# Bar graphs, pie charts and histograms

Ethan D. Bolker
Maura B. Mast

October 23, 2007

## Plan

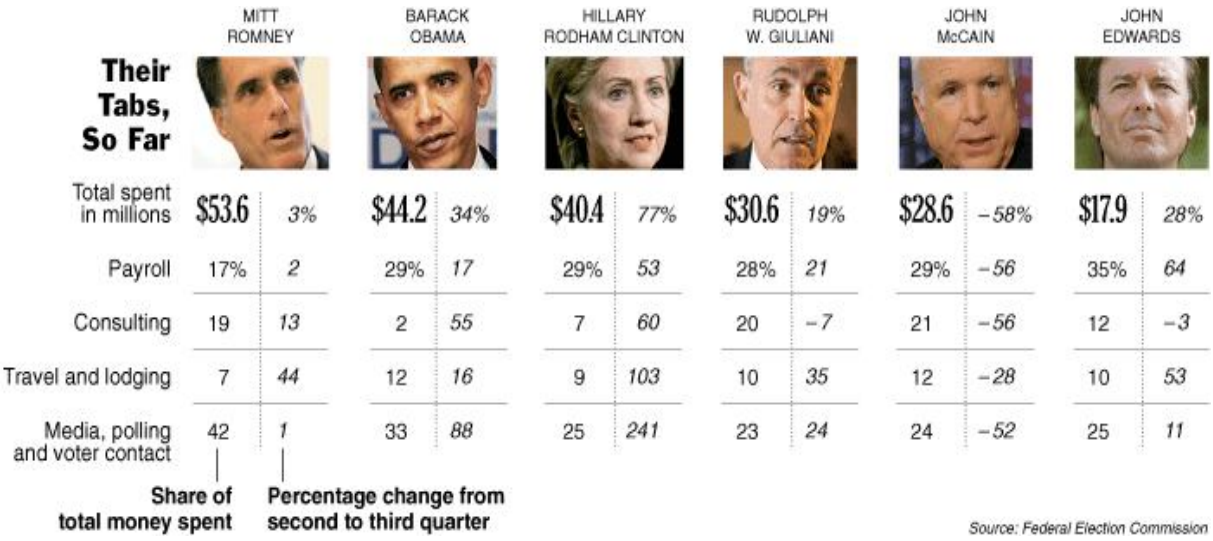
- Bar graphs and pie charts
- Histograms - a special type of bar chart

## Lecture notes

### Bar Charts

A bar chart is a type of graph that uses the heights of the bars to convey information. It's easy to take a quick look at a bar chart and pick out the highest and lowest bars, and therefore get some information about the data. The distribution of the bars (are they all the same height? Is one very high and the rest small?) can also give us some information about how the data are distributed.

We use bar charts when our data fall into categories. The height of each bar represents the number of data values in that category (called the *frequency*). Here is an example. In *The New York Times* on October 22, 2007, the following graphic showed the total amounts spent by the leading presidential candidates.



**Their Tabs, So Far**

| | MITT ROMNEY | | BARACK OBAMA | | HILLARY RODHAM CLINTON | | RUDOLPH W. GIULIANI | | JOHN McCAIN | | JOHN EDWARDS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total spent in millions | $53.6 | 3% | $44.2 | 34% | $40.4 | 77% | $30.6 | 19% | $28.6 | −58% | $17.9 | 28% |
| Payroll | 17% | 2 | 29% | 17 | 29% | 53 | 28% | 21 | 29% | −56 | 35% | 64 |
| Consulting | 19 | 13 | 2 | 55 | 7 | 60 | 20 | −7 | 21 | −56 | 12 | −3 |
| Travel and lodging | 7 | 44 | 12 | 16 | 9 | 103 | 10 | 35 | 12 | −28 | 10 | 53 |
| Media, polling and voter contact | 42 | 1 | 33 | 88 | 25 | 241 | 23 | 24 | 24 | −52 | 25 | 11 |

Share of total money spent

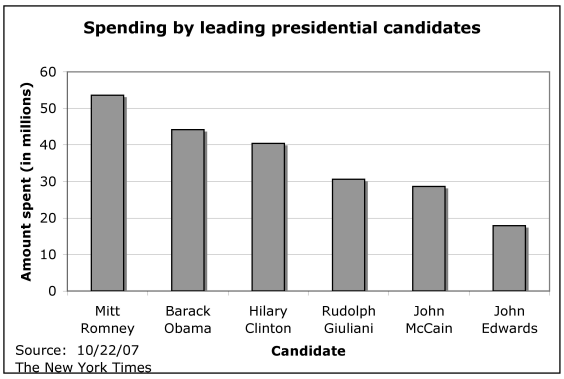
Percentage change from second to third quarter

Source: Federal Election Commission

We can use Excel to make a bar graph with this information. The first step is to construct a table in Excel. Your table should look something like this:

