

# 3 Numerical Descriptive Measures

## 3.1 Measures of Central Tendency

Summarize the data by taking an average.

**Example:** Suppose that a student has the following 6 quiz grades: 95, 100, 85, 89, 10, 97.

In general, the *sample mean* (arithmetic mean) of  $n$  observations on  $X$  is denoted

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}.$$

The **sample median** is the middle value when the measurements are arranged from smallest to largest. For an even number of data, the median is the average of the two middle values.

When outliers exist or at least one tail of the distribution is heavy, then the **sample median** typically is preferred over the **sample mean**, as a measurement of center.

Otherwise, the **sample mean** typically has less variability and is preferred over the **sample median**.

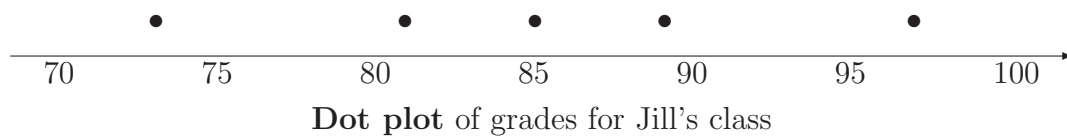
## 3.2 Variation and Shape

The **sample range** is the maximum value minus the minimum value; this measure is reasonable for small data sets, but not large ones.

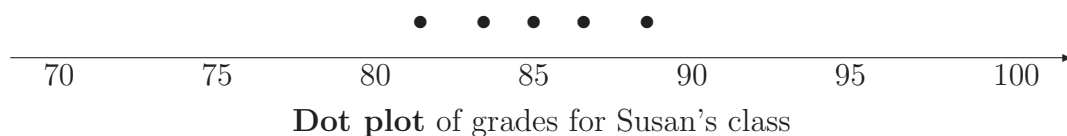
**Example:** Suppose a sample of 100 incomes among employed Virginians might have the smallest value of \$10,000 and the largest value of \$300,000.

Consider examples with light-tailed distributions and no outliers in the data set.

**Example:** Grades by instructor **Jill** for 5 students on a chemistry exam are {81, 85, 97, 73, 89}.



**Example:** Grades by instructor **Susan** for 5 students on a chemistry exam are {84, 88, 85, 86, 82}.



Which instructor do you prefer?

$$\text{sample variance} = s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$
$$\text{sample standard deviation} = s = \sqrt{s^2}$$

**Example:** Jill's class

**Example:** Susan's class

**Remark:** Adding a constant to the data does not affect  $s$ . For example, adding 10 points to everyone's grade increases  $\bar{X}$  and the sample median by 10 points but does not affect  $s$ , the spread.

**Remark:** Multiplying data by a constant  $c$  changes  $s$  by a factor of  $|c|$ .

**Example:** At a company the salaries are \$50,000, \$55,000, and \$60,000.

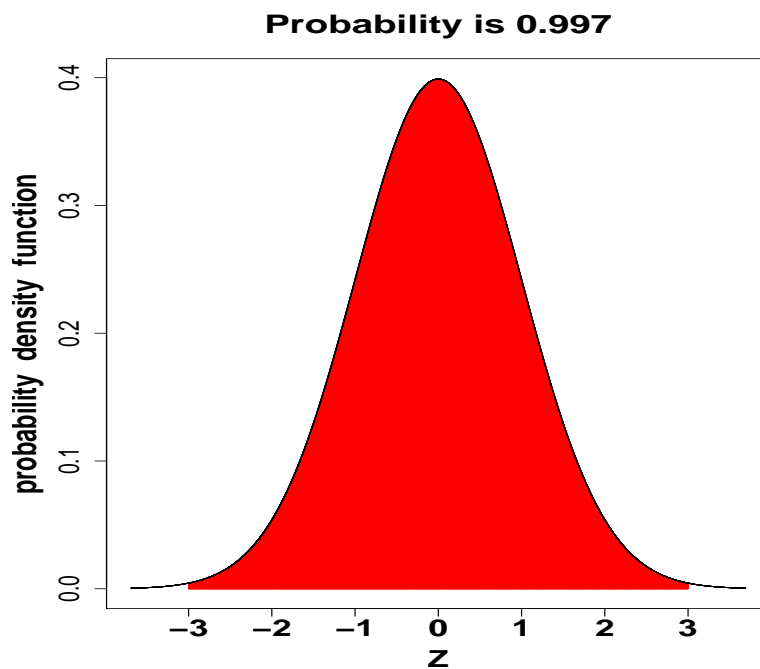
**Example:** Consider the data  $\{8, 8, 8, 8, 8, 8\}$ .

## Identifying potential outliers using the z-score

**criterion:** An observation is an **outlier** if its **z-score** is outside the interval  $-3$  to  $3$ , where the z-score is

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}.$$

Equivalently, an observation is an **outlier** if it is outside the interval  $(\text{mean} \pm 3 \text{ SD})$ .



**Example:** Return to the grades in Jill's class:  $\{81, 85, 97, 73, 89\}$ .

□

## The SHAPE of a distribution

A histogram might be described as

1. unimodal
2. bimodal
3. multimodal

A **sample or population histogram** might be described as

1. symmetric
2. skewed to the right
3. skewed to the left

## 3.3 Numerical Descriptive Measures for a Population

The **population mean** is

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

the **population variance** is

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

and the **population standard deviation** is

$$\sigma = \sqrt{\sigma^2}.$$

$\bar{X}$  “converges” to  $\mu$  as  $n$  gets large.

Likewise,  $s^2$  “converges” to the population variance,  $\sigma^2$ , as  $n$  gets large.

Similarly,  $s$  “converges” to the population standard deviation,  $\sigma$ , as  $n$  gets large.

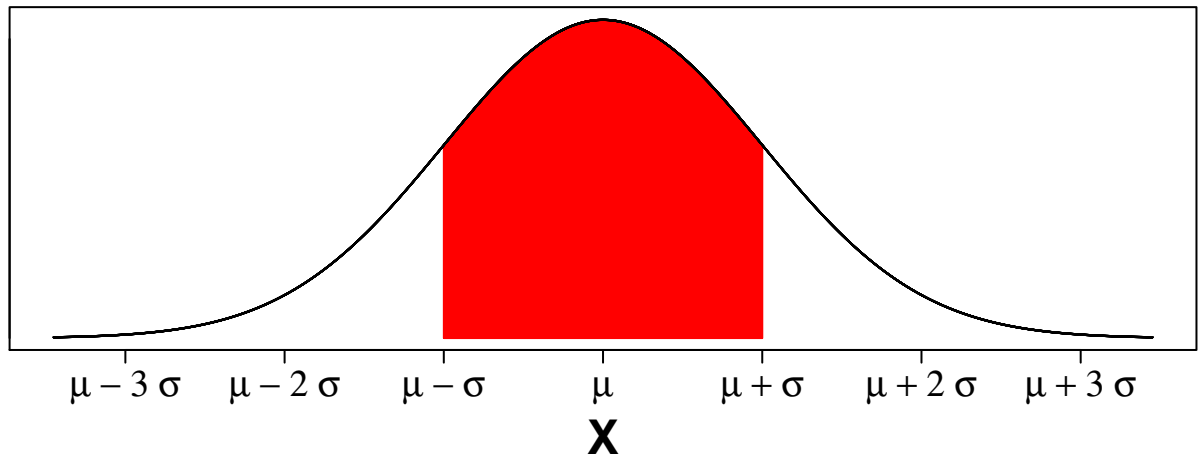
## Empirical Rule

If a large number of observations are sampled from an approximately **normal** distribution, then (usually)

1. Approximately 68% of the observations fall within **one** standard deviation,  $\sigma$ , of the mean,  $\mu$ .
2. Approximately 95% of the observations fall within **two** standard deviations,  $\sigma$ , of the mean,  $\mu$ .
3. Approximately 99.7% of the observations fall within **three** standard deviations,  $\sigma$ , of the mean,  $\mu$ .

