Exploring Data

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Exploring Data

1. Introduction
2. Plots of Data Distributions
   - The Histogram
   - The Boxplot
   - The Stem-and-Leaf Display
   - The Q-Q plot
3. Measures of Location and Spread
   - The Arithmetic Mean
   - Trimmed Mean
   - The Sample Variance and Standard Deviation
4. z-Scores and Their Properties
5. Summary Shape Statistics
   - Skewness
   - Kurtosis
6. Comparing Two Data Sets
   - Side-by-Side Boxplots
   - Displaying Means with Error Bars
   - Bar Graphs for Qualitative Independent Variables
   - Line Plots for Quantitative Independent Variables

Introduction

Descriptive statistics are justified in RDASA3 as stemming from three basic needs:

1. *To find aspects of the data worthy of inferential analysis.* What "catches the eye"? What is important?
2. *As a goal in itself*. Descriptive analysis can yield important findings. For example, when a principal analyzes test results within a school, descriptive analysis can assess things like (a) How the average student is performing, (b) Whether a non-trivial number of students is performing well below standard. As a *hypothesis generation device*. While sifting through data, you may discover important questions about your experimental phenomena that you would not have considered without the exploratory analysis.

Plots of Data Distributions

Graphical analysis allows us to get a quick look at "what's going on in the data."

MWL give a very brief description of 4 basic techniques:

1. The histogram.
2. The boxplot.
3. The stem-and-leaf display.
4. The Q-Q plot.

MWL illustrate the techniques with the `Royer` data set.

R scripts are provided by MWL for many of the textbook calculations. I've adapted them slightly for our lecture style.

Plots of Data Distributions

Throughout the course, we shall, on occasion, use packages recommended by MWL in the support materials for RDASA3. A current installation file *Function Install.r* for those packages is in the R Support Materials directory on the course website.

Periodically, the lab machines may be reset by Vanderbilt ITS, in which case you may need to reinstall any package that will not load. You can do this with the `library()` command.

Plots of Data Distributions

The Histogram

The histogram is a familiar member of the standard group of frequency distribution displays.

We shall demonstrate the histogram with the *RoyerY* data set described on page 21 of RDASA3.

The data are percent correct addition scores for 28 male second-grade students.

Plots of Data Distributions

The Histogram

The histogram is a familiar member of the standard group of frequency distribution displays.

We shall demonstrate the histogram with the *RoyerY* data set described on page 21 of RDASA3.

The data are percent correct addition scores for 28 male second-grade students.

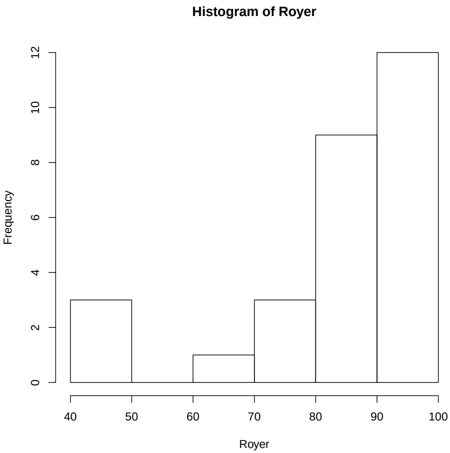
```
> RoyerY <- read.csv("Royer_Y.csv", header = T)
> attach(RoyerY)
```

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)
```

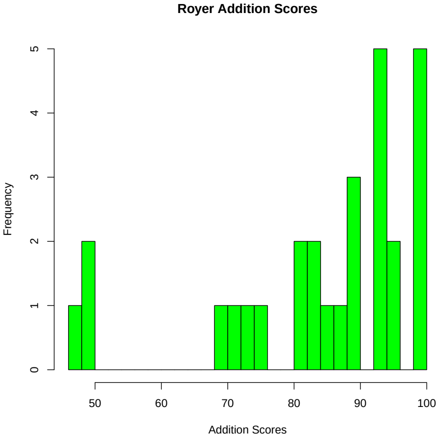


Plots of Data Distributions

The Histogram

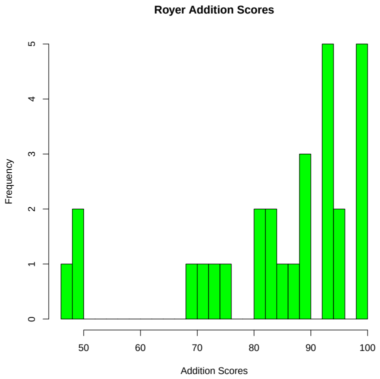
We can do better by increasing the number of breaks and adding a correct axis label.

```
> 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.

Plots of Data Distributions

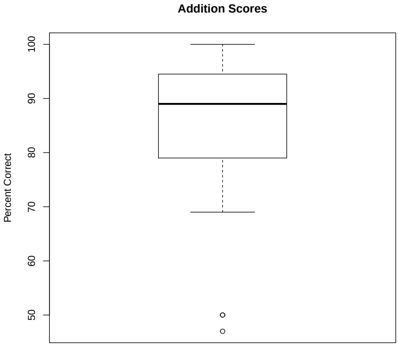
The Boxplot

The boxplot provides a more compact view of summary aspects of a data set.

It offers less detail than the histogram, but can be especially useful for comparing several data sets at once.

Here is a boxplot of the Royer data.

```
> boxplot(Royer, main = "Addition Scores", ylab = "Percent Correct")
```



Plots of Data Distributions

The Boxplot

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.

Plots of Data Distributions

The Boxplot

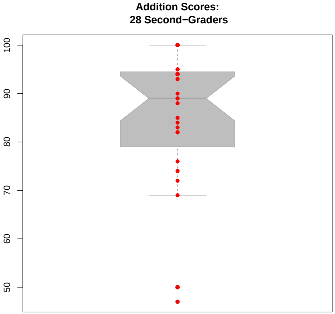
Boxplots may be enhanced in numerous ways, by adding information and using color creatively.

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!

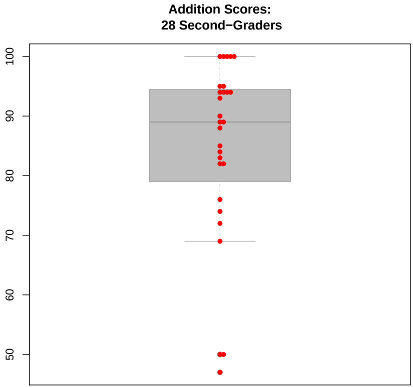
```
> boxplot(Royer, main = "Addition Scores:\n 28 Second-Graders",
+     boxfill = "grey", border = "darkgrey", notch = T)
> points(x = rep(1, 28), y = Royer, pch = 16, col = "red")
```



Plots of Data Distributions

The Boxplot

```
> 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 Boxplot

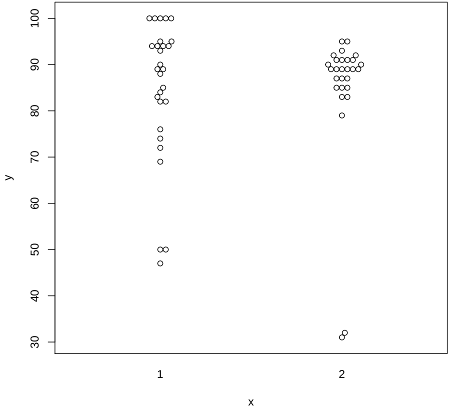
A idea is elaborated in the *beeswarm chart*.

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 Boxplot

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

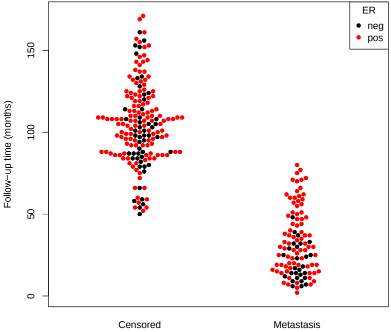


Plots of Data Distributions

The Boxplot

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 Stem-and-Leaf Display

As mentioned in Psychology 310, the stem-leaf display was very popular in the late 1970's.

This method has the advantage of displaying more numerical information than a histogram, while still displaying the shape of the distribution.