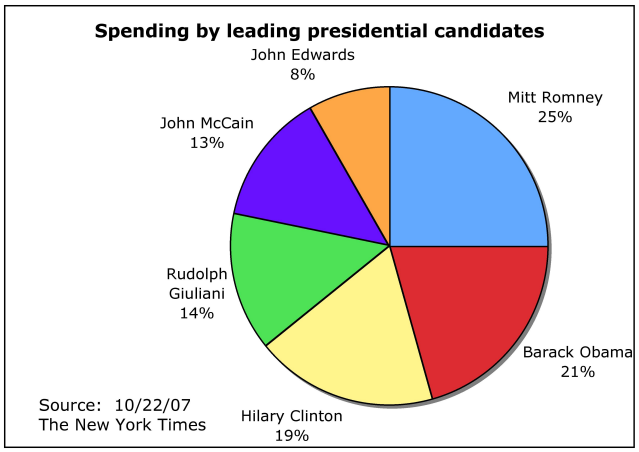
| Presidential Candidate | Amount spent (in millions of $) |
|---|---|
| Mitt Romney | 53.6 |
| Barack Obama | 44.2 |
| Hilary Clinton | 40.4 |
| Rudolph Giuliani | 30.6 |
| John McCain | 28.6 |
| John Edwards | 17.9 |

Select the data in the two columns and click on the Chart Wizard (or go to the Insert menu and click on Chart). Select Column as the Chart Type (Excel interprets a Bar Graph as a graph with horizontal bars so in fact you must select Column to get the vertical bars). Click "Next" or hit return and you'll get a picture of the graph. It is looks right then click "Next" again, put in a title, and label the axis. Finish, then adjust the graph. You don't need a label in this case, since there is only one thing that we are measuring (amount spent). Insert a text box giving the source. Your graph should look something like this:




In this graph, the bars are arranged from highest (representing the largest spender) to lowest (representing the smallest spender). You could rearrange the bars to change the emphasis if you wanted to. Looking at the graph, it's clear that Mitt Romney has outspent the other candidates and that he has spent more than twice as much as John Edwards. We also see a big drop in spending between Hilary Clinton and Rudolph Giulianai, and note that Giuliani and McCain have each spent about the same amount.

Excel will also turn this information into a pie chart for us. To do this, right-click on the white space surrounding the graph and click on Chart Type. Click on Pie as the chart type, then click OK. You may need to reformat your graph to include the labels. To do this, right-click on the white space around the graph again and this time click on Chart Options. Click on the Data Labels tab and click one of the boxes. If you click the "Show label and percent" box, your graph will look something like this:



What happened here? We didn't ask Excel to calculate any percentages. In fact, our original data table didn't have any percentages in it. The reason there are percentages in the pie chart is because that's how pie charts are made. They show proportional relationships among the data. If Mitt Romney has spent $53.6 million dollars, then this represents $53.6 million out of a total of $215.3 million spent by these six candidates, or about 25%. Excel did this calculation for each of the candidates and used these percentages to construct the wedges. The wedge of the graph corresponding to Mitt Romney represents 25% of the area of the circle. For John Edwards, in contrast, the wedge only represents 8% of the circle.

A word of warning: Excel will do exactly what you tell it to do, whether it is really correct or not. In this case, it made a guess about how to construct the percentages, and that guess was correct. If you have percentage data in your table and the percentages (for whatever reason) do not add up to 100%, Excel will still make a pie chart. It's also always a good idea to double-check that the percentages in any pie chart you see (especially in a newspaper or advertisement) add up to 100%. Errors in the construction of pie charts are all too common.

We could have asked Excel to directly calculate the percentage spent by each candidate. Here's what we could do. Put a label at the head of a third column, calling it "Percentage spent".

| Presidential Candidate | Amount Spent (millions of $) | Percentage Spent |
|---|---|---|
| Mitt Romney | 53.6 | |
| Barack Obama | 44.2 | |
| Hilary Clinton | 40.4 | |
| Rudolph Giuliani | 30.6 | |
| John McCain | 28.6 | |
| John Edwards | 17.9 | |
| | 215.3 | |

In cell C2, type the formula

```
=A2/215.3
```

and then hit return. Excel will calculate 0.248 and put it into this cell. Now copy this cell and paste it into the remaining columns. Excel will update

the calculation. Highlight all of the cells in this column and click on % in the Toolbar, and Excel will convert them to percentages. Now you could highlight the names and the percentage data and construct a bar graph or a pie chart. Note that the bar graph will look essentially the same as the first bar graph you constructed. The only difference will be the vertical scale.

## Histograms

A histogram is a special type of bar chart. We use histograms when we can put our data into numerical categories. There are some special rules about how histograms are set up. The categories are ranges of values, with no overlaps and no gaps between the values. The range should be the same for all of the categories. Generally speaking, you can choose the number of categories you want, and this will change the way the graph looks.