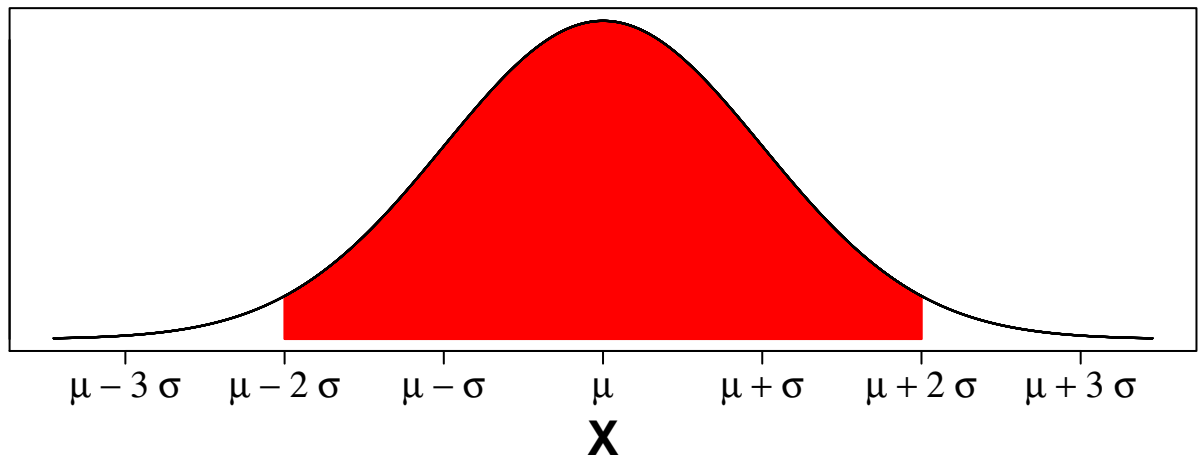
probability density function

**Probability is 0.68**



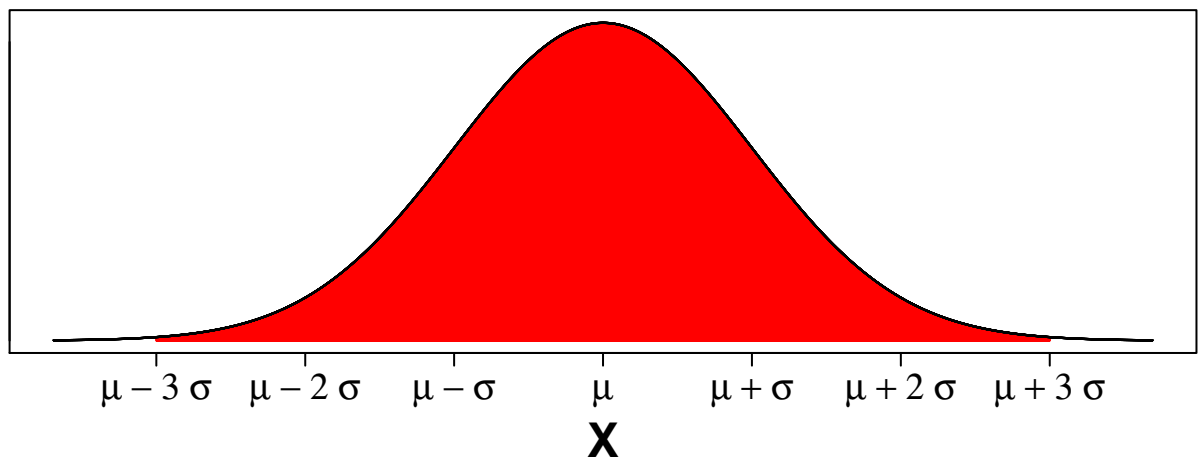
probability density function

**Probability is 0.95**



probability density function

