Soci708 – Statistics for Sociologists\textsuperscript{1}
Module 1 – Looking at Data: Distributions

François Nielsen

University of North Carolina
Chapel Hill

Fall 2009

\textsuperscript{1}Adapted in part from slides from courses by Robert Andersen (University of Toronto) and by John Fox (McMaster University)
Introduction
What is Statistics? The *Challenger* Disaster

- Statistics may be defined as *the science of learning from data* (IPS6e)
- The *Challenger* Accident was a tragic example of the consequences of poor statistical analysis.\(^2\)
  - On 28 January 1986 The U.S. pace shuttle *Challenger* exploded shortly after blastoff, killing the seven astronauts.
  - The cause of the explosion was the failure of rubber O-rings sealing two sections of one of the booster rockets attached to the shuttle.
  - This failure, in turn, was caused by the low temperature at the time of launch which made the O-rings lose their elasticity.

Introduction

What is Statistics? The *Challenger* Disaster

---

**Plots of O-ring damage and launch temperature—Space shuttle Challenger example.** Damage refers to total depth of erosion in O-ring seals.

(a) Plot for Flights with O-Ring Damage

(b) Plot for All Flights

---

Introduction

What is Statistics? The Challenger Disaster

On the day before launch, engineers at Morton Thiokol, the company that built the boosters, recommended that the launch be postponed because of the low forecast temperature for the following day. Officials at NASA and Thiokol examined data on O-ring damage that had occurred on previous launches.

- Engineers plotted a measure of damage to O-rings against temperature at launch time, including only launches with non-negligible damage.
- The plot showed no association of damage with temperature.
- Had they included all the cases they would have seen a clear association: lower temperature → greater damage and would have postponed the launch, avoiding the accident.
Introduction
Data Sets

▶ A data set is a collection of facts assembled for a particular purpose.
▶ We will mainly use rectangular data sets where information is organized in an individual (row) by variable (column) format
  ▶ an individual is a unit of observation – e.g. a person, an organization, a country
  ▶ a variable is a characteristic of the individual – e.g. a depression score, score on a scale of centralization of decision-making, Gross Domestic Product per capita
  ▶ a case is the information on all variables for one individual (corresponding to one row of the data set)
  ▶ an observation is the value of a single variable for a given individual
Introduction
Levels of Measurement

- A **categorical** (*nominal, qualitative*) variable is an exclusive & exhaustive set of attributes
  - E.g. sex or gender, religion, region of the U.S.
- An **ordinal** variable adds an ordering of the categories
  - E.g. Mohr scale of hardness – categories ordered from graphite to diamond by relation “A scratches B”; Likert scales with categories Strongly Agree to Strongly Disagree
- An **interval** variable adds a constant interval between categories
  - E.g. temperature in °F or °C degrees; IQ
- A **ratio** variable adds an absolute zero
  - E.g. temperature in °K degrees, income of individual, GDP of country, age, percentages
  - Thus one can say “$25,000 is half as much as $50,000”
Introduction
Levels of Measurement

- The level of measurement determines the kinds of analysis that can be carried out with a variable
- In practice one can simplify the four-fold typology into two categories:
  - **Qualitative variables:**
    - Includes categorical variables + ordinal variables treated as categorical – e.g. age in years recoded into YOUNG, ADULT, SENIOR categories
    - Analyzed using contingency tables (tabular analysis)
  - **Quantitative variables:**
    - Includes interval variables + ratio variables + ordinal variables treated as interval variables – e.g. “How well do you speak Spanish?” coded from 1 to 5
    - Analyzed using scatterplots & regression analysis

- There are advanced analytical techniques for ordinal data that are beyond the scope of this class
Introduction

Three Central Aspects to Statistics

There are three central aspects (tasks) of statistics:

1. Data Production
   - Designing research (e.g., a survey or an experiment) so that it produces data that help answer important questions
   - These issues will be topic of Module 3 – Producing Data

2. Data Analysis
   - Describing data with graphs and numerical summaries
   - Displaying patterns and trends
   - Measuring differences

3. Statistical Inference
   - Using information about a sample of individuals, drawn at random from a larger population, to establish conclusions about characteristics of the population
Displaying Qualitative Data
Counts and Percentages

- The distribution of a categorical variable can be displayed simply by reporting the counts (or percentages, which make seeing patterns easier) for each category in a table.
- Alternatively they can be presented in bar graphs or pie charts.
- Category percentages are calculated simply as

\[ \text{percent for category } j = 100 \times \frac{\text{count in category } j}{n} \]

where \( n \) is the total sample size (i.e., the number of cases in all categories).
Displaying Qualitative Data
Counts and Percentages

- The example below shows voting intentions for the 1988 Chilean Plebiscite from a survey conducted by FLACSO/Chile; data are from a dataset called Chile in the car package for R.

- Respondents were asked whether they intended to support Pinochet.

<table>
<thead>
<tr>
<th>Intended vote</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Abstain)</td>
<td>187</td>
<td>6.9</td>
</tr>
<tr>
<td>N (Vote ‘No’ against Pinochet)</td>
<td>889</td>
<td>32.9</td>
</tr>
<tr>
<td>U (Undecided)</td>
<td>588</td>
<td>21.8</td>
</tr>
<tr>
<td>Y (Vote ‘Yes’ for Pinochet)</td>
<td>868</td>
<td>32.1</td>
</tr>
<tr>
<td>Total</td>
<td>2700</td>
<td>100</td>
</tr>
</tbody>
</table>

Percent for ‘Yes’ vote = 100 × \( \frac{868}{2700} \) = 32.1
Displaying Qualitative Data
Simple Tabulation in Stata

. * use the File menu to find the Chile.dta file
. use "D:soci708\data\data_from_car_Stata\Chile.dta", clear
. tab vote

<table>
<thead>
<tr>
<th>vote</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>187</td>
<td>6.93</td>
<td>6.93</td>
</tr>
<tr>
<td>N</td>
<td>889</td>
<td>32.93</td>
<td>39.85</td>
</tr>
<tr>
<td>NA</td>
<td>168</td>
<td>6.22</td>
<td>46.07</td>
</tr>
<tr>
<td>U</td>
<td>588</td>
<td>21.78</td>
<td>67.85</td>
</tr>
<tr>
<td>Y</td>
<td>868</td>
<td>32.15</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>2,700</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
Displaying Qualitative Data
Simple Tabulation in R

> library(car) # Loads car package
> data(Chile) # Loads Chile dataset
> attach(Chile) # Attach data so they can be manipulated
> ?Chile # Gives information on the dataset (in window)
> table(vote) # Gives counts of categories of vote
 vote
   A  N  U  Y
187 889 588 868
> length(vote) # Tells number of cases
[1] 2700
> 100*(table(vote)/length(vote)) # Calculates percentages
 vote
   A       N       U       Y
6.925926 32.925926 21.777778 32.148148
Displaying Qualitative Data

