Histograms – mean, median, mode

Ethan D. Bolker
Maura B. Mast

November 8, 2007

Plan

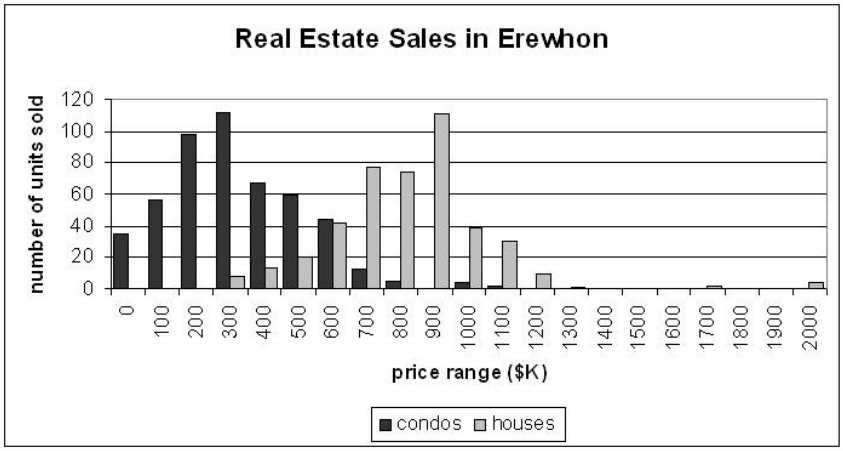
Lecture notes

Shopping for housing in Erewhon

The spreadsheet `http://www.cs.umb.edu/~eb/m114/lectureNotes/1108/HousingPrices.xls` contains (fictional) data for recent housing transactions in Erewhon, MA. We'll use it to study means, medians and modes in histograms, and to learn to use some of the power of Excel to save us lots of typing.

We recommend that you work through the material below in Excel as you read it. But if you're impatient you can find the results of all the computations in `http://www.cs.umb.edu/~eb/m114/lectureNotes/1108/HousingPricesComplete.xls`

1. Create a double histogram (start by selecting range B6:D27, including the empty cell B6, so that Excel knows that the data in column B are labels, not numbers.)



It's clear from the data and from the picture that "on the average" condos cost less than houses, although some condos cost more than some houses.

The mode cost of a condo is \$300K to \$400K, of a house \$900K to \$1M. Those are the highest columns in the respective histograms.

2. Finding the mean and median costs is harder, since we don't have access to the raw data.

   To compute the mean condo price we need to know the total number of condos sold and the total amount paid for those condos. To find the total number of condos sold, just add up the values in column C by typing the Excel formula

   ```
   =SUM(C7:C27)
   ```

   in cell C28. (Be sure to label row 28 `total` in column B.

   To compute the total amount paid for condos we will assume (for lack of any better information) that (for example) the 112 condos that sold for between \$300K and \$400K each sold for exactly \$350K. To make a similar computation for all the rows, we start by entering 50 (the midpoint of the first interval, \$0 to \$100K) in cell E7. Then we need to add 100 to that value as we read down column E. Since that's just what happens in column B we can copy the formula `=B7+100` to cells

E8:E27. Excel automatically adjusts the cell reference from B7 to E8, E9, and so on.

Then we estimate the total spent on all the condos in each range by multiplying the number of houses in the range by the middle of the range, putting the results in column F. That means we would like to see

$$50 * 35 = 1750$$

in cell F7. We can make that computation by entering

```
=C7*E7)
```

in that cell. Then copy the formula to the rest of column F to complete the calculations.

3. Sum column F by copy the formula from C28 or D28 to F28: the result tells us that 184800 hundred thousand dollars was spent on condos. To find the mean condo price we divide that number by the total number of condos sold: the 494 in cell C28. So label cell B29 `mean` and put

   ```
   =F28/C28
   ```

   in cell C29. You should see 374.0890688: the average (mean) cost of a condo was about $400K.

