Statistics 120
Histograms and Variations

Graphics for a Single Set of Numbers

- The techniques of this lecture apply in the following situation:
  - We will assume that we have a single collection of numerical values.
  - The values in the collection are all observations or measurements of a common type.
- It is very common in statistics to have a set of values like this.
- Such a situation often results from taking numerical measurements on items obtained by random sampling from a larger population.

Example: Yearly Precipitation in New York City

The following table shows the number of inches of (melted) precipitation, yearly, in New York City, (1869-1957).

43.6	37.8	49.2	40.3	45.5	44.2	38.6	40.6	38.7	46.0
37.1	34.7	35.0	43.0	34.4	49.7	33.5	38.3	41.7	51.0
54.4	43.7	37.6	34.1	46.6	39.3	33.7	40.1	42.4	46.2
36.8	39.4	47.0	50.3	55.5	39.5	35.5	39.4	43.8	39.4
39.9	32.7	46.5	44.2	56.1	38.5	43.1	36.7	39.6	36.9
50.8	53.2	37.8	44.7	40.6	41.7	41.4	47.8	56.1	45.6
40.4	39.0	36.1	43.9	53.5	49.8	33.8	49.8	53.0	48.5
38.6	45.1	39.0	48.5	36.7	45.0	45.0	38.4	40.8	46.9
36.2	36.9	44.4	41.5	45.2	35.6	39.9	36.2	36.5	

The annual rainfall in Auckland is 47.17 inches, so this is quite comparable.

Data Input

As always, the first step in examining a data set is to enter the values into the computer. The R functions scan or read.table can be used, or the values can be entered directly.

```
> rain.nyc =
    c(43.6, 37.8, 49.2, 40.3, 45.5, 44.2, 38.6, 40.6, 38.7,
      46.0, 37.1, 34.7, 35.0, 43.0, 34.4, 49.7, 33.5, 38.3,
      41.7, 51.0, 54.4, 43.7, 37.6, 34.1, 46.6, 39.3, 33.7,
      40.1, 42.4, 46.2, 36.8, 39.4, 47.0, 50.3, 55.5, 39.5,
      35.5, 39.4, 43.8, 39.4, 39.9, 32.7, 46.5, 44.2, 56.1,
      38.5, 43.1, 36.7, 39.6, 36.9, 50.8, 53.2, 37.8, 44.7,
      40.6, 41.7, 41.4, 47.8, 56.1, 45.6, 40.4, 39.0, 36.1,
      43.9, 53.5, 49.8, 33.8, 49.8, 53.0, 48.5, 38.6, 45.1,
      39.0, 48.5, 36.7, 45.0, 45.0, 38.4, 40.8, 46.9, 36.2,
      36.9, 44.4, 41.5, 45.2, 35.6, 39.9, 36.2, 36.5)
```

Plots for a Collection of Numbers

- Often we have no idea what features a set of numbers may exhibit.
- Because of this it is useful to begin examining the values with very general purpose tools.
- In this lecture we'll examine such general purpose tools.
- If the number of values to be examined is not too large, stem and leaf plots can be useful.

Stem-and-Leaf Plots

```
> stem(rain.nyc)
  The decimal point is at the |

  32 | 7578
  34 | 147056
  36 | 1225778991688
  38 | 3456670034445699
  40 | 1346684577
  42 | 4016789
  44 | 2247001256
  46 | 0256908
  48 | 552788
  50 | 380
  52 | 025
  54 | 45
  56 | 11
```

Stem-and-Leaf Plots

```
> stem(rain.nyc, scale = 0.5)
  The decimal point is 1 digit(s) to the right of the |

  3 | 344444
  3 | 55666667777777888889999999999
  4 | 000000011112222334444444
  4 | 55555666677778999
  5 | 000011334
  5 | 666
```

The argument `scale=.5` is use above above to compress the scale of the plot. Values of `scale` greater than 1 can be used to stretch the scale.

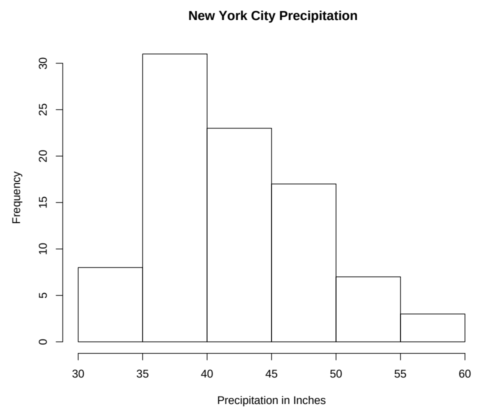
(It only makes sense to use values of `scale` which are 1, 2 or 5 times a power of 10.