```
> stem(Royer)


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

   4 | 7
   5 | 00
   6 | 9
   7 | 246
   8 | 22345899
   9 | 03444455
  10 | 00000
```

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.

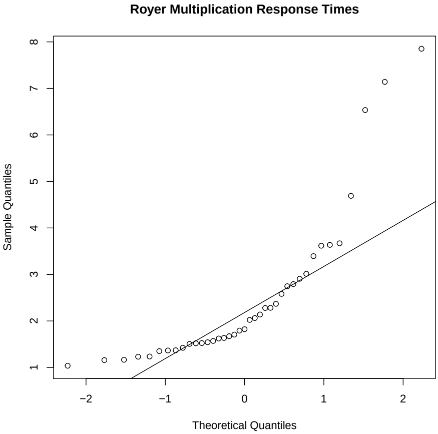
In the QQ plot on the following slide, the reaction time data for males in grades 5–8 are compared to a normal distribution.

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("Royer rt_speed data.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)
```

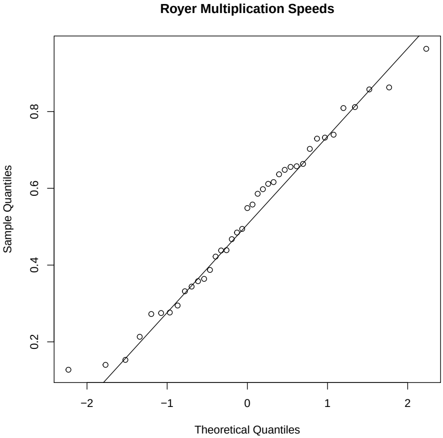


Plots of Data Distributions

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)
```



Measures of Location and Spread

The Arithmetic Mean

The sample mean of a set of numbers X_i is defined as $\overline{X}_\bullet = (1/n)\sum_{i=1}^n X_i$. It has a number of important properties that concern its behavior under *linear transformations* and linear combinations of the data. These properties were dealt with extensively in Psychology 310. If you have not taken Psychology 310, you should immediately review the lecture notes and handouts on linear transformation and linear combination theory.

Measures of Location and Spread

The Arithmetic Mean

Given a set of n scores X_i, a *listwise linear transform* is any transformation of all the numbers that can be written in the form $Y_i = aX_i + b$, where a and b are constants. Then

1. If $Y_i = aX_i + b$, then the mean of the transformed scores follows the relationship
$$\overline{Y}_\bullet = a\overline{X}_\bullet + b \tag{1}$$
2. This implies that *listwise addition, subtraction, multiplication and/or division come straight through in the mean.*

Measures of Location and Spread

Trimmed Mean

One way of making the mean more resistant to outliers is simply to trim a percentage of the outlying cases from the sample.

A simple version of this idea is the *trimmed mean*.

The scores are ordered, and a percentage of cases is trimmed from the upper and lower end of the distribution.

The trimmed mean is calculated on the remaining scores.

Measures of Location and Spread

The Sample Variance and Standard Deviation

The sample variance of a set of numbers X_i is defined as $s^2 = (1/(n-1))\sum(X_i - \overline{X}_\bullet)^2$.

The sample standard deviation is the square root of the sample variance, i.e., $s = \sqrt{s^2}$.

The properties of the sample variance and standard deviation were discussed in detail in Psychology 310.

Measures of Location and Spread

The Sample Variance and Standard Deviation

If $Y_i = aX_i + b$, then the standard of the transformed scores follows the relationship

$$s_Y = |a|s_X \tag{2}$$

This implies that *listwise addition and subtraction have no effect on the standard deviation, while multiplication and/or division by a positive constant come straight through in the standard deviation*.

Likewise

$$s_Y^2 = a^2 s_X^2 \tag{3}$$

z-Scores and Their Properties

We transform a set of scores into *z*-score form by the formula

$$z_i = \frac{X_i - \overline{X}_\bullet}{s_X} \tag{4}$$

In Psychology 310, we prove, using the laws of linear transformation, that a set of *z*-scores must have a mean of 0, a standard deviation of 1, and the same shape as the original numbers.

RDASA3 provides an algebraic proof on page 36.

Summary Shape Statistics

Summary measures of shape of a distribution are of interest, often in the context of assessing adherence of the data to the assumption of normality.

If the data have high kurtosis due to long tails, then the sample mean may be an unreliable estimator of the population mean.

Moreover, standard formulas for the variance of the sample correlation and sample variance can be seriously in error if the data come from a population with high kurtosis.

Summary Shape Statistics

Skewness

Skewness statistics are designed to assess departures from symmetry.

The standard definition of skewness in a population is the *average cubed z-score*

Summary Shape Statistics

Kurtosis

The basic idea of kurtosis is that it is the *average 4th power of the z-scores*.

However, the kurtosis of a distribution is typically expressed relative to that of a normal distribution.

The normal distribution has a kurtosis of 3. So, the most common measure of kurtosis is the average 4th power of the z-scores minus 3.

This can be computed in a number of ways. The simplest version is biased for a normal distribution, but is reported by some programs.

SPSS reports a more complex estimate that is unbiased *if the population is normal*.

Summary Shape Statistics

Kurtosis

Here is code for the simple statistic.