**Probability is 0.997**



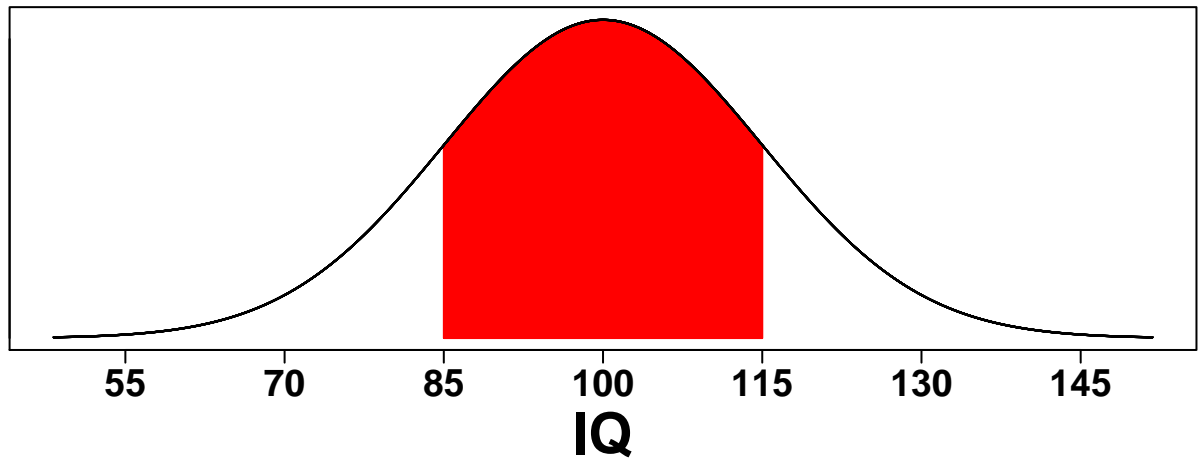
**Example:** IQ scores of normal adults on the Weschler test have a symmetric bell-shaped distribution with a mean of 100 and standard deviation of 15.