Bar Chart

- The distribution of a categorical variable can also be represented as a bar graph or a pie chart.
- In R, a bar graph is created simply by the `plot` function
  ```r
  > plot(vote)
  ```
- Note that in a bar chart the bars do not touch each other.
In Stata, it is simple to create a pie chart with the `graph pie` function.

```
. graph pie, over(vote)
```
Histograms

- Quantitative data can take on many different values
- It is not helpful to display the distribution of the variable directly
- Few cases will have exactly the same value of the variable (i.e., tables of counts and bar graphs are far too cumbersome)
- This would involve listing nearly every case
- Instead we use histograms, stemplots & kernel density curves
A histogram is a bar graph that shows the count or percentage of cases falling in each of the bins

- *Horizontal axis*: The range of the variable
- *Vertical axis*: The count or percent of cases in each bin

Histograms are easily made:

1. Divide the range of the variable into intervals of equal width (called bins)
   - Each case must fit into only one bin
2. Count the number of individuals falling in each bin and draw a bar representing them
   - Unlike with a bar graph, the bars of a histogram touch each other – i.e., there are no spaces between them – reflecting the fact that we are displaying information about a quantitative variable
Histograms

Infant mortality example using the Leinhardt data

- Below are a few cases from the Leinhardt dataset in the car package for R (there are 105 cases in total)
- Because it takes on so many values and there are many cases, we must construct a histogram to view the distribution of infant mortality (infant)

<table>
<thead>
<tr>
<th></th>
<th>income</th>
<th>infant</th>
<th>region</th>
<th>oil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>3426</td>
<td>26.7</td>
<td>Asia</td>
<td>no</td>
</tr>
<tr>
<td>Austria</td>
<td>3350</td>
<td>23.7</td>
<td>Europe</td>
<td>no</td>
</tr>
<tr>
<td>Belgium</td>
<td>3346</td>
<td>17.0</td>
<td>Europe</td>
<td>no</td>
</tr>
<tr>
<td>Canada</td>
<td>4751</td>
<td>16.8</td>
<td>Americas</td>
<td>no</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper.Volta</td>
<td>82</td>
<td>180.0</td>
<td>Africa</td>
<td>no</td>
</tr>
<tr>
<td>Southern.Yemen</td>
<td>96</td>
<td>80.0</td>
<td>Asia</td>
<td>no</td>
</tr>
<tr>
<td>Yemen</td>
<td>77</td>
<td>50.0</td>
<td>Asia</td>
<td>no</td>
</tr>
<tr>
<td>Zaire</td>
<td>118</td>
<td>104.0</td>
<td>Africa</td>
<td>no</td>
</tr>
</tbody>
</table>
Histograms

Construction of histogram for infant mortality rate

- Start by dividing the data into bins:
  - In this case I divided the data into only 14 bins to simplify the example
- Then determine the number of cases in each bin and draw the histogram accordingly

<table>
<thead>
<tr>
<th>Mortality Rate</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–49.9</td>
<td>42</td>
<td>40.0</td>
</tr>
<tr>
<td>50–99.9</td>
<td>24</td>
<td>22.9</td>
</tr>
<tr>
<td>100–149.9</td>
<td>16</td>
<td>15.2</td>
</tr>
<tr>
<td>150–199.9</td>
<td>14</td>
<td>13.3</td>
</tr>
<tr>
<td>200–249.9</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>250–299.9</td>
<td>2</td>
<td>1.9</td>
</tr>
<tr>
<td>300–349.9</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>350–399.9</td>
<td>1</td>
<td>1.0</td>
</tr>
<tr>
<td>400–449.9</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>450–499.9</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>500–549.9</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>550–599.9</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>600–649.9</td>
<td>1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Total: 105 cases

\[ \frac{18}{102} \]
Histograms

Histogram in R

```r
> # In R
> data(Leinhardt)
> attach(Leinhardt)
> hist(infant, nclass=14, 
   main="Distribution of Infant Mortality Rates",
   xlab="Infant Mortality Rate",
   ylab="Count",
   col="red")
```

![Distribution of Infant Mortality Rates](image)
1. Look for the overall pattern of the data described by the shape, center, and spread of the distribution:
   ▶ In a *symmetric distribution*
     ▶ Most cases are near the center
     ▶ Half the cases will be on one side, the other side will be its mirror image
   ▶ In a *skewed distribution* one tail of the distribution is longer than the other:
     ▶ Positive skew: the tail to the right is longer
     ▶ Negative skew: the tail to the left is longer

2. Look for departures from the overall pattern, such as *outliers*:
   ▶ individual values that fall outside the general pattern of the data
A stemplot (a.k.a. stem-and-leaf display) is a special histogram that uses the numerical data themselves to sort the data into bins, and to form the bars of the histogram.

- It works best for small numbers of observations all greater than zero.

To construct a stemplot:

1. Divide each data value into a stem and a leaf
   - Sometimes this can be done by simply using the last digit of the data value as the leaf; for example, for the infant-mortality data:

<table>
<thead>
<tr>
<th>Country</th>
<th>Infant Mortality Rate</th>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden</td>
<td>9.6 →</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Mexico</td>
<td>60.9 →</td>
<td>60</td>
<td>9</td>
</tr>
<tr>
<td>Pakistan</td>
<td>124.3 →</td>
<td>124</td>
<td>3</td>
</tr>
</tbody>
</table>
Stemplots
Constructing a stemplot

- Alternatively, we could round the data, dividing stems and leaves between *tens and units* digits:

<table>
<thead>
<tr>
<th>Country</th>
<th>Infant Mortality Rate</th>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden</td>
<td>9.6 (\approx) 10 (\rightarrow)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mexico</td>
<td>60.9 (\approx) 61 (\rightarrow)</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Pakistan</td>
<td>124.3 (\approx) 124 (\rightarrow)</td>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

- Or, rounding the data, dividing stems and leaves between *hundreds and tens* digits:

<table>
<thead>
<tr>
<th>Country</th>
<th>Infant Mortality Rate</th>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sweden</td>
<td>9.6 (\approx) 10 (\rightarrow)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mexico</td>
<td>60.9 (\approx) 61 (\rightarrow)</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Pakistan</td>
<td>124.3 (\approx) 124 (\rightarrow)</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Stemplots
Constructing a stemplot

2. Write the stem in a vertical column with the smallest at the top and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

   ▶ Note: In a wordprocessor make sure to use tabular numbers or a monospaced font such as Courier so that each digit uses the same amount of space.
Stemplots – In R

> stem(infant)

