Displaying and Summarizing Univariate Data

Graphical displays of data are important as a way of gaining insight and communicating results. We will look first at displaying data on just one variable at a time.

We aim to get a good picture of the *distribution* of the variable — how often the variable takes on its various possible values.

We can also try to summarize the distribution with a few numbers. Of particular interest are measures of *location* and *spread*, such as the *mean* (average) and the *standard deviation*.

Tables for Univariate Categorical Data

Consider the data we have on the 2201 people on the *Titanic*.

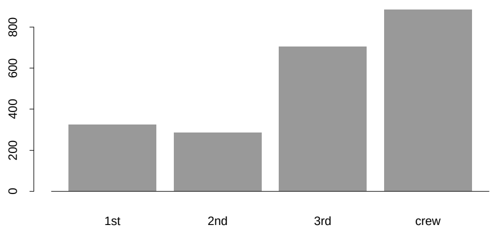
We can use a table to show how many people are in each class (the *frequencies*), and the proportions in each class (*relative frequencies*):

class	frequency	relative frequency
1st	325	0.148
2nd	285	0.129
3rd	706	0.321
crew	885	0.402
Total	2201	1.000

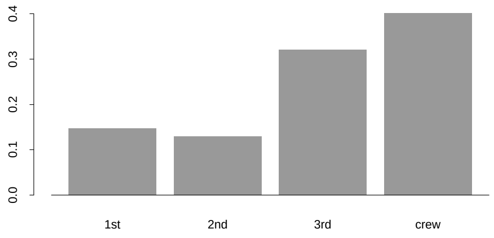
Note: The relative frequencies won't always add up 1.000. How can this be?

Bar Graphs

The frequencies can be plotted in a bar graph:



As can the relative frequencies:



Stem-and-Leaf Plots

Moderate amounts of numerical data can be displayed in a *stem-and-leaf plot*.

Consider the 21 "before" measurements for the calcium and blood pressure data:

```
107 110 123 129 112 111 107 112 136 102
123 109 112 102  98 114 119 112 110 117 130
```

We split each number into a "leaf" — the last digit — and a "stem" — the earlier digits. We then plot the leaves against the stems, as on the left below:

```
 9 : 8               9 : 8
10 : 77292          10 : 22779
11 : 0212249207     11 : 0012222479
12 : 393            12 : 339
13 : 60             13 : 06
```

On the right, the leaves have been ordered within each stem.

Splitting the Stems

Sometimes a stem plot is more informative if we split each stem in two — eg, we might plot 10, 11, 12, 13, 14 on one stem, and plot 15, 16, 17, 18, 19 on another.

Here's the blood pressure data with and without splitting:

```
 9 : 8                  9 : 8
10 : 22779             10 : 22
11 : 0012222479        10 : 779
12 : 339               11 : 00122224
13 : 06                11 : 79
                       12 : 33
                       12 : 9
                       13 : 0
                       13 : 6
```

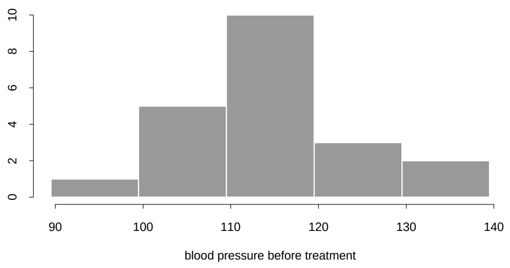
Which plot seems more informative?

Frequency Histograms

A *histogram* is like a stem plot, except leaves aren't distinguished, and the stems needn't correspond to digits. Also, it's generally plotted horizontally.

We divide the data range into equal intervals, and plot bars with heights proportional to the number of data points in each interval.

Here's a histogram of the blood pressure data:

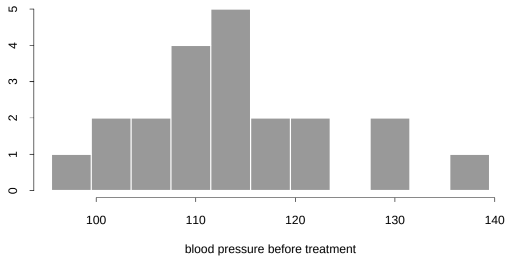


The intervals were (89.5,99.5), (99.5,109.5), etc.

The Effect of Using Different Intervals

With the intervals used on the last slide, the histogram looks much like the stem plot without splitting.

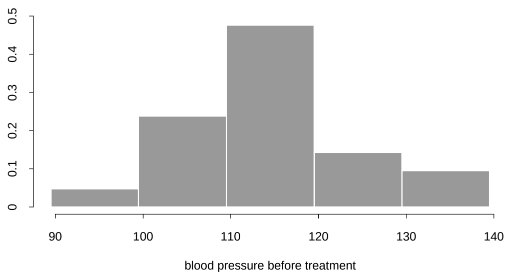
The histogram looks different if we choose intervals of (95.5,99.5), (99.5,103.5), (103.5,107.5), and so forth:

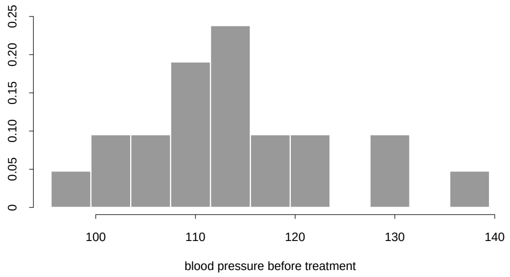


Do you believe the gaps above are at values for blood pressure that are particularly uncommon in the general population?

Relative Frequency Histograms

We can label the bars with relative frequency instead of frequency:





Probability Density Histograms

We'd rather not have the histogram look different just due to a different interval width. Probability density histograms fix that, using a vertical scale chosen so the total area is one: