gmeth: Graphic Methods

Lab 1

Author(s): Eric A. Cohen

This material is part of the ***statsTeachR*** *project*
Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported License: http://creativecommons.org/licenses/by-sa/3.0/deed.en_US

Overview

Why Graphic Methods?

When you have gathered data, all the information is there in the numbers. Why would you create pictures from your data, pictures which present only some of the information, possibly inaccurately?

We create graphic representation of data because they can convey so much, so quickly. Humans have amazing innate skills at processing visual information — likely developed by evolution for other purposes — but we can draft these skills in the service of statistics.

Use

Graphic methods are most used, first, in exploratory data analysis: getting an idea of what information you have and what patterns it might hold. And, second, at the other end of the research process, in presentation: conveying your results to your audience.

One-Dimensional Data

Often you will want to look at a collection of datapoints which vary in only one dimension. A list of the heights of undergraduate students at UMass Amherst in April, 2013; or age at death of 100 laboratory mice given a certain experimental treatment. Graphic methods are an easy way to see

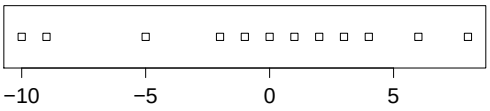
- Where do most of the values lie? (Mean, median, interquartile range. . . .)
- How high do the values go, how low? (Minimum and maximum.)
- Do the values cluster around one center, do they cluster around several dense spots, or not at all?
- Are they distributed symmetrically? And in what shape?
- Are there values that are particularly far from most of the others (outliers)?

Stripcharts

A stripchart simply shows the value of each datapoint, plotted along a single axis. It can be useful for seeing where values cluster, where most of the values lie, and whether there are any outliers.

Exercise 1

1. Using *File>Open* in RStudio, load the dataset `data_for_graphing.RData`
2. Output a stripchart of `data_1`. You should get something like this:



3. Output a stripchart of `data_2`. Then have a look at the contents of `data_2`, to check whether it has any repeated values. Where have those repeated values gone on your stripchart? Is this what you want? Figure out a way to better display a stripchart of data with repeated values.[1]

(Hint: for most of the exercises in this unit you may need to figure out which R function does what you want, and then learn what you need about the arguments and settings the function takes.)

Histograms

A histogram is a bar chart showing how many datapoints are in each of a number of ranges of values. For instance, with height data, how many students have heights of between 160 and 162 centimeters? Between 162 and 164 centimeters? And so on, for ranges up and down.

Exercise 2

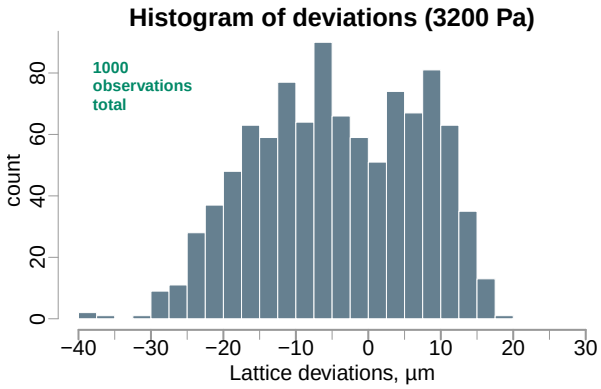
1. Output a histogram of `data_3`.
2. How many bins (ranges) did R automatically divide the data into? Plot a few more histograms of this data, with more and with fewer bins. (Hint: you want the `breaks` argument.) What are your concerns about a histogram plotted with too few bins, and what about a histogram plotted with too many?[2]
3. Using the histograms you just created,
   a) Is the distribution of `data_3` mound-shaped?
   b) Is it symmetric?
   c) Might it be a normal distribution?
4. Answer the same questions for `data_4`.

[1] A little harder, **question 3a**: output a stripchart of `data_3`. This has 1000 values, over a range of only about 60. Where are the repeated values? Does the technique of question 3, above, display them in a useful way? How could you show the shape of this data, in a useful way, using a stripchart?