```
> # Standard deviation with n in the denominator
> S <- function(x) {
+     n <- length(x)
+     return(sqrt((n - 1)/n * var(x)))
+ }
> # z.scores using revised S
> z.score <- function(x) {
+     (x - mean(x))/S(x)
+ }
> # Power sum
> power.sum <- function(x, power) {
+     sum(z.score(x)^power)
+ }
> # Kurtosis relative to normal
> simple.kurtosis <- function(x) {
+     power.sum(x, 4)/length(x) - 3
+ }
```

Summary Shape Statistics

Kurtosis

```
> ## Compare to kurtosis reported by routine from the
> ## 'moments' library
> library(moments)
> kurtosis(Royer)

[1] 3.727

> ## another way to do it
> s4 <- sum((Royer - mean(Royer))^4)
> s2 <- sum((Royer - mean(Royer))^2)
> n <- length(Royer)
> n * s4/s2^2

[1] 3.727
```

Summary Shape Statistics

Kurtosis

Here is code for the SPSS statistic

```
> SPSS.kurtosis <- function(x) {
+     s4 <- sum((x - mean(x))^4)
+     s2 <- sum((x - mean(x))^2)
+     n <- length(x)
+     return(n * (n + 1) * s4/((n - 1) * (n - 2) * (n - 3))/var(x)^2 -
+         3 * ((n - 1)^2)/((n - 2) * (n - 3)))
+ }
> SPSS.kurtosis(Royer)

[1] 1.124
```

Comparing Two Data Sets

Side-by-Side Boxplots

The RoyerY data set includes a second variable, Y, designed to have many summary statistics in common with the Royer variable.

However, it has substantially larger skew and kurtosis, as shown in Table 2.5 of RDASA3.

```
> summary(RoyerY)

     Royer              Y        
 Min.   : 47.0   Min.   :31.0  
 1st Qu.: 80.5   1st Qu.:85.0  
 Median : 89.0   Median :89.0  
 Mean   : 84.6   Mean   :84.6  
 3rd Qu.: 94.2   3rd Qu.:91.0  
 Max.   :100.0   Max.   :95.0  
```

Comparing Two Data Sets

Side-by-Side Boxplots

```
> mean(Royer)

[1] 84.61

> mean(Y)

[1] 84.61

> sd(Royer)

[1] 15.3

> sd(Y)

[1] 15.43
```

Comparing Two Data Sets

Side-by-Side Boxplots

```
> skewness(Royer)

[1] -1.254

> skewness(Y)

[1] -3.021

> SPSS.kurtosis(Royer)

[1] 1.124

> SPSS.kurtosis(Y)

[1] 9.664
```

Comparing Two Data Sets

Side-by-Side Boxplots

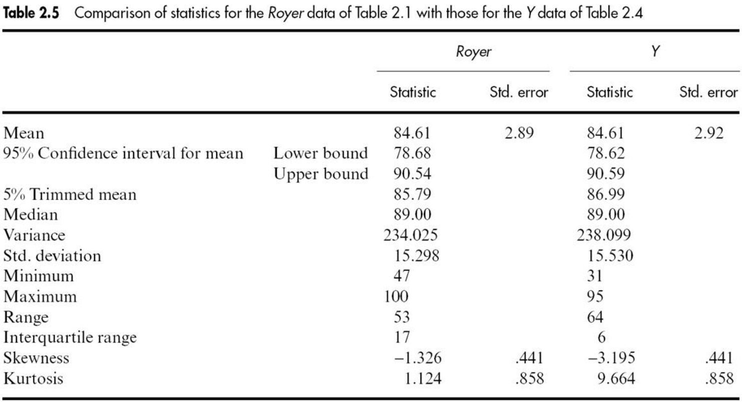


Table 2.5 Comparison of statistics for the *Royer* data of Table 2.1 with those for the *Y* data of Table 2.4

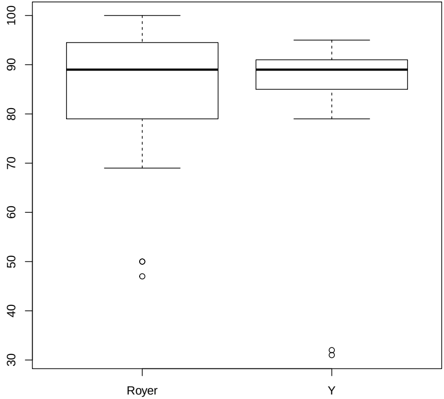
		Royer		*Y*	
		Statistic	Std. error	Statistic	Std. error
Mean		84.61	2.89	84.61	2.92
95% Confidence interval for mean	Lower bound	78.68		78.62	
	Upper bound	90.54		90.59	
5% Trimmed mean		85.79		86.99	
Median		89.00		89.00	
Variance		234.025		238.099	
Std. deviation		15.298		15.530	
Minimum		47		31	
Maximum		100		95	
Range		53		64	
Interquartile range		17		6	
Skewness		−1.326	.441	−3.195	.441
Kurtosis		1.124	.858	9.664	.858

Comparing Two Data Sets

Side-by-Side Boxplots

The side-by-side boxplot makes the situation pretty obvious.

