# Univariate Descriptive Statistics for One Sample

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

1. Introduction
2. The Grouped Frequency Distribution Table
3. The Frequency Histogram
4. The Cumulative Frequency Distribution Plot
5. The Cumulative Relative Frequency Plot
6. The Stem-Leaf Diagram
7. The Box-and-Whisker Plot
8. The Q-Q plot

# Introduction

- Our first step in descriptive statistics is to characterize the data in a single group of observational units, with only one variable measured per unit.
- For example, suppose we measure the weights in pounds of a group of 50 adult males, and obtain the following data.

Table 1 : Weights of 50 Adult Males

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 165 | 105 | 147 | 170 | 169 | 195 | 170 | 162 | 178 | 187 |
| 168 | 166 | 195 | 128 | 186 | 138 | 205 | 142 | 90 | 115 |
| 147 | 143 | 159 | 142 | 134 | 166 | 201 | 108 | 123 | 164 |
| 139 | 143 | 163 | 178 | 111 | 165 | 191 | 194 | 173 | 165 |
| 165 | 127 | 131 | 157 | 110 | 146 | 156 | 151 | 171 | 117 |

- We can begin by constructing a *grouped frequency distribution table* from these data.

# The Grouped Frequency Distribution Table

- The purpose of this kind of a table is to select a set of intervals on the number line, then count the number of values that fall into each interval.
- Connected with a grouped frequency distribution table are a number of glossary terms that we will define and illustrate with the current example.

# The Grouped Frequency Distribution Table

- The first task in constructing a grouped frequency distribution table is to examine the numbers, and divide the range of the values into a reasonable number of intervals.
- Generally, we want to have between 8 and 12 intervals, but there are no hard and fast rules (although there are computer algorithms to select the number of intervals automatically).
- You want enough intervals to display a reasonable picture of distribution shape, and not so many intervals that there are many intervals with low counts. Generally, you also want all intervals to be the same width.
- Connected with a grouped frequency distribution table are a number of glossary terms that we will define and illustrate with the current example.

# The Grouped Frequency Distribution Table

- Examining the data in Table 3, we see that the recorded values range from 90 to 205. If we desire nice even interval endpoints, we might select an interval width of 10.
- This would result in 12 intervals each of width 10.
- We can count the number of numbers in each of the 12 intervals by hand, or we can let R do it for us.

# Grouped Frequency Distribution Table

## Nominal vs. Real Intervals

- In setting up our interval limits, it is important to remember the distinction between *nominal limits* and *real limits* for the intervals
- Suppose that our first interval includes weights that are recorded as being between 90 and 99 pounds. These weights are rounded to the nearest whole pound.
- If your scale is perfectly accurate, but rounds to the nearest whole pound, then a weight of 99 pounds might stand for any value up to but not including 99.5, because, for example, a weight of 99.34 would be rounded down to 99.
- In a similar vein, we realize that a person whose weight is 90 pounds could be as light as 89.5, in which case the weight would be rounded up to 90.
- Putting these facts together, an interval with *nominal limits* of 90 and 99 has real limits of 89.5 and 99.5, and has a real width of 10.

# Grouped Frequency Distribution Table

## R Code

- The data can be retrieved online from a file called *weights.csv*. This file has one variable called `wts`. This code uses a function, `grouped.frequency.table`, which I have created for you:

```
> weights <- read.csv("http://www.statpower.net/310Data/weights.csv")
> wts <- weights$wts
> lower <- seq(90,200,10)
> upper <- seq(99,209,10)
> grouped.frequency.table(wts,lower,upper,round.off=1)
   lower upper lower.real upper.real  f cum.f rel.f cum.rel.f
1    200   209      199.5      209.5  2    50  0.04      1.00
2    190   199      189.5      199.5  4    48  0.08      0.96
3    180   189      179.5      189.5  2    44  0.04      0.88
4    170   179      169.5      179.5  6    42  0.12      0.84
5    160   169      159.5      169.5 11    36  0.22      0.72
6    150   159      149.5      159.5  4    25  0.08      0.50
7    140   149      139.5      149.5  7    21  0.14      0.42
8    130   139      129.5      139.5  4    14  0.08      0.28
9    120   129      119.5      129.5  3    10  0.06      0.20
10   110   119      109.5      119.5  4     7  0.08      0.14
11   100   109       99.5      109.5  2     3  0.04      0.06
12    90    99       89.5       99.5  1     1  0.02      0.02
```

