Descriptive Statistics

Descriptive and Inferential Statistics

Recall that statistical methods are broadly divided into:

- **Descriptive** methods, which focus on the characteristics of a particular set of data;
- **Inferential** methods, which place the particular data set in a broader context, and allow us to relate what we see in the data to that broader context.

Descriptive statistical methods can also be broadly divided into:

- **Graphical displays**, such as bar charts;
- **Numerical summaries**, such as the mean and standard deviation.
Graphical Displays

Stem-and-Leaf Display

Introduced by John Tukey:

`FundRsgn <- scan("Data/Example-01-01.txt")
stem(FundRsgn)`

The numbers on the left are the “stems”, and the digits on the right are the “leaves”.

So for instance the leaf “8” on the row with stem “4” stands for a data value of 48.

The original data are given to one decimal place, so the display contains almost the same information.
Histogram

The *histogram* is another graphical display that summarizes a set of data:

```
hist(FundRsng)
```

The *heights* of the blocks in this histogram shows the “frequency” of various ranges of the data:

- 36 values between 0 and 10, inclusive;
- 18 values between 10 and 20 (strictly, $10 < x \leq 20$);
- and so on.
Optionally, the histogram can display the “density” of the data:

```
hist(FundRspng, freq = FALSE)
```

In this version, the *areas* of the blocks (width $\times$ height) are the *fractions* of the data that lie in the same ranges.

The total area is therefore 1.

The *heights* of the blocks are

$$\text{height} = \frac{\text{fraction}}{\text{width}}$$

which is called the “density” of the observations.
Numerical Summaries

Measures of Location

The average price of a new house sold in April, 2013, was $337,000 (http://www.fedprimerate.com/new%5Fhome%5Fsales%5Fprice%5Fhistory.htm)

The median price in the same month was $279,300.

When we are looking at a large number of values, we may want to know only what is a “typical” value.

The mean and the median are two candidates for measuring the “typical” value.
Mean

Given $n$ values $x_1, x_2, \ldots, x_n$, the *sample mean*, or arithmetic average, is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
Median

The median is defined loosely as the value \( \tilde{x} \) such that half the \( x \) values lie below \( \tilde{x} \) and half lie above.

The simplest way, but not the only way, to find the median is to order the data as

\[
x(1) \leq x(2) \leq \cdots \leq x(n).
\]

If \( n \) is odd, \( n = 2m + 1 \), then the median \( \tilde{x} \) is the unique middle value in the ordered list:

\[
\tilde{x} = x(m+1).
\]

If \( n \) is even, \( n = 2m \), then \( \tilde{x} \) is the average of the two middle values:

\[
\tilde{x} = \frac{x(m) + x(m+1)}{2}.
\]
Quantile

More generally, the $p^{th}$ quantile divides the data into a fraction $p$ that lie below the quantile, and a complementary fraction $1 - p$ that lie above it.

A quantile is usually calculated from the ordered values

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$  

but various specific rules have been proposed.

The quartiles correspond to $p = .25, .75,$ and the deciles correspond to $p = .1, .2, \ldots, .9.$
Why Do We Need Both Mean and Median?

Suppose 5 houses sold for $260,000, $270,000, $280,000, $290,000 and $1,000,000.

```
prices <- c(260, 270, 280, 290, 1000)
mean(prices) # 420
median(prices) # 280
```

The median is typical of the moderate priced houses, but the mean is not representative of either the moderate prices (all below $300,000) or the high priced house.
But the mean is less affected by small variations:

```r
prices <- c(260, 270, 285, 290, 1000)
mean(prices) # 421
median(prices) # 285
```

The mean and median are two *measures of location*.

Many others have been devised to meet various objectives.

The *trimmed mean* (drop the highest and lowest values, and average the rest) is used in some athletic events, and in finance.
Measures of Variability

The prices

260, 270, 280, 290, 300

and

278, 279, 280, 281, 282

both have mean $280,000, but are very different in variability.

Once we know the typical value (a measure of location), the next most interesting aspect of a set of data is usually how much they vary around that typical value (a measure of variability).
Standard Deviation

It is natural to measure variability by looking at how much the individual data values $x_i$ differ from a location measure such as the mean $\bar{x}$.

The sample variance is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

and the sample standard deviation is $s = \sqrt{s^2}$.
Examples

prices <- c(260, 270, 280, 290, 1000)
sd(prices)  # 324.4
prices <- c(260, 270, 280, 290, 300)
sd(prices)  # 15.8
prices <- c(278, 279, 280, 281, 282)
sd(prices)  # 1.58

As with measures of location, many other measures of variability have
been devised to meet various objectives.

For example, the *interquartile range* (IQR) is the difference between
the upper and lower quartiles.
The Boxplot

The **boxplot** (or box-and-whisker plot) is a graphical summary that displays both a measure of location (the median) and a measure of variability (the interquartile range):

```r
boxplot(FundRsgn)
```

The central box extends from the lower quartile to the upper quartile, and the median is shown by the bold line within the box.

A data value that differs from the nearer quartile by more than $1.5 \times \text{IQR}$ is shown individually as a possible “outlier”.