- (a) If 1000 adults are sampled, approximately how many have IQs between 85 and 115?
  
- (b) If 1000 adults are sampled, approximately how many have IQs between 70 and 130?
  
- (c) If 1000 adults are sampled, approximately how many have IQs between 55 and 145?
  
- (d) If 1000 adults are sampled, approximately how many have IQs greater than 130?



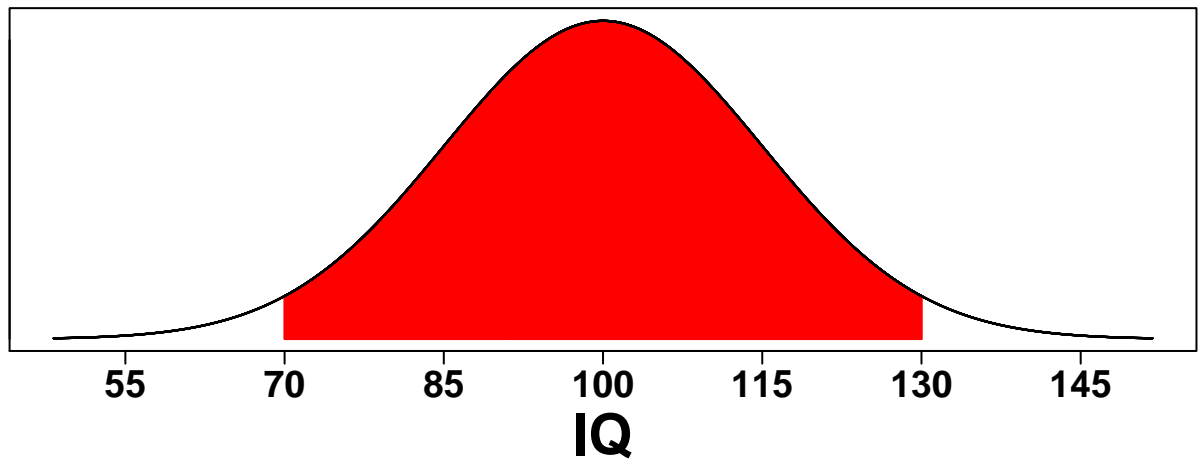
probability density function

**Probability is 0.68**



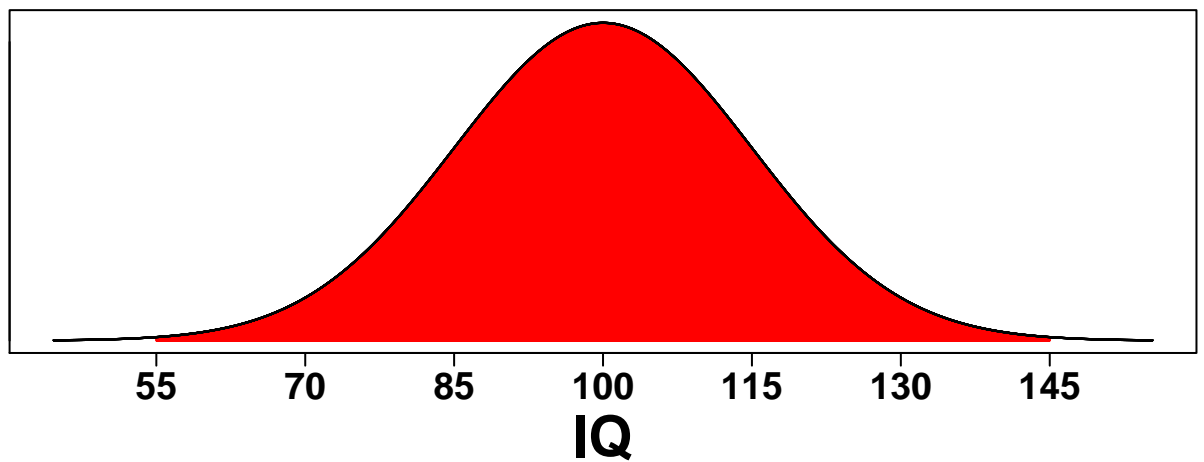
probability density function

**Probability is 0.95**



probability density function

**Probability is 0.997**



## 3.5 Quartiles and the Boxplot

### Interquartile Range

To determine the **interquartile range**, first determine the **first quartile** ( $Q_1$ ) and the **third quartile** ( $Q_3$ ).

$Q_1 = (n + 1)/4$  ranked value, and  $Q_3 = 3(n + 1)/4$  ranked value.

The **interquartile range** or **midsread** is **third quartile** minus **first quartile**.

The **5-number summary** is {minimum, first quartile, median, third quartile, maximum}.

The **5-number summary** divides the data into four (roughly) equal sections (fourths).

**Example:** Consider the following 13 observations:

{45, 48, 53, 103, 160, 10, 63, 68, 70, 55, 58, 75, 77}.

How do the above results change if we replace 160 by 1,000,000?

□

### Identifying potential outliers using the IQR

**criterion:** An observation is an **outlier** if it is at least 1.5 IQR from its nearer quartile.

### The Box Plot (allowing for outliers)

Procedure:

1. Draw rectangle with edges at lower and upper quartiles.
2. Draw a line through the box at the sample median.
3. Draw **astericks** to represent **outliers**.
4. Draw **whiskers**; i.e., lines from the edge of the box to the most extreme observation which is not an outlier.

**Previous example:** Consider the following 13 observations:

{45, 48, 53, 103, 160, 10, 63, 68, 70, 55, 58, 75, 77}.

Check for outliers on the **left**.

Check for outliers on the **right**.