[2] There are some rules of thumb about what the "right" number of bins is, but generally you will be better off looking at your data and trying a few alternatives.

5. Label your best histogram from question 2. Give it a title, make up labels and units for the axes, and add anything else you think makes the plot more explanatory. (It doesn't need to be as fancy as the one to the right, although you'll learn by trying.)

We haven't done it so far, but labeling plots is essential. This example uses hypothetical data, but any real-world plots that you present, for homework or for publication, are meaningless without information about what is being presented.



Stem and Leaf Plots

These present much the same information as histograms, indicating the density of data points in different ranges; however, they also show the numeric values of the data.

Example

The contents of `data_5`:

```
>data_5
 [1] 25 46 73 61 10 18 1 20 85 8 87 2 25 6 86 46 80 95 8 9
```

Maybe this would be clearer if the values were sorted:

```
>sort(data_5)
 [1] 1 2 6 8 8 9 10 18 20 25 25 46 46 61 73 80 85 86 87 95
```

Stem and leaf plot:

```
>stem(data_5, scale=2)
The decimal point is 1 digit(s) to the right of the |
0 |126889
1 |08
2 |055
3 |
4 |66
5 |
6 |1
7 |3
8 |0567
9 |5
```

A stem and leaf plot:

1. Rounds each value to two significant digits;
2. Shows the left digit of each rounded value as the "stem";
3. For each stem, shows the set of all following digits as the leaves.

So, the presence of 10 and 18 in data_5 accounts for the row `1 |08` in the plot. The presence of 46 twice in data_5 accounts for the row `4 |66`.

Exercise 3

1. Create a stem and leaf plot of the data set (1, 13, 2, 23, -10, 4, -19, 14, -11, 2, 15, 13, 0). (You can do this by hand or in R.)
2. The "scale" argument to `stem` is similar to setting the number of bins in a histogram. Output a stem and leaf plot for data_5 without the scale parameter. How is it different from the plot shown above? Which one gives a better idea of the shape of the distribution?

Kernel Density Plots

A kernel density plot is similar to a histogram in showing the shape of a distribution, but it displays a smooth curve instead of stairstep bars. The math behind it is hairy, but essentially a kernel density plot says "Given these values, if they *were* generated as the sum of a few normal distributions,[3] what would the those distributions be?" It then plots the output of that sum.

The "bandwidth" of a kernel density plot is, to some extent, analogous to the bin size of a histogram.