# Grouped Frequency Distribution Table

- In the table, we see each interval's nominal and real limits.
- There are a number of other quantities connected with the $i^{th}$ interval.
  1. The *frequency*, $f_i$, i.e., the number of values that occur in the $i^{th}$ interval.
  2. The *cumulative frequency*, $cum.f_i$, the number of values that occur *at or below* the $i^{th}$ interval.
  3. The *relative frequency*, $rel.f_i$, the proportion of values that occur in the $i^{th}$ interval. The relative frequency is obtained by dividing the frequency by $n$, the total number of cases.
  4. The *cumulative relative frequency*, $cum.rel.f_i$, the proportion of values that occur *at or below* the $i^{th}$ interval. The cumulative relative frequency is obtained by dividing the cumulative frequency by $n$, the total number of cases.

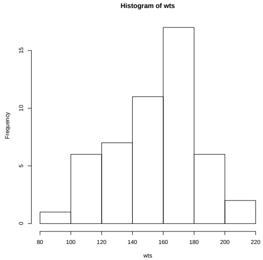
## Grouped Frequency Distribution Table

- Notice that the uppermost cumulative frequency (50 in this case) is always equal to $n$, the total number of observations.
- Notice also that the uppermost cumulative relative frequency is always equal to 1, and, of course, the sum of relative frequencies is equal to 1.

# The Frequency Histogram

- The grouped frequency distribution table can provide some quick summary information about how values are distributed on the number line.
- The *frequency histogram* provides an accompanying visual representation.
- Here is a histogram of the weights data, produced by a defaul call to the R function `hist`.

```
> hist(wts)
```



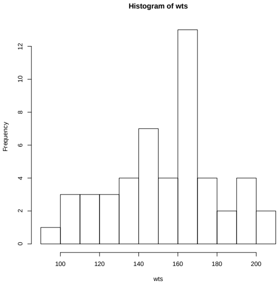
# The Frequency Histogram

- In this case, a default call produced a histogram with just 7 intervals.
- The histogram is slightly misleading in that the uppermost interval extends to 220, well past the largest value actually seen in the data.
- On the next slide is another histogram with an interval width of 10.

# The Frequency Histogram
## Matching Histogram and Frequency Table Intervals

Notice that the frequencies *do not match* those in the frequency distribution table. For example, the frequency for the interval from 160 – 170 does not match that for the frequency distribution table for the interval from 159.5 to 169.5.

```
> hist(wts, breaks=seq(90,210,by=10))
```



# The Frequency Histogram

## Matching Histogram and Frequency Table Intervals

- This is because the histogram intervals are, by default, closed on the right and open on the left. This means that setting an interval from 160–170 with the `hist` function causes numbers greater than 160 and less than or equal to 170 to be included in that interval.
- In order to get the interval to match the previous frequency distribution table, we have two options.

# The Frequency Histogram

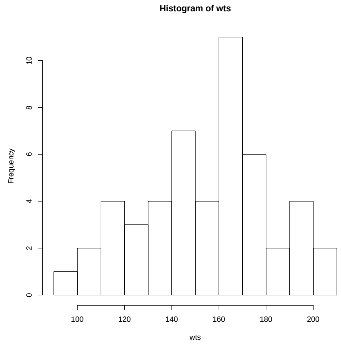
## Matching Histogram and Frequency Table Intervals

- One option is to include the option `right=FALSE`.
- This makes the intervals closed on the left and open on the right, meaning that setting an interval from 160–170 means that values greater than or equal to 160 and less than 170.
- When numbers are rounded to the nearest whole number, "less than 170" is equivalent to "less than or equal to 169."
- With this change, the histogram (shown on the next slide) now matches our previous frequency distribution table.