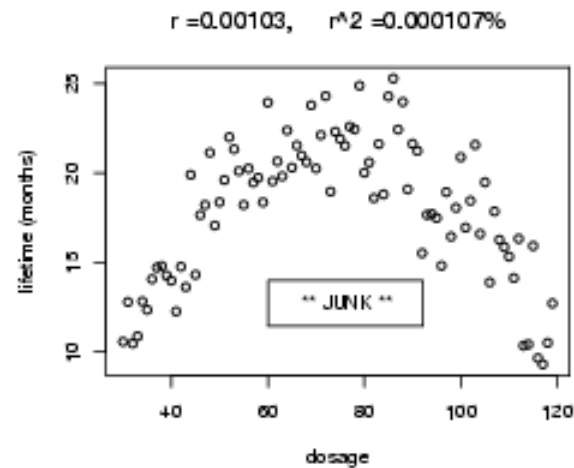
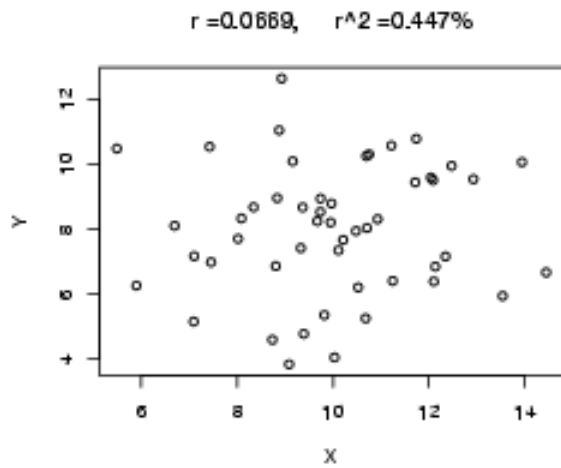
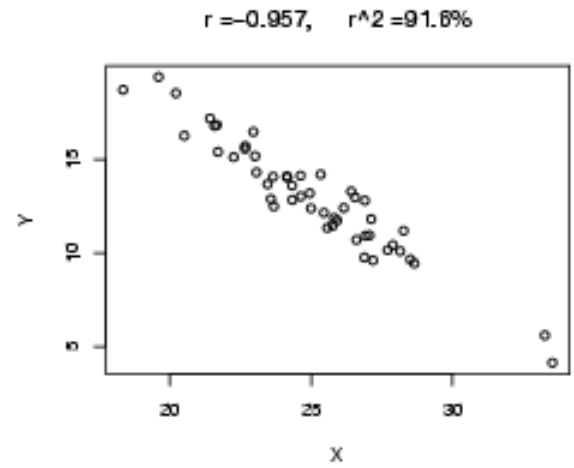
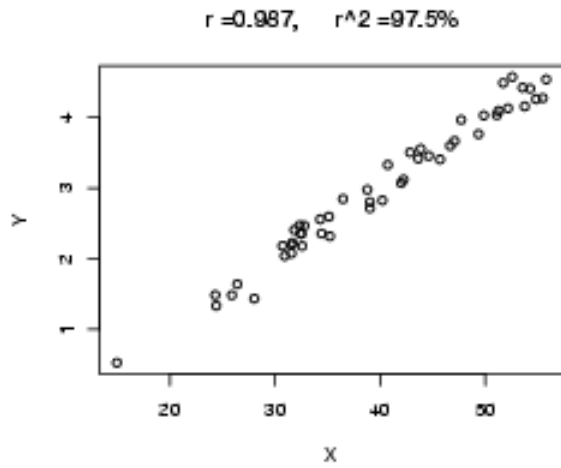
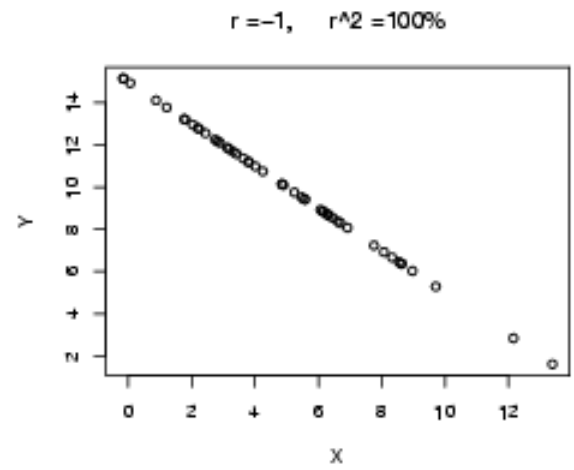
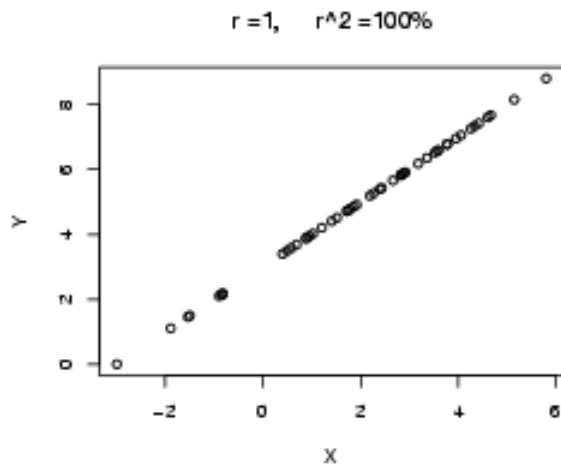
## Which measures of center and spread should one use?