4. To make the same computations for houses we can simply copy and paste with very little thought. Start by copying the formulas in column F to column G, in order to find the total amount paid for houses in each price range. If you try that you'll notice an error: there's a 0 in cell G27 where you should see 8200, the amount paid for the 4 houses at 2050 each. Look at the formula in cell G27 to see why. It's `=D27*F27` where we want `=D27*E7`. When we copied from cell F27 Excel correctly changed the D to an E and incorrectly changed the E to an F. To fix that problem, edit the formula in G7 and copy the correct one to the rest of column G.[1]

   Now sum column G and find the average house price by copying the formulas from F28 to G28 and C29 to D29. The average house price is $868.372093 or about $870K.

[1] There's a better way to prevent this error, which we'll learn later.

5. To find the median condo price we want to compute successively the number of condos that sold for less than \$100K, less than \$200K, and so on. When we reach half the total number of condos we will have found the median.

   So start column H with the contents of C7 (35) in cell H7. Then put the formula `=C8+H7` in H8, in order to compute $35 + 56$, the number of condos that sold for less than \$200K. Copy that formula to the rest of column H. Note that starting at row 18 all the entries are 494, since no condo sold for more than \$1100.

   Before we take the next step, copy column H to column I to do for houses what we've just done for condos.

   There were 494 condos sold, so the halfway point would be at about 250. But we can make Excel find the halfway point for us. Use column J to convert the number of condos in each row of column H to a fraction, by dividing that number by the total. The formula we want is

   ```
   =H7/C$28
   ```

   Put that in cell J7 and copy it to the rest of column J. We see that the condos costing less than \$200K account for 38% of the total (row 9) while those costing less than \$300K account for 61% (row 10). So the median condo price is about halfway between, or about \$250K. That's less than the mean condo price, because there are a few high price condos that skew the distribution.

   Wait a minute! What's the dollar sign doing in the previous formula? Why did we use

   ```
   =H7/C$28
   ```

   and not

   ```
   =H7/C28   ?
   ```

   Because we are smarter than Excel. If we used C28 in the formula and then copied it from cell H7 to cell H8 the result would be

```
=H8/C29
```

when what we want is

```
=H8/C28
```

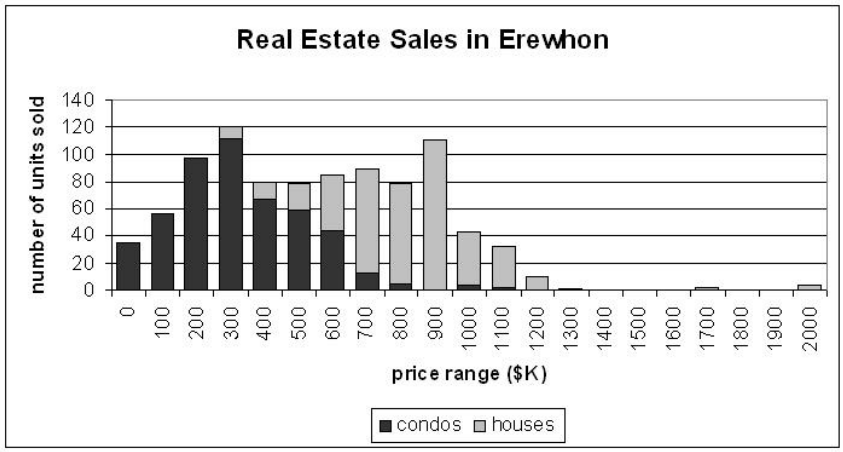
where the index of the H7 has been changed to H8 but C28 stays C28. The dollar sign in front of the 28 tells Excel *not* to increase the 28 when the formula is copied to the next row.[2]

If we copy column J to column K we find that the median price for houses is about \$750K – less than the mean because this distribution is right skewed too.

We can copy column J to column K and get the right answer for houses because we used `C$28`, *not* `$C$28`, in the condo formulas. So when we copied them over one row Excel changed the C to a D.

The combined housing market

To study the sales of condos and houses without distinguishing one from the other we can start by simply changing the chart type from one with adjacent columns to one with stacked columns.



[2]Using a `$` to prevent Excel from changing a reference is a trick we could have used earlier in this discussion.

