do the histogram as before *except*, use the cutoffs you just created for the "Bin Range" values (just below the "Input Range" you used for the variable values).

See the file Default-hist-bins.xlsx in the Excel section of the course webpage for an example Workbook.



2. Excel: Working with Subsets of Data

We often want to select subsets of our data.

Let's use the Default data and do the histograms of the balance variable with default equal to No and Yes.

Let's try this using the /Data/Sort & Filter/Filter tool.

What we will do is copy the balance values to a new column using *only the values such that default=Yes*.

Then we do the same using only the *default=No* values.

Then we can make histograms using our two new columns.

To copy the balance[default="Yes"] values:

- click a cell in the data.
- click /Data/Sort & Filter/Filter. Drop down menus will appear beside each variable name.
- click the drop down menu beside default and select the boxes so that only Yes is checked, and then hit OK.

Now you have only the rows with default=Yes showing!!

Now just copy the balance column to a new column, for example:

- click C (column C), Ctrl-C.
- click F1, Ctrl-V. Esc.

Then turn off the filter by clicking the Filter funnel again.

Change F1 to a new name (I put balanceY, instead of balance).

Now do the same again but use default=No. Copy these values into column G and relabel the column balanceN.

I put my bin cut-offs in a separate worksheet and then made histograms of balanceN and balanceY using the same bins.

See hist-balance-defYN.xlsx.

3. Excel: Tables with PivotTables

The PivotTable facility in Excel is *nice*.
To make a table using PivotTables:

- Click on a cell in your data.
- /Insert/PivotTable.
- A dialogue should come up with your data range already selected, so just click OK.
- Now you should have the "PivotTable Field List" (on the right of your sheet).
- *Drag a variable to the "Row Labels" Pane for a one-way table. For a two-way table drag an additional variable to the "Column Labels" Pane. (For example, with the midcity data, I put Brick in the Row Labels Pane and Nbhd in the Column Labels Pan).

*To drag a variable, click on the variable name in "Choose fields to add to report", and drag it with the mouse while holding your click.

- The variables in "Row Labels" and "Columns Labels" determine what subsets of the data we look at. To choose a variable to summarize in the subsets, drag a variable to the Values pane. (For example, with the midcity data, we put Price in the Values pane).
- To choose the summary method (mean, median, ..) click the drop-down menu beside the variable name in the "Values" pane, and then click "Value Field Settings". You will get a list of summary functions to choose from. (for example with the midcity data, we chose Average).

That's it!!!

To bin up a continuous variable you need another step.
Let's use the midcity data and put SqFt in the "Row Labels" pane and Nbhd in the "Columns Labels" pane. This will give you a table with a lot of rows for a lot of SqFt values.

Then

- click on one of the "Row Labels" (e.g. 1710) and then right-click to get a menu.
- Click on "group".
- You can then choose a grid by choosing the Start grid value, the Ending grid value, and the increment. (I tried start=1450, end=2590, increment=200).

Note:
If you click on the table you created with PivotTable, you get the "PivotTable Field List" back, *and* "PivotTable Tools" on the Ribbon.

Obviously, there is a lot of stuff here.

Try (for example) /PivotTable Tools/Options/PivotChart.
If you then click on the chart you get, there are again lot's of options.

To get tables of counts (and Row/Columns percentages)

- Put a variable in "Row Labels" and a variable in "Column Labels". These two variables will define your table.
- Put any variable in "Values" and then use the drop-down menu beside the variable name in Values and choose "Value Field Settings" and then choose Count. This will give you a table of counts.
- To change the table display to percentages, row percentages ... click on a count in the table and then right-click to get a menu. Then do /Value Field Settings/Show Values As and then click the drop-down menu to get your choices.