```
> boxplot(RoyerY)
```



Comparing Two Data Sets

Displaying Means with Error Bars

When we measure a parameter with a statistic, there is almost inevitably a sampling error.

The *standard error* of a statistic is the standard deviation of the sampling distribution of that statistic over repeated samples.

With n independent observations, the standard error of the sample mean $\overline{X}_{\bullet}$ is $\sigma/\sqrt{n}$, where σ is the standard deviation in the population of scores.

MWL define SEM as $s/\sqrt{n}$. This is, of course, an *estimate* of the standard error of the mean, not the actual standard error. (There is a long history of applied statisticians blurring this distinction.)

If we knew σ, and had normally distributed populations, we could construct a 95% confidence interval for the sample mean with error bars roughly 2 standard errors in length. A 68% confidence interval would have error bars roughly one standard error in length.

We can, in more typical circumstances, still construct approximate confidence intervals based on estimated standard errors and the t-distribution.

Comparing Two Data Sets

Bar Graphs for Qualitative Independent Variables

Bar graphs can be used to
compactly present data on means of several groups on two or more variables.

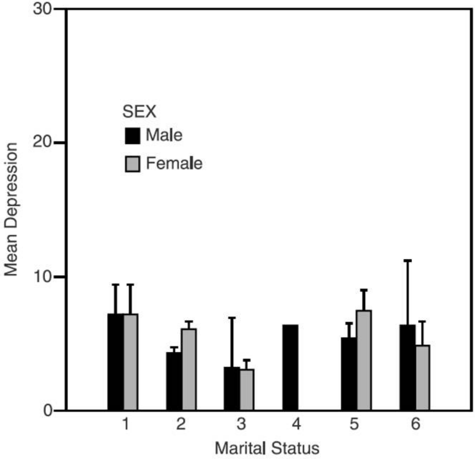


Fig. 2.6 Bar graph of mean depression scores as a function of sex and marital status.

Comparing Two Data Sets

Line Plots for Quantitative Independent Variables

Line plots can be used to display means with error bars.

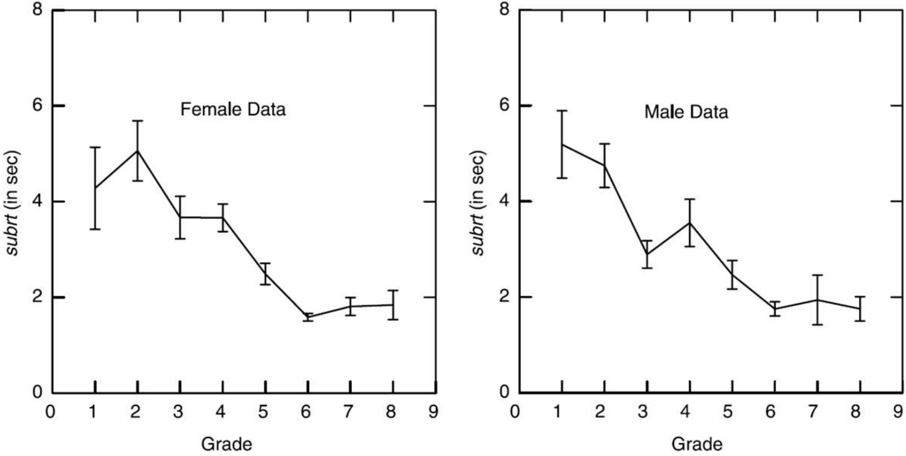


Fig. 2.7 Line graph of subtraction response times (*subrt*) as a function of grade.