Lying with statistics

Ethan D. Bolker
Maura B. Mast

October 30, 2007

Plan

- Double bar graphs
- Using graphs to lie with data
- Frequency and Relative Frequency Histograms (they look the same)

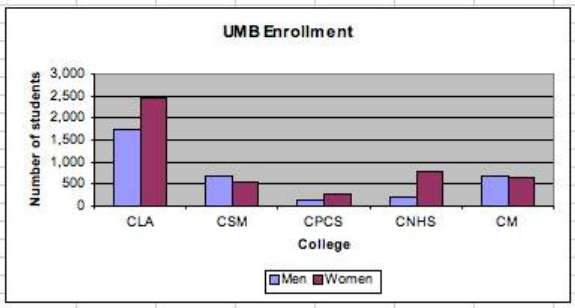
Lecture notes

College Enrollment bar graph

On the homework we asked you to look at the UMB website and gather some data about college enrollment. We first asked for a bar graph displaying the actual number of men and women in each college, then we asked you to construct a bar graph showing the percentage of each college that consists of men and the percentage of college enrollment consisting of women. Here is the data table and the first bar graph that we built:

UMB Enrollment by College
Ethan Bolker, October 30, 2007
Source: www.oirp.umb.edu/2006/stat_portrait/index.html, table 30

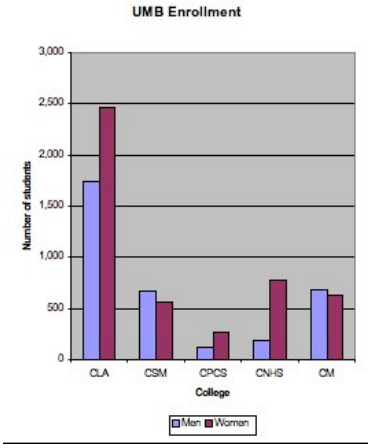
College	Number of Men	Women	Total	Percent Men	Women
CLA	**1,742**	**2,469**	4,211	41.36785	58.63215
CSM	**665**	**562**	1,227	54.19723	45.80277
CPCS	**122**	**270**	392	31.12245	68.87755
CNHS	**192**	**771**	963	19.93769	80.06231
CM	**682**	**633**	1,315	51.86312	48.13688



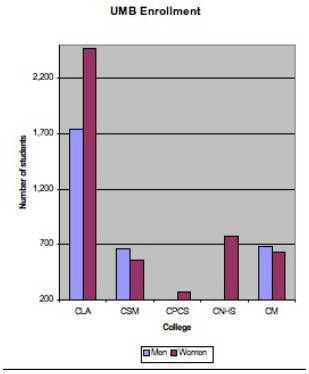
We can make some immediate observations by looking at this graph.

- The College of Liberal Arts has the highest overall enrollment, with just over 4000 students. In fact, the enrollment of CLA is larger than the enrollments of the other colleges combined.
- CPCS has the smallest enrollment of the colleges, with a total enrollment below 500.
- In CLA, CPCS, and CNHS there are more women than men. In fact, we see that there are many more women than men in CNHS, perhaps 4 times as many.
- In CSM and CM there are more men than women, although the proportions are fairly close.

With a little bit of effort, we could skew this graph to exaggerate some claims (for example, perhaps the Dean of CLA wants to emphasize this college's enrollments compared to the other colleges). By re-sizing the window, for example, we can make the bars look taller (and therefore make the enrollments appear larger). Here is one example:

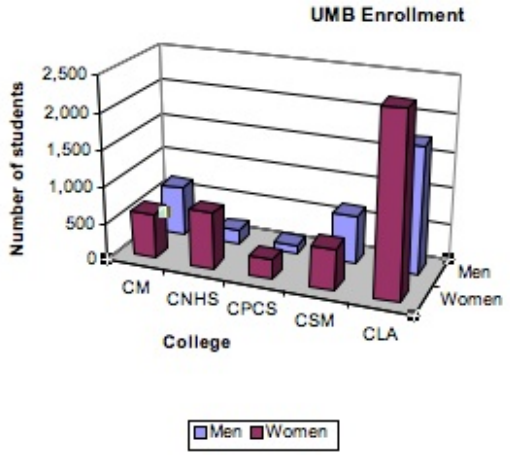


We could also chop off the scale to further skew the picture that the graph gives us. This is not entirely legitimate, since we are losing data and we are very deliberately altering the picture. Here is what it could look like if we started the vertical scale at 200 and put in 2500 as the maximum:



With these adjustments (which you can make by right-clicking or double-clicking on the vertical scale and then clicking on the SCALE tab at the top of the window that appears) we have seriously distorted the picture. It appears now (to the casual observer) that CPCS has few students (and no men) and the CLA enrollments are soaring about the enrollments of the other colleges.

Of course we could adjust our graph in other ways. We could take the scale off the vertical axis completely, or color the graph to emphasize a particular piece of data. We could also turn the graph into a three-dimensional picture. In this example, I've rotated the graph so that the CLA bars are to the right (and so appear a bit larger).



Are we lying when we do any of these things to the graph? No, in the sense that we have not changed the data. What we have done is change the perspective of the graph (which then changes how the viewer may interpret it) or modify the scale to make a particular emphasis. As a result, we are being deceptive because we are not giving the full picture, or the mathematically precise picture. We doubt you will, in your life, be in a position to construct and distort graphs (well, we hope not). We are sure that you will, in your life, encounter graphs that are poorly drawn and, intentionally or not, deceptive.

Relative frequency bar graphs

The previous examples all used what statisticians call *frequencies*; that is, the heights of the bars correspond to the number of people in each group. We can also construct a bar graph using relative data, or percentages. The first step is to ask Excel to calculate the percentages for us. If you set up the first calculation correctly, you can copy and paste it into the rest of the cells and Excel will do the work for you.

