6

Income Distribution

This chapter covers a lot of ground — two new kinds of average (median and mode) and ways to understand numbers when they come in large quantities rather than just a few at a time: bar charts, histograms, percentiles and the bell curve. To do that we introduce spreadsheets as a tool.

Chapter goals:

Goal 6.1. Work with mean, median and mode of a dataset.

Goal 6.2. Introduce the normal distribution, with its mean and standard deviation.

Goal 6.3. Understand how skewed distributions lead to inequalities among mean, median and mode.

Goal 6.4. Make routine calculations in Excel.

Goal 6.5. Use a spreadsheet to ask and answer "what-if" questions.

Goal 6.6. Create bar charts and other types of charts with a spreadsheet.

Goal 6.7. Use histograms to group and explore data.

Goal 6.8. Calculate averages for grouped data.

Goal 6.9. Understand the basics of descriptive statistics, including bell curve, bimodal data and margin of error.

6.1 Salaries at Wing Aero

Table **??** shows the distribution of workers' salaries at Wing Aero, a small hypothetical company. In keeping with our *Common Sense Mathematics* philosophy we should work with real data. But most companies keep this kind of information private. Any similarity between our imagined Wing Aero and any real company is purely coincidental. In Exercise **??** we'll apply the lessons we learn to look at income distribution in society at large.

The company has about 30 employees. We want to understand the salary distribution. A natural place to begin is with the average. But adding thirty numbers by hand (even with a calculator) is tedious and error-prone. A spreadsheet on your computer can do the arithmetic faster and more accurately. We have worked the examples in this text with Microsoft's Excel and with the free spreadsheet LibreCalc, available from `www.libreoffice.org/`. Online applications like Google sheets offer most of the features you will need. Use the examples here to work out how to access them.

We've organized this chapter as a spreadsheet tutorial — you can follow it step by step in Excel as you read it. If you are using LibreCalc you will find the same features. Sometimes we will provide screenshots from both.

See Section **??** for some general software tips and information about alternatives to Excel.

If you're online you can save typing time by downloading the Wing Aero spreadsheet from `WingAero.xlsx`. That spreadsheet and all the others you'll need live at file:///home/eb/Documents/csmlatest/
If you build it for yourself, use column `A` for the employees and column `B` for the salaries. Put the labels in row `7` and the data in rows `8:37`, not side by side as in the table. You should see Figure **??**.

What do you notice?

The employees are listed in decreasing order of importance (or prestige), but only approximately in decreasing order of salary. Some supervisors earn more than some managers, and some workers more than some supervisors. We can make those discrepancies visible by sorting the data.

Select the rectangular block of data in rows `8` through `37`, columns `A` and `B`. Be sure to select both columns so they will be sorted together. Choose the `Sort` dialog box from the `Data` tab, select sorting by Salary, `Largest to Smallest`, as in Figure **??**.

Often Excel offers you more than one way to do a job. This is one way to sort, in Excel 2013. There are others. Other versions of Excel may use different menus. But the ability to sort will be available in any spreadsheet program you use. Figure **??** shows how to do it in LibreCalc.

Employee	Salary (thousands of $)	Employee	Salary (thousands of $)
CEO	299	Supervisor	43
CTO	250	Supervisor	51
CIO	250	Supervisor	38
CFO	290	Supervisor	33
Manager	77	Supervisor	42
Manager	123	Supervisor	49
Manager	84	Worker	25
Manager	63	Worker	19
Manager	68	Worker	41
Manager	49	Worker	17
Manager	82	Worker	26
Manager	87	Worker	25
Supervisor	42	Worker	21
Supervisor	37	Worker	28
Supervisor	29	Worker	27

Table 6.1. Wing Aero salary distribution

If you sort the data again alphabetically (by Employee, `A to Z`) the table returns (nearly) to its original state, because the employee categories were alphabetical at the start. But each category is now sorted by salary.

To find the average salary at Wing Aero we tell Excel to add the entries in column `B` and divide by the number of employees.

Enter the label `Total` in cell `A38`. Then go to cell `B38`. Make sure the `Formula Bar` is visible. (Use the `View` menu to find it if it's not.) In the formula box type an equals sign `=`, to tell Excel you want it to do some arithmetic, and then the name of the operation,

```
=SUM(
```

since you are about to add up some numbers. The open parenthesis asks Excel to prompt you for information. It suggests

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

as in Figure **??**.

Select cells `B8:B37`, close the parentheses and type `enter` or click the check icon on the `Formula Bar`. You should see `Total 2315` in cells `A38` and `B38`. Wing Aero's total annual payroll is $2315K — about $2.3 million.

To find the average salary we must divide the total $2315K payroll by the number of employees. Rows `8` through `37` contain employee records so there are $37 - 8 + 1 = 30$

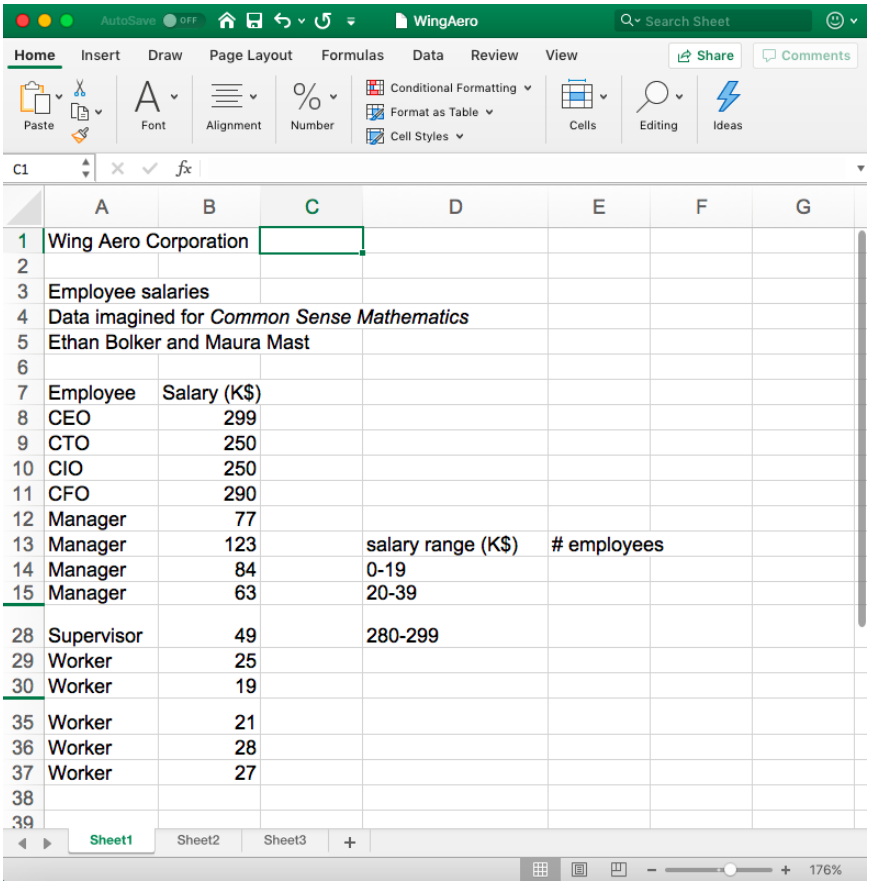


Figure 6.2. Wing Aero spreadsheet (some hidden rows)

employees. But it's better to ask Excel to count the rows for you. Type the label `Count` in cell `A39` and begin formula `=COUNT(` in cell `B39`. Finish the formula by selecting cells `B8:B37` or by typing the addresses of those cells and closing the parentheses. You should see

```
Count    30
```

Finally, type the label `Average` in cell `A40` and put formula `=B38/B39` in cell `B40`. You should see

```
Average    77.16667
```

so the average annual salary at Wing Aero is about $77,000. Excel rounded the exact 77.1666... to 77.16667 when it ran out of space. We rounded to two significant digits because that's all the precision we have in almost all the data. In Exercise **??** you'll learn how to tell Excel to round for you.

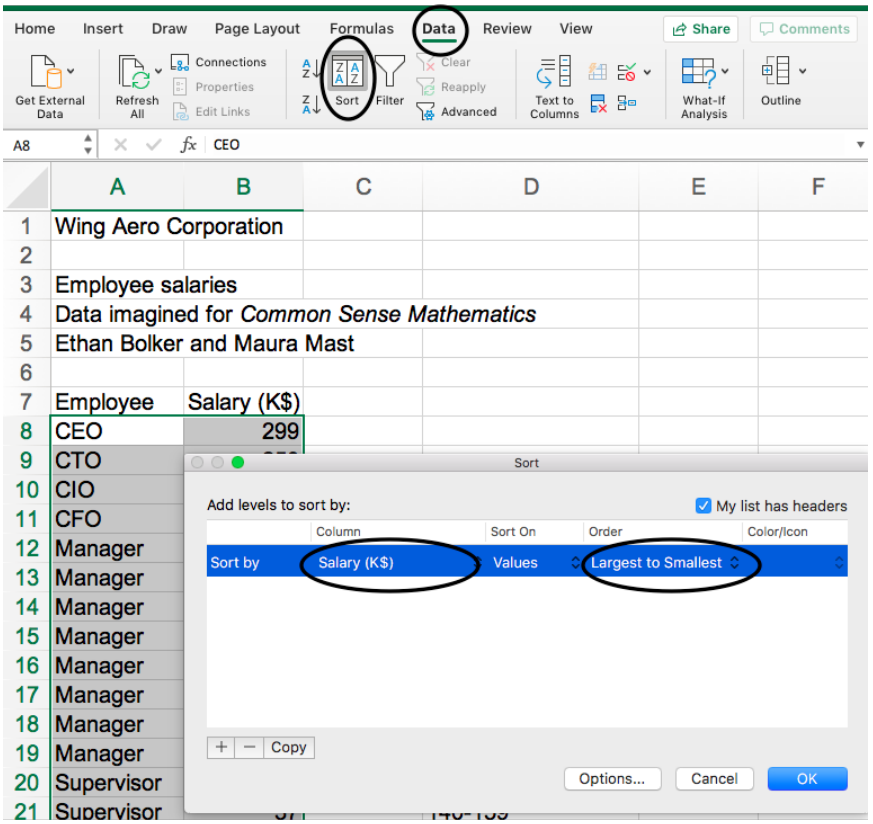


Figure 6.3. Sorting Wing Aero salaries (Excel)

Excel's built-in functions `SUM` and `COUNT` are separately useful, which is part of why we showed them to you, but if it's the average you want Excel can find it in one step.

Enter `=AVERAGE(B8:B37)` in cell `B41`, click the check icon and Excel tells you again that the average is 77.16667. Put "`computed using SUM/COUNT`" in cell `C40` and "`computed using AVERAGE function`" in cell `C41`.

6.2 What if?

Suppose that the CEO convinced the Board of Directors to double his salary, to \$598K (even though the company lost money). To see how that would affect the payroll statistics, go to cell `B8` and change the 299 there to 598. Excel automatically updates all your computations, increasing the total annual payroll to \$2,614 thousand and the average annual salary by about \$10,000 to more than \$87,000.

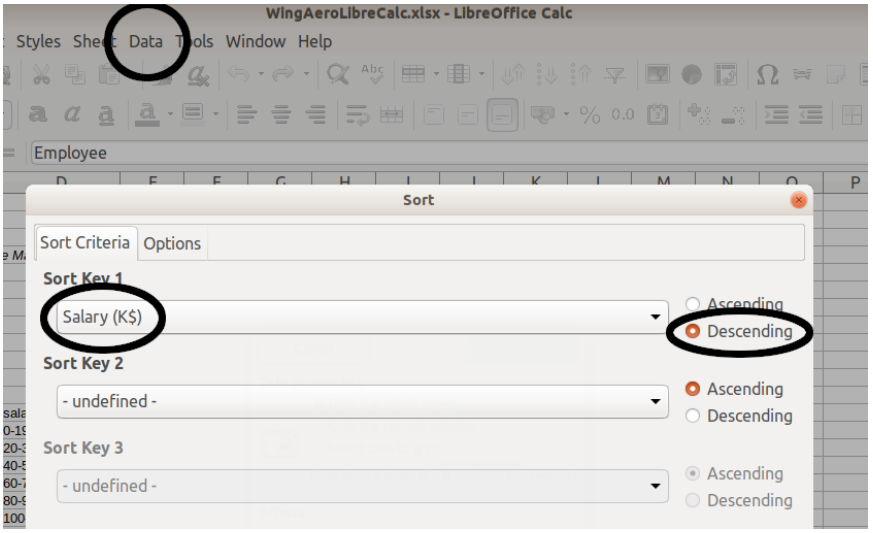


Figure 6.4. Sorting Wing Aero salaries (LibreCalc)

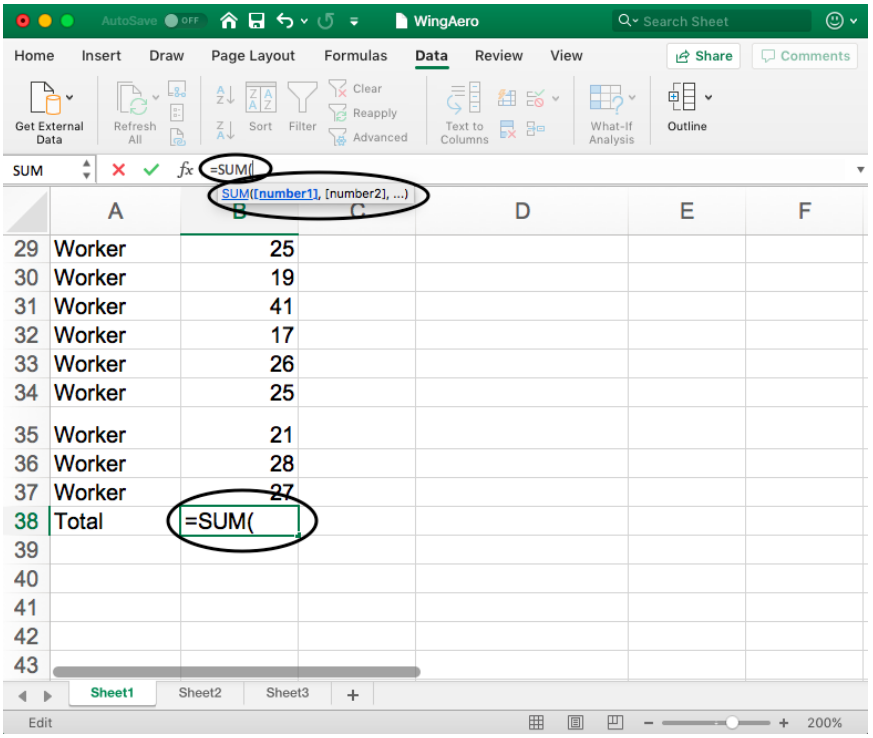


Figure 6.5. Summing Wing Aero salaries

We can learn several useful lessons from our work so far.

- Excel is faster and more accurate at arithmetic than we are.

- Excel is an excellent tool for answering what-if questions, because it automatically updates its computations when the data change.
- The average we calculated with `SUM/COUNT` and then again with `AVERAGE` is a terrible way to summarize the salary structure at Wing Aero. The CEO's 100% raise increases the average salary by about $10,000, or 13% — but he's the only employee who's actually better off!

After you've tried out these changes, restore the original values by clicking the `Undo` icon on the toolbar. This very useful feature allows you to undo the last few changes you've made. Use it to fix mistakes or to get back to where you were before you asked a "what-if" question.

6.3 Using software

Why use a spreadsheet at all? There are several reasons.

- to carry out large tedious calculations rapidly and correctly.
- to answer "what-if?" questions without having to redo arithmetic.
- to draw charts.
- to learn a tool you may use long after you've finished this course.

Here are some tips for working with Excel, and with software packages in general.

- To figure out something new, you can use the application's built in help, search the web, ask a friend or teacher, or just play around. Which you try first depends on your personality.
- Many applications provide several ways to do the same task. That means you may get different advice or instructions from different sources. Choose a way that suits your style.
- Use the menus for things you do rarely, but learn the keyboard shortcuts for things you do often — in particular, `control-C` for copy and `control-V` for paste can save time.
- Learn about `undo`. **Save your work often**.
- When you are about to make significant changes to a spreadsheet or a document make a copy of what you have, so that you can return to it if you change your mind about what you should have done. In the Wing Aero study we created several, calling them `WingAero1.xlsx`, `WingAero2.xlsx` and so on.
- Create backup copies of important documents often — off your computer. Use a thumb drive (flash drive), an external hardrive, or the cloud.

- Some things work the same way in different applications (browsers, Word, Excel) — for example, selecting with the mouse, cut and paste with keyboard shortcuts.
- In many software applications placing the mouse over a feature you are interested in and right clicking often lets you view and change the properties of that feature. "Do the right thing" is a good mnemonic.
- Applications often try to guess what you intend to do. That can be good or bad. We'll see soon that Excel can adjust cell references automatically — that's usually, but not always, what you want. Word processors may try to fix your spelling — perhaps correctly, perhaps not.

Excel and LibreCalc are full function spreadsheet programs. With Google Sheets (`www.google.com/sheets/about/`) you can create spreadsheets in the cloud. That software is powerful enough to do the arithmetic we need for *Common Sense Mathematics* but it has far fewer chart formatting features than full fledged programs. Spreadsheet applications on tablet computers lack those feature too. Don't even think of trying to do spreadsheet work on your phone.

6.4 Median

In Section **??** we found that the average salary at Wing Aero was $77,000. This is a pretty good annual salary. If you saw that in a job advertisement you'd think it was a pretty good company to work for. Maybe, maybe not. In Section **??** we saw how it's skewed by the CEO's earnings. When his (or her) salary increased from $299,000 to $598,000, the average salary jumped by $10,000 to $87,000. But no one else's salary changed!

The $77,000 "average" is misleading in other ways too. Most of the employees — 26 out of 30 — have salaries less than the average. That contradicts what we like to think "average" means. To find a salary that's "in the middle", sort the 30 line table again so that salaries are increasing. Since the table starts in row `8` and has 30 entries, rows `22` and `23` are the middle rows and the entries in cells `B22` and `B23` are the middle salaries. That means half the employees make $42,000 or less (the entry in cell `B22`) and half make $43,000 or more. So we might want to say that the "average" salary is $42,500. There's a name for this kind of "average" — it's the *median*. The first "average" we computed above is called the *mean*.

On the spreadsheet, change the `Average` label in cell `A41` to `Mean`.

Then put

```
Median   42.5   computed by finding middle of sorted list
```

into cells `A45`, `B45` and `C45`.

In some ways the median is a fairer "average" than the mean for describing the Wing Aero salary structure. It tells you more about the way salaries are distributed. In particular,

Job	Number	Total salary ($K)
Executive	4	1,089
Manager	8	633
Supervisor	9	364
Worker	9	229

Table 6.6. Wing Aero salary distribution by job category

the median salary isn't affected by the CEO's big raise. Try changing that salary again in Excel: the mean changes, as it did before, but the median stays the same.