# The Frequency Histogram

## Matching Histogram and Frequency Table Intervals

```
> hist(wts, breaks=seq(90,210,by=10),right=FALSE)
```



# The Frequency Histogram

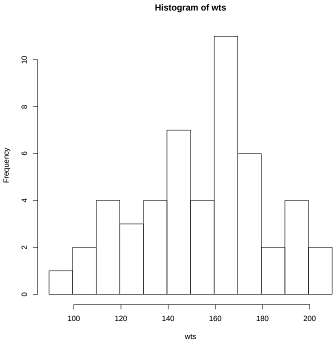
## Matching Histogram and Frequency Table Intervals

- An alternative option is to use the *real limits* from the frequency distribution table as the limits for the histogram. Let's try that.
- We need to begin the histogram at the smallest lower real limit (89.5 in this case), and end at the largest upper real limit (209.5 in this case), making the interals 10 units wide.
- Notice that the numbers on the horizontal axis of the plot *do not match* the actual interval breakpoints.

# The Frequency Histogram

## Matching Histogram and Frequency Table Intervals

```
> hist(wts, breaks=seq(89.5,209.5,by=10))
```



# The Frequency Histogram

## Matching Histogram and Frequency Table Intervals

- The bottom line on histograms is that the choice of the number of intervals and the precise breakpoints is partly an artistic decision.
- You should know exactly what the breakpoints are.
- With the R function `hist`, you can always determine precisely what breakpoints were used to construct the histogram by using the command `plot=FALSE` as shown below.
- However, you also must recall whether the intervals are closed on the right (the default) or on the left.

# The Frequency Histogram

## Matching Histogram and Frequency Table Intervals

```
> hist(wts, breaks=seq(89.5,209.5,by=10),plot=FALSE)

$breaks
 [1]  89.5  99.5 109.5 119.5 129.5 139.5 149.5 159.5 169.5 179.5 189.5 199.5
[13] 209.5

$counts
 [1]  1  2  4  3  4  7  4 11  6  2  4  2

$density
 [1] 0.002 0.004 0.008 0.006 0.008 0.014 0.008 0.022 0.012 0.004 0.008 0.004

$mids
 [1]  94.5 104.5 114.5 124.5 134.5 144.5 154.5 164.5 174.5 184.5 194.5 204.5

$xname
[1] "wts"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
```

# The Frequency Histogram

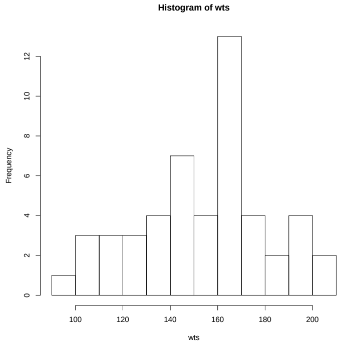
## Matching Histogram and Frequency Table Intervals

- There are several other ways you can alter the number of intervals in the histogram.
- For example, suppose you wanted the histogram to have exactly 10 intervals.
- You could use the command `breaks=10` as shown below.
- However, R has used your command only as a suggestion. Using a complicated algorithm, it has decided that the "best" histogram that comes close to that number of intervals has 12 intervals, rather than 10.

# The Frequency Histogram

## Matching Histogram and Frequency Table Intervals

```
> hist(wts, breaks=10)
```



# Plots of Data Distributions

## The Histogram

- Here is another histogram demonstration with the *RoyerY* data set described on page 21 of RDASA3.
- The data are percent correct addition scores for 28 male second-grade students.
- We begin by loading in the data directly from the course website.

```
> RoyerY <- read.csv(
+   "http://www.statpower.net/310Data/Royer_Y.csv")
```

- In order to make it easier to reference the variables in the file directly, we **attach** it.