Look back at the CollegeEnrollment spreadsheet. To calculate the percentage of students in the College of Liberal Arts who are men, we would set up the fraction $\frac{1742}{4211}$. Thinking ahead to how we will use the power of Excel to do this, we instead write the following formula in to cell E7:

```
=B7/D7
```

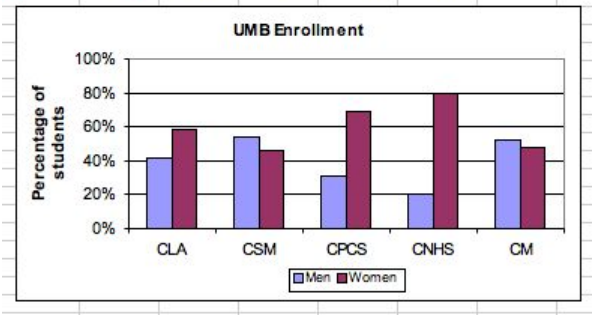
and then hit Return. Excel will calculate this as 0.54197. You can copy this cell and paste it into the remaining cells in this column and Excel will

automatically calculate the fraction of men in each of the colleges. Now go to cell F7 and ask Excel to calculate $= C7/D7$. Excel will return 0.45802. You can then copy and paste this formula into the remaining cells in this column and Excel will update the calculation. [1]Now highlight all of the cells and click on the % button in your toolbar, and Excel will convert these numbers to percentages (of course, you could always adjust your calculation to include multiplication by 100 and then skip this step).

Now that we have calculated the percentage of men and women for each college's enrollment, we can make a graph with this data. To highlight non-adjacent columns, you need to hold down the Apple key and then highlight the non-adjacent cells. Then construct the bar graph following the usual steps and your table and graph should look something like this:

UMB Enrollment by College
Ethan Bolker, October 30, 2007
Source: www.oirp.umb.edu/2006/stat_portrait/index.html, table 30

	Number of			Percent	
College	Men	Women	Total	Men	Women
CLA	**1,742**	**2,469**	4,211	41%	59%
CSM	**665**	**562**	1,227	54%	46%
CPCS	**122**	**270**	392	31%	69%
CNHS	**192**	**771**	963	20%	80%
CM	**682**	**633**	1,315	52%	48%



In this graph we have lost the actual data, but we can now make some comparisons using the percentages. Here are some things we could say.

- The college with the largest percentage of women in CNHS; in fact, women outnumber men in this college by 4 to 1.
- In CPCS, there are just over twice as many women as men.

[1] There is a way to do these two sets of calculations at once. In cell E7, type the formula $= B7/\$D7$ and then copy and paste this into the rest of the column *and* into cells F7 through F11 in the next column. Because of the \$ in front of the D, Excel stays in column D for all of the calculations. Try it without the \$ and see what happens. Excel calls this a *mixed reference*.

- CSM and CM are fairly evenly split in enrollment between men and women, with a slightly higher percentage of men than women.
- In CLA, the percentage of women is higher than the percentage of men by almost 20 percentage points.

Histograms of Percentages

When we constructed the histogram of salary data for the Wing Aero Corporation, we used the frequencies for each salary range; that is, we used the number of employees in each salary range. Another way to look at this data is to calculate the *percentage* of employees who are in each salary range.

To find this information, we use Excel. Load the Wing Aero spreadsheet `http://www.cs.umb.edu/~eb/m114/lectureNotes/1025/WingAeroStudy.xls`.

This spreadsheet shows the frequency histogram for the salary data. We will add a new column, representing relative frequency, and make a new histogram. Put a new column heading in cell F14, labeled Percentage of employees. To calculate the percentage for each salary range, you will have to tell Excel to divide the number of employees in that range by the total number of employees. You can type

```
=E15/30
```

into cell F15, then hit return. Excel should calculate 0.066667. Then copy and paste this into the remaining cells in this column.[2]

Highlight all of these cells and click on the % button in the Toolbar and Excel will convert them to percentages. How highlight the salary ranges and the percentages and ask Excel to make a histogram. Remember to push the bars together after you have constructed the graph, by double clicking on one of the bars, then clicking on the "Options" tab and then setting the Gap Width to 0. When you are done, your graph should look something like this:

[2] Another way to do this calculation is to type $= E15/\$E\$30$ into cell F15. Then you can copy and paste it into the remaining cells in the column as before. Because of the dollar signs on both sides of E, Excel does not adjust this cell as it moves through the rows and columns. This is called an *absolute reference*.

salary range (K$)	# employees	Percentage of employees
0-19	2	7%
20-39	10	33%
40-59	7	23%
60-79	3	10%
80-99	3	10%
100-119	0	0%
120-139	1	3%
140-159	0	0%
160-179	0	0%
180-199	0	0%
200-219	0	0%
220-239	0	0%
240-259	0	0%
260-279	2	7%
280-299	2	7%
total	30	100%

If you compare this graph to the original frequency histogram, you should notice something interesting: the graphs have the same shape. The only aspect that is different is the scale on the vertical axis, since in one graph we are measuring the actual number of employees in each salary range and in the other we are measuring the percentage of employees in each salary range. Is one graph better than the other? It depends on the information you want to provide. Using the new graph, you could note that "Almost 35% of Wing Aero employees earn a salary that is between $20,000 and $40,000." If you used the actual numbers, you would say: "10 Wing Aero employees earn a salary that is between $20,000 and $40,000." This second statement only really makes sense if we know how many employees there are (that is, if we have a context for the statement). That's one example in which the percentage data is more useful to use.