Excel knows how to compute medians. Enter `=MEDIAN(B8:B37)` in cell `B46` and check that you get the same value: 42.5. Enter "`computed using MEDIAN function`" in `C46`.

Finding the median with the `MEDIAN` function is better than finding the middle of the sorted list yourself because it works even when the data aren't sorted. Suppose the supervisor making $43K gets a raise to $50,000. Enter that new value as `50` in the spreadsheet. Then see that Excel has recalculated the median in cell `B46`: it's now 45.5.

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.

6.5 Bar charts

Often pictures are better than numbers when we wish to convey information convincingly. A *bar chart* is one such picture.

We use bar charts when we have data that fall naturally into categories. The height of each bar represents the value for that category. When you want general understanding rather than numerical detail it's easier to compare the heights of bars visually (in both relative and absolute terms) than the values of numbers.

In order to understand the Wing Aero income distribution better we will use Excel to draw two bar charts with four columns each, one showing the number of employees in each of the four job categories executive, manager, supervisor and worker, the other the total earnings in each category. We will use the original Wing Aero salary data from Section **??**.

You can find the data from Table **??** in the range `D7:F11` in the copy of the Wing Aero salary distribution spreadsheet at `WingAeroBarCharts.xlsx`. Excel calculated the values in columns `E` and `F` using the `COUNT` and `SUM` functions. For Figure **??** we asked Excel to show us how it made those calculations.

Figure **??** shows the first two pictures from that spreadsheet.

These side by side bar charts dramatically demonstrate that the executives make up a small part of the workforce but enjoy a large part of the salary expenses!

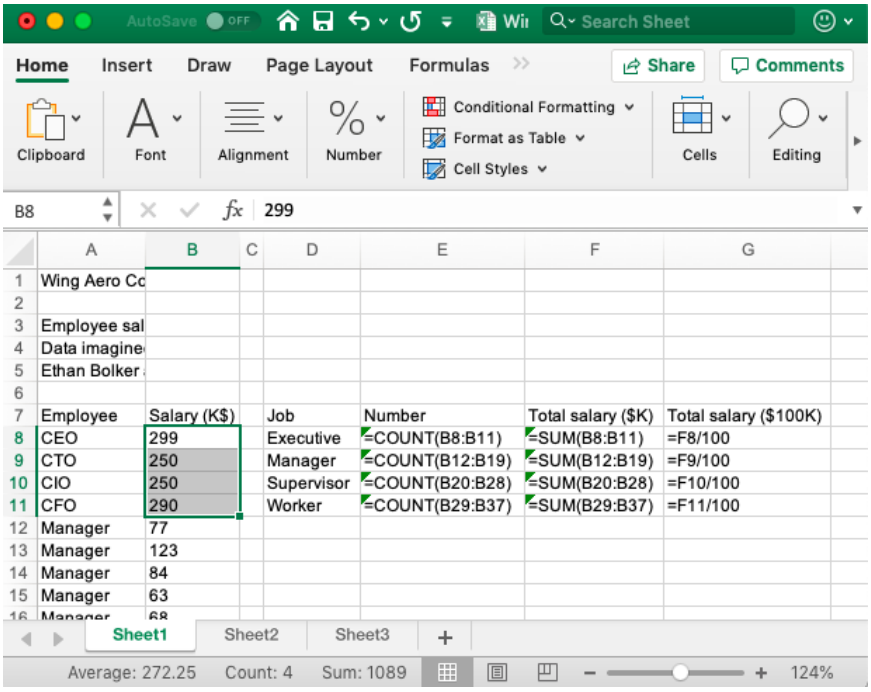


Figure 6.7. Showing the formulas used in a spreadsheet

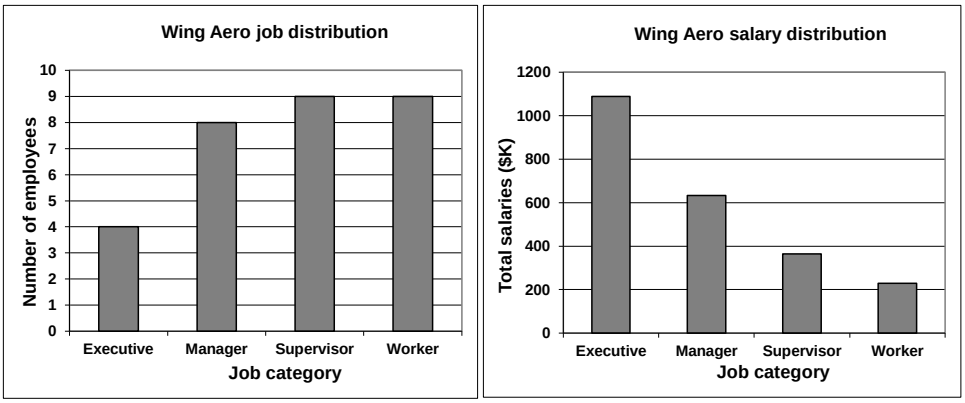


Figure 6.8. Wing Aero employee information by category

Delete the charts from your copy of the spreadsheet, so that you can learn how to build them for yourself. For the first one, select the data in columns `D` and `E`, rows `8` through `12` — the range `D8:E12`. Then ask for a new chart by selecting `Column` and `2D Column` from the `Insert` tab, as in Figure **??**. (Don't ask for a `bar` chart.)

Use the mouse to move the chart so that it does not hide any of the data. Then add the appropriate labels and change the colors so that the result is suitable for black and white printing. There are several ways to go about these tasks; experiment until you find ones

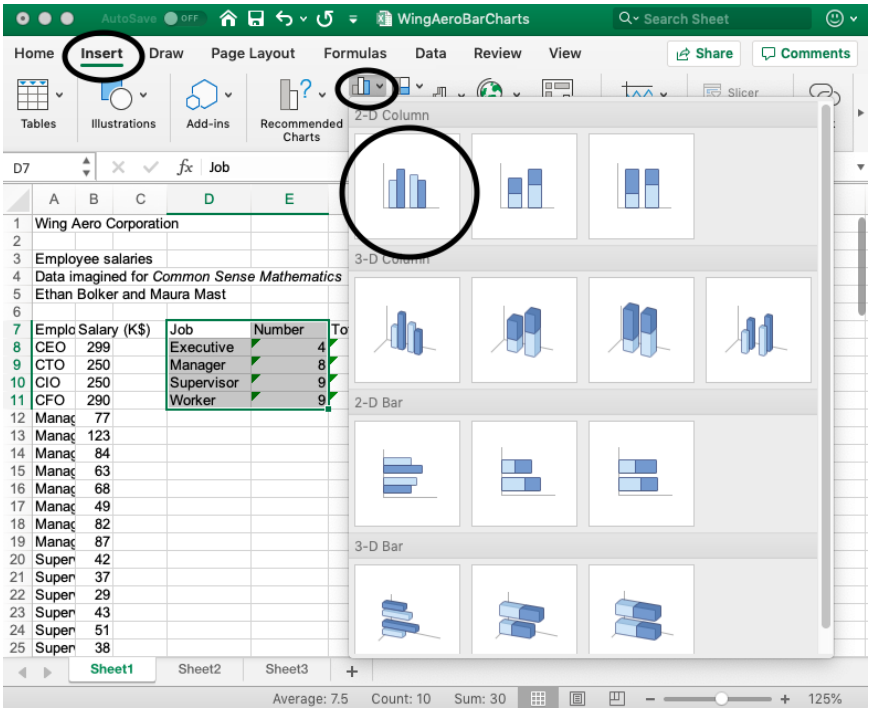


Figure 6.9. Inserting a chart in a spreadsheet

that work for you. Section **??** has a tip about how the right mouse button can help with these tasks.

To build the second picture the same way you must select the data in columns `D` and `F` without selecting column `E`. To do that, select rows `8:12` in column `D`. Then hold down the control (PC) or apple or command (Mac) key and use the mouse to select those rows in column `F`.

It would be even better to combine the two pictures. We can do that in Excel by building a column chart for the full range `D8:F12`. The result (after adding titles and fixing colors) is the chart on the left in Figure **??**.

The problem with that picture is that the bars for the employee numbers are nearly invisible, because data values for the numbers vary from 4 to 9 while those for the total salaries vary from \$200K to \$1200K. We can fix that by reporting the salary totals in hundreds of thousands of dollars rather than just in thousands of dollars (that is, by changing the units for measuring salary). Column `G` contains those numbers; we used it to draw the second picture. There you see clearly the opposing trends in the categories: total wages decrease as the number of employees increases.

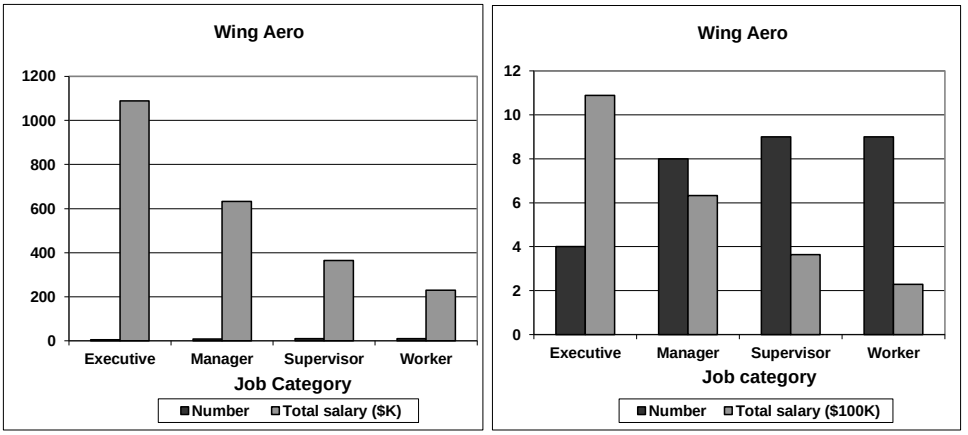


Figure 6.10. Wing Aero: side by side bar charts

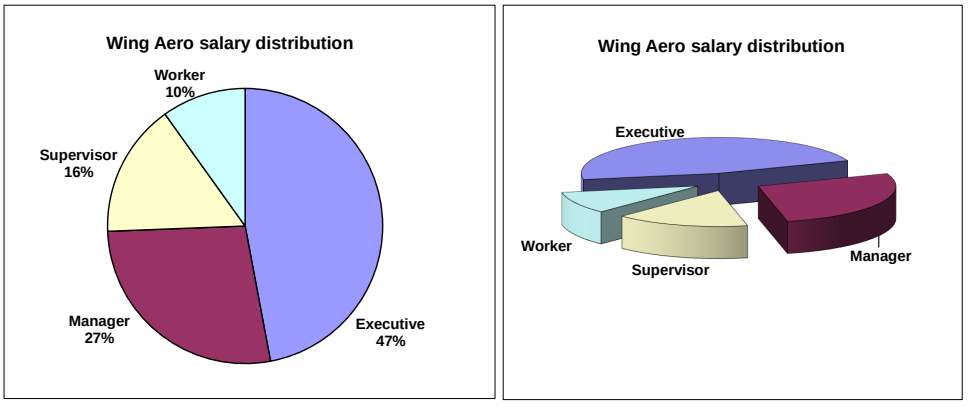


Figure 6.11. Wing Aero salary distribution pie charts

6.6 Pie charts

Excel allows you to change the chart type on the fly. Select the column chart showing the total salary by job category. Right click on that chart. Select `Change Chart Type ...` and then the first `Pie`. Adjust labels and colors to create the first of the charts in Figure **??**.

The pie chart shows clearly in yet another way that the executives are the winners at Wing Aero. They take home nearly half the salary total. If you wanted to make that look a little less dramatic, you could omit the percentages and ask Excel to show a three dimensional version of the chart, as in the second picture. There we've rotated the picture so that the executive wedge is at the back, so it looks even smaller in perspective.

Salary range ($K)	Number of 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	2
260-279	0
280-299	2

Table 6.12. Wing Aero salary distribution by salary range

6.7 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. The four categories in the previous section might not provide enough detail.

Another useful way to summarize the data is to divide the salaries into ranges and count the number of employees whose salary falls in each range. Then we can use the ranges as the categories in a bar chart. In this example we'll use $20K ranges. That means we think of two employees who make $29K and $33K as having approximately the same salary, 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. That's easy when the data are sorted in increasing order. Table **??** shows what we found.

To save you typing, we've listed the categories in cells `D15:D29` in `WingAero.xlsx`. You should check our work and enter the data in column `E`.

To see that you haven't missed anyone, `SUM` the range `E15:E29` to make sure the answer is 30, the known number of employees. (The sum isn't a perfect check. Although the total is correct, we might have put some employees into the wrong categories.)

We can use the data to draw a *histogram* — a bar chart where the categories on the x-axis specify data ranges and the y-axis counts or percentages for each range. You can see the resulting histogram in Figure **??**.

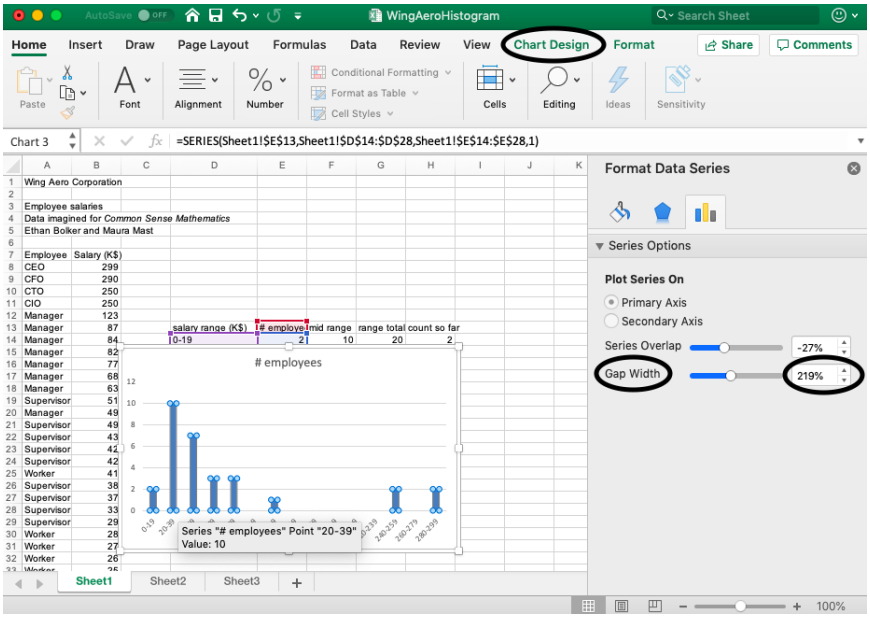


Figure 6.13. Formatting a histogram

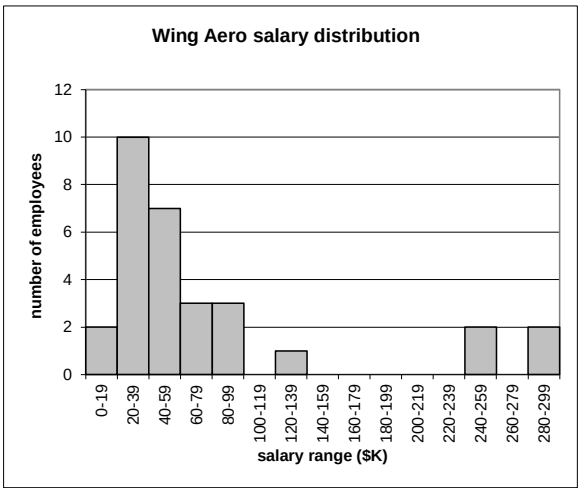


Figure 6.14. Wing Aero salary histogram

Start as usual by building a column chart from the data in cells `D13:E28`. In a histogram it's conventional to make adjacent bars touch. To do that in Excel, right or double click on one of the columns and look at the menu that appears. There you can change the `Gap Width` to 0, as in Figure **??**. While you're there, figure out how to change the colors of the bars, and fix the labels. You can see the full spreadsheet with all the computations we've done so far at `WingAeroHistogram.xlsx`.

It's worth taking some time to study this histogram, which shows how the data are distributed. Most of the salaries are less than $100K and there's a large gap in salaries between $139K and the executives who make more than $200K. Although this information is all in the table, the histogram makes it visible and dramatic.

We chose $20K for the size of the salary ranges so that we would have enough categories to show what was going on but few enough to make the graph understandable. In Exercise **??** you can explore what happens with different choices.

One of the themes of this chapter has been to use Excel for tedious repetitive calculations. So you might wonder whether Excel could build Table **??** for the grouped data if we told it the ranges we were interested in. Then the numbers would automatically update when we asked "what-if" questions. The good news is that Excel can do this job. The bad news is that its histogram building tools are rather clumsy. We will content ourselves with doing the counting by hand. If you're ambitious, try Excel help or search the internet for "excel histogram" to find out how to group data for an Excel histogram.

6.8 Mean, median, mode

There's a third kind of "average" — the mode — that's sometimes informative. The *mode* is the most common value. The histogram from the previous section 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 category with the highest bar.

The mode is most useful for data aggregated into ranges, as in a histogram. In the raw Wing Aero data there is no well-defined mode. Each of the values $250K, $49K, $42K and $25K appears twice. In Excel the function `MODE(B8:B37)` reports $250K when column B is sorted in descending order and $25K when the column is sorted the other way.

Each of "mean," "median" and "mode" can legitimately be called an "average." That ambiguity makes it easy to lie with statistics without actually lying. The CEO at Wing Aero may brag that workers at his company earn an average $77K per year, while the union argues that the average salary is $30K per year.

A cynic would advise you to use the "average" that tells the story your way and hope your listener won't know the difference.

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 cluster toward the left of the histogram with a long tail off to the right. For data that's right skewed, as this is:

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

In a histogram the mode is the peak, the median splits the area in half and the mean is where the graph would balance if it were a cardboard cutout.

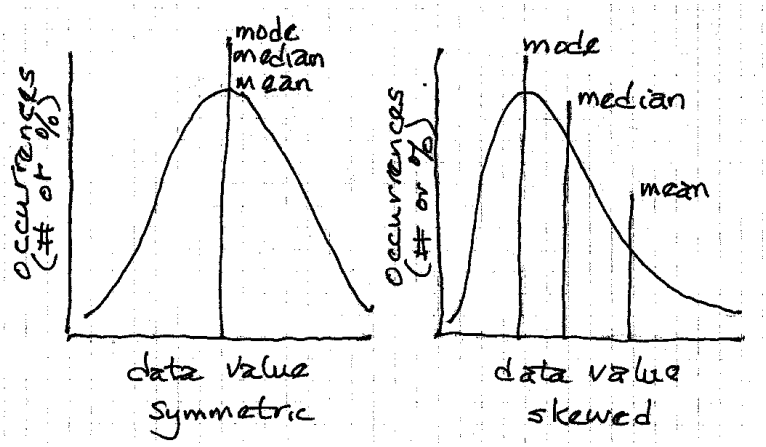


Figure 6.15. Symmetric and skewed distributions

If you're learning about these several kinds of averages for the first time you may want mnemonics to help you remember which is which. The "med" in median suggests correctly that it's in the middle. You can remember mode because there are "mo' of them than anything else."