The decimal point is 2 digit(s) to the right of the |

0 | 1111111112222222222223333333344444
0 | 55555566666666666667778888889
1 | 000011222333344
1 | 555566666778899
2 | 002
2 | 6
3 | 0
3 |
4 | 0
4 |
5 |
5 |
6 |
6 | 5
Stemplots – In Stata

use "D:soci708\data\data_from_car_Stata\Leinhardt.dta", clear
stem infant Stem-and-leaf plot for infant -- infant rounded to integers

0** | 10,10,11,12,12,13,13,14,14,15,16,17,17,18,18,18,19
0** | 20,20,21,22,22,24,26,26,27,27,28,28,32,34,39,39
0** | 40,43,45,46,49,50,52,54,55,58,58,58
0** | 60,60,61,61,63,64,65,68,72,76,78,79,79
0** | 80,84,86
1** | 00,00,02,04,10,14
1** | 20,24,24,25,27,29,33,37,38
1** | 48,49,50,53,58,59
1** | 60,60,63,70,70
1** | 80,80,87,90
2** | 00,00,16
2** |
2** | 59
2** | 
2** | 
3** | 00
3** | 
3** | 
3** | 
3** | 
4** | 00
4** | 
4** | 
4** | 
4** | 
5** | 
5** | 
5** | 
5** | 
5** | 
5** | 
6** | 
6** | 
6** | 
6** | 
6** | 
6** | 
6** |
Time Plots
Time series data & time plots

▶ **Time Series Data**
- Data collected for a single unit of observation over successive periods of time
- Usually data are evenly spaced through time (not necessarily, however) and there is only one observation per time point

▶ **Time Plots**
- Line graph of the variable (on $y$-axis) against time (on $x$-axis)
- The points on the graph are joined by a line
- Look for trends, and deviations from the general pattern
Time Plots

- Example from the Hartnagel data in the car package for R
- The data represent male theft conviction rates (per 100,000) in Canada from 1931 to 1968
Time Plots
R-Script for previous slide

```r
> data(Hartnagel)
> attach(Hartnagel)
> # next 2 lines optional, to make plot square
> op<-par(no.readonly = TRUE)
> par(pty="s")
> plot(year, mtheft, type='o',
+ main="Male conviction rate (per 100,000), 1931-68",
+ ylab="Conviction Rate",
+ xlab="Year",
+ cex=1.5, col="red")
> # type='o' joins points with a line
> # cex=1.5 makes the points 1.5 times as large
```
Time Plots
Divorce Rate (per 1000 Married Females) in the U.S.
Time Plots

Another Example – Average Age of Menarche in 5 Western Countries

The decline in the average age of menarche in five Western industrial nations

Time Plots
Early (1785) Time Plot by William Mayfair (Tufte 1983)
Time Plots
1786 Time Plot by William Mayfair Showing Improved Design (Tufte 1983)
Describing Distributions with Numbers

- Measures of Center
  - Mode
  - Mean
  - Median

- Measures of Spread
  - Range
  - Quartiles
  - Five Number Summaries & Boxplots
  - Variance and Standard Deviation
Measuring Center: The Mean

Problems with the mode

- The *mode* is the value interval or category with the most observations.
  - It is most useful to describe qualitative data.
  - When examining histograms the mode is sensitive to bin definition and to random data fluctuations.

A bimodal distribution  
A unimodal distribution
Measuring Center: The Mean

The Mean $\bar{x}$

- The *mean* is denoted by the symbol $\bar{x}$ (pronounced “xbar”)
- It is the most common measure of center, but not always the best.
- The mean is calculated as the arithmetic average:
  - Add all the scores, and divide by the total number of cases
- If the distribution of the variable is skewed, the mean is a poor measure of the center of the distribution
  - The mean is pulled in the direction of the skew
Measuring Center: The Mean
Calculating the Mean

▶ The formula for the mean is:

\[ \bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \]

\[ = \frac{1}{n} \sum x_i \]

▶ If we have five observations (3, 3, 4, 2, 10) the mean is calculated as follows:

\[ \bar{x} = \frac{3 + 3 + 4 + 2 + 10}{5} \]

\[ = \frac{22}{5} = 4.4 \]

> #In R:
> x<-c(3,3,4,2,10)
> mean(x)
> [1] 4.4
Measuring Center: The Mean
Effects of skewness and outlying observations

- In the previous example, the observation with a score of 10 was an outlier.
  - It pulled the distribution so that it has a positive skew.
  - It also pulls the mean toward itself.
- If we omit this observation, the mean becomes much smaller:
  \[
  \bar{x} = \frac{3 + 3 + 4 + 2}{4} = \frac{12}{4} = 3
  \]
- There is a large difference between the two means, indicating that:
  - The observation is *influential*, i.e., including it or not makes a big difference on the mean.
  - The mean is *not a robust measure*.
  - In the presence of influential cases the median is a better measure of center.
Simply speaking, the median is the midpoint of the distribution, i.e. the number such that half the observations are smaller and half are larger.

To find the median of $n$ observations:

1. Order all observations from lowest to highest.
2. Find the median location as $(n + 1)/2$.
3. The median is the value of the observations(s) at position $(n + 1)/2$ from the beginning of the list.

- Odd number of observations: the median is the middle observation.
- Even number of observations: the median is the mean of the two middle observations.
Measuring Center: The Median

Simple Examples

- **Start with the odd $n$ case:**
  - Begin by ordering the observations from lowest to highest: 2.1, 2.3, 3.7, 4.1, 10
  - Find the median location: $(n + 1)/2 = (5 + 1)/2 = 3$
  - Therefore, $M = 3.7$ (the observation in position 3)

- **Assume now that $n$ is even:**
  - The ordered observations are 2.1, 2.3, 3.7, 4.1, 10, 11
  - The median location is $(n + 1)/2 = (6 + 1)/2 = 3.5$
  - Now $M = 3.9$ (the mean of 3.7 and 4.1):

```
> # In R
> x<-c(2.1, 2.3, 3.7, 4.1, 10, 11)
> median(x)
[1] 3.9
```
Assume again we have 5 observations: 2.1, 2.3, 3.7, 4.1, 10

The mean is:

\[ \bar{x} = \frac{2.1 + 2.3 + 3.7 + 4.1 + 10}{5} = 4.44 \]

As the median location \((n + 1)/2 = (5 + 1)/2 = 3\), the median is the value of the third observation of the ordered data, so that \(M = 3.7\)

It is clear here that the mean (4.44) is more influenced by the outlying value (10) than the median (3.7)

With skewed distributions or in presence of outliers, the median is a better measure of center than the mean.
Mean Versus Median

Davis data example

- This example uses the Davis data in the car package for R
- Here we are interested in weight of 200 teenagers
- A histogram of the distribution indicates an outlier

```r
# In R
> library(car)
> data(Davis)
> attach(Davis)
> hist(weight)
```
Mean Versus Median
Davis data example