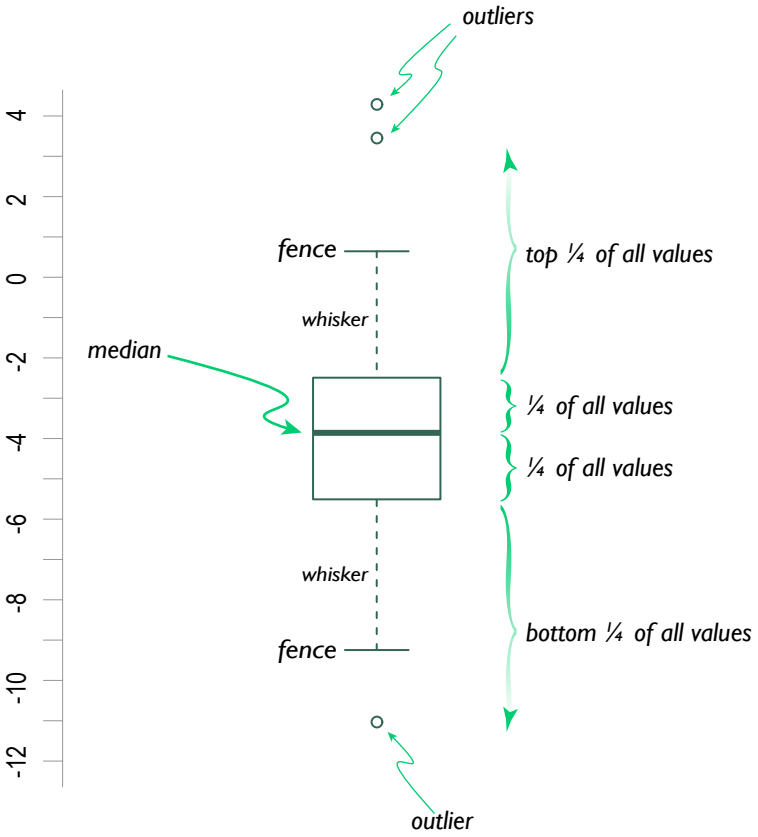
Exercise 4

1. Output kernel density plots of `data_3` and `data_4` . Try a couple of different bandwidths. Do these plots change your answers to questions 3 and 4 in exercise 2?

[3] Or square distributions, or triangle distributions.

Box Plots

A box plot offers a visual summary of a data set along with some details about its range and mean:



- The central line shows the median of the data.
- The box shows how far up and down the middle half of the values go, giving an idea of how spread out the central mass of the data is.
- The fences show the highest and lowest values, excluding outliers.
- Outliers are plotted as points outside the fences.

The distance between the top and bottom of the box is the *interquartile range* (IQR). Values further from the median than 1.5 times the IQR are considered outliers.[4]

[4]This is the most common usage, but some box plots use other values to define fences and outliers. In your own plots it is best to be clear about your usage, and keep an eye on the fine print around box plots that you encounter in reading or research.

Exercise 5

Referring to the plot above:

1. What is the median of this data set?
2. How many outliers does this data have (by the definition above) and what are their values?
3. Give three different ranges, each of which contains half of the data points in this set.

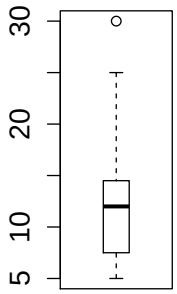
Exercise 6

In R, it is easy to create quick vectors of made-up data. For instance:

command	x contains
`x <- c(2, 3, 5, 7, 11, 100)`	`2, 3, 5, 7, 11, 10`
`x <- (5:10)`	`5 6 7 8 9 10`
`x <- rep(10, times=3))`	`10 10 10`
`x <- c(1, 2, 3, 5, 10:15, rep(100, times=3)`	`1 2 3 5 10 11 12 13 14 15 100 100 100`

Then, it's easy to make boxplots:

```
>x <- c(rep(5, times=3), 10:15, 25, 30)
>boxplot(x)
```



1. Enter and run the commands given above. Make sure that `x` contains what you expect, and that you can generate the box plot.

Then, using these (or other) commands, make data sets to create:

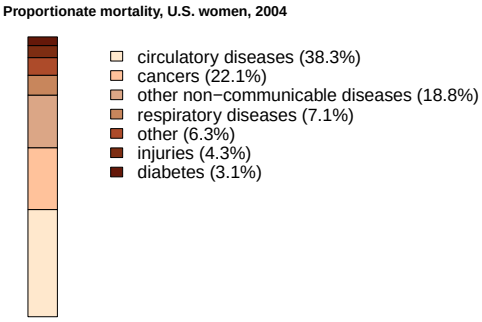
2. A box plot with a tall box and very short whiskers.
3. A box plot with most of the box above the mean.
4. A box plot with a short top whisker and a long bottom whisker.

Pie Charts and Stacked Bar Charts

A pie chart is used to show how a whole is divided into proportions. Pie charts are generally inappropriate in scientific communication, as people are bad at intuitively grasping relative angles or areas, so a pie chart can easily mislead.

In most contexts this is the only good pie chart:[5]

A stacked bar chart is a more accurate way of presenting data in which components sum to a fixed total: [6]



A clear introduction to stacked bar charts in R is here, http://www.cs.grinnell.edu/~rebelsky/Courses/MAT115/2008S/R/stacked-bar-graphs.html (retrieved 5 June 2013).