6.9 Computing averages from histograms

Often all we know about data is a summary like that presented in a histogram. We'll see here how to estimate the mode, median and mean from that information. We will use the Wing Aero histogram in Figure **??** as an example. Since we know the correct averages we can see how good our estimates are.

We've already found the mode. It's the highest bar in the histogram — the range with the largest entry: \$20,000 — \$39,000. Note carefully — the mode is the range, not the height of the bar. We could report that range, or report the mode as the middle of the range: about \$30,000. The raw data do not have a mode that makes sense, so there's nothing to compare this estimate to.

To estimate the median salary from the histogram we need to find the salary such that half the employees make less and half more. The first bar tells us that two make less than \$19K. Adding the number of employees in the first two ranges tells us that $2+10=12$ make less than \$39K. Looking at the next range we see that $2+10+7=12+7=19$ make less than \$59K. Since there are 30 employees, it's the 15th and 16th whose salaries are closest to the median. They are clearly pretty much in the middle of the third category, so we can estimate the median as the midpoint of that category, say \$50,000. The correct value (from the raw data) is \$42,500.

Since there are just eight ranges with data and the median occurred in the third one, we didn't need Excel to do the arithmetic. In a more complicated example things might not be so easy, so let's see how to make Excel do the work. Label column `H` in cell `H15` as

"`count so far`". Then copy the value from `E15` to `H15` by typing `=E15` in H15. Then enter `=H15+E16` in `H16`, to add the number of employees counted so far (in `H15`) to the number in this range (in `E16`).

Now select that formula and copy it to `H17:H29`. A miracle has happened! Cell `H29` contains the value 30. If you click on that cell you can see that it's the result of the formula `=H28+E29`. Excel read your mind, and automatically updated the row references to cells in columns `H` and `E` with each copy down the column. It knew you wanted to add the value in the cell above to the value in the cell three over to the left.

The last entry should be 30 since all the employees make less than $299K. Now it's easy to see that the count so far passes the midpoint of 15 in the middle of the third range.

Estimating the mean salary is the hardest. Since the only Wing Aero data we have is what we used to draw the histogram we can't expect to find it exactly. All we can say about the 10 employees in the $20K — $39K range is that they earn somewhere in the neighborhood of $30K. So our best guess is to assume they all earn exactly that. If we make a similar assumption for each of the other ranges then our estimate for the mean is the weighted average

$$\frac{2 \times \$10K + 10 \times \$30K + \cdots + 2 \times \$290K}{\text{total number of employees}}.$$

That's too much arithmetic to do by hand so we'll use Excel.

Put the label `mid range` in cell `F13`. We want to use cells `F14:F28` to hold the values 10, 30, ..., 290 that are the middles of the ranges in cells `D14:D28`. There's a quick trick for that. Enter the 10 in cell `F14` and enter the formula

```
=F14+20
```

in cell `F15` . Excel will display `30` there. That's because it reads the formula as

add 20 to the contents of cell `F14`.

Now we want to add 20 each time you move down a row. To do that, copy the formula in `F15` and paste it into cells `F16:F28`. (This takes advantage again of Excel's correct guess about what we are trying to do.)

Next label column `G` by typing `range total` in cell `G13`. Then put the formula

```
=E14*F14
```

in cell `G14`. That asks Excel to multiply the numbers in cells `E14` and `F14`. You should see 20. That's the first number to add in the weighted average computation we're working on.

When you copy that formula to cells `G15:G28` you should see 580 at the end of the list. That's the miracle yet again.

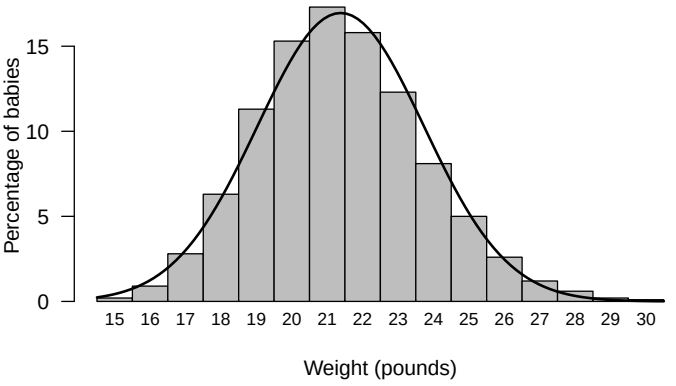


Figure 6.16. Baby weights

To compute the weighted average you must sum the values in column `G`. Since cell `E29` contains the sum of the values in column `E`, just copy the formula from that cell to cell `G29`. Excel will automatically change the column reference, turning the formula `=SUM(E14:E28)` into `=SUM(G14:G28)`. The sum is 2360. To find the mean, enter the formula `=G20/E29` in cell `G30`. Excel shows you 78.66667. Label that value as the mean.

Our estimate of the mean salary from the histogram is $79,000. We shouldn't report more precision than that, since we made many approximations along the way. Do note that the estimate is not very far from the true mean of $77,167 computed from the raw data.

6.10 The bell curve

We used the data on baby weights in Table **??** to study percentiles. We constructed the histograpm Figure **??** shows the histogram we constructed from that table. The highest bar shows that the mode — the most common weight for male babies one year old in 2019 was about 21.5 pounds. In Section **??** we found the median to be about the same. Using the techniques from Section **??** we computed the mean as a weighted average. It is 21.4 pounds.

That suggests that the distribution is just little bit skewed to the right.

Many factors contribute to a baby's weight at one year: birth weight, heredity, nutrition, ... In general, when many small effects combine to give a total, the distribution of values forms a *bell curve*. The one shown in the figure is a mathematically correct bell curve that approximates the real data. The mathematically correct name for the bell curve is *normal distribution*.

value	percentile
$\mu - 3\sigma$	0.1
$\mu - 2\sigma$	2.3
$\mu - \sigma$	15.9
μ	50.0
$\mu + \sigma$	84.1
$\mu + 2\sigma$	97.7
$\mu + 3\sigma$	99.9

Table 6.17. Percentiles for the normal curve, mean μ, standard deviation σ

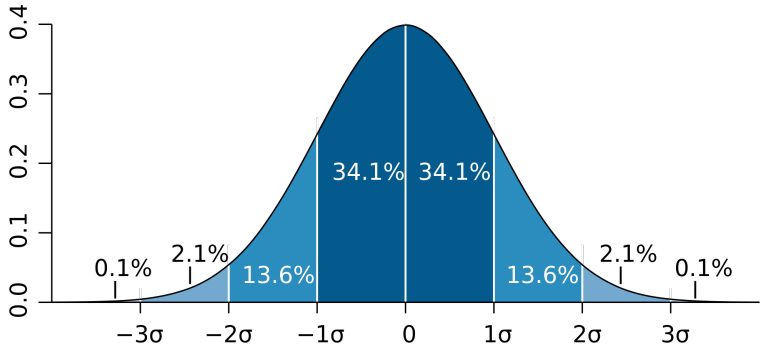


Figure 6.18. How the normal distribution spreads out [R**??**]

A normal distribution is always symmetrical. Its mean, median and mode are in the same place. To construct the one in the figure we need one more number besides the mean — a measure of how fast the curve spreads out. That measure is called the *standard deviation*. It's usually written with the Greek letter sigma: σ; the mean is usually written with the Greek letter mu: μ. For the baby weight data the standard deviation is about 2.6 pounds.

There's a nice way to use the standard deviation to describe how a bell curve spreads out.

- 2/3 of the values are less than one standard deviation away from the mean,
- 95% of the values are less than two standard deviations away,
- 99.7% are less than three standard deviations away.

Figure **??** illustrates these percentages. Table **??** summarizes them in terms of percentiles.

Since the baby body weight distribution is very close to normal, with mean μ about 21.4 pounds and standard deviation σ about 2.4 pounds, we know about 2/3 of the babies weighed between $\mu - \sigma = 21.4 - 2.4 = 19.0$ and $\mu + \sigma = 21.4 + 2.4 = 23.8$ pounds. Approximately 95% weighed between 16.6 and 26.2 pounds. 2.5% weigh more than 28.6 pounds and 2.5% weighed less than 14.2 pounds.

Figure **??** shows three bell curves with the same mean $\mu = 21.4$, but three different standard deviations, $\sigma = 1.2, 2.4$ and 4.8. The middle one matches the baby weight data.

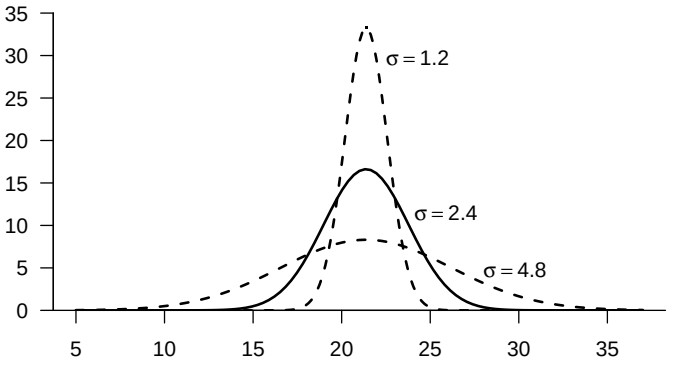


Figure 6.19. Three bell curves

The mathematics needed to calculate standard deviations and draw bell curves is more than we will present here. You will learn about it if you take an introductory course in statistics. All you should remember now is the rough relationship between standard deviation and spread we sketched above and that the official name for the bell curve is *normal distribution*.

Sometimes normal distributions are hidden in data. The black points in Figure **??** plot how the rate of diagnosis (in cases per 100,000 people) of Hodgkin lymphoma (a kind of cancer) for white females depends on the age of the woman diagnosed. This graph is *bimodal*: there are two peaks, one at about 20 years, the other at 75 years. There is no single value that can legitimately be called the mode. The mean and the median would each be about 45 years, but they make no sense at all. Even though they are "averages" there are few 45 year olds with the disease. The disease probably has two very different causes, one of which occurs more often in young people, the other in old people.

We can understand this distribution as a combination of two normal distributions. The left bell curve for the early onset of Hodgkin lymphoma has a mean of about 24 years with a standard deviation of 11 years. The right bell curve for late onset has a mean of about 79 years with a standard deviation of 19 years — it spreads out more slowly. The smooth curve is the sum of the two normal distributions. It matches the data very well.

6.11 Margin of error

The Pew Research Center conducted a study in July of 2012 that asked about support for President Obama's tax position. Their report said (in part)

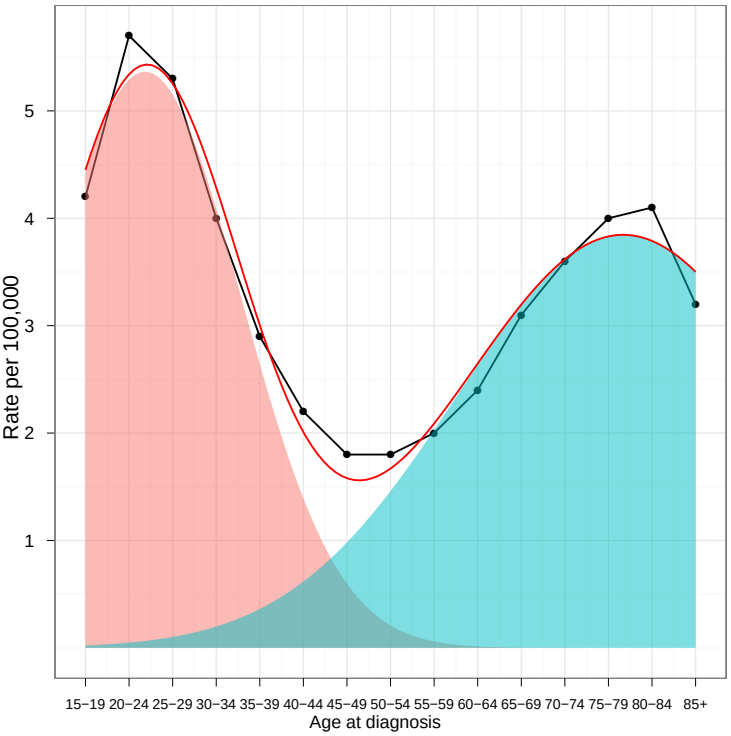


Figure 6.20. Hodgkin lymphoma incidence (white females) [R**??**]

> By two-to-one (44% to 22%), the public says that raising taxes on incomes above $250,000 would help the economy rather than hurt it, while 24% say this would not make a difference.
>
> [The poll reached 1,015 adults and has a margin of sampling error of plus or minus 3.6 percentage points.] [R**??**]

The first paragraph quoted above reports the results of a survey. The second tells you something about how reliable those results are. It's clear that the smaller the margin of error the more you can trust the results. Understanding the margin of error quantitatively — seeing what the number actually means — is much more complicated. A statistics course would cover that carefully; we can't here. Since the term occurs so frequently, it's worth learning the beginning of the story. Even that is a little hard to understand, so pay close attention.

The survey was conducted in order to discover the number of people who thought the tax increase would benefit the economy. If everyone in the country offered their opinion then we would know that number exactly. If we gave the survey to just three or four people we could hardly conclude anything. The people at the Pew Research Center decided to survey the opinions of a *sample* of the population — 1,015 people chosen at random. Of the particular people surveyed, $0.44 \times 1,015 = 447$ people thought the tax increase would benefit the economy. If they'd surveyed a different group of 1,015 people, they would probably see a different number, so a different percent.

The 3.6 percentage point margin of error says that if they carried out the survey many times with different samples of 1,015 people, 95% of those surveys would report an answer that was within 3.6 percentage points of the true value.

There's no way to know whether this particular sample is one of the 95%, or one of the others. About five of every 100 surveys you see in the news are likely to be bad ones where the margin of error surrounding the reported answer doesn't include the true value. Survey designers can reduce the margin of error by asking more people (increasing the sample size). But all that can do is reduce the margin of error.

The report doesn't explicitly mention 95%. That's just built into the mathematical formula that computes the margin of error from the sample size. Even that conclusion may be too optimistic. The margin of error computation only works if the sample is chosen in a fair way, so that everyone is equally likely to be included. If they asked 1,015 people at random from an area where most people were Democrats (or Republicans) or rich (or poor) the result would be even less reliable. The report describes the efforts taken to get a representative sample.

6.12 Exercises

A spreadsheet is just a tool. It doesn't answer questions, it provides numbers and pictures you use to answer questions. So keep on writing complete sentences explaining the meaning and context of the numbers you report. The numbers in cells in a spreadsheet are no more use by themselves than the numbers in a calculator display.

If an exercise asks for hard copy of a spreadsheet, format it well. Be sure to look a preview before you print to make sure that charts fit on one page and don't cover important numbers. Label the data columns, cells containing important computations, axes and legends in charts. Numbers are useless when they can't be understood.

Exercise 6.12.1. [S][R][Section **??**] CXO.

What do "CEO", "CTO", "CIO" and "CFO" stand for in the Wing Aero salary table in Section **??**?

In CXO the C and the O stand for Chief and Officer. The middle letter stands for Executive, Technology, Information (or sometimes Investment) and Financial. Sometimes you see COO (Operating).

"CIO" also stands for "Congress of Industrial Organizations," a federation of labor unions, founded in 1935.

Exercise 6.12.2. [S]Section **??**[Section **??**][Goal **??**][Goal **??**] What if...?

Open up the original Wing Aero spreadsheet and use your spreadsheet to calculate the mean and median, as we did in Section **??**. For each exercise that follows, write a clear statement - using the numbers - that summarizes what you found.

(a) Suppose all the managers get a $10K raise. Change their salaries and see how Excel updates the mean and median.

(b) Go back to the original spreadsheet (use the `Undo` feature if you can). Experiment with changing salaries (of any of the workers — your choices) so that the mean and the median increase. Explain how you made your choices.

(c) Reset back to the original spreadsheet. How would you change salaries so that the mean decreases but the median stays the same?

(d) Reset back to the original spreadsheet again. How would you change salaries so that the mean stays the same but the median decreases? Which salaries did you change?

(a) Suppose all the managers get a $10K raise. Change their salaries and see how Excel updates the mean and median calculation.

The mean increased to about $79.8K. The median stayed the same, since all the people who got this raise were already in the top half of the list.

(b) Go back to the original spreadsheet (use the `Undo` button if you can). Experiment with changing salaries (of any of the workers — your choice) so that the mean and the median increase.

One way to do this is to increase the salary of any of the workers making less than the median to a value greater than the median. For example, give the one making $25K a raise to $75K. Now the mean is $78.8K and the median is $46K

(c) Reset back to the original spreadsheet. How would you change salaries so that the mean decreases but the median stays the same?

Easy. Just cut the CEO salary in half.

(d) Reset back to the original spreadsheet again. How would you change salaries so that the mean stays about the same but the median decreases?

To decrease the median I need to move someone from below $42.5K to above. I can do that with the worker making $19K — raise his salary by $50K to $69K. If I then take $50K away from the CEO the mean will be the same.

Exercise 6.12.3. [S][R][Section **??**] [Goal **??**] Formatting in Excel.

Format the cells containing averages in the Wing Aero spreadsheet so that the numbers displayed for the various averages are rounded to the nearest thousand dollars (no decimal places).

There are several ways to find the menu that allows you to set the precision your spreadsheet uses for numbers. One way in Excel is to right click on the cell with the number (or select a range of cells and then right click). Choose `Format Cells ...` from the menu, then `Number` from the `Category:` menu on the left. Then enter `0` for the number of decimal places.

Exercise 6.12.4. [S][Section **??**][Goal **??**][Goal **??**] [Goal **??**] Practice finding the median.

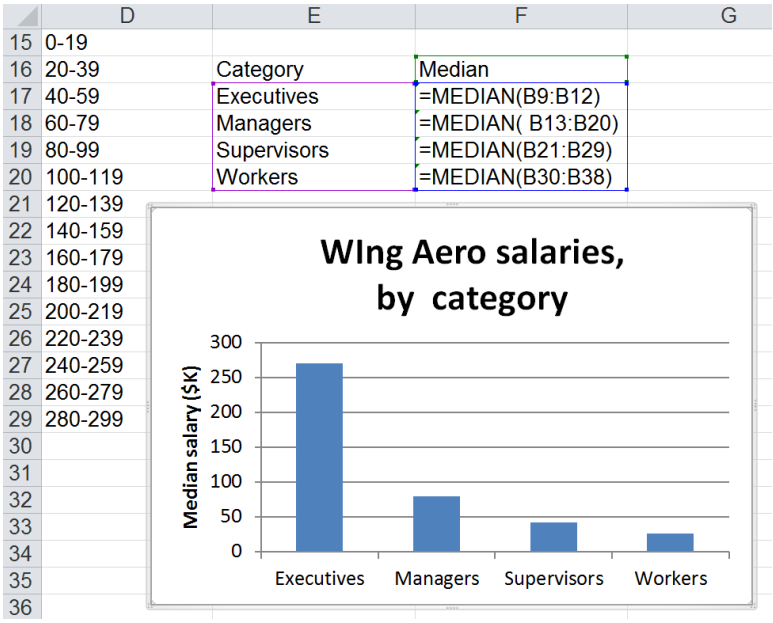
Open up the original Wing Aero spreadsheet from `WingAero.xlsx`.

