Chapter 2
Location, Variability and Process

Section 2.2
Numerical Summaries Based on Position

David W. Gerbing

The School of Business
Portland State University

- Numerical Summaries Based on Position
  - Location in terms of Position
  - Variability in terms of Position
  - Distribution Shape

2.2a
Location in terms of Position

Outliers

Always evaluate data for outliers

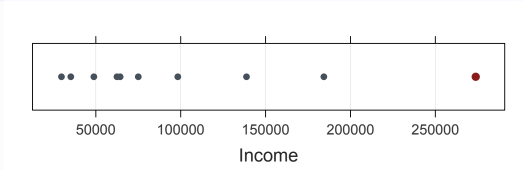
- **Outlier**: Value considerably different from most remaining values of the distribution
- Ex: Consider the following distribution of 10 Incomes at

  `http://lessRstats.com/data/outlier.csv`

  $29,750 $35,250 $48,840 $62,440 $64,320 $75,000 $98,280 $138,750 $184,330 $273,800

- lessR `ScatterPlot()`, or `sp()`, identifies the outlier in red

```
> sp(Income)
```



Impact of Outliers

Some statistics considerably impacted

- Outliers can have a sizable impact on the value of a statistic, such as drawing the mean closer to the outlier

  $m = (\$29,750 + \ldots + \$273,800) \,/\, 10 = \$101,076$

- The sensitivity of the mean to an outlier is more clearly demonstrated by a more extreme example, so replace the largest value, $273,800, with $2,738,000

  $m = (\$29,750 + \ldots + \$2,738,000) \,/\, 10 = \$347,496$

- Dramatically increasing just one value dramatically increased the mean of these 10 data values from $101,076 to $347,496
- The resulting mean is not representative of any of the data values, not the smaller data values or the one large data value, the outlier

Interpretation of Outliers

Data is only meaningful when generated by the same process

- Not always, but often, the process that generates an outlier is different from that which generated the remaining values
- Extreme example of different processes: The mean of five SAT scores and five annual GNP values can be correctly calculated, but this mean has no meaningful interpretation
- The mean is not meaningful because there is no single concept or entity that all the data values have in common
- **Key Concept**: A summary statistic should summarize data sampled from a single population, that is, data generated by a single process
- Identify and then analyze outliers from a different population as a separate group, and then generalize the results to the population of interest

Order Statistics

Divide a distribution into different groups depending on order

- In the previous example of outliers, all 10 values were sorted from smallest to largest
- **Order statistic**: A statistic calculated from a distribution in which the values are ordered from the smallest to the largest
- The values of statistics such as the mean and standard deviation change dramatically in the presence of outliers
- **Key Concept**: Order statistics are more resistant to outliers than statistics such as the mean because the extreme values in a distribution are ignored
- Examples of order statistics include the trimmed mean and median for location and the range and interquartile range for variability

Trimmed Mean

One approach for dealing with outliers

- **20% trimmed mean**: Mean of remaining values after discarding smallest 10% and largest 10% of the values, rounded down to nearest integer
- To trim 10% of 10 values: round-down$(.10 * 10) = 1$
- Lopping off one value on each side, calculate the 20% trimmed mean as the mean of *remaining* eight values

$$m_{20\%} = \frac{\$35,250 + \$48,840 + \ldots + \$138,750 + \$184,330}{8} = \$88,401.25$$

- To calculate a trimmed mean, use the R function `mean`, with the option `trim` set to the amount to trim from *each* tail, and `na.rm` set to `FALSE` to ignore any missing data
- 20% trim: `> mean(Y, trim=.1, na.rm = FALSE)`

Median

The extreme trimmed mean

- **Median** of a distribution: Value that divides a distribution of sorted values into two sections with the same number of values in each
- For a distribution with an even number of values, the median is the average of the two values closest to the middle

	1	2	3	4	5
1st 5 values:	$29,750	$35,250	$48,840	$62,440	$64,320
	6	7	8	9	10
2nd 5 values:	$75,000	$98,280	$138,750	$184,330	$273,800

$$Y_{median} = \text{average of values with ranks of 5 and 6} = (\$64,320 + \$75,000)/2 = \$69,660$$