The mean and median weight are as follows:

```
> mean(weight)
[1] 65.8
> median(weight)
[1] 63
```

It was determined that one observation was miscoded 166 instead of 66; after this is fixed, the mean = 63.8 and the median = 63

Again, we see that the mean is unduly pulled by a single observation, but the median is unaffected

The median may be the best measure of central tendency for skewed distributions, or in the presence of outliers, although it allows for less powerful statistical methods.
Measuring Spread: The Quartiles
Problem of the Range as a Measure of Spread

- The *range* is the difference between the maximum (highest) and minimum (lowest) values of the observations
- It is a simple measure of spread
- However the range has *limited utility* because
  - It depends on the smallest and largest values, and is therefore greatly affected by outliers
  - It is an unstable measure if the sample is drawn from a large population because it uses the most atypical values

```r
# In R
> x1<-c(3,3,4,2,4,4,4,4,4,2,6,7,100)
> x2<-c(3,3,4,2,4,4,4,4,4,2,6,7)
> range(x1)
[1] 2 100
> range(x2)
[1] 2 7
```
Measuring Spread: The Quartiles
Finding the First, Second & Third Quartiles

- To find Q1, Q2, and Q3:
  1. Order the observations from smallest to largest
  2. Find the median; this is Q2
  3. Use the median to divide the distribution in half
  4. The observation in the middle of the first half of the data (i.e., the median of those data) is Q1
  5. The observation in the middle of the second half of the data (i.e., the median of those data) is Q3

- Simple example:
  - Assume again the following *ordered* observations: 2.1, 2.3, 3.7, 4.1, 10, 11
  - We know that Q2 (the median) is 3.9 (between 3.7 and 4.1), thus Q1 is the median of 2.1, 2.3, 3.7 (Q1=2.3) and Q3 is the median of 4.1, 10, 11 (Q3=10)
  - Therefore, the *five number summary* is: 2.1, 2.3, 3.9, 10, 11
  - Of course, with so few cases, the five number summary is not really needed!
The Five-Number Summary & Boxplots

- The *five-number summary* is both a measure of spread & a measure of center
- It consists of a list of five numbers:
  1. Minimum
  2. First Quartile (Q1): The point that has 25% of observations below it
  3. Median (M or Q2): 50% of observations are below it
  4. Third Quartile (Q3): The point that has 75% of the observations below it
  5. Maximum
- The *inter-quartile range* or IQR (a.k.a. *hinge spread*) is the difference between Q3 and Q1. It is a useful measure of spread
- We display the five-number number summary simply in ascending order (min, Q1, M, Q3, max), or graphically using a boxplot
The Five-Number Summary & Boxplots

Davis data example using R

- Again using the Davis data in the car package
- Recall that each observation indicates the weight of a teenagers (200 observations)
- There are two simple ways to find the five number summary in R:

```r
> # In R
> library(car)
> data(Davis)
> attach(Davis)
> summary(weight)

   Min. 1st Qu.  Median    Mean 3rd Qu.   Max. 
  39.00  55.00   63.00   65.80  74.00  166.00

> fivenum(weight)

[1] 39 55 63 74 166
```
The Five-Number Summary & Boxplots

- A boxplot displays the information from the five number summary in graphical form

```
> # In R:
> data(Davis)
> attach(Davis)
> boxplot(weight)
```
The Five-Number Summary & Boxplots

The $1.5 \times \text{IQR}$ rule for suspected outliers
Side-by-side boxplots are useful to quickly compare the distributions of two or more variables.

Using the Davis data again, I constructed boxplots of weight for both men and women.

It is clear that men are heavier. There is also an extreme outlier for women.

```r
> boxplot(weight~sex,
main="Boxplots of weight by sex",
ylab="Weight", col="yellow")
```

![Boxplots of weight by sex](image-url)
Measuring Spread: The Standard Deviation

The variance $s^2$

- The variance $s^2$ can be seen as (very nearly) the average squared distance of the observations from the mean.
- The variance of $n$ observations $x_1, x_2, \ldots, x_n$ is:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum (x_i - \bar{x})^2$$

- As the variance $s^2$ is measured in squared units of the original variable, it is not a very intuitively meaningful number.
  - For example, the variance of weight in the Davis data (both sexes combined) is $s^2 = 227.9 \text{ kg}^2$. 

Typically we are more interested in the \textit{standard deviation}, \( s \) which is simply the square root of the variance, rather than the variance itself.

This is the most commonly used measure of spread.

\[
s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}
\]
Measuring Spread: The Standard Deviation
Calculating $s^2$ and $s$

- To calculate $s^2$ break down the formula

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

into its individual components and make a table.

Steps:

1. List all the individual observations
2. Calculate the mean
3. Subtract the mean from each individual observation
4. Square each result and sum them all together
5. Divide by $n - 1$ to get $s^2$
6. Take the square root to get $s$
Measuring Spread: The Standard Deviation
Calculating $s^2$ and $s$

- Infant-mortality rates for the 10 South American countries

<table>
<thead>
<tr>
<th>Country</th>
<th>$x_i$</th>
<th>$x_i - \bar{x}$</th>
<th>$(x_i - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecuador</td>
<td>78.5</td>
<td>7.99</td>
<td>63.84</td>
</tr>
<tr>
<td>Venezuela</td>
<td>51.7</td>
<td>−18.81</td>
<td>353.82</td>
</tr>
<tr>
<td>Argentina</td>
<td>59.6</td>
<td>−10.91</td>
<td>119.03</td>
</tr>
<tr>
<td>Brazil</td>
<td>170.0</td>
<td>99.49</td>
<td>9898.26</td>
</tr>
<tr>
<td>Chile</td>
<td>78.0</td>
<td>7.49</td>
<td>56.10</td>
</tr>
<tr>
<td>Columbia</td>
<td>62.8</td>
<td>−7.71</td>
<td>59.44</td>
</tr>
<tr>
<td>Peru</td>
<td>65.1</td>
<td>−5.41</td>
<td>29.27</td>
</tr>
<tr>
<td>Uruguay</td>
<td>40.4</td>
<td>−30.11</td>
<td>906.61</td>
</tr>
<tr>
<td>Bolivia</td>
<td>60.4</td>
<td>−10.11</td>
<td>102.21</td>
</tr>
<tr>
<td>Paraguay</td>
<td>38.6</td>
<td>−31.91</td>
<td>1018.25</td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{\sum x_i}{n} = \frac{705.1}{10} = 70.51
\]