(a) Use Excel to find the median salaries for each category of employees (workers, managers, etc.).

(b) Show the data in a properly labeled bar chart.

(c) Do you think the median is a representative "average" for each category? Explain.

Here's a screen shot showing my work, from the spreadsheet `../Answers/WingAeroMedianSolution.xls`



The median is a reasonable representation of the "average" salary in each category.

Exercise 6.12.5. [S][Section **??**][Goal **??**][Goal **??**] Averaging averages.

(a) Find the average (mean) salary at Wing Aero for each of the four categories of employees (use the original data set, from Section **??**).

(b) Show the data in a properly labeled bar chart.

(c) Compute the weighted average of these averages to check that you get the correct mean for the whole payroll.

[See the back of the book for a hint.] You may draw your bar chart neatly by hand, or use Excel.

(a) Find the average (mean) salary at at Wing Aero for each of the four categories of employees.

The following table shows the average salaries by category. I built the spreadsheet at `../Answers/WingAeroAverageByCategorySolution.xlsx` to find these numbers.

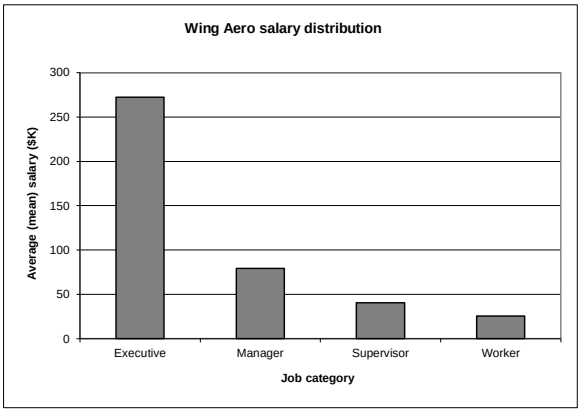


Figure 6.21. WingAero average salaries, by category

Job	Number	Total salary ($K)	Average salary ($K)
Executive	4	1089	272.25
Manager	8	633	79.13
Supervisor	9	364	40.44
Worker	9	229	25.44

(b) Show the data in a properly labeled bar chart.

See Figure **??**.

(c) Compute the weighted average of these averages to check that you get the correct mean for the whole payroll.

The weighted average is

$$\frac{4 \times 272.25 + 8 \times 79.13 + 9 \times 40.44 + 9 \times 25.44}{30} = 77.61.$$

That's the same value as the ordinary average of the original data.

Exercise 6.12.6. [S][Section **??**][Goal **??**] [Goal **??**] Cash-strapped T proposes 23 percent fare increase.

On March 29, 2012 the Massachusetts Bay Transportation Authority (MBTA) provided the fare data in Table **??**. Riders can pay with a stored-value Charlie Card or with a Charlie Ticket bought on the spot.

(a) What are the relative and absolute changes in the Charlie Card bus fare?

You don't have to do this in your head, without a calculator, but you should be able to.

(b) Create a spreadsheet for these data, with eight rows (one for each of the eight categories) and three columns, for the category name, the existing fare and the proposed fare. Create a properly labeled bar graph to display the data.

(c) Label the next two columns appropriately to hold the relative and absolute changes. Fill those columns with Excel formulas to compute the correct values. Do not compute the values elsewhere and enter them in the spreadsheet as numbers.

Fare Category	Current	Proposed
Charlie Card		
Bus	$1.25	$1.50
Subway	$1.70	$2.00
Senior Bus	$0.40	$0.75
Senior Subway	$0.60	$1.00
Student Bus	$0.60	$0.75
Student Subway	$0.85	$1.00
Charlie Ticket		
Bus	$1.50	$2.00
Subway	$2.00	$2.50

Table 6.22. MBTA fare increases [R**??**]

[See the back of the book for a hint.] If you want the absolute change in column `E` then put formula

```
=D7-C7
```

and copy the formula to the rest of the rows.

For the relative increase, divide instead of subtracting. Then subtract 1 to get the percentage increase.

(d) Imagine that you are addressing a public meeting about these fare increases. How would you argue that an unfair burden is being placed on people who pay using a Charlie Ticket? How would you argue that senior citizens are most hard hit?

(e) Find the mean of the relative percent increases. Explain why the answer is not the 23 percent quoted in the headline.

(f) What extra information would you need to check that the correct mean is 23 percent?

(g) (Optional) Find out why the stored value card is called a Charlie Card.

(h) (Optional, and difficult) Create a column for the percentage of MBTA revenue for each of the categories and fill in some values that sum to 100% and give a weighted average fare increase of about 23%.

There's an important and subtle distinction here between weights as a percentage of revenue and weights as a percentage of trips.

(a) What are the relative and absolute changes in the Charlie Card bus fare?

The absolute change is $0.25; the relative change is 20%. I did them in my head.

(b) Create a spreadsheet for these data, with eight rows (one for each of the eight categories) and three columns, for the category name, the existing fare and the proposed fare. Create a properly labeled bar graph to display the data.

My spreadsheet with its chart is at `../Answers/MBTAFareIncreasesSolution.xlsx`.

(c) Label the next two columns appropriately to hold the relative and absolute changes. Fill those columns with Excel formulas to compute the correct values. Do not compute the values elsewhere and enter them in the spreadsheet as numbers.

(d) In my solution the absolute and relative changes are in columns `E` and `F`. I put these formulas in cells `E7:F7`:

```
=D7-C7          =D7/C7
```

I put the percent increase in cell `G7` with the formula

```
=F7-1
```

and formatted the result as a percent.

Then I copied these formulas to the rest of the rows.

(e) Imagine that you are addressing a public meeting about these fare increases. How would you argue that an unfair burden is being placed on people who pay using a Charlie Ticket? How would you argue that senior citizens are most hard hit?

The Charlie Ticket riders have the largest absolute change among all the categories: 50 cents. That's not fair.

The senior citizens have the largest two relative changes among all the categories: 67% for subway fares and a whopping 88% for the bus. That's not fair.

(f) Find the mean of the relative percent increases. Explain why the answer is not the 23 percent quoted in the headline.

According to Excel, the mean of the relative percent increases is 37%. The correct mean is a weighted average of the percentages in each category, with weights the percentage of revenue in each category.

(g) What extra information would you need to check that the correct mean is 23 percent?

To check that the *Globe* reported the correct mean I would need to know the percentage of rides in each category, or the percentage of revenue for each category.

(h) (Optional) Find out why the stored value card is called a Charlie Card.

In 1949, the MBTA proposed a 5 cent increase in the charge for customers exiting trains above ground. This fare increase prompted the composition of the song "M.T.A." (or "Charlie on the M.T.A." as most people refer to it) about a fictional subway rider named Charlie who was doomed to ride forever because he didn't have that nickel to get off of the T. The MBTA website `www.mbta.com/about_the_mbta/history/?id=19582` has a short description. When the MBTA introduced plastic fare cards they named them "Charlie cards".

The Kingston Trio made the song famous in 1959.

(i) (Optional, and difficult) Create a column for the percentage of MBTA revenue for each of the categories and fill in some values that sum to 100% and give a weighted average fare increase of about 23%.

They are in the spreadsheet.

Class	Percentage
Freshman	40
Sophomore	25

Table 6.23. Freshman and sophomore enrollments

Exercise 6.12.7. [S][Section **??**][Goal **??**] Why not pie charts?

Do some internet research to discover why bar charts are usually better than pie charts. Write the reasons in your own words (don't just cut and paste). Identify the sources of your information and comment on why you think those sources are reliable.

Searching for

are pie charts bad

in October 2014 finds these first five websites

- `www.businessinsider.com/pie-charts-are-the-worst-2013-6`
- `www.stevefenton.co.uk/Content/Pie-Charts-Are-Bad/`
- `www.datavis.ca/gallery/evil-pies.php`
- `www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00018S`

Exercise 6.12.8. [A][U][Section **??**][Goal **??**] Misleading pie charts.

Table **??** shows some student enrollment data at a college. It's clearly incomplete: Juniors and Seniors are missing.

Flashy pie charts are common in the media, but staid and boring bar charts are usually more informative and less deceiving. Because Excel makes it so easy to switch to a pie from a bar, designers may be tempted by the glitz. You should not succumb to that temptation. Here's an exercise where you can find out why.

(a) Construct a bar chart displaying these data, with columns for Freshman and Sophomore enrollments.

(b) Change the chart type to `pie` in your spreadsheet.

(c) Explain what is wrong with the new chart.

Exercise 6.12.9. [S][W][Section **??**][Goal **??**] [Goal **??**] What if?

Add five workers each earning \$18K to the original Wing Aero payroll by inserting some rows in the table. Comment on what happens to the mean, median and modal incomes.

Note that Excel automatically recomputes these averages, but not the charts for which you created data by hand.

When you've done this exercise, undo your changes and check that the three kinds of averages revert to their old values.

When I add five workers earning \$18K each the mean falls from \$77.16K to \$66.71K, the median from \$42.5K to \$41K. The mode is now \$18K. (The value of the mode before the new workers was ambiguous. What Excel says it is depends on the order in which the employees are listed. In general you should report modes only for grouped data.)

Exercise 6.12.10. [S][Section **??**][Goal **??**][Goal **??**] Population pyramids.

At `www.census.gov/data-tools/demo/idb/informationGateway.php` you can choose a country and a year, then ask for a kind of bar chart known as a *population pyramid*. You can also download the data used to build the chart.

(a) Construct population pyramids for the United States and for Sudan for the year 2010.

(b) Estimate the number of people in the United States in 2010 between 0 and 9 years old. Do the same for Sudan. What fractions of the populations do these numbers represent?

(c) Find the modal age range for the United States population. Do the same for Sudan.

(d) Find the age range for the United States that has the smallest population. Do the same for Sudan.

(e) Compare the population distributions of the United States and Sudan. Write several sentences that highlight aspects of each distribution that you think are quantitatively significant. Use the results of the previous part of the problem.

(a) Construct population pyramids for the United States and for Sudan for the year 2010.

See Figures **??** and **??**.

(b) Estimate the number of people in the United States in 2010 between 0 and 9 years old. Do the same for Sudan. What fractions of the populations do these numbers represent?

I did Sudan first.

To find that number I need to add the bottom two bars on each side, for the boys and girls age 0-4 and 5-9. That looks to be about $3.6+3.2+3.5+3.0 = 13.1$ million.

Then I downloaded the underlying data tables as an Excel spreadsheet and checked my estimate. The number of 0-9 year olds in Sudan is $7,018,343+6,128,436$ which is indeed about 13.1 million.

The total Sudan population is 43.9 million (from the data table). The fraction of children is thus $13.1/43.9 = 0.298405467 \approx 30\%$.

The corresponding figures for the United States are about $11+11+10+10 = 42$ million children. That's about $42/310 \approx 13.5\%$.

(c) Find the modal age range for the United States population. Do the same for Sudan.

The modal age is the largest bar. For the United States that's 45-49 years. For Sudan it's 0-4 years!

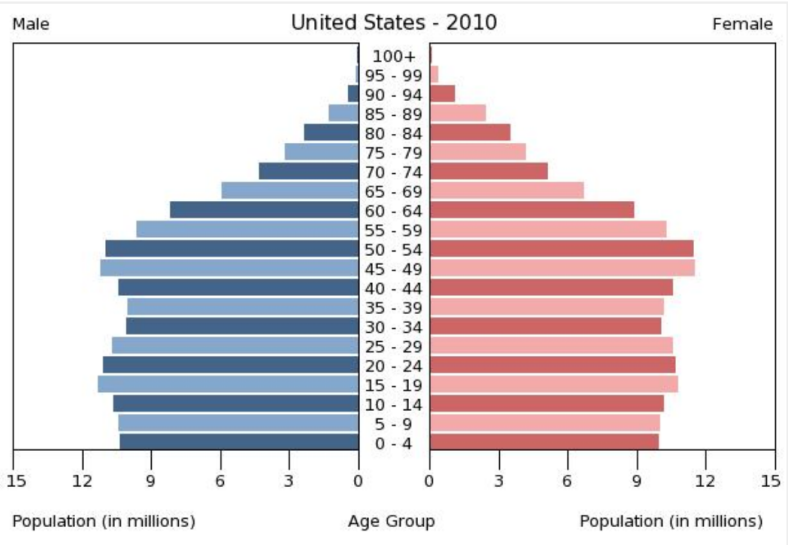


Figure 6.24. U. S. population pyramid [R**??**]

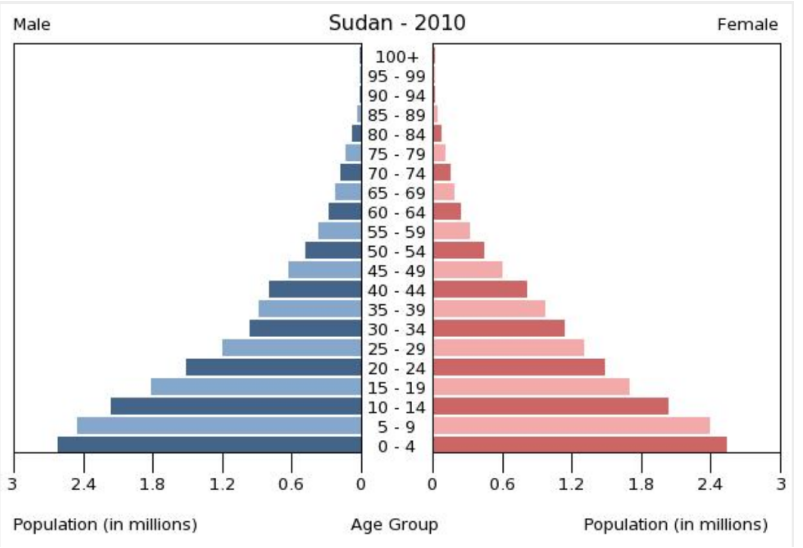


Figure 6.25. Sudan population pyramid [R**??**]

(d) Find the age range for the United States that has the smallest population. Do the same for Sudan.

In each population the smallest age range is 100+. That's not surprising, and also not very interesting.

(e) Compare the population distributions of the United States and Sudan. Write several sentences that highlight aspects of each distribution that you think are quantitatively significant. Use the results of the previous part of the problem.

The population pyramid for Sudan really looks like a pyramid. The number of people in each five year cohort decreases steadily with age. That indicates a very large birth rate, with lots of people dying young. It also spells disaster for the society — there aren't lots of adults to take care of the children.

The United States population pyramid looks more like a house: vertical walls with a slanted roof on top. That suggests that the population structure is relatively stable, with few deaths at young ages. It also looks likely that the little bulge at 45-55 years will lead to a large number of senior citizens in not too long. That's a social problem of a different kind.

Exercise 6.12.11. [S][Section **??**][Goal **??**] Lake Wobegon.

(a) Look back at the Wing Aero data and answer the following questions:

- What percentage of the employees make more than the mean salary?
- What percentage of the employees make more than the median salary?
- What percentage of the employees make more than the mode salary?

(b) In his radio show *A Prairie Home Companion* host Garrison Keillor regularly tells his audience about Lake Wobegon, where "all the children are above average".

Is that possible, with any of the meanings of "average"?

(a) Look back at the Wing Aero data and answer the following questions:

- What percentage of the employees make more than the mean salary?
  8 of the 30 employees make more than the mean of \$77K. That's $8/30 = 0.266 = 27\%$.
- What percentage of the employees make more than the median salary?
  50%! That's the meaning of "median" so it's the answer to this question even if I don't know the value of the median.
- What percentage of the employees make more than the mode salary?
  This is a slightly subtle question. The mode only makes sense for grouped data. The figure in the text shows that the mode for Wing Aero is about \$30K. 21 of the 30 employees, or 70%, make more than that.

(b) In his radio show *A Prairie Home Companion* host Garrison Keillor regularly tells his audience about Lake Wobegon, where "all the children are above average".

Is that possible, with any of the meanings of "average"?

No — that's what makes Keillor's description funny.

For the median, exactly half are above average. For the mean and the mode it's possible for (a lot) more than half to be above average, but never "all".

Exercise 6.12.12. [S][C][Section **??**][Goal **??**] Working for Walmart.

On December 2, 2009 Bloomberg News reported on a settlement in which Walmart agreed to pay \$40 million to up to 87,500 employees because the company had failed to pay overtime, allow rest and meal breaks and, in addition, manipulated time cards.

	Physics		English	
	professors	salary	professors	salary
Women	1	\$100K	8	\$50K
Men	9	\$90K	2	\$40K

Table 6.26. Salary structure at a university

Eligible present and former employees would receive \$400 to \$2,500 each — on average \$734.

The lawyers for the employees asked for fees of \$15.2 million from the \$40 million. [R**??**]

(a) Compute the mean compensation, assuming that there are 87,500 eligible employees and that the lawyers have taken their cut.

(b) Compare your answer to the reported minimum compensation of \$400 and the reported \$734 the average worker will receive.

(c) Draft a letter to the editor or the reporter, politely pointing out that both the reported "averages" made no sense, and asking for more detail or a correction.

(a) Compute the mean compensation, assuming that there are 87,500 eligible employees and that the lawyers have taken their cut.

$$\frac{\$40\text{M} - \$15.2\text{M}}{87,500 \text{ workers}} = 283 \frac{\$}{\text{worker}}.$$

(b) Compare your answer to the reported minimum compensation of \$400 and the reported \$734 the average worker will receive.

The \$283 mean compensation can't be less than the \$400 minimum compensation. Something is crazy here.

The \$734 that the "average worker" gets is even crazier. It's much too large to be the median. There's no way half the workers can get more than \$734 if the minimum is \$400 unless the mean is more than halfway from \$400 to \$734. That's not true even before I take out the lawyers' \$15.2 million.

Exercise 6.12.13. [R][A][S][Section **??**][Goal **??**] Is it discrimination?

Table **??** shows the salary structure of two departments in a hypothetical university.

(a) What is the average (mean) salary of the professors? Of the women professors? Of the men?

(b) Answer the same questions for the median.

(c) Answer the same questions for the mode.

(d) Write a few sentences to convince someone that men in this university are paid better than women. Then write a few sentences to convince someone of just the opposite. Explain the contradiction.

