Visualizing Data: Basic Plot Types

Data Science 101

Stanford University, Department of Statistics
Today’s lecture focuses on these basic plot types:

- bar charts
- histograms
- boxplots
- scatter plots
- densities

Which visualization is best depends on

- whether the data are univariate or bivariate data
- whether the variables are discrete or continuous
- context
The importance of context for graphics

Figure 1. Connecticut Traffic Fatalities, 1955-1956

fails to control for the six common threats to the validity of experiments specified below:

1. History. This term denotes specific events, other than the experimental treatment, occurring between the pretest and posttest, which might account for the change. It furnishes a "rival hypothesis" to the experimental hypothesis, a competing explanation of the before-to-after change that must be eliminated as implausible, by one means or another, before full credence can be given to the experimental hypothesis. For instance, 1956 might have been a particularly dry year, with fewer accidents due...
The importance of context for graphics

Figure 2. Connecticut Traffic Fatalities, 1951-1959

A coincident reform of the record keeping, ruling out valid inferences as to effects. The Chicago police reform cited above is a case in point. In the present instance, we have found no evidence of a change in record keeping or index computing of the type that would produce a pseudo-effect.

The likelihood of regression, or of selection for "treatment" on a basis tending to introduce regression, is supported by inspection of the time-series data. The largest change of any year is not the one after the crackdown, but is instead the upswing in the series occurring in 1954-55, just prior to the crackdown. In terms of crude fatality rates, 1955 is strikingly the highest point reached. It thus seems plausible...
Even more context

Figure 4. Traffic Fatalities for Connecticut, New York, New Jersey, Rhode Island, and Massachusetts (per 100,000 persons)
Discrete Variables (a.k.a. Categorical Variables)

Roughly speaking, **discrete** variables take only a few unique values (we might turn them into **factors** in R).

- Win, Lose, Tie
- Treatment vs Control
- Can be numeric, e.g. cars grouped by 4, 6, or 8 cylinders.

```r
by(mtcars$mpg, mtcars$cyl, mean)
```

```
mtcars$cyl: 4
[1] 26.66364

mtcars$cyl: 6
[1] 19.74286

mtcars$cyl: 8
[1] 15.1
```
Bar chart

Bar chart: height of the bar is proportional to some **summary statistic** associated with a discrete variable.

```r
barplot(by(mtcars$mpg, mtcars$cyl, mean),
      main = "Fuel Efficiency of 32 Different Cars",
      xlab = "Cylinders", ylab="MPG",
      sub = "data: mtcars", col="blue")
```

**Fuel Efficiency of 32 Different Cars**

- **MPG**
  - 0
  - 15

- **Cylinders**
  - 4
  - 6
  - 8
In theory, *continuous* variables may take infinitely many values (though in practice resolution is limited by the measuring apparatus’ precision). For example, *mpg* is a continuous variable: a car’s average fuel economy could be any number between 0 and $\infty$.

Histograms and density plots are often used to display continuous univariate data.
Histograms display quantitative data like a bar graph but they allow for unequal block lengths.

```r
hist(mtcars$mpg, main = "Fuel Efficiency of 32 Different Cars", xlab="MPG", col="blue",
     breaks = c(10, 15, 19, 20, 21, 25, 35))
```

### Fuel Efficiency of 32 Different Cars

![Histogram of Fuel Efficiency](image-url)
Area is proportional to frequency so the percentage falling into a block can be discerned without a vertical scale (since the total area equals 100%). But it’s helpful to have a vertical scale (density scale). Its unit is ‘% per unit’, so ‘% per mpg’ in above example.

Histograms show two kinds of information...
Histograms show density

Density (crowding): The bar height tells how many cars are in one unit on the horizontal scale. The highest density ($0.094 = 9.4\%$) is at 20mpg. Even though only 3 cars have 20mpg, adjusting for width, more cars fall in that category than any other. By contrast, the density is only 0.019 for the 6 cars with $mpg > 25$. 
Histograms show percentages

**Percentages** (relative frequencies) are given by:

\[
Area = Height \times Width
\]

For example, 18.75% fall into the most fuel efficient category (between 25 and 35 mpg) because the corresponding area is

\[
Area = (10 \text{ mpg}) \times (0.01875 \text{ % per mpg})
\]

Alternatively, eyeballing shows this area makes up roughly 1/5 of the total area.
Density plots

A density plot is like a “smoothed” histogram.

counts.at.mpg = seq(from=1, to=1, length=length(mtcars$mpg))

plot(mtcars$mpg, counts.at.mpg, main = "Fuel Mileage", xlab="Miles per gallon", ylab="Fuel Mileage")
Density plots

A density plot is like a “smoothed” histogram.

```
plot(density(mtcars$mpg), main = "Fuel Mileage")
```

Fuel Mileage

N = 32   Bandwidth = 2.477
Choice of Bandwidth: How Many Peaks?

**Small bandwidths** capture many local peaks but may be unstable (‘wiggly’) elsewhere. **Big bandwidths** are ‘smoother’ (fewer peaks).
Scatter plot

For **bivariate** data where *both* variables are *continuous*, scatterplots are the standard way to display association.

```r
plot(GaltonFamilies$midparentHeight, 
     GaltonFamilies$childHeight, main="Galton's Famous Data", 
pch=20) # pch denotes point type
```

---

**Galton's Famous Data**

- **Parent Height (midpoint)**
- **Child Height**
A special case of the scatter plot is a ‘time series.’ By convention, time is always set to be the x variable.

Air Quality in New York, 1973

data: airquality

Ozone

May, Jun, Jul, Aug, Sep, Oct

0, 50, 100, 150
Box Plots

Box plots provide a compact summary of a variable—both its median as well as values that the variable typically takes.

```r
boxplot(mtcars$mpg)
```
Details about the box plot

- For each value of the *categorical* variable, it depicts the following information about the *continuous* variable:
  - The median
  - The first and third quartiles; these are the *hinge* values, which represent the extent of the box
  - The smallest observation *greater than* the first quartile *less* 1.5 times the interquartile range *or* the minimum, whichever is larger; this is the extent of the *lower whisker*
  - The largest observation *less than* the third quartile *plus* 1.5 times the interquartile range *or* the maximum, whichever is smaller; this is the extent of the *upper whisker*
  - Points that lie outside 1.5 times the interquartile range from the *hinges*, as *individual points* (sometimes denoted *outliers*)
The box plot is also convenient for displaying **bivariate** data, where one variable is *continuous* and the other is *categorical*. Box plots give a useful summary of the distribution of the *continuous* variable for each value of the *categorical* variable.

```r
boxplot(mtcars$mpg ~ mtcars$cyl)
```
Summary

- Visualization should be chosen based on variable number and type and the research question at hand.

<table>
<thead>
<tr>
<th></th>
<th>Discrete</th>
<th>Continuous</th>
<th>One of Each</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate</td>
<td>Table</td>
<td>Histogram or Density</td>
<td></td>
</tr>
<tr>
<td>Bivariate</td>
<td>Contingency Table</td>
<td>Scatter</td>
<td>Boxplot or two-color Barplot</td>
</tr>
</tbody>
</table>

- Compared with barplots, boxplots convey more information about sampling variability (but typically require more text to explain to the reader).
- Density plots and histograms contain rich information about a variable’s distribution but may require some calibration for the appropriate amount of “wigliness”.