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

\[
= \frac{1}{10-1} \times 12606.83
= 1400.76
\]

\[
s = \sqrt{1400.76} = 37.4
\]
Measuring Spread: The Standard Deviation

Calculating $s^2$ and $s$

- Calculate $s$ in R

```
> # In R, infant mortality for 10 Latin American countries
> x<-c(78.5,51.7,59.6,170.0,78.0,62.8,65.1,40.4,60.4,38.6)
> var(x)
[1] 1400.759
> sd(x)
[1] 37.42671
```
Measuring Spread: The Standard Deviation

Properties of the Standard Deviation

- The standard deviation $s$ measures the *spread of the distribution around the mean*
- Unlike the variance $s^2$ the standard deviation $s$ is measured in the original units of the variable, not squared units
  - E.g., the $s$ for weight in the Davis data (both sexes) is 15.1 kg, for a mean weight of 65.8 kg
- Since $s$ is also calculated from the mean it should be *used only when the mean is the appropriate measure of center* (in particular, for symmetric distributions)
  - Like the mean, the standard deviation is *not resistant*: it can be strongly influenced by outliers
- When there is no spread (i.e., when all observations are equal), $s = 0$; When there is spread, $s > 0$
Choosing Measures of Center & Spread

- The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or one with outliers.
- Use $\bar{x}$ and $s$ only for reasonably symmetric distributions that are free of outliers.
- For some purposes $\bar{x}$ is needed, even though the distribution is skewed.
  - E.g., we need to calculate the average value of homes in the county to estimate the total value of homes and thus the overall impact of a change in the real estate tax rate.
Changing the Unit of Measurement

- A *linear transformation* changes the original variable $x$ into the new variable

$$x_{\text{new}} = a + bx$$

Examples:
- If $x$ is a distance in kilometers, the same distance in miles is

$$x_{\text{new}} = 0.62x$$

with $a = 0$ and $b = 0.62$
- If $x$ is a temperature in degrees Fahrenheit, the same temperature in degrees Celsius is

$$x_{\text{new}} = \frac{5}{9}(x - 32) = -\frac{160}{9} + \frac{5}{9}x$$

where $a = -\frac{160}{9}$ and $b = \frac{5}{9}$. Thus last Saturday’s high of 91 F translates into 32.8 C.

- *Linear transformations do not change the shape of a distribution*, although the center and spread will change.
Density Curves

- Histograms are useful to describe the distribution of a variable, but the appearance of the graph can differ dramatically depending on the number of bins employed.
- This problem can be overcome using nonparametric density estimation (a.k.a. density curves or kernel density estimation).
  - The density curve can be thought of as smoothed histograms.
  - The density curve is defined as enclosing an area of 1. The y-axis is rescaled so that the total area under the smoothed line (the area within the bins) equals 1.
  - In other words, density curve shows the density of cases around specific points in the histogram rather than the frequency counts in predefined bins.
Density Curves

Density curve in R – Vertical scale is adjusted so total area under the curve is 1

```
> data(Leinhardt)
> attach(Leinhardt)
> # density function works only
> # if there are no missing cases
> infant2<-na.omit(infant)
> hist(infant2, probability=TRUE)
> lines(density(infant2),
>      col="red", lwd=3)
```
Density Curves
Density curve in Stata – Vertical scale not adjusted

```
. use "D:\soci708\data\data_from_car_Stata\Leinhardt.dta", clear
. histogram infant, bin(14) frequency kdensity xtitle(Infant Mortality Rate per K Live Births)
   (bin=14, start=9.6000004, width=45.742857)
```
We usually scale a histogram so that the height of the bars tells us either the *number* of observations or the *percentage* of observations in each bin:
An alternative is to scale (that is, label) the vertical axis so that the area of each bar gives the proportion of the data falling in the corresponding bin.

- In this case, the total area of the bars adds up to 1.0.
- The vertical axis now gives densities rather than numbers or percentages.
- The calculation of the density is shown for one bar in the following histogram:
  - 42 of the 193 countries have infant-mortality rates in the first bar (0.10, with width 10), so the proportion in the first bar is 42/193 = 0.218, and the density in the first bar is 0.218/10 = 0.0218.
Density Curves

$10 \times 0.0218 = 0.218$
Density Curves

- Suppose that we collect a great deal of data for a quantitative, continuous variable such as height.
  - We take advantage of the large size of the dataset by making a histogram with many bars, each of which is very narrow.
  - In most cases, the histogram will get smoother and smoother as the number of bars grows.
  - We can imagine that “in the limit” areas under the histogram can be closely approximated by a smooth curve (as in the following graphs, drawn for 50,000 observations; the darker bars represent heights between 130 and 135 cm.)
Density Curves

- This smooth curve is called a *density curve* (or *density function*), which has the following properties:
  - The total area under the density curve is 1.0.
  - The proportion of the data lying between any two values of the variable is simply the area under the curve between those two values.
  - The density curve never goes below the horizontal axis.
- Finding areas under density curves usually requires advanced mathematics – integral calculus – but we need simply to understand the basic idea (that the area represents the proportion of cases).
Center and Spread for Density Curves

Center: Mean and Median

- Just as we learned to calculate the mean, median, and standard deviation of a set of observations, we can also define the mean, median, and standard deviation of a density curve.
- The median of a density curve is the value of the variable that has half the area below it and half the area above it.
- The formal definition of the mean of a density curve requires advanced mathematics (calculus), but you can think about the mean in the following simple way: If the density curve is cut out of a thick piece of solid material (like metal), then the mean is the point at which the curve balances.
  - The mean of a density curve is symbolized by the Greek letter \( \mu \) (“mu”).
  - Greek letters are traditionally used for characteristics of density curves, to help us keep separate the mean \( \bar{x} \) of a set of observations, and the mean \( \mu \) of an idealized density curve.
Center and Spread for Density Curves

Center: Mean and Median

- If the density curve were made of some solid material (like steel), the mean would be the point on the density curve where it could be perfectly balanced.
- It is not the point that divides the area in half (that is the median).

Adapted from Moore & McCabe (2006, Figure 1.25 p. 68)
Center and Spread for Density Curves

Center: Mean and Median

- In a symmetric distribution, the mean and the median coincide.
- In a skewed distribution, both the mean and the median are pulled in the direction of the skew – the mean more so than the median.
It is also possible to define the standard deviation of a density curve, which is symbolized by the Greek letter \( \sigma \) (“sigma”).