These calculations are so straightforward that they're easier with pencil and paper (maybe a calculator) than with Excel.

(a) What is the average (mean) salary of the professors? Of the women professors? Of the men?

The mean professor salary (in thousands of dollars) is

$$\frac{1\times\$100+9\times\$90+8\times\$50+2\times\$40}{20}=\$69.5.$$

The mean for the women is

$$\frac{1\times\$100+8\times\$50}{9}=\$55.6.$$

The mean for the men is

$$\frac{9\times\$90+2\times\$40}{11}=\$80.9.$$

(b) Answer the same questions for the median.

The median salaries for all, women and men are \$70K (halfway between \$90K and \$50K), \$50K and \$90K respectively.

(c) Answer the same questions for the mode.

The mode salaries for all, women and men are \$90K, \$50K and \$90K respectively.

(d) Write a few sentences to convince someone that men in this university are paid better than women. Then write a few sentences to convince someone of just the opposite. Explain the contradiction.

Obviously the men are paid better than the women. Each of the three averages is higher for men than women!

Obviously the women are better paid than the men: all the women do better than all the men in both departments.

This seeming contradiction is possible because the women outnumber the men in the English (poorer) department, while the opposite is true in the physics (richer) department. There the lone woman makes the highest salary — perhaps because the Dean (a woman) insisted that an all male physics department would discourage women students.

Exercise 6.12.14. [S][Section **??**] [Goal **??**][Goal **??**] Choosing data ranges for a histogram.

When you change the widths of the intervals in a histogram you get a (slightly) different picture of the data.

(a) Redo the Wing Aero histograms in the text using salary ranges of size \$10K and then of size \$50K.

[See the back of the book for a hint.] Start with a copy of `WingAeroHistogram.xlsx`. Create new data in columns `D` and `E` below the existing data there for the two new histograms. Fill in column `F` appropriately. Then you can copy and paste from what's there to fill in columns `G` and `H` and the sums and the mean.

If Excel wants to treat your salary ranges as dates, try formatting the cells as text.

To make the charts, copy the histogram that's there, then find the place in Excel where you can change the source data to the new rows in columns `D` and `E`.

(b) In each case use the techniques from Section **??** to estimate the mean, median and mode.

(c) Discuss the advantages and disadvantages of these possible choices for the salary range, comparing them to our choice of $20K.

(a) Redo the Wing Aero histograms in the text using salary ranges of size $10K and then of size $50K.

See `../Answers/WingAeroHistogramRangesSolution.xlsx` for the new histograms.

(b) Estimate the mean, median and mode.

Here's a table showing all three averages, computed from the data directly and from the three histograms.

Data	Mean	Median	Mode
complete	$77.2K	$42.5K	nonsense
$20K intervals	$79K	$48K	$20K-$39K
$10K intervals	$77K	$45K	$20K-$29K
$50K intervals	$73K	$45K	$0K-$50K

The smaller the data ranges in the histograms the closer the approximated averages are to the true values.

(c) Discuss the advantages and disadvantages of these possible choices for the salary range, comparing them to our choice of $20K.

None of these histograms seems particularly "better" than the other two. Each gives a good visual picture of how Wing Aero salaries are distributed. The various averages are close enough to the correct values so that they are not misleading.

Exercise 6.12.15. [S][Section **??**][Goal **??**] Texting teens.

We drew the chart in Figure **??** using data from *The Boston Globe* on April 15, 2012.

(a) Estimate the mode, median and mean number of text messages sent by teenagers each day.

(b) In total, approximately how many text messages are sent by the 23 million American teens each day?

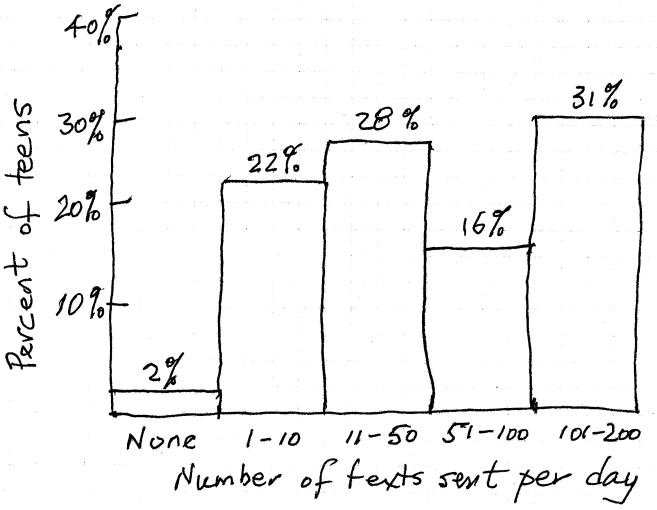


Figure 6.27. Teen texting [R**??**]

(c) The percentages don't add up to 100%. Why might that have happened?

(d) If you asked a random teenager how many text messages she sent yesterday what are the chances (what is the probability) that it was more than 50? More than 100? More than 25?

(e) What percent of teenagers text more than the median amount?

(f) Does the figure display a histogram?

(g) Create an Excel chart that reproduces the figure.

(a) Estimate the mode, median and mean number of text messages sent by teenagers each day.

The mode is the tallest column: more teenagers text 101-200 messages a day than any other category. It's also reasonable to say that the mode is about 150 messages per day.

Adding the percentages for the first three columns gives $2\% + 22\% + 28\% = 52\%$. That means just over half the teenagers text up to 50 messages per day. That's a good estimate for the median.

I will compute the mean as a weighted average, using the midpoints of the ranges for each range.

$$0.02 \times 0 + 0.22 \times 5 + 0.28 \times 30 + 0.16 \times 75 + 0.31 \times 150 = 68,$$

so the mean number of daily teen text messages is about 70.

I could also have computed using the exact midpoints of the ranges. The arithmetic is uglier:

$$0.02 \times 0 + 0.22 \times 5.5 + 0.28 \times 30.5 + 0.16 \times 75.5 + 0.31 \times 150.5 = 68.705,$$

which is still about 70. The extra precision isn't worth the effort.

(b) In total, approximately how many text messages are sent by the 23 million American teens each day?

Since I have the mean, 70 messages per teenager on average, all I have to do is multiply by 23 million to get about 1.6 billion messages per day.

(c) The percentages don't add up to 100%. Why might that have happened?

The percentages sum to 99%. The missing one percent is probably roundoff error.

(d) If you asked a random teenager how many text messages she sent yesterday what are the chances (what is the probability) that it was more than 50? More than 100? More than 25?

There's a 47% chance that a random teenager texts more than 50 messages in a day and a 31% chance that s/he texts more than 100.

For 25 messages I have to use a fraction of the 11-51 message column. Since 25 is about one third of the way from 11 to 50 I will suppose that about two thirds or 18% of the teens in that category text more than 25 messages a day. Adding that 18% to the 47% for the two top columns gives me an estimate of 65% for the chances that a teenager texts more than 25 times a day.

(e) What percent of teenagers text more than the median amount?

By definition that's 50%.

(f) Does the figure display a histogram?

Yes, but not a good one since the ranges are not all the same width.

(g) Create an Excel chart that reproduces the figure.

See `../Answers/TeenTextingSolution.xlsx`.

Exercise 6.12.16. [S][Section **??**][Goal **??**] [Goal **??**] Websites are often confusing.

Jakob Nielsen evaluated the usability of voter information websites for the 2008 election for each of the fifty states and the District of Columbia. You can read his analysis at `www.nngroup.com/articles/aspects-of-design-quality/` . His article includes the histogram in Figure **??**.

(a) What is the modal usability score for these 51 home pages?

(b) Reproduce this histogram in Excel.

[See the back of the book for a hint.] When you enter the data in two columns put the categories (usability scores) on the left since they are the labels for the x-axis. Put the numbers of websites on the right since they are the values that go with the categories, and should plot vertically, on the y-axis.

If your spreadsheet is anything like ours, it may well think that you want to display both columns on the y-axis, since both columns are numbers. If it does that it will label the x axis with the numbers 1 to 10.

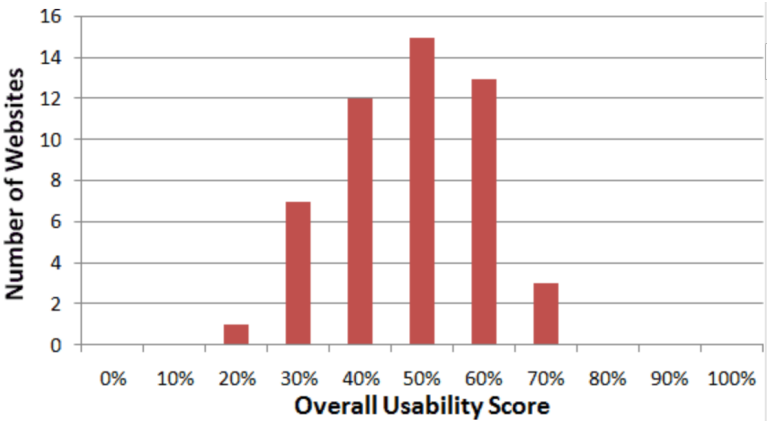


Figure 6.28. Website usability [R**??**]

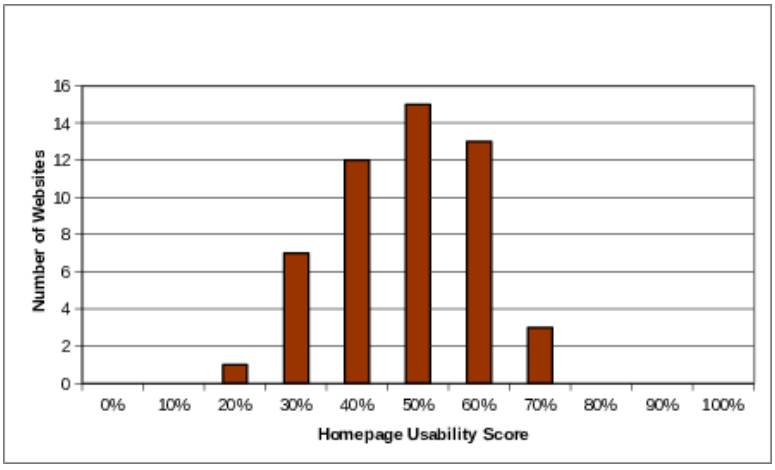


Figure 6.29. Website usability

If that happens, delete the data series corresponding to the bars you don't want (the percentages). Then right click on the chart and explore until you find the place that allows you to enter the fields you want to use as x-axis category labels.

(c) Estimate the median usability score for these 51 home pages.

(d) How many of them have a usability score less than the median score?

(e) Estimate their mean usability score.

(a) What is the modal usability score for these 51 home pages?
The 50% bar is the tallest, so that's the modal score.

(b) Reproduce this histogram in Excel.
See Figure **??**. The spreadsheet is at `../Answers/HomePageQualitySolution.xlsx`.
I matched the color as best I could.

(c) Estimate the median usability score for these 51 home pages.

I will add up the numbers in each category starting from the left, to see when I reach half the websites — that would be 25 or 26. The sum of the first five categories is $0+0+1+7+12=20$; the sum of the first six is $20+15=35$. So the 25th site in the 50% category.

That means $20/51 \approx 40\%$ of the websites have usability scores of 40% or less while $35/51 \approx 70\%$ have scores 50% or less. Since I have to report a single number for the median, I'll say it's about 47% — nearer 50% than 40%.

(d) How many of them have a usability score less than the median score?

Half the home pages have a score less than the median. That will be true even if I made a mistake computing the median!

(e) Estimate their mean usability score.

The nice round numbers suggest that there were 10 usability tests and eich site either succeed or fail in each one. If that's how the score was computed, then every website passed at least two of the tests and none passed more than seven. With that interpretation the mean usability score is the weighted average

$$\frac{1\times 20+7\times 30+12\times 40+15\times 50+13\times 60+3\times 70}{51}=0.48,$$

which is about 48%.

I did the computation in Excel, since I had already entered the numbers to build the histogram.

But there may have been a more precise accounting, so that the bar marked "50%" counts all the scores between 50% and 60%. Then I wouuld use the midpoint of the range for each column. That would just add 5 percent to each score in the above weighted average, which would lead to 48% + 5% = 53% as my estimate for the mean.

In fact the article reports that the highest score was 77%, so the second interpretation is correct.

Exercise 6.12.17. [U][S][Section **??**][Goal **??**] Doublethink.

In January 2012 one could read this at `www.businessinsider.com/where-the-one-percent-live-the-15-richest-counties-in-america-2012-2` :

> Living in Arlington isn't cheap, so you'd better be making at least the median household income to live in this county just outside Washington, D.C. [R**??**]

The "median household income" in the quote is the median income in Arlington County. How does this statement contradict itself?

The definition of "median" means that half the households in Arlington make less than the median. But they live there anyway.

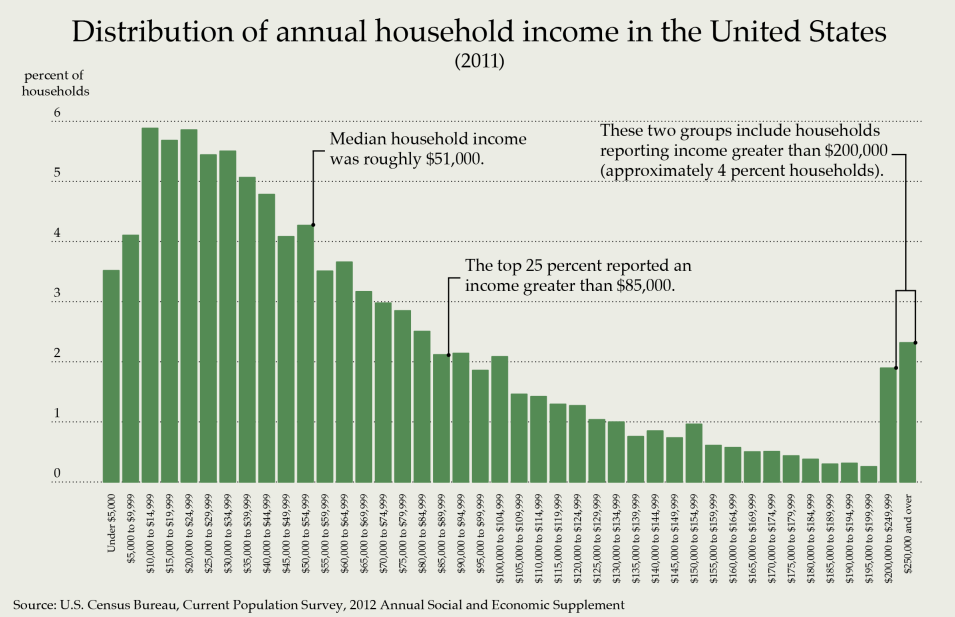


Figure 6.30. United States household income — 2012 Estimate [R**??**]

Exercise 6.12.18. [S][C][W][Section **??**][Goal **??**][Goal **??**] Household income in the United States.

The histogram in Figure **??** shows the estimated percentages of households in income groups $5,000 increments apart, except for the two farthest right columns.

We've put the data (and a copy of the figure) in the spreadsheet `Households2012.xlsx`.

(a) Check the quantitative assertions in the text in the Wikipedia chart.

(b) Build a histogram in Excel that comes as close as possible to matching the one from Wikipedia. Create the same chart and axis titles. Change the grid lines. Put in the comments as text boxes. Match the fonts.

(c) Do the percentages sum to 100%? If not, what might explain the discrepancy?

(d) To estimate the mean household income you will need an estimate of the mean for the households with incomes greater than $250,000. There's no top to this range, so you can't use the middle of the range.

What value for the mean for the last category makes the mean for the whole population equal to the median?

(e) Search for an estimate of the mean household income for the whole population. What mean for the last category results in this overall mean?

See `../Answers/Households2012Solution.xlsx`.

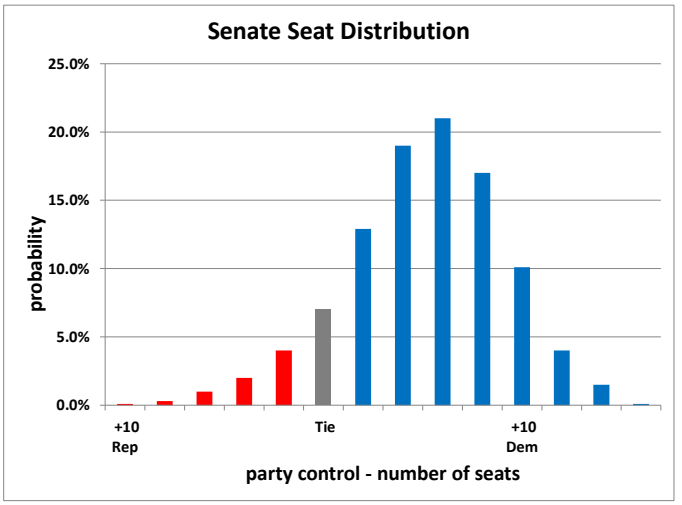


Figure 6.31. The fight for the Senate [R**??**]

Exercise 6.12.19. [S][Section **??**] [Goal **??**][Goal **??**] Fight for the Senate.

A graph like the one in Figure **??** appeared in Nate Silver's Five Thirty Eight column in *The New York Times* on October 31, 2012. The x-axis displays the number of seats held by each party: the Tie in the middle is 50 Democrats, 50 Republicans. The +10 Dem corresponds to 55 Democrats, 45 Republicans.

Nate Silver constructed this histogram by imagining (simulating) many thousands of elections and recording the percentage of time each Democratic/Republican split occurred. We estimated the percentages in his chart and entered them in the spreadsheet `Oct31SenateProjection.xlsx` so you don't have to type them yourself. (We rounded the really tiny percentages to zero.) Use Excel whenever it's most convenient for you.

(a) What is the most likely number of Democratic senators?

(b) What number of Democratic senators represents the mode of this distribution?

(c) What is the probability that there are more than 50 Democratic senators?

(d) What number of Democratic senators is the median of this distribution?

(e) If you had the complete list of all Nate Silver's imagined elections and sorted it by the number of Democratic senators, how many Democratic Senators would there be in the middle election on that list?

(f) Use Excel to compute the (weighted) average number of Democratic senators for these imagined elections.

(g) What actually happened in the election?

(a) What is the most likely number of Democratic senators?

53 — the highest bar.

Spending on Fire/EMS services

Study from the Boston Globe, March 30, 2009
2007 population figures from Wikipedia

City	Population	Fire/EMS spending per resident	Fire/EMS personnel per 1000 residents
Boston	608,352	$452.15	3.4
San Francisco	799,183	$315.81	2.2
Columbus, OH	747,755	$255.70	2.1
Seattle	594,210	$247.75	1.8
Baltimore	637,455	$225.98	2.7
Memphis	674,028	$220.22	2.5
Detroit	916,952	$201.54	1.6
Nashville	590,807	$194.43	1.9
Philadelphia	1,449,634	$187.63	1.6
Jacksonville	805,605	$179.99	1.5
New York	8,274,527	$157.56	1.7
Los Angeles	3,834,340	$137.80	0.9

Figure 6.32. Fire protection spending [R**??**]

(b) What number of Democratic senators represents the mode of this distribution?