Quartiles and Percentiles

Divide a distribution into more than two equal parts

- Define order statistics that divide a distribution into any number of equal parts, not just two parts as with the median
- **Quartile**: Values that divide a distribution into 4 equal parts
  - 1^{st} or lower quartile: The smallest $\frac{1}{4}$ of the values are below it and the largest $\frac{3}{4}$ above it
  - 2^{nd} or middle quartile: The median
  - 3^{rd} or upper quartile: The smallest $\frac{3}{4}$ of the values are below it and the largest $\frac{1}{4}$ above it
- **Percentile**: The values that divide a distribution into 100 equal parts
  - Most often used to interpret test scores
  - A score at the 80^{th} percentile indicates that 80% of the values of the distribution are below the score

Quantiles

Divide the ordered distribution into any specified percentage

- Each median, quartile and percentile divides a distribution into two sections, such as the 3^{rd} quartile with $\frac{3}{4}$ of the sorted values below it
- More generally, define a value that splits a sorted distribution into two sections with *any* specified proportion of values in the lower part
- **Quantile**: The q^{th} quantile, Y_q, for a distribution of values of variable Y has proportion q of the values below it
  - Median: .5 quantile
  - 1^{st} Quartile: .25 quantile
  - 3^{rd} Quartile: .75 quantile
  - Minimum: 0.0 quantile
  - Maximum: 1.0 quantile

Quantiles from lessR

Specified quantiles

- **5-number summary** of a distribution: Minimum and maximum and the three quartiles (or their approximations)

- The `lessR` `SummaryStats` function provides the 5-number summary, plus other descriptive statistics of a distribution of values for a variable

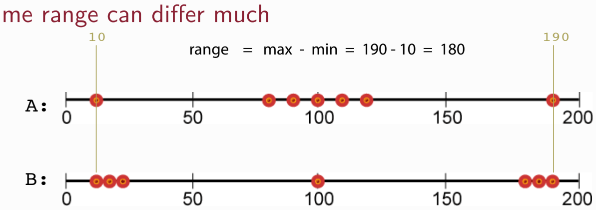
```
> d <-
   Read("http://lessRstats.com/data/outlier.csv")

> SummaryStats(Income)
  [excerpted]

    min      Qrt1       mdn      Qrt3       max
29750.0   52240.0   69660.0  128632.5  273800.0
```

2.2b
Variability in terms of Position

Range

The simplest indicator of variability

- **Range:** $+$ difference between minimum and maximum values
- The range is based on only the minimum and maximum, with no assessed impact of other values on variability
- The problem is that the variability of two distributions with the same range can differ much



- The range for both distributions is 180, but their standard deviations are considerably different, $s_A = 53.5$ and $s_B = 85.1$

Interquartile Range (IQR)

Compare the IQR with the Standard Deviation

- A useful index of the variability of a distribution, a type of range, is defined in terms of quartiles
- **Interquartile Range** or IQR: Positive difference between the first and third quartiles
- Calculate the IQR from the 1^{st} and 3^{rd} quartiles, and so is not influenced by values beyond those two boundaries, in particular, outliers
- The sample standard deviation, s, is calculated from all the squared mean deviated values of the distribution, and so is even more sensitive to outliers than is the mean

- For variable Y, `SummaryStats(Y)` returns many statistics, including the interquartile range of Y, labeled as IQR

BoxPlot: Plotting Quartiles, Detecting Outliers

The box plot was introduced by John Tukey in 1977

- Construct the box plot from only five values of the distribution, the 5-number summary, the minimum and maximum and the three quartiles (or their approximations)
- **Box plot**: The body of the box extends from approximately the 1st to the 3rd quartiles, with a line through the median and perpendicular lines extending out from the edges
- The boxplot is particularly useful to identify outliers
- **Potential Outlier**: Values between 1.5 IQR's and 3.0 IQR's from the edges of the box
- **Outlier**: Values more than 3.0 IQR's from either box's edge
- **Whisker**: A line from a box's edge that extends to the most extreme data value that is *not* a potential outlier, that is, within 1.5 IQR's of the edges

Constructing the Box: The Hinges

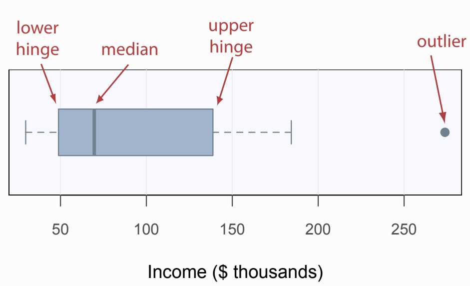
A little history

- **Hinges**: The box edges, approximately equal to or equal to the 1st and 3rd quartiles
- Why do the box edges in general only *approximate* the 1st and 3rd quartiles?
- Tukey developed the box plot in the early 1970's, before computer graphics, and so wanted easy, fast computations
- The procedures Tukey developed for calculating the hinges were apparently approximations of the true 1st and 3rd quartiles to simplify the computation
- Probably better if modern computer software would compute the boxplot with the true quartiles, though R uses hinges
- Conceptually, think of the box as spanning 50% of the data values, the IQR, that is, between the 1st and 3rd quartiles