[5] Laszlo Thoth, licensed under a Creative Commons Attribution-Noncommercial-Share Alike 2.5 License. http://tongodeon.livejournal.com/583338.html, retrieved 3 June 2013.

[6] World Health Organization Global Infobase, Mortality: https://apps.who.int/infobase/Mortality.aspx?l=&Group1=RBTCntyByRg&DDLCntyByRg=AMR&DDLCntyName=1002&DDLYear=2004&TextBoxImgName=go, retrieved 4 June 2013.

Comparing One-dimensional Data

Since these graphical techniques provide quick ways of summarizing one-dimensional data, they can also be used as quick ways to see the similarities and differences among two or more data sets.

Exercise 7

1. Using *File>Open*, load the dataset `temperature_data.RData`.

This contains a single list, named `temperatures`. Each item in that list is a vector of numbers

These vectors are combined in a list, rather than a dataframe, because the vectors are of different lengths. A dataframe is like a table of data and a table has columns all of the same length. A list is a series of items, and the items don't have to have anything in common. In this case, the items are all vectors of numbers but of different lengths

2. Create a boxplot of `temperatures`.

R automatically plots three boxplots, one for each vector of numbers, within a single set of axes. This default behavior is just what we want right now.

3. This data is of maximum daytime temperatures, recorded over a period of four years, at three different locations.[7] So make the boxplot again, adding a label to the y-axis saying "Temperature (°C)" and a title "Maximum daytime temperatures, 2005–2008".
4. One location is Volcanoes National Park, Hawaii; one is Oklahoma City, Oklahoma; and one is Amherst, Massachusetts. Which is which, and how do these plots help you decide? (Hint: having trouble being sure? Which location would you rather vacation at, in winter *and* summer?)

It can similarly be informative to compare histograms of various data sets. Plotting several histograms at once in R takes a bit more work than plotting several box plots, but until you learn that you can always plot your histograms independently and then compare them.

Two-dimensional Data

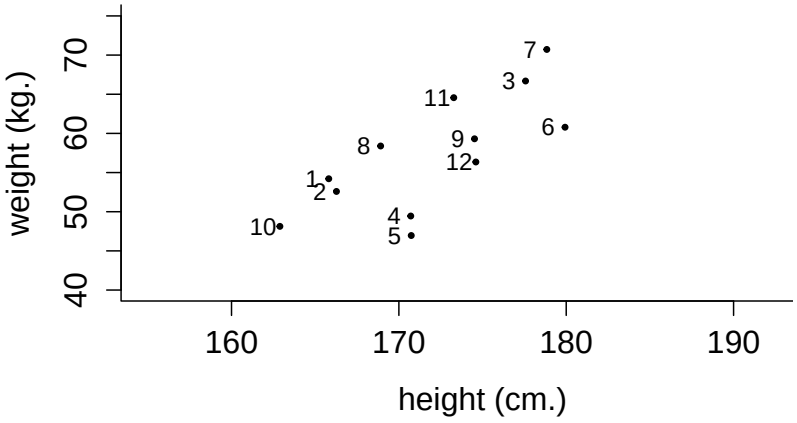
So far we have been visualizing data sets that vary along only one axis. But of course most research involves measurements of several quantities for each data point. For instance,

- Height and weight of each undergraduate at the University of Massachusetts, Amherst.
- Age, BMI, blood pressure, and forced expiratory volume of each of 100 patients hospitalized with COPD.

[7] Data obtained from U. S. National Oceanic and Atmospheric Administration historical weather station data, available for download at http://www.ncdc.noaa.gov/cdo-web/#t=secondTabLink, retrieved 2 June 2013.

Scatter plots

A scatter plot shows the value of two quantities for each data point:



In this scatter plot of the heights and weights of 12 people, we can see that subject 10 had a height of ~163 cm., and a weight of ~48 kg.; and so on. We can also see (which we couldn't have known from a plot of height or weight alone) that weight seems to increase as height increases.

