Averages and Histograms – Income Distribution

Ethan D. Bolker
Maura B. Mast

October 25, 2007

Plan

- Averages: mean, median, mode
- Histograms
- Computing averages from histograms

Lecture notes

Mean and Median

Consider the following salary information for a (hypothetical) small company.[1]

[1] In keeping with our philosophy for this course we should find real data from a real company for this exercise. But most companies keep this kind of information private.

Employee	Salary (K$)
CEO	299
CTO	250
CIO	250
CFO	290
Manager	77
Manager	122
Manager	84
Manager	63
Manager	68
Manager	49
Manager	84
Manager	84
Supervisor	42
Supervisor	37
Supervisor	29
Supervisor	43
Supervisor	51
Supervisor	38
Supervisor	33
Supervisor	42
Supervisor	49
Worker	25
Worker	19
Worker	41
Worker	17
Worker	26
Worker	25
Worker	21
Worker	28
Worker	27

What does this table tell us? We'll study it, using Excel to help us out with some of the tedious work. So load that spreadsheet now, from `http://www.cs.umb.edu/~eb/m114/lectureNotes/1025/WingAero.xls`.

What do you notice?

The table is arranged in decreasing order of importance (or prestige), but only approximately in decreasing order of salary. Some Supervisors seem

to make more than some Managers, and some Workers more than some Supervisors. We can make those discrepancies visible by sorting the data by salary.

Select the data on rows 9 and 38. Be sure to select all three columns A,B and C so they will be sorted together.

Choose `Data->Sort` from the menu bar and sort the data with salaries descending.[2] You can then clearly see the anomalies.

If you sort the data again alphabetically (by employee, column A) you can return the table to (nearly) its original state, because the employee categories were (nearly) alphabetical (by coincidence) at the start.

What's the *average* salary of Wing Aero employees? The usual meaning of "average" is "add up, then divide by the total number of items." We'll tell Excel how to do that for us.

Enter the word "Total" in Cell A40. Then go to Cell B40. Make sure the formula toolbar is visible. (Use the `view` menu to find it if it's not.) In the formula box type

```
=SUM(
```

to tell Excel you are about to add up some numbers. It will prompt you by asking for

```
SUM(number1, [number2], ...)
```

Select Cells B9:B38, close the parentheses and click the green check icon. Excel displays

```
Total   2315
```

in cells A40 and B40. Wing Aero's total annual payroll is \$2315K – about \$2.3 million.

To find the average employee salary we must divide this by the total number of employees. That's easy: the 30 rows from 9 to 38 contain employee records. But it's better to ask Excel to count the rows for us. Type `Count` in Cell A41 and then `=COUNT(` in Cell B41. Finish the formula in B41 by selecting the relevant cells B9:B38 or by typing the addresses of those cells and close the parentheses. You should see

[2] If you selected just the salary column B, Excel will warn you that you may be making an error, and even offer to try to correct it for you. That's kind of it.

`Count  30`

Finally, label Cell A42 `Average` and put the formula `=B40/B41` in Cell B42. You should see

`Average  77.16666667`

so the average annual salary at Wing Aero is about $77,000.

It's important to understand how to compute averages. But once you do understand you can ask Excel to do the work for you. Go to cell B43 and enter the formula `=AVERAGE(B9:B38)`, click the green check icon and see that Excel tells you again that the average is 77.16666667.

$77 thousand is a pretty good annual salary. If you read that it was the average at Wing Aero you'd think it was a pretty good company to work for. Maybe, maybe not.

Suppose that Wing Aero lost only a little money last year. The CEO convinced the Board of Trustees that he deserves to have his salary doubled, to $598K. To see what that does to the payroll statistics, go to Cell B9 and change the 299 there to 598. Excel automatically updates all the computations you've made, increasing the total annual payroll to $2,614 million and the average annual salary by about $10K to more than $87,000.

We can learn two useful things from our work so far.

- Excel is an excellent tool for asking "what-if" questions, because when the data change it automatically updates computations it's made.[3]
- The average salary is not a very good number to use to summarize the salary structure at Wing Aero. The CEO's big raise increases the average salary – but he's the only employee who's actually better off!

Suppose we sort the 30 line table so that salaries are increasing. Since the table starts at Cell A9 and has 30 entries, the entries in Cells B23 and B24 are the middle ones. That means half the employees make $42K or less (the entry in Cell B23) and half make $43K or more. In that sense we might want to say that the "average" salary is $42.5K. There's a name for this kind of "average" – it's the "median". The usual "average" we computed above is the "mean".

[3] You can even hire a new employee by inserting a new line in the table, or fire one by deleting a line. In either case Excel will adjust the computations of SUM, COUNT and AVERAGE appropriately. Try that out.