53 — the highest bar.

(c) What is the probability that there are more than 50 Democratic senators?

85.6% — it's `=SUM(C14:C21)` in the spreadsheet.

(d) What number of Democratic senators is the median of this distribution?

The cumulative sum passes 50% at 53 Democrats. So the median is 53.

(e) If you had the complete list of all Nate Silver's imagined elections and sorted it by the number of Democratic senators, how many Democratic Senators would there be in the middle election on that list?

There would be 53 Democrats. This is just the median again.

(f) Use Excel to compute the (weighted) average number of Democratic senators for these imagined elections.

The weighted average is 52.6 Democratic senators.

See `../Answers/Oct31SenateProjectionSolution.xlsx` for the calculation.

(g) What actually happened in the election?

The 2012 election produced a Senate with 53 Democrats.

Exercise 6.12.20. [S] [W][Section **??**] [Goal **??**][Goal **??**] What cities pay for fire protection.

On Monday, March 30, 2009 *The Boston Globe* published an article comparing the amount various cities spent on Fire and EMS services. Figure **??** is a screenshot of a spreadsheet where we entered some of the data, along with population figures. You can download that spreadsheet from `FireSpending.xlsx`. Use the data to answer these questions.

(i) Population

Population range	Number of cities
500K-600K	
600K-700K	
700K-800K	
800K-900K	
900K-1000K	
1000K-2000K	
2000K-3000K	
3000K-4000K	
$>$ 4000K	

Table 6.33. City populations

(a) What is the mean population of the twelve cities for which data are presented?
(b) What is the median population of the twelve cities for which data are presented?
(c) Create Table **??** in Excel. Fill in the second column there. Then create a properly labeled histogram for the data.
(d) Use your histogram to estimate the mode population for these cities.
(e) What percent of the U.S. population lives in these twelve cities?

(ii) Fire/EMS spending per person

(a) What is the mean amount spent for Fire/EMS services per person in these twelve cities?
(b) Estimate the median amount spent for Fire/EMS services per person in these twelve cities.
(c) Estimate the mode amount spent for Fire/EMS services per person in these twelve cities.

(iii) What do firefighters earn?

There is enough information in the spreadsheet to calculate the average (mean) earnings of Fire/EMS personnel in each of the twelve cities. Do that, in a fresh column in your spreadsheet.

(a) In which city do Fire/EMS personnel have the highest average salary? How much is it?
(b) In which city do Fire/EMS personnel have the lowest average salary? How much is it?
(c) Where does Boston rank in the list of Fire/EMS personnel salaries?
(d) Explain how Boston can be at the top of the list in Fire/EMS expenses per resident although it does not pay the highest salaries.

(iv) Correction the next day!

On Tuesday, the next day, *The Boston Globe* published a correction, which said that Boston's fire department expenses were $285 per resident in the last fiscal year. [R**??**]

Look at the answers to the questions above and indicate which have changed (and how) and which stayed the same.

[See the back of the book for a hint.] The mean amount spent for Fire/EMS services per person is not the Excel `AVERAGE` of the amounts spent per person by each city. It's wrong to average those numbers since they are already averages. You must weight them by the city populations in order to compute the total amount spent by all the people in all the cities. Then divide by the total population.

You can't compute the median with Excel's `MEDIAN` function for the same reason. Further hint: almost everyone lives in New York.

You can't answer the question "What do firefighters earn?" by finding the mean of the twelve numbers in column C. Compute the mean correctly as a weighted average. You will probably want to start by creating a column labeled

```
total Fire/EMS expenses
```

and fill in the value for each city.

You can find the work in the spreadsheet at `../Answers/FireSpendingSolution.xlsx`.

(i) Population

(a) What is the mean population of the twelve cities for which data are presented?
A straightforward use of `=AVERAGE(B7:B18)` in Excel tells me that the mean population is 1,661,071. The last few digits really don't mean much since it's unlikely that the populations listed for the cities are really accurate down to the last person. A better answer is about 1,660,000.

(b) What is the median population of the twelve cities for which data are presented?
A straightforward use of `=MEDIAN(B7:B18)` in Excel tells me that the median population is 773,469. That's halfway between the populations of San Francisco and Columbus. It's much too precise. A better answer is that the median is about 770,000.

(c) Here is the table:

Population range	Number of cities
500K-600K	2
600K-700K	3
700K-800K	2
800K-900K	1
900K-1000K	1
1000K-2000K	1
2000K-3000K	0
3000K-4000K	1
$>$ 4000K	1

See the histogram in the spreadsheet. Note the labels for the axes and the title, that the bars touch one another, and that the vertical scale does not list halves of cities.

(d) The highest bar in the histogram corresponds to the largest number in the table: the mode population is 600K-700K.

(e) The `SUM` function in Excel tells me that the total population of these twelve cities is 19,932,848.

That represents $19,932,848/300,000,000 \approx 7\%$ of the people in the United States

(ii) Fire/EMS spending per person

(a) Column `C` reports the average (mean) spending in each of the twelve cities. To find the overall average I can't just average these averages, since the cities are different sizes. I need to use those as the weights. So I created column `E` to hold `total Fire/EMS expenses` and filled each row with the product of the values in columns `B` and `C`: the number of residents times the spending per resident. Then I summed column `E` to find out that these cities spent a total of $3,707,134,657 (almost four billion dollars!) to provide Fire/EMS services for their total population of 19,932,848 people. Dividing, I found that on average they spent

$$\frac{\$3,707,134,657}{19,932,848 \text{ people}} \approx 186 \frac{\$}{\text{person}}$$

(b) New York accounts for more than 40% of the 20 million people in these cities. If you add in Los Angeles then you get more than half. Since those two are the two cities at the bottom of the list, the median Fire/EMS cost per resident is somewhere between the values for those cities — probably about $150.

(c) The most common amount is what the New Yorkers pay, so the mode is $157.56.

(iii) What do firefighters earn?

There is enough information in the spreadsheet to calculate the average (mean) earnings of Fire/EMS personnel in each of the twelve cities. Do that, in a fresh column in your spreadsheet.

I computed that two ways, just to make sure I was right. The first way uses the contents of columns `C` and `D` to fill column `H` this way: `=1000*Cn/Dn` for each of the rows. That works because the units for column `C` are $ per resident while for column `D` they are firefighters per 1000 residents. When I divide I get $ per firefighter, which is just what I want.

The second way is to compute the number of firefighters in each city — fill column `H` with `=B7*D7/1000`. Then divide the total fire/EMS expenses (already computed in column `E`) by that number and put the result in column `I`. Columns `H` and `I` match.

(a) Firefighters in Los Angeles have the highest average salary: $153,111.

(b) Firefighters in Baltimore have the lowest average salary: $83,696.

(c) Boston ranks fourth from the top, at $132,985.

(d) Boston spends more per resident on Fire/EMS personnel than other cities even though it pays them less because it has a lot more of them in proportion to population — it's at the top of that list.

(iv) Correction the next day!

When I change the Boston Fire/EMS expense per resident from $452.15 to $285 the following figures change:

- The mean Fire/EMS expense becomes $181/resident, down from $186 per resident.
- The mean Fire/EMS salary becomes $107,600, down from $110,635.

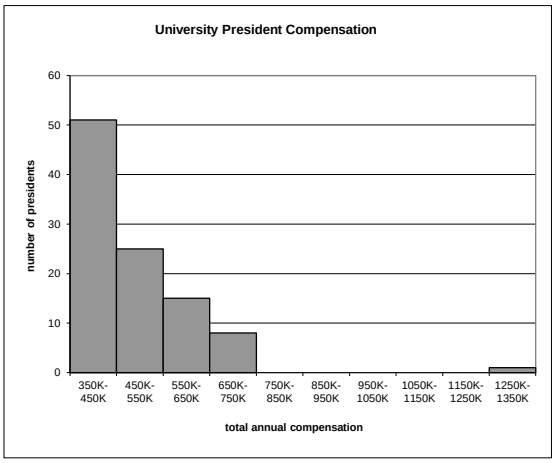


Figure 6.34. College presidents' pay [R**??**]

- The average Boston Fire/EMS salary falls from $132,985 to $83,824 — second lowest, only about $100 above the minimum.

Exercise 6.12.21. [S][Section **??**] [Goal **??**] College presidents' pay.

Figure **??** is a histogram showing the total compensation for the 100 best paid presidents of public universities. The data are from an article in the April 3, 2011 issue of *The Chronicle of Higher Education* (`chronicle.com/article/Presidents-Defend-Their/126971` .)

To save you staring at the picture, the number of presidents in each of the ranges (reading from left to right) is 51, 25, 15, 8 and then 1 in the last range.

You may use Excel for this exercise if you wish, but you don't have to.

(a) What is the mode of this distribution?

(b) Estimate the median compensation.

(c) Estimate the mean compensation.

(d) Write two arguments one of which indicates that the average president is paid appropriately, one that presidential pay is too high.

(a) What is the mode of this distribution?

The mode is $350K-$450K — the most common salary range.

(b) Estimate the median compensation.

51 presidents are in the $350K-$450K range and 49 make more than that, so the median is $450K — right at the top of the most common range.

(c) Estimate the mean compensation.

weight range (frams)	percent of sample
20-40	10
40-60	10
60-80	20
80-100	10
100-120	50

Table 6.35. Xorlon fleegs

I computed a weighted average of the middles of the ranges, with weights the number of presidents in each range:

$$\frac{51\times 400+25\times 500+15\times 600+8\times 700+1\times 1300}{100}=488,$$

so the mean compensation is approximately \$490K. Perhaps that should be rounded to \$500K.

I redid the computation in the spreadsheet at `../Answers/CollegePresidentsSolution.xlsx`.

(d) Write two arguments one of which indicates that the average president is paid appropriately, one that presidential pay is too high.

Most of the 100 best paid presidents of public universities make about \$400K per year. They run large organizations in the public interest and make much less money than CEOs of companies of comparable size. They are fairly paid.

The average salary of these presidents of public universities is nearly half a million dollars a year. That is much more than faculty make at those institutions. They should be paid more like public servants.

Exercise 6.12.22. [A][S][W][Section **??**] [Goal **??**][Goal **??**][Goal **??**] Xorlon fleegs.

Table **??** shows the distribution of weights of a sample of fleegs from the planet Xorlon, where weight is measured in frams.

We made up the numbers in this artificial problem so that the arithmetic would be easy. We made up the words, too. "frams" is one Prof. Bolker could have used in a Scrabble game with his wife — for 57 points on a triple word score.

(a) Sketch a neat properly labeled histogram that displays the data.

(b) Create a properly labeled histogram in Excel that matches the one you just drew. You can start by downloading `XorlonFleegs.xlsx`.

Answer the following questions. You may work in Excel or with a calculator or do mental arithmetic.

(c) Estimate the mode fleeg weight.

(d) Estimate the median fleeg weight.

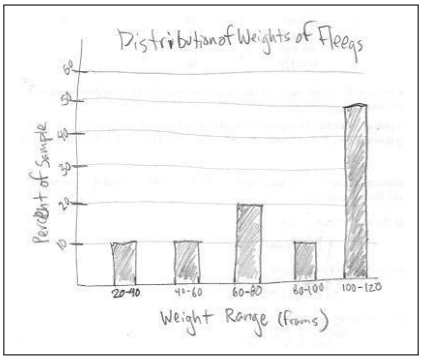


Figure 6.36. Xorlon fleegs

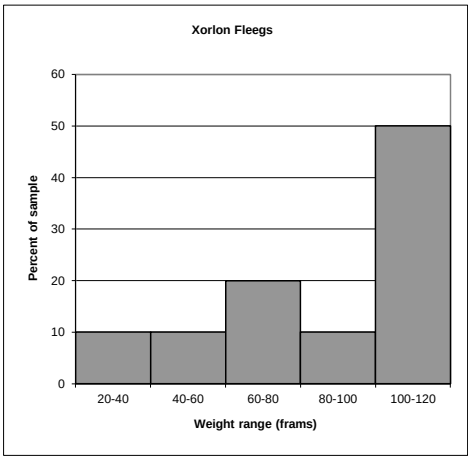


Figure 6.37. Fleegs on Xorlon

(e) Estimate the mean fleeg weight.

(f) Estimate the percentage of fleegs that weigh more than the median.

(g) Estimate the percentage of fleegs that weigh more than the mean.

(h) Estimate the percentage of fleegs that weigh more than the mode.

(a) Figure **??** shows a student's neat properly labeled histogram that displays the data.

(b) Figure **??** shows an Excel histogram that displays this data. The source is the spreadsheet `../Answers/XorlonFleegsSolution.xlsx`.

(c) Estimate the mode fleeg weight.

The highest bar is 100-120 frams, so that, or 110 frams, is a good estimate for the mode.

(d) Estimate the median fleeg weight.

Half the fleegs in the sample weigh less than 100 frams and half weigh more, so the median weight is 100 frams.

(e) Estimate the mean fleeg weight.

Averaging the midpoints of each weight range using the percentages as weights is the best way to estimate the mean. The result is

$$0.10 \times 30 + 0.10 \times 50 + 0.20 \times 70 + 0.10 \times 90 + 0.50 \times 110 = 86$$

so the mean weigh is 86 frams.

(f) Estimate the percentage of fleegs that weigh more than the median.

The meaning of "median" tells me that just 50% of the fleegs are heavier than the median weight.

(g) Estimate the percentage of fleegs that weigh more than the mean.

86 frams is 6/20 or three tenths of the way into the 80-100 fram range, so I'd estimate that the percentage of fleegs weighing more than that is

$$\frac{7}{10}10\% + 50\% = 57\%.$$

(h) Estimate the percentage of fleegs that weigh more than the mode.

I estimated the mode as 110 frams. Half the fleegs in the highest range, or 25%, weigh more than that.

Exercise 6.12.23. [S][Section **??**] [Goal **??**] Erdős numbers.

Paul Erdős (1913-1996) was the most prolific mathematician of the twentieth century. He was famous (in mathematical circles) for the way he worked — he travelled from school to school, writing joint papers with mathematicians at each.

From Wikipedia:

> An Erdős number describes a person's degree of separation from Erdős himself, based on their collaboration with him, or with another who has their own Erdős number. Erdős alone was assigned the Erdős number of 0 (for being himself), while his immediate collaborators could claim an Erdős number of 1, their collaborators have Erdős number at most 2, and so on. [R**??**]

You can think of the Erdős Number Project as a description of the social network of mathematicians. Its home page is `www.oakland.edu/enp` ; there you can find out that

> ...the median Erdős number is 5; the mean is 4.65, and the standard deviation is 1.21.

and see the raw data in Table **??**.

> [T]he lists of coauthors and the various other statistics on this site are updated about once every five years. The current version was posted on July 14, 2015 and includes all information listed in MathSciNet and DBLP through mid-2015.

Erdős number	mathematicians
0	1
1	504
2	6,593
3	33,605
4	83,642
5	87,760
6	40,014
7	11,591
8	3,146
9	819
10	244
11	68
12	23
13	5

Table 6.38. Erdős numbers [R**??**]

(a) Use Excel to draw a histogram for the distribution of Erdős numbers.

(b) What is the mode of this distribution?

(c) Verify the claims for the median and mean.

(d) Professor Bolker (one of the authors of this book) has an Erdős number of 2 (he wrote a paper with Patrick O'Neil, who wrote a paper with Erdős).

Professor Bolker also wrote a paper with his granddaughter Eleanor and his son, her uncle Benjamin. So their Erdős numbers and that of Professor Mast (the other author of this book) are at most 3? Might any of them be less? Might they decrease in time?

(e) How many mathematicians have a finite Erdős number?

(f) There are some mathematicians whose Erdős number is infinite. How can that be?

(g) (Optional) Check to see whether the statistics at `www.oakland.edu/enp` have been updated, and update this question if they have.

(a) Use Excel to draw a histogram for the distribution of Erdős numbers.

See `../Answers/ErdosSolution.xlsx`.

Making the values in cells `A6:A13` the labels on the bars is tricky. To do that, build the chart using only `B6:B13` for data. Then select the data (right click) and edit the field on the right for category axis labels.

(b) What is the mode of this distribution?

The mode is 5; it's the most common Erdős number.

(c) Verify the claims for the median and mean.

There are 268,015 mathematicians connected (directly or indirectly) to Erdős. Arranging them in order of their Erdős numbers, the middle one has number 5. So that's the median.

I used the spreadsheet to calculate the mean: it's 4.65.

(d) Professor Bolker (one of the authors of this book) has an Erdős number of 2 (he wrote a paper with Patrick O'Neil, who wrote a paper with Erdős).

Professor Bolker also wrote a paper with his granddaughter Eleanor and his son, her uncle Benjamin. So their Erdős numbers and that of Professor Mast (the other author of this book) are at most 3? Might any of them be less? Might they decrease in time?

Each of those three might have an Erdős number of 1 or 2 (but they don't). If one of them some day writes a paper with someone with an Erdős number of 1 their Erdős number will drop from 3 to 2.

Since Erdős is dead, no one can now earn an Erdős number of 1.

(e) How many mathematicians have a finite Erdős number?

This is just the total in the data: 268,015.

(f) There are some mathematicians whose Erdős number is infinite. How can that be?

Mathematicians who have never coauthored a paper are not connected to Erdős at all. That's also true even if they have written joint papers but none of their coauthors is connected to Erdős. We say they have an infinite Erdős number. It's not really infinite; that's a kind of in joke among mathematicians.

(g) Check to see whether the statistics at `www.oakland.edu/enp` have been updated, and update this question if they have.

Not yet in January, 2019.

Exercise 6.12.24. [S][A][Section **??**] [Goal **??**] Ruritania.

Find the mean, median and mode age for male residents of Ruritania using the histogram or the data in the spreadsheet `Ruritania.xlsx`. (Ruritania is a fictional country in central Europe which forms the setting for *The Prisoner of Zenda*, a fantasy novel written by Anthony Hope.)

The mode is easy — it's 0-9 years, since that's the longest bar in the histogram.

Adding the percentages in each category I see that the first three come to exactly 50% (in Excel, where the numbers are rounded for display but not for computation). So the median Ruritanian age is about 30 years (half the people are younger, half older).

I used Excel to compute the weighted average of the midpoints of the categories (5, 15, ...) and found the mean age to be about 36 years.

Exercise 6.12.25. [A][S][Section **??**][Goal **??**] Wing Aero percentiles.

In a histogram, data are grouped into ranges of the same numerical width. When finding percentiles the ranges contain known percentages of the data items — the widths vary. Use your sorted list of Wing Aero salaries to answer the following questions.