Stem-and-Leaf Plots

- Stem and leaf plots are very "busy" plots, but they show a number of data features.
  - The location of the bulk of the data values.
  - Whether there are outliers present.
  - The presence of clusters in the data.
  - Skewness of the distribution of the data .
- It is possible to retain many of these good features in a less "busy" kind of plot.

Histograms

- Histograms provide a way of viewing the general distribution of a set of values.
- A histogram is constructed as follows:
  - The range of the data is partitioned into a number of non-overlapping “cells”.
  - The number of data values falling into each cell is counted.
  - The observations falling into a cell are represented as a “bar” drawn over the cell.



Types of Histogram

Frequency Histograms

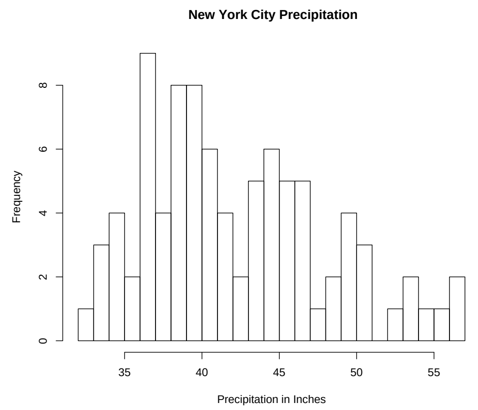
The height of the bars in the histogram gives the number of observations which fall in the cell.

Relative Frequency Histograms

The area of the bars gives the proportion of observations which fall in the cell.

Warning

Drawing frequency histograms when the cells have different widths misrepresents the data.



Histograms in R

- The R function which draws histograms is called `hist`.
- The `hist` function can draw either frequency or relative frequency histograms and gives full control over cell choice.
- The simplest use of `hist` produces a frequency histogram with a default choice of cells.
- The function chooses approximately $\log_2 n$ cells which cover the range of the data and whose end-points fall at “nice” values.

Example: Histogram Options

Optional arguments can be used to customise histograms.

```
> hist(rain.nyc, breaks = seq(30, 60, by=3),
       prob = TRUE, las = 1, col = "lightgray",
       main = "New York City Precipitation",
       xlab = "Precipitation in Inches")
```

The following options are used here.

1. `prob=TRUE` makes this a *relative frequency* histogram.
2. `col="gray"` colours the bars gray.
3. `las=1` rotates the *y* axis tick labels.

Example: Simple Histograms

Here are several examples of drawing histograms with R.

(1) The simplest possible call.

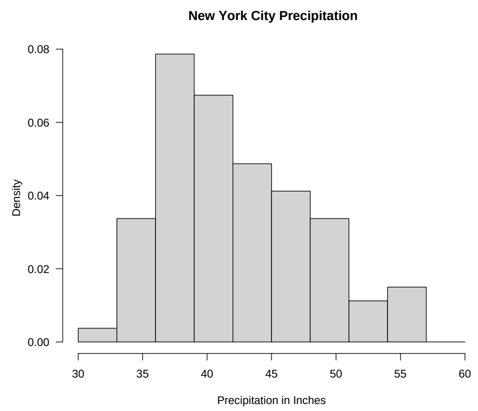
```
> hist(rain.nyc,
       main = "New York City Precipitation",
       xlab = "Precipitation in Inches" )
```

(2) An explicit setting of the cell breakpoints.

```
> hist(rain.nyc, breaks = seq(30, 60, by=2),
       main = "New York City Precipitation",
       xlab = "Precipitation in Inches")
```

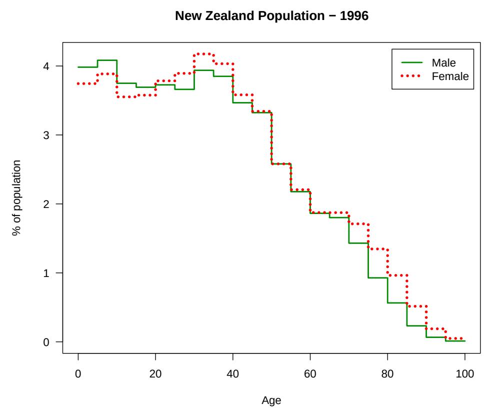
(3) A request for approximately 20 bars.

```
> hist(rain.nyc, breaks = 20,
       main = "New York City Precipitation",
       xlab = "Precipitation in Inches" )
```



Histograms and Perception

1. Information in histograms is conveyed by the heights of the bar tops.
2. Because the bars all have a common base, the encoding is based on "position on a common scale."