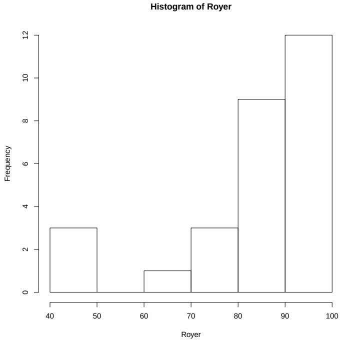
```
> attach(RoyerY)
```

# Plots of Data Distributions

## The Histogram

- A direct call to R's `hist` function produces a plot that is a bit lacking in detail.

```
> hist(Royer)
```

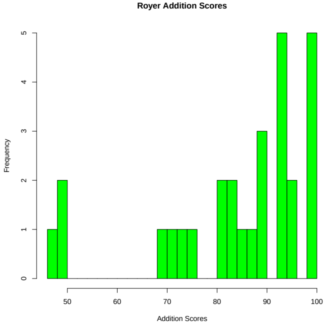


# Plots of Data Distributions

## The Histogram

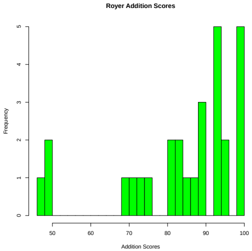
- We can do better by increasing the number of breaks and adding a correct axis label.

```
> hist(Royer, breaks=20, main="Royer Addition Scores",
+  xlab="Addition Scores", ylab="Frequency", col = "green")
```



# Plots of Data Distributions

## The Histogram



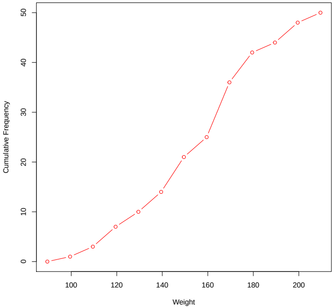
- Now we can clearly see a number of key aspects of the data:
  1. Most of the students did very well. Indeed, many scored 100%.
  2. The distribution is skewed, typical of test scores where the test “lacks headroom.”
  3. There is a gap in the distribution, with 3 students doing very poorly.
- A number of authors (e.g., Micceri (1989)) have commented that such characteristics are quite common in empirical data sets.

# The Cumulative Frequency Distribution Plot

- Instead of plotting a frequency distribution, we might choose instead to plot a cumulative frequency distribution.
- This is usually accomplished using points instead of bars.
- The cumulative frequency is plotted against the *upper real limits*.
- Why? Because we know that, in spite of the rounding off, if measurements are accurate, then the cumulative frequency plotted in the graph is correct at a point infinitesimally below the upper real limit.
- So although grouping data introduces uncertainty about where individual numbers are, some aspects of the distribution can be charted with perfect accuracy.

# The Cumulative Frequency Distribution Plot

```
> cum.f.plot(wts,lower,upper,round.off=1,
+            col="red",xlab="Weight",
+            ylab="Cumulative Frequency")
```

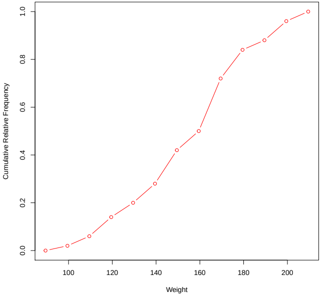


# The Cumulative Relative Frequency Distribution Plot

- The cumulative relative frequency plot shows cumulative relative frequencies versus upper real limits of the intervals.
- A *quantile* in a distribution is point in the distribution at or below which a certain proportion of cases falls.
- So, for example, if half the cases in a distribution fall at or below 100, then 100 is the .50 quantile for that distribution.
- Notice then, that the quantile value for any point in a distribution is also the cumulative relative frequency at that point.
- Consequently, a cumulative relative frequency plot is also a quantile plot.