This distribution has no unambiguous mode. The interval $200K - $300K shows the highest bar, but the one at the interval $800K - $900K is nearly as high. It's best to call this kind of distribution *bimodal*, an adjective that describes its shape.

For bimodal distributions the mean and the median are often particularly misleading.

To find the mean housing price we divide the total spent on housing (184800 + 373400, from cells F28 and G28) by the total number of units (494+430 from cells C28 and D28) to find a mean value of 604.11, or about $600K. But the graph shows us that not so many units actually sold for the average price.

This is a good place to expose another kind of fallacy. If you try to compute the average price by averaging the averages for condos and houses you see

$$\frac{374.0890688 + 868.372093}{2} = 621.2305809$$

which is about $620K – larger than the correct value of about $600K. That's because the correct calculation takes into account the fact that there are more condos than houses to reach the correct smaller answer. The moral of this story is that you can't average averages.

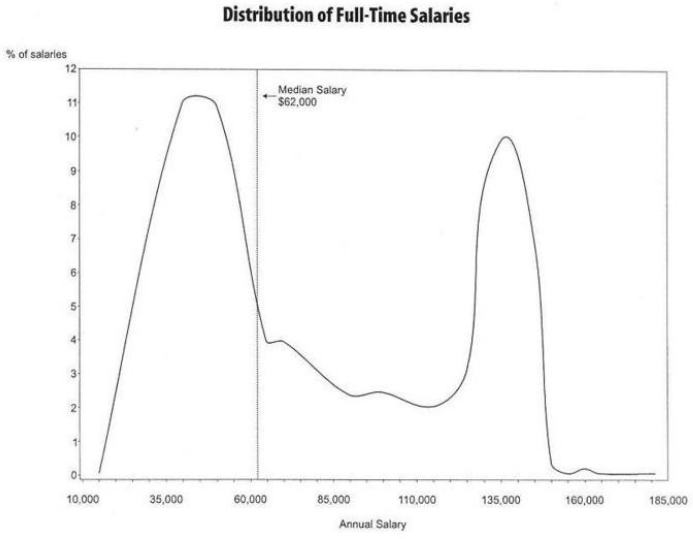
Finally, the median unit price for the combined data is about $500K (see column L in the spreadsheet). That average, like the mean, corresponds to relatively few units.

Bimodal distributions

On September 4, 2007 the Empirical Legal Studies blog at `http://www.elsblog.org/the_empirical_legal_studi/2007/09/distribution-of.html` contained the post

> Distribution of 2006 Starting Salaries: Best Graphic Chart of the Year
>
> The most recent edition of NALP's serial publication, *Jobs & JD's*, includes the chart ... below. It is the distribution of full-time salaries for all members of the Class of 2006 who reported income data to their respective law school (22,665 graduates). If you were looking for a single graphic to illustrate the most vexing problems facing law firms, law students, and law schools,

this would be it. A more dramatic bimodal distribution you will not find.



Note: The graph is based on 22,665 salaries. A few salaries above $200,000 are excluded for clarity. The first peak in this graph reflects salaries of both $40,000 and $50,000 (each about 11% of reported salaries). The second peak reflects salaries of $135,000 (10% of reported salaries) and $145,000 (7% of reported salaries).

The sample includes – in order of size – private practice (55.8%), business (14.2%), government (10.6%), judicial clerks (9.6%), public interest (5.4%), and other (2.8%). Half of the graduates make less than the $62,000 per year median – but remarkably, there is no clustering there. Over a quarter (27.5%) make between $40k-$55k per year, and another quarter (27.8%) have an annual salary of $100K plus.

If the chart were a flipbook of the last twenty years, the first mode would be relatively stationary, barely tracking inflation, while the second mode would be moving quickly to the right – i.e., the salary wars. In fact, because of the recent jump to $160K in the major markets, the second mode has already moved even more to the right.

For bimodal distributions like this none of the three kinds of averages we've studied makes sense. There is no reasonable single mode, and the mean and median don't correspond to any particular reality. Bimodal distributions often occur when data for two distinct populations have been combined. The best way to analyze them is to separate the two populations. Read the blog for an analysis of the two populations that contribute to this distribution.