(a) How many Wing Aero employees are in the bottom tenth percentile in salary?

(b) What is the salary cutoff for the bottom tenth percentile?

(c) Answer the same questions for the top tenth percentile.

(a) How many Wing Aero employees are in the bottom tenth percentile in salary?

By definition, 10 percent of the Wing Aero employees are in the bottom tenth percentile in the salary distribution. Since there are 30 employees, 3 are in the bottom tenth percentile.

(b) What is the salary cutoff for the bottom tenth percentile?

The salaries of the three workers in the bottom tenth percentile are \$17K, \$19K and \$21K. The next highest salary is \$25K, so any number between that and \$21K can be thought of as the cutoff point for the bottom tenth percentile.

(c) Answer the same questions for the top tenth percentile.

The three employees in the top tenth percentile are the CEO, the CFO and one of the other two executives. The salary cutoff for the top tenth percentile is \$250K.

Exercise 6.12.26. [U][Section **??**][Goal **??**] [Goal **??**] Sick-leave proposal.

An article in the *Orlando Sentinel* on August 6, 2012 discussed a ballot initiative that would require employers with 15 or more workers to provide paid time off for employees for illness-related issues. The article polled voters to gauge support for placing this question on the ballot for the November election, and noted that among likely voters, 67 percent supported the initiative while 26 percent opposed it.

The article reported that the poll surveyed 500 people and had a margin of error of 4.4 percentage points. [R**??**]

(a) Why don't the percentages from this poll add up to 100%?

(b) Explain why this statement is not true: "67% of the residents likely to vote in November support the measure."

(c) Explain what the 4.4 percentage point margin of error means for this poll.

Exercise 6.12.27. [S][Section **??**][Goal **??**][Goal **??**] The Boston Marathon. Table **??** contains data for the numbers of men and women who finished the 2012 Boston marathon, grouped by finishing times. For example, 26 men and one woman finished with a time between two and two and a half hours. (That one woman was a wheelchair racer.)

We've entered the data in the spreadsheet `Marathon2012.xlsx`.

Answer the following questions. Do as much of the arithmetic in Excel as possible.

(a) Sketch a neat histogram for this data.

(b) Draw your histogram with Excel. Does it match your sketch?

(c) How many men finished the marathon? How many women?

Finishing time	Men	Women
2:00-2:30	26	1
2:30-3:00	444	27
3:00-3:30	1,844	260
3:30-4:00	3,389	1,714
4:00-4:30	2,819	2,833
4:30-5:00	1,861	1,966
5:00-5:30	1,068	1,013
5:30-6:00	607	609
6:00-6:30	323	339
6:30-7:00	160	162

Table 6.39. The 2012 Boston marathon

(d) Use the data to estimate the mode, median and mean for the men's finishing times. Mark these times on the handwritten histogram sketch.

(e) Suppose my friend ran the marathon and finished ahead of half the men. What was his finishing time (approximately)?

(f) About what percentage of the women finished ahead of half the men?

[See the back of the book for a hint.] Warning: 2:30 is 2 hours and 30 minutes. That's 2.5 hours, not 2.3 hours.

(a) Sketch a neat histogram for this data.
(not done)

(b) Draw your histogram with Excel. Does it match your sketch?
See spreadsheet.

(c) How many men finished the marathon? How many women?
Men: 12541, women: 8924, computed with Excel `=SUM()`.

(d) Use the data to estimate the mode, median and mean for the men's finishing times.

- Mode:
  The men's mode finishing time is three to three and a half hours (the highest bar), or about 3:15.
- Median:
  I used Excel to discover that half the men finished in just over four hours (45% were faster than 4:00 and 67% faster than 4:30). I'll estimate 4:10 as the time for the middle runner.
- Mean:
  Using the midpoints of the intervals and doing the computations in Excel, I estimate the mean as about 250 minutes, which is 4 hours and 10 minutes — about the same as the median.

I marked these averages in the chart in the spreadsheet at `../Answers/Marathon2012Solution.xlsx`.

(e) Suppose my friend ran the marathon and finished ahead of half the men. What was his finishing time (approximately)?

His finishing time must have been the median time: about 4:10.

(f) About what percentage of the women finished ahead of half the men?

I need to find the fraction of women who finished faster than 4:10. Only 22% finished in less than four hours, and about 54% finished in less than 4:30, so I'll estimate that 40% finished ahead of half the men.

This is really interesting — the women's median is about 4:20, which isn't much larger than the men's.

Exercise 6.12.28. Income growth.
Moved to EXTRAEXERCISESURL.

Exercise 6.12.29. [U] Income growth.

On April 26, 2014 *The Boston Globe* reported that 2012 per capita income in Massachusetts grew to $49,354, up 3.2% from 2008, after adjusting for inflation.

(a) How much was Massachusetts per capita income in 2008, in 2012 dollars?

(b) How much was Massachusetts per capita income in 2008, in 2008 dollars?

(c) This income figure is an average. Is it a mean, a median or a mode? Explain how you know.

(d) Estimate the total 2012 income for Massachusetts.

Exercise 6.12.30. [S][Section **??**][Goal **??**][Goal **??**] Scrabble.

The Wikipedia page `en.wikipedia.org/wiki/Scrabble_letter_distributions` shows the point value of each of the 100 Scrabble tiles.

(a) Draw a bar chart illustrating the number of tiles with each of the point values from 0 to 10. The x-axis labels should be

0 1 2 3 4 5 6 7 8 9 10

The heights of the bars should correspond to the number of tiles with each value.

This is a difficult chart to create in Excel. Before you try, draw it by hand, so you know what you want the end result to look like. What Excel shows you first is likely to be far from your goal.

(b) What are the mode, median and mean point values? Show them on your (hand written) chart.

(c) What percentage of the tiles are worth more than 1 point?

(d) What percentage of the tiles are worth less than the median number of points?

(e) Answer these questions for Scrabble in some other language (your choice) and discuss the differences between that language and English.

(a) Draw a bar chart illustrating the number of tiles with each of the point values from 0 to 10. (Do this by hand first, then in Excel.)

For the spreadsheet, see `../Answers/ScrabbleSolution.xlsx`. You'll find instructions there for creating the proper kind of chart — Excel didn't guess correctly what was wanted.

(b) What are the mode, median and mean point values? Show them on your (hand written) chart.

mode	1	more tiles have this value than any other value
median	1	half the tiles are worth at most 1, half at least 1
mean	1.87	this is the "average" value of a tile — the sum of `D8:D18` divided by the number of tiles (the sum of `B8:B18`)

(c) What percentage of the tiles are worth more than 1 point?

23%.

(d) What percentage of the tiles are worth less than the median number of points?

Usually, 50% are less than the median by definition. In this particular case when the median, 1, is so common (in fact it's the mode) only 2% of the tiles are less valuable.

(e) Answer these questions for Scrabble in some other language (your choice) and discuss the differences between that language and English.

I haven't done this.

Exercise 6.12.31. [S][Section **??**][Goal **??**][Goal **??**][Goal **??**] Many flights arrive early!

The spreadsheet at `ArrivalDelays.xlsx` contains data on how many minutes late American Airlines flights to Boston's Logan airport were in January, 2014.

(1) What does a "negative delay" mean?

(2) Later you'll be asked to draw a histogram of this data in Excel. Sketch a neat approximate version first, with proper titles and reasonable scales for both axes and a proper title for the whole chart. You don't need to draw all the bars!

(3) Draw your histogram with Excel. Does it match your sketch?

(4) How many flights were counted in this data?

(5) What percentage of the flights arrived on time?

(6) Use the data to estimate the mode, median and mean arrival delay. Show these values on your histogram sketch.

(7) Flights that are more than two hours late are *outliers* — the delay is probably not American Airlines' fault. Estimate the mode, median and mean arrival delays if you don't include the outliers.

(1) What does a "negative delay" mean?

A negative arrival delay means the flight arrived early.

(2) Later you'll be asked to draw a histogram of this data in Excel. Sketch a neat approximate version first, with proper titles and reasonable scales for both axes and a proper title for the whole chart. You don't need to draw all the bars!

Not done here.

(3) Draw your histogram with Excel. Does it match your sketch?

See the spreadsheet `../Answers/ArrivalDelaysSolution.xlsx`

(4) How many flights were counted in this data?

There were 861 American Airlines flights to Logan in January 2014.

(5) What percentage of the flights arrived on time?

$534/861 = 62\%$ of the flights were on time (or early).

(6) Use the data to estimate the mode, median and mean arrival delay. Show these values on your histogram sketch.

(a) Mode: -10 to 0 minutes (highest bar).
(b) Median: -10 to 0 minutes (flight number 430 out of 861).
(c) Mean: 6.5 minutes late, computed in the spreadsheet as a weighted average.

(7) Flights that are more than two hours late are *outliers* — the delay is probably not American Airlines' fault. Estimate the mode, median and mean arrival delays if you don't include the outliers.

I redid the calculations in the spreadsheet using only the flights whose delays were less than 120 minutes. That omitted 20 flights.

(a) Mode: unchanged at -10 to 0 minutes (highest bar)
(b) Median: unchanged -10 to 0 minutes (flight number 420 out of 841)
(c) Mean: 2.0 minutes late, computed in the spreadsheet as a weighted average.

Exercise 6.12.32. [S][Section **??**][Goal **??**][Goal **??**] Quotes in *Common Sense Mathematics*.

The spreadsheet `CSMquotes.xlsx` contains data on the number of words in quotes used in an early draft of this text.

(a) Create a properly labeled histogram displaying the data. You may sketch the histogram with pencil and paper, or use Excel.

(b) Calculate the total number of quotes.

(c) Estimate the total number of words in the quotes.

(d) Estimate mode, median and mean quote sizes, and mark them on your histogram.

(e) Explain why the mean is the largest of the three averages.

(f) Estimate the total number of words in the text.

(g) Estimate the percentage of words in the text that are in quotes.

(a) Create a properly labeled histogram displaying the data. You may sketch the histogram with pencil and paper, or use Excel.

See `../Answers/CSMquotesSolution.xlsx`.

(b) Calculate the total number of quotes.

From Excel: 192.

(c) Estimate the total number of words in the quotes.

From Excel: $11740 \approx 12,000$.

(d) Estimate mode, median and mean quote sizes, and mark them on your histogram.

From Excel:

mode	20-30 words per quotation
median	just over 50 words
mean	61

(e) Explain why the mean is the largest of the three averages.

The skew in the data is just like that for income distribution. The few long quotes increase the mean without affecting the mode or the median.

(f) Estimate the total number of words in the text (you can use the online copy of the text in the lab if you wish).

(g) Estimate the percentage of words in the text that are in quotes.

The text has about 300 pages. I estimate 30 lines per page (low, but there are blank pages and figures), 20 words per line for a total of 180,000 words. With those estimates the fraction of words in quotes is 0.06522, so about 7 percent.

Exercise 6.12.33. [S][W][Section **??**][Goal **??**][Goal **??**][Goal **??**] Ricky's tacos.

A story in *The Boston Globe* on February 7, 2015 stated that

> Food prices over the past year have increased at four times the rate of overall inflation, with fresh products, such as meat, vegetables, and dairy, soaring even faster. Ground beef prices, for example, are up about 20 percent from a year ago. Shoppers at local grocery stores have felt the sharp rise in prices, but for Ricky Reyes, owner of the taqueria on Dorchester Avenue, costlier ingredients mean it is getting harder to keep the price of his signature beef taco down. [R**??**]

Use Table **??** to answer the following questions. We've entered the data in the spreadsheet `tacos.xlsx`.

(a) Is the *Globe* correct about the percent increase in the cost of beef?

(b) Fill in the column showing the percent increase in cost of each of the ingredients.

Ingredient	2013 cost per pound	2014 cost per pound	percent increase in cost	percent of taco filling
Beef	$3.46	$4.16		45
Tomato	$1.73	$2.19		20
Cheese	$5.39	$5.44		20
Lettuce	$0.99	$1.11		15

Table 6.40. Taco costs

[See the back of the book for a hint.] Your spreadsheet can do the arithmetic using the formula

$$\frac{\text{new value} - \text{original value}}{\text{original value}}$$

(c) Find the cost of a pound of taco filling in 2013 and 2014. Then find the percent increase in the cost of the filling.

(d) One way for Mr. Reyes to reduce the cost increase would be to change the percentages of meat and cheese, keeping the lettuce and tomato the same. What would the percent of each be if he wanted to keep the increase in a pound of filling to just 10%?

(e) Do you think customers would notice if Mr. Reyes changed the recipe using your answer to (d)?

[See the back of the book for a hint.] For part (d), use guess-and-check in Excel. Set up the computation so that when you change the percentage of meat the percentage of cheese and the total change automatically.

My solution is in the spreadsheet `../Answers/tacoSolution.xlsx`. I've copied the answers here.

(a) Is the *Globe* correct about the percent increase in the cost of beef?

Yes. 4.16/2.46 = 1.2023, which is "about a 20% increase".

(b) Fill in the column showing the percent increase in cost of each of the ingredients.

I did that in column `D`.

(c) Find the cost of a pound of taco filling in 2013 and 2014. Then find the percent increase in the cost of the filling.

Column `G` (`H`) has the 2013 (2014) cost of $3.13 ($3.56). Increase is 13.9%.

(d) One way for Mr. Reyes to reduce the cost increase would be to change the percentages of meat and cheese, keeping the lettuce and tomato the same. What would the percent of each be if he wanted to keep the increase in a pound of filling to just 10%?

I played with the meat percent in cell `I7`, adjusting cheese percent in I9 automatically, until `J12` was about 1.1. The answer: 54% meat instead of 45%, 11% cheese instead of 20%.

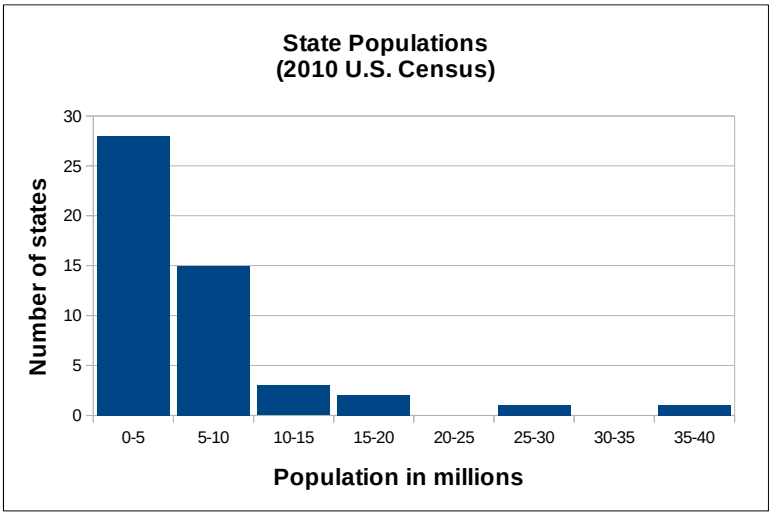


Figure 6.41. State populations

(e) Do you think customers would notice if Mr. Reyes changed the recipe using your answer to (d)?

I think they'd notice that there's only about half as much cheese. I think they would rather pay a little more for their favorite taco.

Exercise 6.12.34. [S][Section **??**][Goal **??**][Goal **??**] State populations.

Figure **??** shows the U. S. population distribution among the 50 states based on the 2010 U.S Census. The data are in `StatePopulations2010.xlsx`.

(a) Recreate the histogram in Excel.

(b) What is the modal population of states?

(c) Estimate the median population of states from the histogram. Compare that to the median population Excel calculates.

(d) Estimate the mean population of states from the histogram. Compare that to the mean population Excel calculates.

(a) Recreate the histogram in Excel.

See `../Answers/StatePopulations2010Solution.xlsx`.

(b) What is the modal population of states?

The highest bar is the first one, counting states with fewer than 5 million people, so the modal population is 0-6 million. I'd count 2.5 million as a correct answer.

(c) Estimate the median population of states from the histogram. Compare that to the median population Excel calculates.

There are fifty states so the median population is the 25th in size. There are 28 states with a population less than 5 million, so I will estimate that the 25th has population about 4.5 million.

Excel says the median is 4436369.5 million, so my estimate is pretty good.

(d) Estimate the mean population of states from the histogram. Compare that to the mean population Excel calculates.

I used the midpoint of each 5 million range as the typical population for the states in that range and calculated the weighted average (with Excel). The result was 6.4 million. The `AVERAGE` function found the true mean to be 6.16 million. That's an error of just about 4%.

Review exercises.

Exercise 6.12.35. [A] [R][Section **??**][Goal **??**][Goal **??**]

Create an Excel spreadsheet and put the following numbers in the first column.

14 15 22 50 0 33 16 18 23 40 47

(a) Use Excel to find the mean, median and mode of these numbers

(b) Change the first number from 14 to 23. How do the Excel calculations change?

(c) Click the "undo" button and confirm that Excel reverts back to the original set of numbers.

(d) Change the last four numbers to 0 (so that the data now read

14 15 22 50 0 33 16 0 0 0 0

How do the different averages change? Explain how the data are skewed.

Exercises added for the second edition.

Exercise 6.12.36. [S] Jonathan Dushoff's beer.

Jonathan Dushoff says he averages two beers a week but drinks just one beer in an average week.

Invent a months worth of data that explains this seeming contradiction, with one of the averages the mean and the other the mode.

Jonathan might drink one beer in each of the first three weeks of the month and five in the last week. Then his mean beer consumption is

$$\frac{8 \text{ beers}}{4 \text{ weeks}} = 2 \frac{\text{beers}}{\text{week}}$$

while in a typical week — the mode — he drinks just one beer.

Exercise 6.12.37. [U][Section **??**][Goal **??**] Car and truck prices.

The article on new car and truck prices that we studied in Section **??** first asserts that

> . . . the average price of a new vehicle in the second quarter [of 2008] fell 2.3 percent from a year earlier to $25,632 . . .

and later

> The result is the average new vehicle now costs less than 40 percent of an average household's median annual income, the analysts said, whereas from 1991 to 2007, it would cost more than half of the median income. [R**??**]

Verify as much of this last assertion as you can.

Exercise 6.12.38. [U][A][Section **??**][Goal **??**] [Goal **??**][Goal **??**] Enrollments.

The final enrollment report for the past year at an unnamed small college provided the following information about students: 450 freshmen, 421 sophomores; 400 juniors and 511 seniors.

(a) Create an Excel spreadsheet containing this data. Label the columns appropriately.

(b) Ask Excel to calculate the total number of students enrolled during the past year. Label this result.

(c) Create a properly labeled bar chart of the student data.

(d) A corrected enrollment report noted that there were 419 juniors. Make that adjustment in your spreadsheet and check that the other information (total number of students, bar chart) is correctly updated.

(e) Using this new information, ask Excel to calculate the percentage of students who are freshmen, sophomores, juniors and seniors. Copy and paste so that you type as few formulas as possible.