# The Cumulative Relative Frequency Distribution Plot

```
> cum.f.plot(wts,lower,upper,round.off=1,
+           relative=TRUE,col="red",
+           xlab="Weight",
+           ylab="Cumulative Relative Frequency")
```



# The Cumulative Relative Frequency Distribution Plot

## Percentiles

- A *percentile point* in a distribution is a point at or below which a certain percentage falls.
- For example, the $90^{th}$ percentile is that point at or below which 90 percent of the cases fall.
- Any numerical value has a *percentile value* in a distribution.
- For example, if a value 100 has 20 percent of the cases at or below it, then 100 has a percentile value of 20.
- Percentile values are very useful for telling you "where you stand" on some measure of performance.
- Cumulative relative frequencies can be converted into percentiles simply by multiplying them by 100.

# The Stem-Leaf Diagram

- The stem-leaf diagram is a very popular quick “by hand” technique for displaying a frequency distribution in grouped form without necessarily throwing away any information from the numbers
- The stem-leaf diagram has its primary use with small-to-moderate sized lists of numbers, and, as we shall see, it is less flexible than the histogram in terms of interval limits.

# The Stem-Leaf Diagram

- For example, consider again the weights data.
- Here is a stem-leaf diagram produced by a basic call to the R function `stem`.
- Note that the `stem` command in R is defective, because it chooses the wrong number of intervals and represents some numbers incorrectly.

```
> stem(wts)

  The decimal point is 1 digit(s) to the right of the |

   8 | 0
  10 | 580157
  12 | 3781489
  14 | 22336771679
  16 | 2345555668900138 8
  18 | 671455
  20 | 15
```

# The Stem-Leaf Diagram

- By changing the `scale` parameter, you can obtain a correct basic stem-leaf plot with the `stem` function.
- Here we have the correct diagram. The first row represents a 90, the second row 105 and 108.

```
> stem(wts,scale=2)

  The decimal point is 1 digit(s) to the right of the |

   9 | 0
  10 | 58
  11 | 0157
  12 | 378
  13 | 1489
  14 | 2233677
  15 | 1679
  16 | 23455556689
  17 | 001388
  18 | 67
  19 | 1455
  20 | 15
```

# A Better Stem-Leaf Diagram

- The `aplpack` library has a `stem-leaf` diagram function that is superior to the `stem` command in base R.
- Besides providing the stems and leaves, the *depths* are also plotted on the left side of the chart.

# A Better Stem-Leaf Diagram

- Depths of the intervals are the cumulative frequencies counting down from the top, or up from the bottom.
- For example, the third interval down from the top contains the numbers 110,111,115,117 and the cumulative frequency down from the top is 7.

```
> library(aplpack)
> stem.leaf(wts)
```

```
1 | 2: represents 12
 leaf unit: 1
            n: 50
   1     9 | 0
   3    10 | 58
   7    11 | 0157
  10    12 | 378
  14    13 | 1489
  21    14 | 2233677
  (4)   15 | 1679
  25    16 | 23455556689
  14    17 | 001388
   8    18 | 67
   6    19 | 1455
   2    20 | 15
```

# A Better Stem-Leaf Diagram

- The third interval up from the bottom contains the numbers 86 and 87, and the cumulative frequency up from the bottom is 8 for that interval.
- The depths “meet in the middle,” where we find a number surrounded by parentheses. This interval is not given a depth, because it contains the median, or middle value. Rather, the parentheses contain the frequency for that interval.

```
1 | 2: represents 12
 leaf unit: 1
            n: 50
   1     9 | 0
   3    10 | 58
   7    11 | 0157
  10    12 | 378
  14    13 | 1489
  21    14 | 2233677
  (4)   15 | 1679
  25    16 | 23455556689
  14    17 | 001388
   8    18 | 67
   6    19 | 1455
   2    20 | 15
```

# The Box-and-Whisker Plot

- The *box-and-whisker plot* (often referred to simply as a *boxplot*) is a device for visually conveying several key aspects of a list of numbers.
- The boxplot offers less detail than a well designed histogram, but can be especially useful for comparing several distributions at once.