Before you even begin to make the histogram graph, you will have to do some work. You can go through the steps on the handout `http://www.cs.umb.edu/~eb/m114/lectureNotes/1023/HistogramTutorial.pdf` or look at the Lecture Notes for Oct. 25 to get a reminder for the procedure.

As an example, we will look at the data on the spreadsheet `http://www.cs.umb.edu/~eb/m114/lectureNotes/1025/BostonTemperatures.xls`. This spreadsheet shows the actual high and low temperatures along with the actual average temperatures for the first 22 days in October, along with the normal highs, lows and averages. We could certainly make a scatter plot of (some or all) of this data; instead, let's make a histogram using the actual average temperature data.

The first step is to set up temperature intervals. It helps to sort the Average temperatures. To do this, select cells C5 to C26 then click on the AZ↓ button on the Toolbar (or go to the Data menu and click on Sort). Now we can see that the average temperature during this period in October ranged from a low of 52 to a high of 74. The next step is to set up intervals. I'd like to compare this data to the normal temperature averages, and I notice they start at a low of 50 degrees, so I'll use that as the starting point for the intervals. I'll set the first interval as 50 - 54 degrees.There is nothing special about choosing this range. In general, you want to choose intervals that are not too small and not too big. This seems about right given the variation of temperatures.

Start a new column in Excel and set up the intervals in that column. It should look like this:

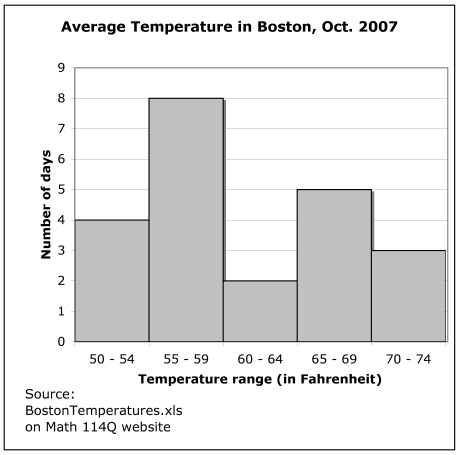
```
Temperature ranges
50 - 54
55 - 59
60 - 64
65 - 69
70 - 74
```

Notice a few things about these intervals. They are all the same size (in fact, they all include five possible temperatures); there are no gaps between intervals; and they go far enough to capture all of the data.

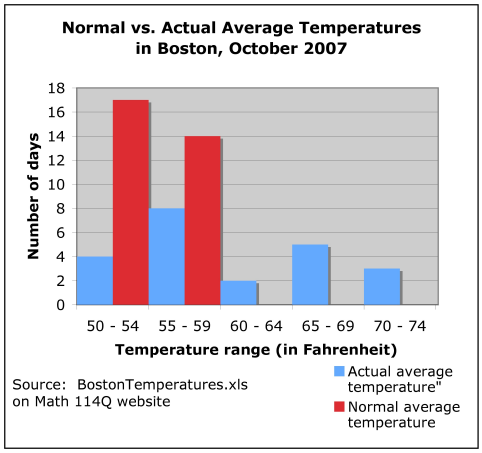
Now move over one more column and count the number of times the average temperature fell into each interval. You can do this by hand (another reason why it's helpful to sort the data). Your table should look something like this:

| interval | frequency |
|---|---|
| 50 - 54 | 4 |
| 55 - 59 | 8 |
| 60 - 64 | 2 |
| 65 - 69 | 5 |
| 70 - 74 | 3 |

Now select the both columns of data (intervals and frequencies) and click on the Chart Wizard. Choose a column graph again, and go through the steps of labeling it as usual. When you are done, double click (or right-click) on one of the bars. You should get a window with OPTIONS as one of the tabs. Click on that tab, then set the Gap width to 0. Click OK and the bars should be pushed together. You may need to change the formatting of the colors and lines so that the bars are easily recognizable. Your graph should look something like this:




From this graph, we can get a visual sense of the *distribution* of average temperature data in October. We can see that most of the average temperatures were below 60 degrees, but that there were a significant number of days in which the average temperature was warm. In fact, we see that for 3 days in October, the average temperature was above 70 degrees. This seems unusual for October in New England. We can verify this by making a double histogram with the average temperature data. Since we have already set up the intervals for temperature, we can sort the normal average temperature data and count the number of days in which the temperature fell into one of those intervals. Then we highlight all of the data and ask Excel to make a column chart. After pushing the bars together we get the following graph.




Now we can make a direct comparison between the normal average temperature and the actual average temperature. We see that in fact it was quite

unusual for there to have been so many warm average temperatures in October. In fact, on average the temperature in October does not go above 60 degrees Fahrenheit.