The standard deviation of a density curve is harder to visualize than the mean, but – as in the case of the standard deviation \( s \) of a set of observations – we can think of \( \sigma \) as a kind of average distance to the mean.
Normal Distributions

- *Normal curves* are a family of density curves that are symmetric, single-peaked, and bell-shaped.
- A specific characteristic of normal distributions is that any particular normal distribution is completely described by its mean $\mu$ and standard deviation $\sigma$.
  - The notation $x \sim N(\mu, \sigma)$ means that $x$ is normally distributed with mean $\mu$ and standard deviation $\sigma$.
- The larger the standard deviation, the more spread out is the curve; curvature changes at a distance $\sigma$ from $\mu$.
Normal Distributions

Importance of Normal Distributions

- Optional – The height of the curve is the value of the normal probability density function (pdf) given at any point \( x \) by

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}
\]

- Normal distributions are exceedingly important in statistics because they are good descriptions of the distributions of many phenomena that are affected by many factors (such as genes and environmental influences) each of which has a small impact on the phenomenon in question, e.g.:
  - Test scores, repeated measurements of the same quantity, biological measurements,...
  - Chance outcomes, such as tossing a coin many times

- The natural mechanism responsible for the convergence to a normal distribution is called the Central Limit Theorem (see later)

- For the same reason, many tests used in statistics are normally distributed.
Normal Distributions

The 68–95–99.7 Percent Rule

- For any normal distribution with mean $\mu$ and standard deviation $\sigma$:
  - 68% of cases fall within one $\sigma$ of the mean;
  - 95% fall within two $\sigma$ of the mean;
  - 99.7% fall within three $\sigma$ of the mean

The 68–95–99.7 percent rule for normal distributions. From Moore & McCabe (2006, Figure 1.27 p. 72).
The 68–95–99.7 percent rule applied to the distribution of heights of young women

Here $\mu = 64.5\text{in}$ & $\sigma = 2.5\text{in}$

The 68–95–99.7 rule applied to the heights of young women.
From Moore & McCabe (2006, Figure 1.28 p. 72).
If we know the mean $\bar{x}$ & standard deviation $s$ we can calculate the standardized value $z$ (called z-score) for each observation:

$$z = \frac{x - \bar{x}}{s}$$

The z-score tells us how many standard deviations the observation falls from the mean.

- A negative z-score indicates that the observation is smaller than the mean; a positive z-score indicate that the observation is larger than the mean.
Standardization

Calculating z-Scores: An Example

- For this example we look at the infant mortality rate (Leinhardt data in the car library)
- For infant mortality, $\bar{x} = 89.0$ and the $s = 90.8$
- Canada has an infant mortality rate of 16.8
- Therefore, the standardized score for Canada is:

$$z = \frac{x - \bar{x}}{s} = \frac{16.8 - 89.0}{90.8} = -0.795$$

- In other words, the infant mortality rate in Canada is .795 standard deviations less than the mean

```r
> # Using R
> library(car)
> data(Leinhardt)
> mean(infant, na.rm = TRUE)
[1] 89.04752
> sd(infant, na.rm = TRUE)
[1] 90.8017
> (16.8-89.05)/90.8
[1] -0.7957048
```
Standardization
Uses of z-Scores

- Contrary to a widespread belief, standardized scores (z-scores) can be calculated for *any* distribution, whether it is normal or not, as long as one knows $\bar{x}$ and $s$

- But it is true that z-scores are particularly useful in the context of normal distributions, as explained later

- When a distribution is skewed or has outliers, z-scores are less meaningful
  - because they are based on $\bar{x}$ and $s$, which are not resistant measures of center and spread, respectively

- Contrary to another widespread belief, transforming a variable into z-scores *does not render the distribution of the variable normal*
  - In fact the transformation does not change the shape of the distribution at all
Standardization
Standard Normal Distribution

- Notation: a normal distribution with mean $\mu$ and standard deviation $\sigma$ is denoted $N(\mu, \sigma)$
  - e.g. the distribution of height of young women with $\mu = 64.5\text{in}$ and $\sigma = 2.5\text{in}$ is $N(64.5, 2.5)$
- The *standard normal distribution*, denoted $N(0, 1)$ is the normal distribution with mean 0 and standard deviation 1.
- If a variable $X$ has any normal distribution $N(\mu, \sigma)$ with mean $\mu$ and standard deviation $\sigma$, then the standardized variable

$$Z = \frac{X - \mu}{\sigma}$$

has the standard normal distribution.
Normal Distribution Calculations

▶ If we assume that a distribution is normal, we can easily determine the proportion of cases that have a score lower or higher than a particular value
  ▶ The area to the left of a z-score is the proportion lying below the observation; the area to the right is the proportion of observations that are larger
▶ We can also think of these areas as probabilities
  ▶ i.e., the probability that a case falls below or above a particular value
▶ It is also possible, then, to determine the proportion of cases falling between two observations
▶ Some hand calculators are able to calculate areas under the standard normal curve; Alternatively, a table can be used (see Table A in IPS6e, inside back cover)
Normal Distribution Calculations

General Strategy for Normal Probability Calculations

1. Transform a problem in terms of a variable X into a standard normal problem in terms of z-score. To do this transform the value of X into a z-score as
   ▶ \( z = (x - \mu)/\sigma \)

2. Resolve problem in terms of z-scores
   ▶ e.g., find the proportion of cases falling below a certain value \( z \)

3. If needed, transform \( z \) back into a value \( x \) of the original variable X using the formula
   ▶ \( x = \mu + z\sigma \)

   For example, suppose one wants to estimate how many people weigh less than 45.3 Kg (100 Lb) in the Davis data where \( \mu = 65.8 \) Kg & \( \sigma = 15.09 \) Kg
   ▶ \( z = (45.3 - 65.8)/15.09 = -1.36 \)
Normal Distribution Calculations
Finding a Cumulative Proportion for z-Scores

- A weight of 45.3 Kg translates to $z = -1.36$. I want to know what proportion of teenagers weigh less.

- Start in the first column of Table A (IPS6e, inside back cover) finding $-1.3$; then go across the row to column that says .06. The value here (.0869) is the area to the left of $z = -1.36$.

- We conclude that only about 8.7% of teenagers weigh less than 45.3 Kg (100 Lb).
Normal Distribution Calculations
How To Calculate a Cumulative Probability?

1. From printed table (see previous slide)
2. From a calculator on the web, e.g. at http://rockem.stat.sc.edu/prototype/calculators/index.php3
3. From Stata:
   ```
   . display normal(-1.36)
   .08691496
   ```
4. From R:
   ```
   > pnorm(-1.36,0,1)
   [1] 0.08691496
   ```
5. From Excel:
   ```
   =NORMSDIST(-1.36) 0.086914962
   ```
Normal Distribution Calculations

An example (1)

- The following example is based on the Davis data from the car package in R
- You’ve been asked to determine what proportion of people have a weight between 60 and 80 kilograms (assume the distribution is normal)

1. Find the mean and the standard deviation of weight (after correcting observation #12)

   \[ \mu = 65.255; \quad \sigma = 13.32282 \]

2. Draw a picture of the normal curve, locating the mean and the values of interest (those between which we want the area)
Normal Distribution Calculations
An example (2)

3. Standardize the two values (60 and 80):
   - For 60: $z = \frac{60-65.255}{13.32282} = -0.39443601$
   - For 80: $z = \frac{80-65.255}{13.32282} = 1.1067477$

4. Use Table A to find the area between $z = -0.39$ and 1.11
   - Area to left of $-0.39$ is 0.3483
   - Area to left of 1.11 is 0.8665
   - Therefore, the area between $z = -0.39$ and $z = 1.11$ is the difference between these two areas calculated above:
     \[ 0.8665 - 0.3483 = 0.5182 \]

5. We conclude that about 52% of people weigh between 60 and 80 kilograms
Normal Distribution Calculations
Change in Ivy League Freshmen IQ, 1926–1964 (1)

In 1926 the average IQ of freshmen in Ivy League and Seven Sisters universities was 117. To what percentile of the IQ distribution does 117 correspond, assuming IQ normally distributed with \( \mu = 100 \) and \( \sigma = 15 \)?