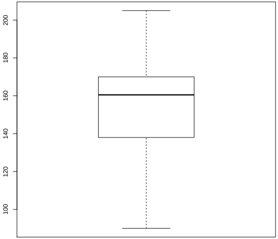
# The Box-and-Whisker Plot

- The fundamental ideas behind the boxplot are that
  1. The *box* displays the 25th, 50th, and 75th percentiles.
  2. The *whiskers* extend to the highest and lowest scores *that are not outliers*.
  3. Outliers are represented with circles and (in the case of extreme outliers) stars.
  4. The *H-Spread* is essentially equal to the interquartile range, i.e., $P_{75} - P_{25}$, or, the height of the box.
  5. An outlier is any observation falling more than 1.5 H-Spreads from either the top or bottom of the box.
  6. Some boxplots define an *extreme outlier* as an observation that is more than 3.0 H-Spreads from either the top or bottom of the box.
- There are numerous possible minor variations of the boxplot, depending on how one defines the sample quantiles that are used to construct the box and compute the H-Spread.
- Should one use the sample quantiles? What about gaps? What about unbiased estimation of the quantiles? Under what assumptions?
- It would take several lectures to cover all the possibilities in detail. We'll just go with the R default.

# The Box-and-Whisker Plot

Here is a basic boxplot of the *weights* data. We see that the $50^{th}$ percentile is around 162, the $75^{th}$ percentile around 170, the $25^{th}$ percentile around 140. The range extends from around 90 to 205.

```
> weights <- matrix(wts,ncol=1)
> boxplot(weights)
```

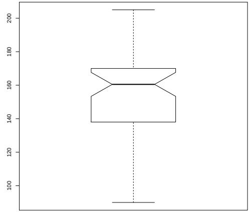


# The Box-and-Whisker Plot

## Adding a Notch

If the command `notch=TRUE` is added, a notch is drawn in each side of the boxes. If two boxplots are plotted for two different samples from two populations, and the notches of two plots do not overlap, this is 'strong evidence' that the two medians differ for the two populations. Here we simply demonstrate what a notched boxplot looks like.

```
> weights <- matrix(wts,ncol=1)
> boxplot(weights,notch=TRUE)
```

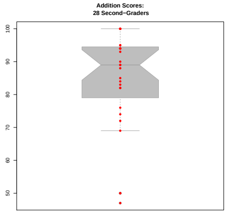


# Plots of Data Distributions

## The Boxplot – Some Enhancements

- Boxplots may be enhanced in numerous ways, by adding information and using color creatively.
- One option, useful if the sample size is not too large, is to add the actual data points to the plot.
- The only problem here is that there is overlap of the points. In the next slide, I show one way around that.
- See if you can figure out what I am doing!

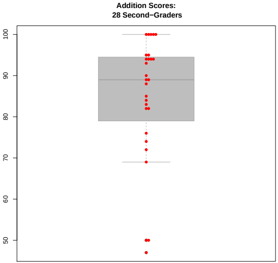
```
> boxplot(Royer, main="Addition Scores:\n 28 Second-Graders",
+         boxfill="grey",border="darkgrey",notch=T)
> points(x=rep(1,28),y=Royer,pch=16,col="red")
```



# Plots of Data Distributions

## The Boxplot – Some Enhancements

```
> boxplot(Royer, main="Addition Scores:\n 28 Second-Graders",boxfill="grey",border="darkgrey")
> points(x=c(1,1,1.01,1,1,1,1,1,1.01,
+ 1,1,1,1,1,1.01,1,1,1,
+ 1.01,1.02,1.03,1,1.01,1,1.01,1.02,
+ 1.03,1.04),y=sort(Royer),pch=16,col="red")
```



# Plots of Data Distributions

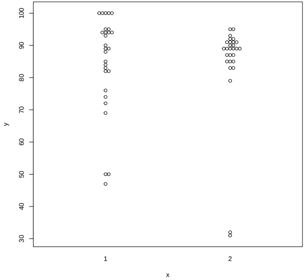
## The Beeswarm Chart

- A idea is elaborated in the *beeswarm chart*.
- Here, we stack the $Y$ and *Royer* data in one column $y$, use a grouping variable $x$, and construct a beeswarm chart using the `beeswarm` package.