Exercise 8

1. Using *File>Open*, load the dataset `height_and_weight_data.RData`.[8]
2. Click the dataframe `ht_wt` in the Workspace pane to get an idea of what the dataframe contains.
3. Create a scatterplot of `ht_wt`, using the `plot` command. Include labels specifying "height (cm.)" for the x axis, and "weight (kg.)" for the y axis.

If it is given a two-column dataframe for its data, the `plot` command assumes that you want a scatterplot, with the second variable plotted against the first variable — this is what you just did.

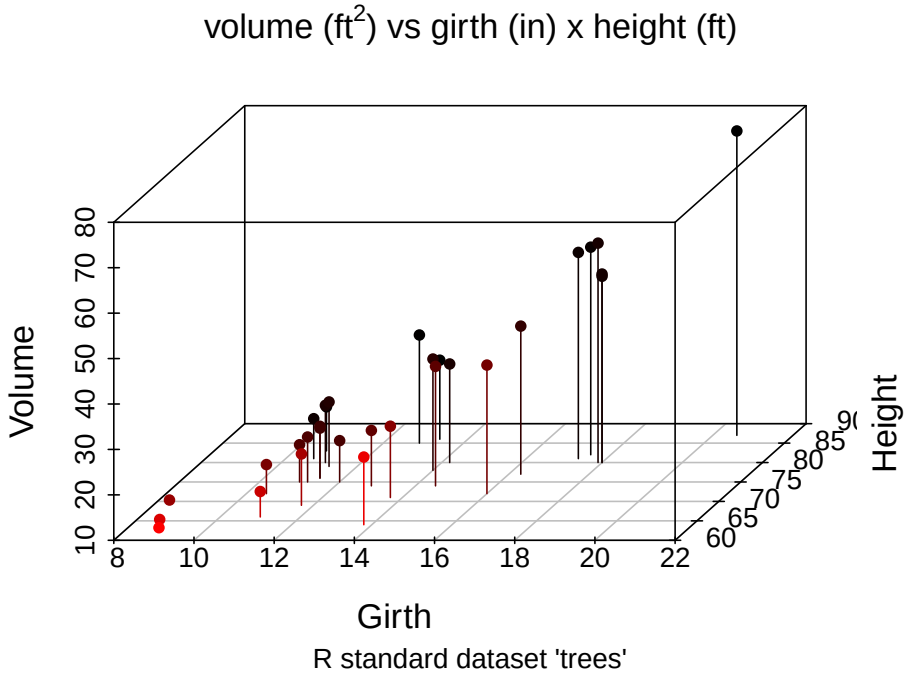
4. Based on this sample of 200 subjects, would you still say that weight increases as height does? Would you want the job of predicting *accurately* a person's weight given only their height?

[8] So, Hung-Kwan, et al., Secular changes in height, weight and body mass index in Hong Kong Children, BMC Public Health. 2008 Sep 21; 8:320., doi: 10.1186/1471-2458-8-320. As abstracted in Statistics Online Computational Resource wiki, http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights, retrieved 1 June 2013.

Multi-dimensional Data

Three-dimensional data

R can also plot three-dimensional data as a scatterplot. Using the optional package *scatterplot3d*, we can obtain plots such as this:[9]