Standardize the value 117:

\[
z = \frac{117 - 100}{15} = 1.133333
\]

Find area to the left of 1.133333 = 0.871463

Conclude 117 corresponds to the 87th percentile of IQ

---

Normal Distribution Calculations

Change in Ivy League Freshmen IQ, 1926–1964 (2)

▶ In 1964 for the same 14 schools the average freshman was at the 99th percentile of IQ. What was the average freshman IQ score?

▶ First calculate the $z$ score corresponding to the 99th percentile as:

$$ z = 2.326348 $$

▶ Then transform the $z$ score back to the IQ scale:

$$ x = 100 + 2.326348 \times 15 = 134.9 $$

▶ Conclude the average freshman IQ in 1964 was $\approx 135$

> # in R
> qnorm(0.99)
[1] 2.326348
> 100 + 2.326348*15
[1] 134.8952

. * In Stata
. display(invnormal(0.99))
2.3263479
. display(100+2.3263479*15)
134.89522
Normal Distribution Calculations
The Four Normal Distribution Functions in Stata

1. Normal *pdf* (probability density function)

2. Normal *cdf* (cumulative density function, a.k.a. *probability* function)

3. Inverse normal cdf (a.k.a. *quantile* function)

4. Normal number generating function

   . * In Stata
   . display normalden(-1.497)
     .13010115

   . display normal(-1.497)
     .06719663

   . display invnormal(.06719663)
     -1.497

   . display rnormal(100, 15)
     131.8627

   . * To draw new data set
   . drawnorm newx, n(500) clear (obs 500)
Comparing Normal Distributions

Threshold Problems

- The normal distribution is an important device to compare normally distributed populations with respect to the proportion of the respective populations above (or below) a given threshold.
- It is a valuable tool for understanding:
  - Issues related to definitions of morbidity – e.g. what are the consequences of setting a certain cutoff of BMI for obesity; of IQ for cognitive deficiency.
  - Outcomes of selection processes for different groups – e.g. college admission; occupational achievement.
Comparing Normal Distributions
Threshold Problems – Cholesterol Levels

- Cutoffs for blood cholesterol levels are:\(^4\)
  - Over 200 but below 240 mg/dl = *borderline high cholesterol*
  - Over 240 mg/dl = *high cholesterol* (needs medical attention)

- Cholesterol levels for women aged 20 to 34 follow an approximately normal distribution with mean 185 mg/dl and s.d. 39 mg/dl

  - What percent of young women have levels 200 mg/dl and above?
    - \( z = (200 - 185) / 39 = 0.3846154 \)
    - \( 1 - \text{area left of} \ 0.3846154 = 0.3502612 \)
    - 35.0% of young women have levels 200 mg/dl and above

  - What percent of young women have levels 240 mg/dl and above?
    - \( z = (240 - 185) / 39 = 1.410256 \)
    - \( 1 - \text{area left of} \ 1.410256 = 0.07923205 \)
    - 7.9% of young women have levels 240 mg/dl and above

\(^4\)Based on Problems 1.110 & 1.111 in IPS6e
Comparing Normal Distributions
Threshold Problems – Cholesterol Levels

- Cholesterol levels for men aged 55 to 64 follow an approximately normal distribution with mean 222 mg/dl and s.d. 37 mg/dl
  - What percent of middle-aged men have levels 200 mg/dl and above?
    - $z = \frac{200 - 222}{37} = -0.5945946$; $1 - \text{area left of } (-0.5945946) = 0.7239428$
    - 72.4% of middle-aged men have levels 200 mg/dl and above
  - What percent of middle-aged men have levels 240 mg/dl and above?
    - $z = \frac{240 - 222}{37} = 0.4864865$; $1 - \text{area left of } 0.4864865 = 0.3133111$
    - 31.3% of middle-aged men have levels 240 mg/dl and above
Comparing Normal Distributions

Threshold Problems – Cholesterol Levels

Practical

- Calculate percent with cholesterol levels 280 mg/dl and above for both groups
  - Women 20–34:
  - Men 55–64:
- What is the risk of levels 280 mg/dl and above for men 55–64 relative to women 20–34?
Comparing Normal Distributions
Threshold Problems – Cholesterol Levels

<table>
<thead>
<tr>
<th></th>
<th>&gt;200 mg/dl</th>
<th>&gt;240 mg/dl</th>
<th>&gt;280 mg/dl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women 20 to 34</td>
<td>35.0</td>
<td>7.9</td>
<td></td>
</tr>
<tr>
<td>Men 55 to 64</td>
<td>72.4</td>
<td>31.3</td>
<td></td>
</tr>
</tbody>
</table>

| Male/female risk | 2.07       | 3.96       |

- Relative male risk of levels 200 and above is about 2 (2.07)
- Relative male risk of levels 240 and above is almost 4 (3.96)
- This illustrates a general pattern of group differences for a normally distributed trait:
  - Difference between groups in the proportion above (or below) a given cutoff *increases* the more extreme the cutoff
Comparing Normal Distributions

Difference in means of two normal distributions: Cohen’s $d$

- Cohen’s $d$ is a standardized measure of the difference between the means of two distributions. It is often called a measure of effect size (i.e., the “effect” of group membership).

- Cohen’s $d$ is calculated as the difference between the two means $\mu_1$ and $\mu_2$ scaled in units of the standard deviation $\sigma$ of $x$ within the groups:

  $$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- When $\sigma$ is not the same within the two groups, one can take $\sigma$ as (1) that of a standard scale (e.g., for IQ scores $\sigma = 15$), or (2) the standard deviation within one of the groups (e.g., the control group in an experiment), or (3) a composite of the standard deviations of the two groups: $\sigma = \sqrt{(\sigma_1^2 + \sigma_2^2)/2}$
Comparing Normal Distributions

Difference in means of two normal distributions: Cohen’s $d$

- Cohen (1988) suggests that values of $d = .2$, .5, and .8 indicate small, medium and large effects, respectively.
- In the cholesterol level example (using the third formula)

$$
d = \frac{222 - 185}{\sqrt{(39^2 + 37^2)/2}} = \frac{37}{38.01} = 0.97
$$

which indicates a large difference (effect size) on Cohen’s scale.