# Plots of Data Distributions

## The Beeswarm Chart

```
> library(beeswarm)
> y <- rbind(Royer,Y)
> x <- rbind(rep(1,28),rep(2,28))
> beeswarm(y~x)
```

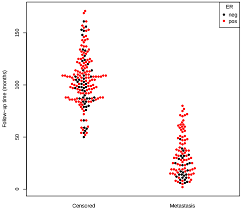


# Plots of Data Distributions

## The Beeswarm Chart

- Here is a more ambitious example:

```
> data(breast)
> breast2 <- breast[order(breast$event_survival, breast$ER),]
> beeswarm(time_survival ~ event_survival, data = breast2, pch = 16,
+          pwcol = as.numeric(ER), xlab = '',
+          ylab = 'Follow-up time (months)',
+          labels = c('Censored', 'Metastasis'))
>          legend('topright', legend = levels(breast$ER), title = 'ER',
+          pch = 16, col = 1:2)
```



# Plots of Data Distributions

## The Q-Q Plot

- The *Q-Q Plot* is a method for evaluating how closely the shape of a distribution adheres to a particular functional form.
- The quantiles of the observed data are plotted against the corresponding quantiles of a theoretical distribution.
- If the shapes of the distributions are the same, then the Q-Q plot should be a straight line.
- When a Q-Q plot is applied to sample data to assess distributional shape, the parameters of the target distribution may have to be estimated from the sample data.

# Plots of Data Distributions

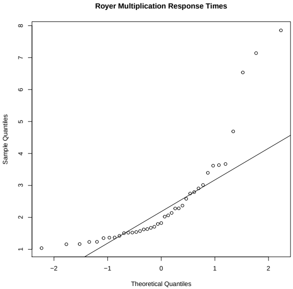
## The Q-Q Plot

- The *Q-Q Plot* is a method for evaluating how closely the shape of a distribution adheres to a particular functional form.
- The quantiles of the observed data are plotted against the corresponding quantiles of a theoretical distribution.
- If the shapes of the distributions are the same, then the Q-Q plot should be a straight line.
- In the QQ plot on the following slide, the reaction time data for males in grades 5–8 are compared to a normal distribution.
- Clearly, the reaction time data are non-normal. Are they positively or negatively skewed? Can you tell from the plot?

# Plots of Data Distributions

## The Q-Q Plot

```
> mult_time <- read.csv("http://www.statpower.net/Content/311/Lecture%20Notes/Royer%20rt_speed%20data.csv",header=T)
> attach(mult_time)
> m58 <- subset(mult_time, gender=="Male" & grade > 4)
> qqnorm(m58$M_RT,main = "Royer Multiplication Response Times")
> qqline(m58$M_RT)
```

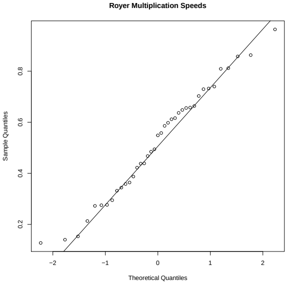


# Plots of Data Distributions

## The Q-Q Plot

- On the other hand, the reaction *speed* data appear to be very close to normally distributed.

```
> qqnorm(m58$M_Speed,main = "Royer Multiplication Speeds")
> qqline(m58$M_Speed)
```


Royer Multiplication Speeds
Sample Quantiles
Theoretical Quantiles