Higher-dimensional data

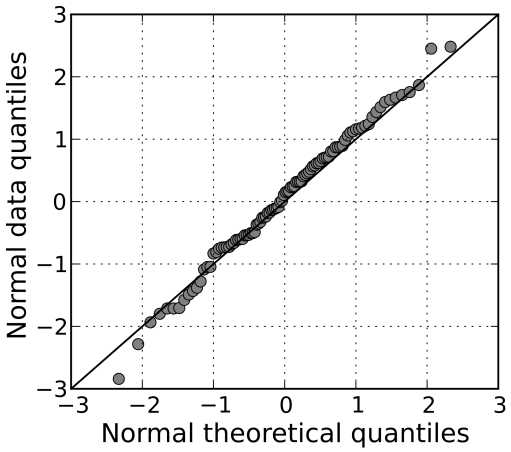
Unfortunately the nature of the space-time continuum and the human mind make it impossible for us[10] to visualize images in more than three dimensions. To investigate possible correlations between more than three variables, it can be valuable to examine several scatterplots, each showing two of the variables. Note that the number of possible two-dimensional scatterplots will increase exponentially with the number of variables, so with more than a few variables you will need to use real-world knowledge to choose which pairs of variables are most interesting.

[9]Code very slightly modified from that here: R-enthusiasts, R Graph Gallery, Sactter Plot 3D, http://gallery.r-enthusiasts.com/graph/Scatter_plot_3D_44, retrieved 23 June 2013

[10]Most of us.

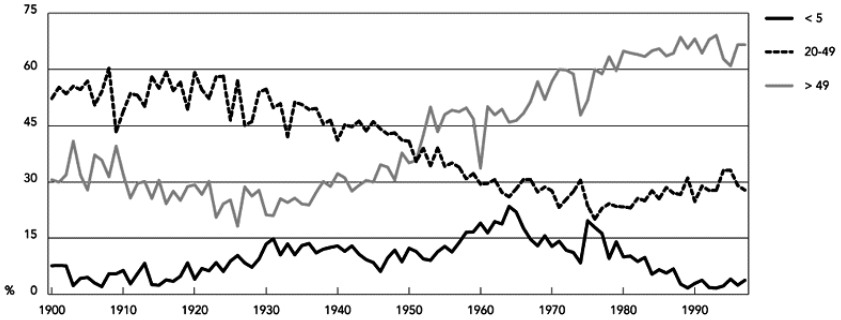
Other techniques

Other graphical techniques are available for more advanced or more specialized analyses, including:

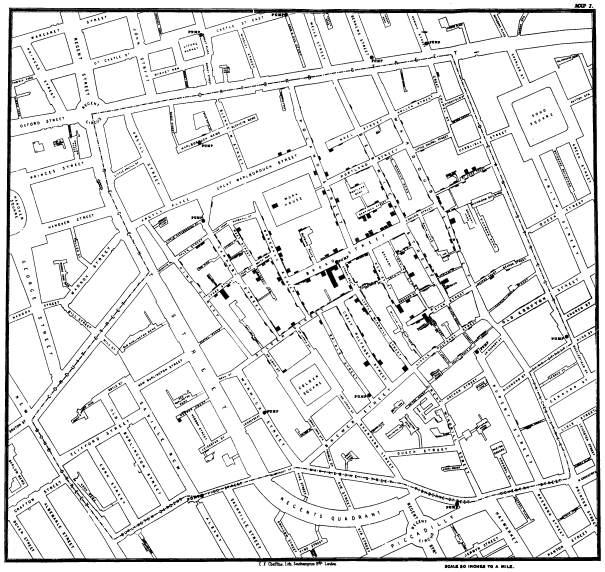


Quantile-quantile (Q-Q) plots[11]

[11] Wikimedia commons, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license, http://commons.wikimedia.org/wiki/File:Normal_normal_qq.svg retrieved 21 June 2013



Time-series plots[12]



Maps (John Snow's original map of London cholera outbreak, 1854.[13])

These types of plots and others, can be valuable tools in investigating or presenting complex data.

[12] Antunes, Jos Leopoldo Ferreira et. al. Tuberculosis in the twentieth century: time-series mortality in So Paulo, Brazil, 1900-97 Cad. Sade Pblica vol.15 n.3 Rio de Janeiro July/Sept. 1999 http://dx.doi.org/10.1590/S0102-311X1999000300003 retrieved 21 June 2013

[13] Wikimedia commons http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg, retrieved 22 June 2013