BoxPlot

- Obtain the boxplot with the lessR BoxPlot, or bx, function

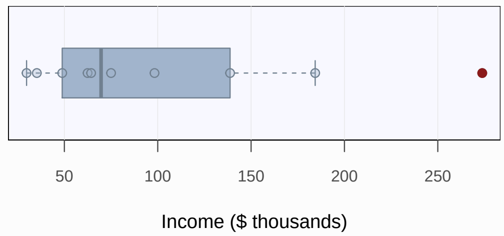
```
> Read("http://lessRstats.com/data/outlier.csv")
If you wish to divide all the data values by 1000:
> d <- Transform(Inc000=Income/1000)
> BoxPlot(Inc000, xlab="Income ($ thousands)")
```



BoxPlot with Data

- The function BoxPlot can also display the data values as a 1-dimensional scatter plot, also called a dot plot

```
> BoxPlot(Inc1000, xlab="Income ($ thousands)",
    add.points=TRUE)
```



Interpretation of the Box Plot

Graphics output

- The maximum value of the distribution, around $270,000, is an outlier according to the definition of being more than 3 IQR's from the box
- Most of the data values lie between approximately $30,000 and $180,000
- The values larger than the median are considerably more spread out than the smaller values, what is called skew, the topic presented next
- Obtain the exact numerical values of the corresponding cutoff values from the following text output

BoxPlot

Comprehensive listing of order statistics

```
Present: 10
Missing: 0
Total  : 10

Minimum      : 29.75
Lower Whisker: 29.75
Lower Hinge  : 48.84
Median       : 69.66
Upper Hinge  : 138.75
Upper Whisker: 184.33
Maximum      : 273.80

1st Quartile : 52.24
3rd Quartile : 128.63
IQR          : 76.39
```

2.2c
Distribution Shape

Basic Properties of a Distribution

Properties to understand regarding any set of data values

- **Key Concept**: There are two types of information of general importance regarding a distribution of data values
  - A description of its shape, such as with a histogram or a verbal description of the histogram
  - Summary statistics calculated from the data values, such as the sample mean, m and sample standard deviation, s
- Each distribution of interest should be described by its shape and at least its mean and standard deviation
- Include order statistics such as the median and the minimum and maximum values to further enhance the description of the distribution of data values

Mean, Median and Distribution Shape

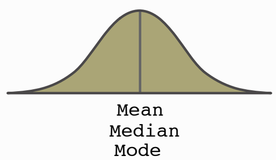
Different distributions assume different shapes

- Outliers, particularly outliers that are not too severe, can occur even when all values are from the same underlying distribution, depending on its shape
- Mathematically, the shape of a distribution could be almost anything, but there are just several different fundamental shapes of distributions of data found in the real world of applications
- These fundamental shapes are presented in terms of idealized data distributions of perfectly smooth curves instead of any one specific histogram of data that will more or less follow the idealized form
- These idealized distributions, such as the smooth bell-shaped normal curve, are discussed some more in the next chapter

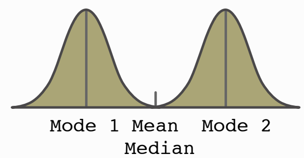
Symmetric Distributions

Symmetry and mode

- **Symmetric Distribution**: Shape of the distribution with one-side of the distribution a mirror image of the other side
- **Mode**: Most frequently occurring value or range of values
- Symmetric distributions can be uni-modal or multi-modal



Symmetric and Unimodal

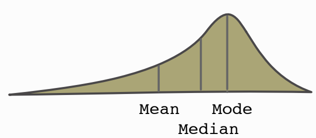


Symmetric and Bi-modal

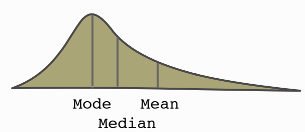
Skewed Distributions

Skewness and center

- **Skewed Distribution**: Shape of the distribution with the skewed side of the distribution having the longer tail
- Distributions can be skewed left or right, referring to the direction of the tail



Skewed Left



Skewed Right

- An example is Income, which is skewed right such that comparatively few people make so much money
- **Key Concept**: Order statistics such as the median are less sensitive to outliers than statistics such as the mean

Index Subtract 2 from each listed value to get the Slide

- The End