In some ways the median is a fairer "average" than the mean for the Wing Aero salary structure. It really tells you more about the company. And note that median salary doesn't change when the CEO gets his big raise.

Return to the spreadsheet and change the `Average` label in Cell A42 to `Mean`.

Then put

```
Median   42.5
```

into Cells A45 and B45.

Excel knows how to compute medians too. Try `=MEDIAN(B9:B38)` in Cell B46 and check that you get the same value: 42.5. The advantage to using Excel (once you understand the meaning of "median" is Excel's ability to recalculate on the fly. Suppose the Supervisor on row 23 gets a raise, to $50. Enter that value instead of the 42 in Cell B23. Then observe that Excel has recalculated the median in Cell B46: it's now 46.[4]

There's a third kind of average, the *mode*. We'll return to that after we've summarized the salary data in a different way.

Histograms

Wing Aero is small enough so that we can see the whole salary table at a glance. But if there were 1000 employees that wouldn't be possible. To understand the numbers we'd have to summarize them. One way to do that is to decide on some appropriate salary ranges, and then count the number of employees whose salary falls in each range. That is, in order to understand the numbers better we will ignore some of the details. In this example we'll use $20K salary ranges. That means we think of two Supervisors who make $29K and $33K as approximately the same, since each falls in the $20K-$39K category.

We need to count the number of employees making less than $20K, then the number making between $20K and $39K, and so on. To save you typing, we've listed these categories in Cells D15:D29. Since the data are sorted in increasing order, it's easy to do the counting. The resulting table in columns D and E, rows 15 to 29 is

[4] It figured this out even though the data were no longer sorted.

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

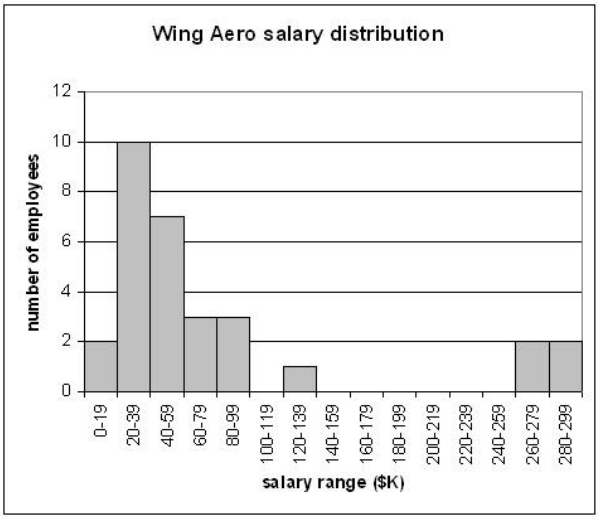
To check that we haven't missed anyone, we `SUM` the range E15:E29 to make sure the answer is 30, the known number of employees.[5]

With this data we can draw a *histogram* – a bar chart that provides a visual representation of numerical data in a table like this where the first column represents categories and the second the frequency of items in that category.

In our histogram the horizontal axis will represent the categories. A column chart is right for this kind of data. In Excel select the range D15:E29, go to the chart wizard, ask for a column chart (that will probably be the default). As you proceed through the wizard, add appropriate titles and legends. When you are done, set the gap width between the bars to 0% [6] and adjust the colors for black and white printing. The result will look something like this:

[5] This isn't a perfect check. We might have the right total but have put some employees into the wrong categories. The right way to build this table is to have Excel do it for you. It can, but it's not easy. If you search for `excel histogram` on the internet you can find out how.

[6] Right click on one of the bars, select Format Data Series, then the Options tab.



The histogram offers visual understanding of the salary structure that's hard to grasp just staring at the numbers or knowing some averages.

You can see the full spreadsheet with all the computations we've done so far at `http://www.cs.umb.edu/~eb/m114/lectureNotes/1025/WingAeroStudy.xls`.

Mode

There's a third kind of "average" that's often informative. The *mode* is the most common value. The histogram shows that there are more employees with salaries in the \$20K-\$40K range than any other. So the mode is about \$30K. It's the highest point in the histogram.

The mode is most useful for data aggregated into ranges, as for a histogram. In the raw Wing Aero data the mode is \$84K – it appears three times since three of the Managers happen (by coincidence?) to make that much. But it's clearly wrong to call that the "most common" salary.

When distributions are symmetric, the mean, median and mode are in the same place. The Wing Aero salary distribution isn't symmetric, it's *skewed* to the right. That's the fancy way to say that the bulk of the data clusters toward the left of the histogram with a long tail off to the right. For data that's right skewed, as this is:

$$\begin{array}{ccccc} \text{mode} & < & \text{median} & < & \text{mean} \\ 30 & < & 42.5 & < & 77.1 \end{array}$$

Use the "average" that tells the story you want to tell and hope your listener won't know the difference.