(f) Create a new bar chart displaying the percentages.

(g) Convert your bar chart to a pie chart.

Exercise 6.12.39. [U][A][Section **??**][Goal **??**] SAT percentiles.

A student received this notification on his college entrance exam:

English Language Arts:	77th percentile
Mathematics:	88th percentile

Explain this report in everyday language. Your answer might begin "More than three quarters of the students taking this test . . . "

Exercise 6.12.40. [U][Section **??**][Goal **??**] Comparing the states.

You can do this exercise using Excel, or with properly documented research. (Your instructor may specify one method or the other.)

Find the mean, median and mode for the populations of the 50 states.

Display the answers to the previous question on a properly labeled histogram. Discuss your findings — is the distribution skewed?

Redo parts (a) and (b) for the areas of the states.

Redo parts (a) and (b) for the population densities (people per square mile).

Exercise 6.12.41. [N][Section **??**][Goal **??**] Jellybean margin of error.

`andrewgelman.com/2011/08/that_xkcd_carto/`

Exercise 6.12.42. [U][Section **??**][Goal **??**] [Goal **??**][Goal **??**][Goal **??**] Reputation on stack exchange.

Stackexchange.com (`stackexchange.com`) is a network of online question and answer websites. Users who post questions and provide answers earn reputation based on community feedback. Table **??** shows the number of users with reputations in certain ranges on January 6, 2013 for all stackexchange sites and for the particular site `tex.stackexchange.com/` where the authors have asked and answered questions about the TEX software used to prepare the manuscript for *Common Sense Mathematics*. You can find current information at `stackexchange.com/leagues` .

Estimate the mode, median and mean for each distribution. This is subtle in several ways. The bucket sizes vary. Data at the top and bottom end of the range are very scarce. Ask about sensitivity to the assumptions made there about the actual means for the top and bottom categories.

Exercise 6.12.43. [U][Section **??**][Goal **??**] Deceptive pie charts.

Build the pie chart from Section **??**.

(a) Experiment with the pie chart features to make it look like the managers' salaries are the largest. You can't actually change the data to do this — you need to use the 3D and other pie chart features to make it look like the managers' salaries are large.

(b) Play around with different types of charts in Excel. Find a chart type and an effect (3D, most likely) that really distorts the data.

Reputation	Users
100,000+	97
50,000+	297
25,000+	938
10,000+	3,249
5,000+	6,874
3,000+	11,150
2,000+	15,650
1,000+	24,867
500+	34,857
200+	45,107
1+	1,478,007

All sites

Reputation	Users
100,000+	2
50,000+	14
25,000+	24
10,000+	60
5,000+	91
3,000+	132
2,000+	179
1,000+	281
500+	395
200+	591
1+	18,417

TEX— LATEX

Table 6.42. Stackexchange reputation

Exercise 6.12.44. [U][C][Section **??**] [Goal **??**][Goal **??**] River lengths.

At `en.wikipedia.org/wiki/List_of_rivers_by_length` Wikipedia offers a chart of 163 major rivers, organized by length.

(a) Construct a bar chart with nine categories, using the first digit of the length of the river to determine the category.

You might expect all the bars to be the same height, since there are nine possible starting digits. But they're not. In fact, there are no rivers longer than 6000 km.

The fact that for the short rivers there are more whose length begins with small digits is an instance of *Benford's law*. You can look up more about it if you're curious.

(b) Use Excel to create new columns with river lengths measured in yards, in feet and in inches. Draw each of those bar charts. Discuss what you see.

Exercise 6.12.45. [Section **??**] [Goal **??**][Goal **??**] Fight for the Senate (2016).

A graph like the one in Figure **??** appeared in Nate Silver's 538 website on November 7, 2015. The *y*-axis displays the number of seats held by each party, the *x*-axis the probability of that outcome.

Nate Silver constructed this histogram by imagining (simulating) many thousands of elections and recording the percentage of time each Democratic/Republican split occurred.

(a) What is the most likely number of Democratic senators?

(b) What number of Democratic senators represents the mode of this distribution?

(c) What is the probability that there are more than 50 Democratic senators?

(d) What number of Democratic senators is the median of this distribution?

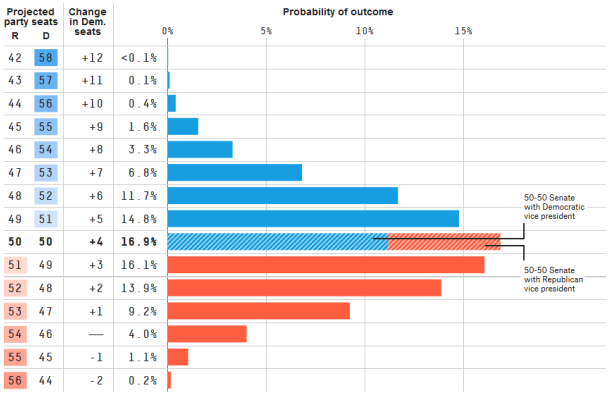


Figure 6.43. The fight for the Senate [R**??**]

(e) If you had the complete list of all Nate Silver's imagined elections and sorted it by the number of Democratic senators, how many Democratic Senators would there be in the middle election on that list?

(f) Use Excel to compute the (weighted) average number of Democratic senators for these imagined elections.

(g) What actually happened in the election?

Exercise 6.12.46. [N] Boston's payroll.

This histogram appeared in *The Boston Globe* on February 14, 2017.

The raw data are available on the web. There's an anonymized copy at `Boston14Payroll.xlsx` [R**??**] .

The histogram and the spreadsheet suggest many possible questions. Some suggestions:

(a) Estimate bar heights from the graphic.

(b) Estimate median and mean, with Excel or otherwise.

(c) Discuss how you would report the mode in order to convey the most information.

(d) Check the estimates of the median and mean by looking at the raw data.

(e) Check estimates of the bar heights by reconstructing the histogram from the raw data.

Ask the usual questions: mode, median, mean. Bimodality is interesting.

Exercise 6.12.47. [N] Votes weighted by IQ.

An answer to the question "What would be the possible issues with an IQ based voting system?" at `worldbuilding.stackexchange.com/questions/83199/what-would-be-the-possible-issues-with-an-iq-based-voting-system` offers this:

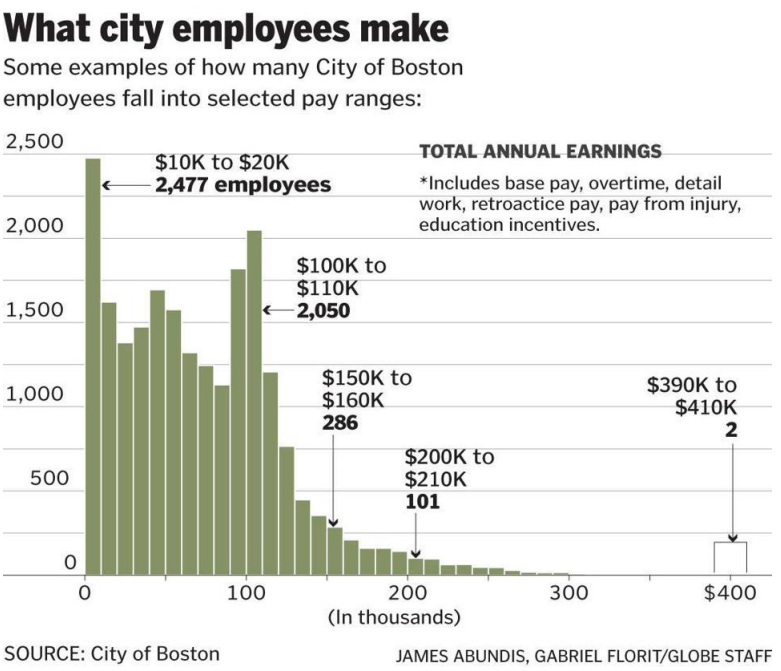


Figure 6.44. Boston Municipal Payroll [R**??**]

IQ	Percentile	% population in bracket	number of people	average IQ	weighted votes	share of weighted votes
150	0.9995709	0.04%	42911.6534	150	6436748.01	0.06%
140	0.9961696	0.34%	340130.8591	145	49318974.57	0.49%
130	0.9772499	1.89%	1891963.691	135	255415098.3	2.55%
120	0.9087887	6.85%	6846121.994	125	855765249.2	8.56%
110	0.7475075	16.13%	16128118.54	115	1854733632	18.55%
100	0.5	24.75%	24750753.29	105	2598829095	25.99%
90	0.2524925	24.75%	24750753.24	95	2351321558	23.51%
80	0.0912113	16.13%	16128118.54	85	1370890076	13.71%
70	0.0227501	6.85%	6846121.994	75	513459149.5	5.13%
60	0.0038304	1.89%	1891963.691	65	122977639.9	1.23%
50	0.0004291	0.34%	340130.8591	55	18707197.25	0.19%
40	3.169E-05	0.04%	39743.0499	45	1788437.246	0.02%
30	1.532E-06	0.00%	3015.3698	35	105537.943	0.00%
20	4.832E-08	0.00%	148.402	25	3710.05	0.00%
10	9.9E-10	0.00%	4.7327	15	70.9905	0.00%
0	0	0.00%	0.099	5	0.495	0.00%

Figure 6.45. Votes weighted by IQ [R**??**]

If you're giving everyone exactly as many votes as their IQ — the effect on the actual vote doesn't appear to be very much at all. See the table in [Figure **??**] for a population of 100 million to illustrate — the vote share for IQ 120 + is slightly higher than with a normal democracy, but the fact there are exponentially fewer people in these higher intelligence brackets means that the linear multiplier on their vote weight has less and less of an effect.

The comments are interesting too.

Exercise 6.12.48. [U][N] How cold is it really?

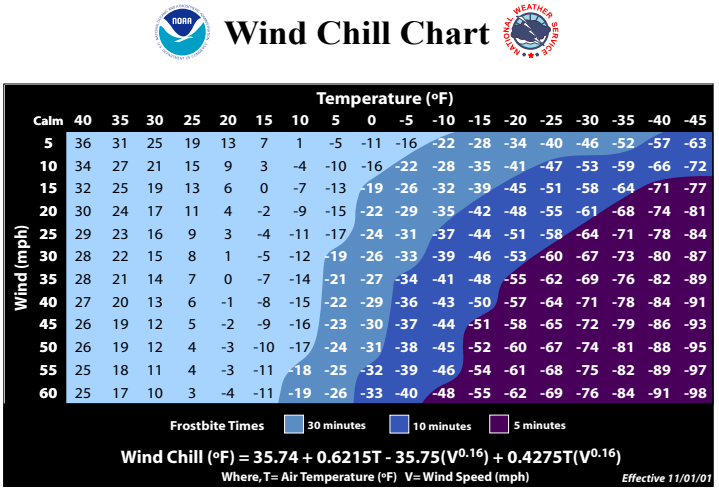


Figure 6.46. Calculating the Wind Chill [R**??**]

Figure!**??** shows the National Weather Service calculation for how cold it feels in terms of the temperature and the speed of the wind.

The formula there came from research done by Maurice Bluestein; you can read the story in his obituary at `www.nytimes.com/2017/09/14/science/maurice-bluestein-who-modernized-the-wind-chill-index-dies-at-76.html`

There's a spreadsheet you can play with at `windchill.xlsx`.

The New York Times

Exercise 6.12.49. [U][W][N] "Average" household wealth.

On December 8, 2017 you could read in a Associated Press Report in *The Boston Globe* headlined "Surging stocks lift US wealth, yet most still trail '07 peak" that

> Surging stock prices and steady increases in home values powered American household wealth to $96.9 trillion this fall, the Federal Reserve said Thursday. The gains, however, aren't widely shared.
>
> . . .
>
> In 2016, the latest figures available, median household wealth was still 34 percent below its prerecession, 2007 level. Average household wealth, meanwhile, fully recovered from the downturn and was 7 percent higher last year. The average figure is pulled up by very wealthy families.
>
> While average household wealth reached $667,600 in 2016, net worth for the median household was just $78,100. [R**??**]

Note: There's even enough information here to determine the number of households - for a sanity check.

Note: Is this a place for a discussion of the difference between wealth and income?

Exercise 6.12.50. [U] Median age vs. mean age?

On September 9, 2018 Dante Ramos wrote in *The Boston Globe* that

> According to research by Portland State University in Oregon, the median age of voters in a Boston mayoral election is 51, more than 14 years older than the average adult in the city. [R**??**]

(a) Why is it correct to use the median age rather than the average (mean) age in this report?

(b) Why is it better to write "older than tha average adult" rather than "greater than the average adult's age"?

(c) Think of a situation where it would be more useful to know that mean age rather than the median age.

R?? en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg (last visited August 3, 2015). Licensed under the Creative Commons Attribution 2.5 Generic License.

R?? Data source: Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 9 Regs Limited-Use, Nov 2008 Sub (1973-2006) , National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2009, based on the November 2008 submission. Graphic drawn by Ben Bolker.

R?? Raising Taxes on Rich Seen as Good for Economy, Fairness, Pew Research Center (July 16, 2012), www.people-press.org/2012/07/16/raising-taxes-on-rich-seen-as-good-for-economy-fairness (last visited July 28, 2015). Quoted with permission.

R?? E. Moskowitz, Cash-strapped T proposes 23 percent fare increase, *The Boston Globe* (March 28, 2012), bostonglobe.com/metro/2012/03/28/mbta-unveils-percent-fare-hike-limited-service-cuts-also-proposed/moCl42rwr0Nf5xyx20ZQGP/story.html (last visited March 29, 2020).

R?? International Programs, United States Census, www.census.gov/population/international/data/idb/informationGateway.php , select Population Pyramid Graph — 2010 — United States, (last visited August 11, 2015).

R?? International Programs, United States Census, www.census.gov/population/international/data/idb/informationGateway.php , select Population Pyramid Graph — 2010 — Sudan, (last visited August 11, 2015).

R?? M. C. Fisk and J, Lawrence, Walmart to Settle Massachusetts Suit for $40 Million (Update2), *Bloomberg News* (December 2, 2009), www.bloomberg.com/apps/news?pid=newsarchive&sid=a2AClc9J8WwE (last visited October 2, 2015).

R?? Graphic redrawn from data from data J. P. Kahn, Missed connections in our digital lives, *The Boston Globe* (April 15, 2012), www.bostonglobe.com/metro/2012/04/14/missed-connections-our-digital-lives/bPHauWdvUl5XAd1ol7SOQL/igraphic.html (last visited July 31, 2019).

R?? Jakob Nielsen, Aspects of Design Quality Nielsen Norman Group (November 2, 2008), retrieved (with permission) from www.nngroup.com/articles/aspects-of-design-quality/ (last visited February 28, 2020).

R?? K. Geldis, The Richest Counties in America, *TheStreet* (February 13, 2012), www.thestreet.com/story/11415107/3/the-richest-counties-in-america.html (last visited July 22, 2015).

R?? Wikipedia, https://upload.wikimedia.org/wikipedia/commons/3/3d/Distribution_of_Annual_Household_Income_in_the_United_States_2011.png (last visited September 22, 2019), (Creative Commons Attribution-ShareAlike License, en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License).

R?? Graphic redrawn from data scraped from a Nate Silver *New York Times* graphic published October 31, 2012. The original seems not to be available.

R?? Data from D. Slack, Boston spends most on firefighters in US, *The Boston Globe* (March 30, 2009), www.boston.com/news/local/massachusetts/articles/2009/03/30/boston_spends_most_on_firefighters_in_us/ , data for graphic at www.boston.com/news/

local/massachusetts/articles/2009/03/30/fire_spending/ , (last visited March 29, 2020).

R?? D. Slack and J. C. Drake, Error made in fire dept. report, *The Boston Globe* (March 31, 2009) www.boston.com/news/local/massachusetts/articles/2009/03/31/error_made_in_fire_dept_report/ (last visited July 22, 2015).

R?? Data from J. Stripling and A. Fuller, Presidents Defend Their Pay as Public Colleges Slash Budgets, *The Chronicle of Higher Education* (April 3, 2011), chronicle.com/article/Presidents-Defend-Their/126971 (last visited August 11, 2015).

R?? Paul Erdős, en.wikipedia.org/wiki/Paul_Erdos (last visited July 22, 2015), (Creative Commons Attribution-ShareAlike License, en.wikipedia.org/wiki/Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License).

R?? The distribution of Erdős numbers, The Erdos Number Project, wwwp.oakland.edu/enp/trivia/ , (last visited October 11, 2015), reproduced with permission.

R?? M. Schlueb and D. Damron, Activists press officials to put sick-leave proposal to voters, *Orlando Sentinel* (August 6, 2012), www.orlandosentinel.com/news/os-xpm-2012-08-06os-sick-leave-ballot-race-20120806-story.html (last visited July 28, 2019).

R?? M. Woolhouse, A Boston taco tells the tale of far-reaching food cost woe, *The Boston Globe* (February 06, 2015). www.bostonglobe.com/business/2015/02/05/food-prices-spike-increasing-cost-taco/vU3c42L99X9fBt25opSKkO/story.html (last visited December 16, 2015).

R?? New car prices fall at fastest rate ever, Associated Press reported in *The Boston Globe* (September 5, 2008) www.boston.com/business/articles/2008/09/05/new_car_prices_fall_at_fastest_rate_ever (last visited July 22,, 2015).

R?? Screen capture, projects.fivethirtyeight.com/2016-election-forecast/senate/?ex_cid=2016-forecast (last visited November 7, 2016).

R?? www.bostonglobe.com/metro/2017/02/14/police-detective-tops-boston-payroll-with-total-over/6PaXwTAHZGEW5djgwCJuTI/story.html (last visited February 14, 2017).

R?? City of Boston, Employee Earnings Report 2014, data.boston.gov/dataset/employee-earnings-report/resource/941c9de4-fb91-41bb-ad5a-43a35f5dc80f (last visited July 31, 2019).

R?? worldbuilding.stackexchange.com/questions/83199/what-would-be-the-possible-issues-with-an-iq-based-voting-system (last visited June 12, 2017), (Creative Commons Share Alike License: creativecommons.org/licenses/by-sa/2.5/legalcode) .

R?? www.nws.noaa.gov/om/cold/wind_chill.shtml (last visited September 15, 2017).

R?? C. Rugaber, "Surging stocks lift US wealth, yet most still trail '07 peak", Associated Press, *The Boston Globe*, December 8, 2017, www.bostonglobe.com/business/2017/12/07/surging-stocks-lift-wealth-yet-most-still-trail-peak/DSb0ih1lSlzlz8tqA6t7QO/story.html (last visited December 8, 2017).

R?? D. Ramos, Young voters, claim your power, *The Boston Globe*, September 9, 2018, www.bostonglobe.com/opinion/2018/09/07/for-millennials-power-there-for-taking/NITLEtmQtk4Ecubk7XhozL/story.html (last visited September 9, 2018)