- In the Davis data (corrected), the distribution of height is:
  - $\mu_1 = 178.0, \sigma_1 = 6.441$ for men
  - $\mu_2 = 164.7, \sigma_2 = 5.659$ for women
  - $d = (178 - 164.7)/6.063 = 2.19$

which indicates a rather big difference.
Comparing Normal Distributions
Difference in means of two normal distributions: Cohen’s $d$ – Real example

- In the aftermath of *The Bell Curve* a blue-chip panel of the American Psychological Association prepared a report reviewing the literature on IQ entitled “Intelligence: Knowns and Unknowns.” (Neisser et al. (1995))

- Conclusions of the report are often expressed using Cohen’s $d$.

- On male-female IQ differences the report concludes that:
  1. There is no evidence of male-female difference in general intelligence ($d \approx 0$).
  2. Average male-female differences are found in some specific cognitive abilities, e.g.
     - Scores on mental rotation tasks: $d = 0.9$ (male higher)
     - Scores on math portion of SAT: $d = 0.33$ to $0.5$ (male higher), “with many more males scoring in the highest ranges”
     - Synonym generation and verbal fluency (e.g., naming words that start with a given letter): $d = 0.5$ to $1.2$ (female higher), “many more males than females are diagnosed with” dislexya, other reading disabilities, and stuttering.

---

*Herrnstein and Murray (1994)*
Comparing Normal Distributions
Role in Public Debate

- Comparisons of normal distributions underlie some public debates such as the one following remarks by Harvard President Lawrence Summers about the small proportions of women faculty in top Physics and Engineering departments.
  - The remarks by Lawrence Summers can be found here: Remarks at NBER Conference on Diversifying the Science & Engineering Workforce, 14 Jan 2005. (Summers’ remarks involve properties of the normal distribution.)
  - The response by the Council of the American Sociological Association can be found here: ASA Council Statement on Causes of Gender Differences in Science and Math Career Achievement, 28 Feb 2005
Normal Quantile Plots
Assessing the Normality of a Distribution

- To assess the normality of a variable one can look at the histogram or kernel density curve and verify that the distribution “looks” normal.
- A more sensitive diagnostic is the normal quantile plot.
  - Basically it is a plot of each observation $X$ (on the $y$-axis) against the value that would be expected if the distribution is normal (on the $x$-axis).
  - If the plot looks approximately like a straight line this indicates the distribution is normal.
  - Gaps at the tails indicate outliers.
  - Curvilinear deviations from the straight line indicate skewness.
Normal Quantile Plots

Steps in Constructing a Normal Quantile Plot

1. Order the observations from smallest to largest
2. Call $i$ the rank of the observation from 1 (smallest) to $n$ (largest)
3. Then $\frac{i-0.5}{n}$ corresponds to a subdivision of the interval (0,1) into equal intervals
4. Calculate the expected value of the observation as the “normal score”, the z-score corresponding to the $100\left(\frac{i-0.5}{n}\right)$th percentile of the $N(0,1)$ distribution; this is the same as “grading on the curve”
5. Finally plot the observations against the normal scores
6. Stata `qnorm` further transforms the normal scores into $X$ scores with mean $\bar{x}$ and standard deviation $s$. 
Normal Quantile Plots

- Infant-mortality rates for the 10 South American countries (sorted)

<table>
<thead>
<tr>
<th>Country</th>
<th>$x_i$</th>
<th>$i$</th>
<th>$\frac{i-0.5}{n}$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraguay</td>
<td>38.6</td>
<td>1</td>
<td>0.05</td>
<td>−1.645</td>
</tr>
<tr>
<td>Uruguay</td>
<td>40.4</td>
<td>2</td>
<td>0.15</td>
<td>−1.036</td>
</tr>
<tr>
<td>Venezuela</td>
<td>51.7</td>
<td>3</td>
<td>0.25</td>
<td>−0.674</td>
</tr>
<tr>
<td>Argentina</td>
<td>59.6</td>
<td>4</td>
<td>0.35</td>
<td>−0.385</td>
</tr>
<tr>
<td>Bolivia</td>
<td>60.4</td>
<td>5</td>
<td>0.45</td>
<td>−0.126</td>
</tr>
<tr>
<td>Columbia</td>
<td>62.8</td>
<td>6</td>
<td>0.55</td>
<td>0.126</td>
</tr>
<tr>
<td>Peru</td>
<td>65.1</td>
<td>7</td>
<td>0.65</td>
<td>0.385</td>
</tr>
<tr>
<td>Chile</td>
<td>78.0</td>
<td>8</td>
<td>0.75</td>
<td>0.674</td>
</tr>
<tr>
<td>Ecuador</td>
<td>78.5</td>
<td>9</td>
<td>0.85</td>
<td>1.036</td>
</tr>
<tr>
<td>Brazil</td>
<td>170.0</td>
<td>10</td>
<td>0.95</td>
<td>1.645</td>
</tr>
</tbody>
</table>

- Normal quantile plot is plot of $x_i$ against $z_i$, e.g. as in Moore et al. & R `qqnorm`

- Stata `qnorm` is plot of $x_i$ against $x'_i = \bar{x} + s \times z_i$
Normal Quantile Plots

> # in R
> x <- c(38.6, 40.4, 51.7, 59.6, 60.4, 62.8, 65.1, 78.0, 78.5, 170.0)
> par(pty="s") # make figure square
> qqnorm(x, main="Normal Q-Q Plot of Infant Mortality",
        col="red", cex=1.5)
> qqline(x, col="blue")

▶ Outlier Brazil appears clearly at upper right corner of the plot
Normal Quantile Plots
Stata examples with the Davis data (corrected) – Why is weight less normal-looking?

```
. qnorm height, xsize(2) ysize(2)
. qnorm weight, xsize(2) ysize(2)
```
Normal Quantile Plots
Simulated Symmetric (N(0,1)), Fat-Tailed (t 5df) & Right-skewed ($\chi^2$ 3df) Distributions
Conclusions

- The mean and standard deviation are very useful summary statistics when the distribution of the variable is normal
  - If not, they are typically not appropriate
  - For example, for skewed distributions we may choose to look at the median (which is less influenced by unusual cases) rather than the mean
- This caution makes it clear that it is important to graphically explore the distributional shape of a variable before calculating and relying on numerical summaries.
  - As will become clear as the course goes on, graphical exploration should be done at all stages of data analysis.
- The normal distribution also has important predictive uses in comparing populations in relation to a threshold