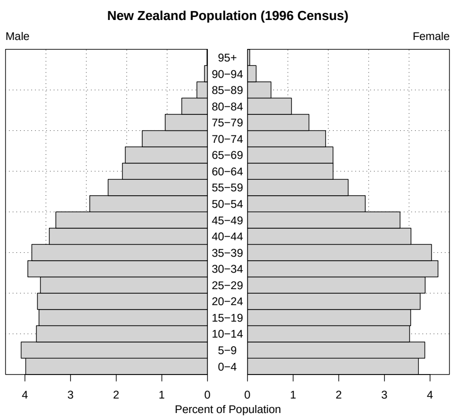


Comparison Using Histograms

- Sometimes it is useful to compare the distribution of the values in two or more sets of observations.
- There are a number of ways in which it is possible to make such a comparison.
- One common method is to use "back to back" histograms.
- This is often used to examine the structure of populations broken down by age and gender.
- These are referred to as "population pyramids."

Superposition and Perception

- Superimposing one histogram on another works quite well.
- The separate histograms provide a good way of examining the distribution of values in each sample.
- Comparison of two (or more) distributions is easy.

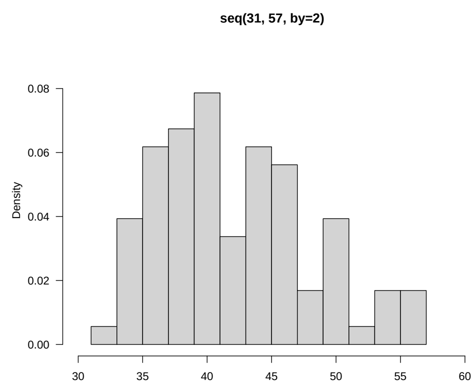


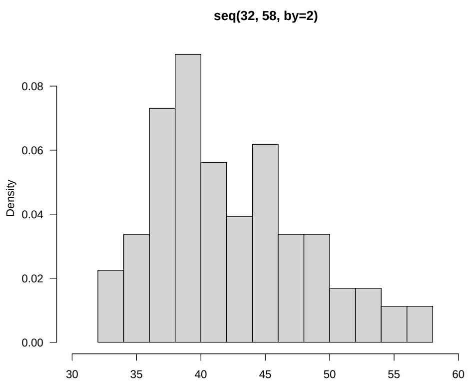
The Effect of Cell Choice

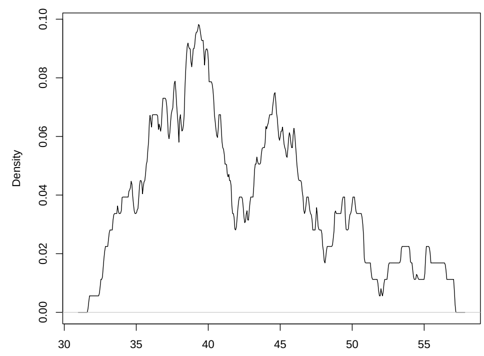
- Histograms are very sensitive to the choice of cell boundaries.
- We can illustrate this by drawing a histogram for the NYC precipitation with two different choices of cells.
  - `seq(31, 57, by=2)`
  - `seq(32, 58, by=2)`
- These different choices of cell boundaries produce quite different looking histograms.

Back to Back Histograms and Perception

- Comparisons within either the "male" or "female" sides of this graph are made on a "common scale."
- Comparisons between the male and female sides of the graph must be made using length, which does not work as well as position on a common scale.
- A better way of making this comparison is to superimpose the two histograms.
- Since it is only the bar tops which are important, they are the only thing which needs to be drawn.







The Inherent Instability of Histograms

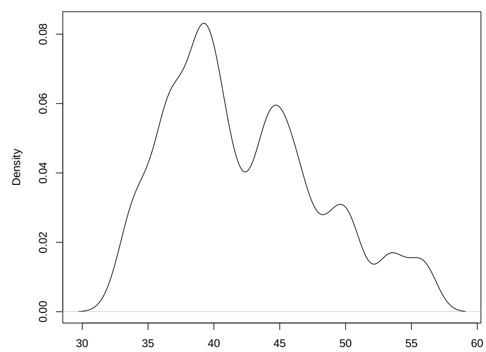
- The shape of a histogram depends on the particular set of histogram cells chosen to draw it.
- This suggests that there is a fundamental instability at the heart of its construction.
- To illustrate this we'll look at a slightly different way of drawing histograms.
- For an ordinary histogram, the height of each histogram bar provides a measure of the density of data values within the bar.
- This notion of data density is very useful and worth generalising.

Stability

- The basic idea of computing and drawing the density of the data points is a good one.
- It seems, however, that using a sliding histogram cell is not a good way of producing a density estimate.
- In the next lecture we'll look at a way of producing a more stable density estimate.
- This will be our preferred way to look at a the distribution of a set of data.

Single Bar Histograms

- We can use a single histogram cell, centred at a point x and having width w to estimate the density of data values near x.
- By moving the cell across the range of the data values we will get an estimate of the density of the data points throughout the range of the data.



Single Bar Histograms

- The area of the bar gives the proportion of data values which fall in the cell.
- The height, $h(x)$, of the bar provides a measure of the density of points near x.