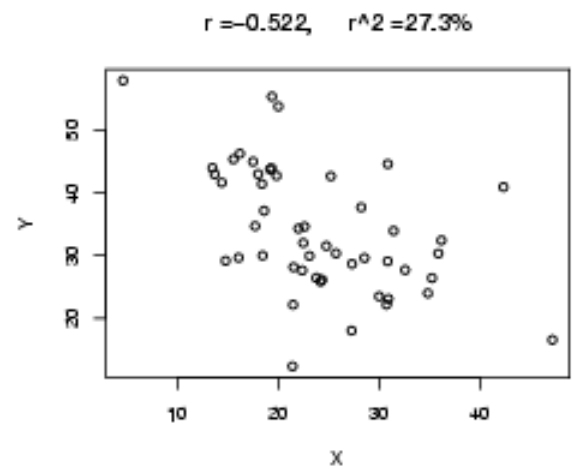
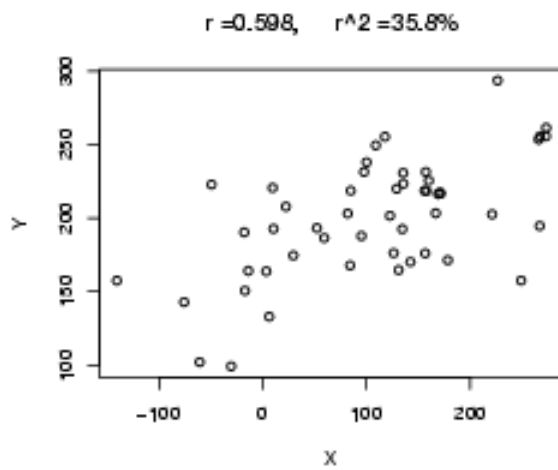
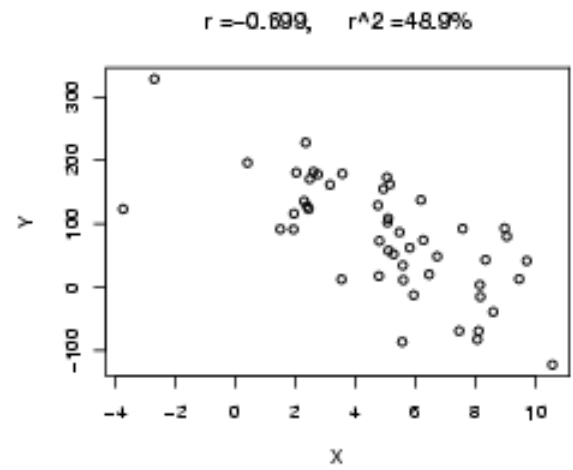
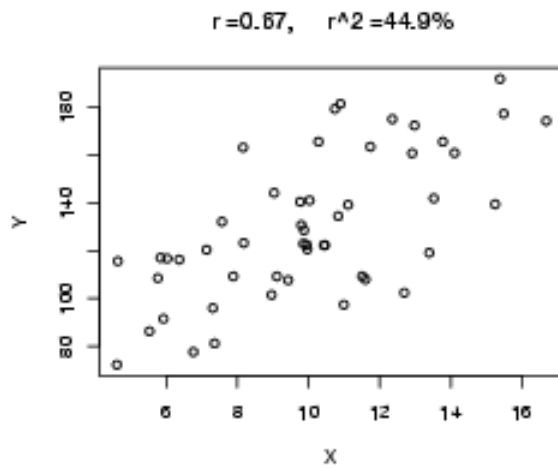
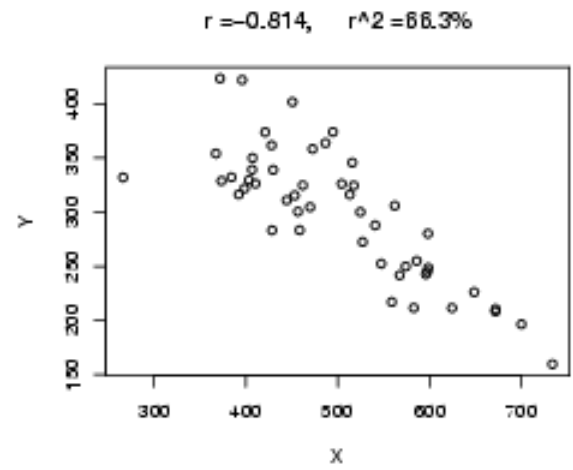
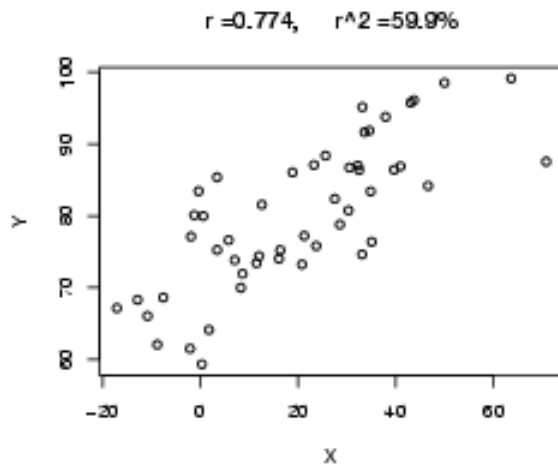
Suppose a data set has outliers (or the distribution has at least one heavy tail).

Suppose a data set seems to be from a distribution which is **normal** (or approximately normal).

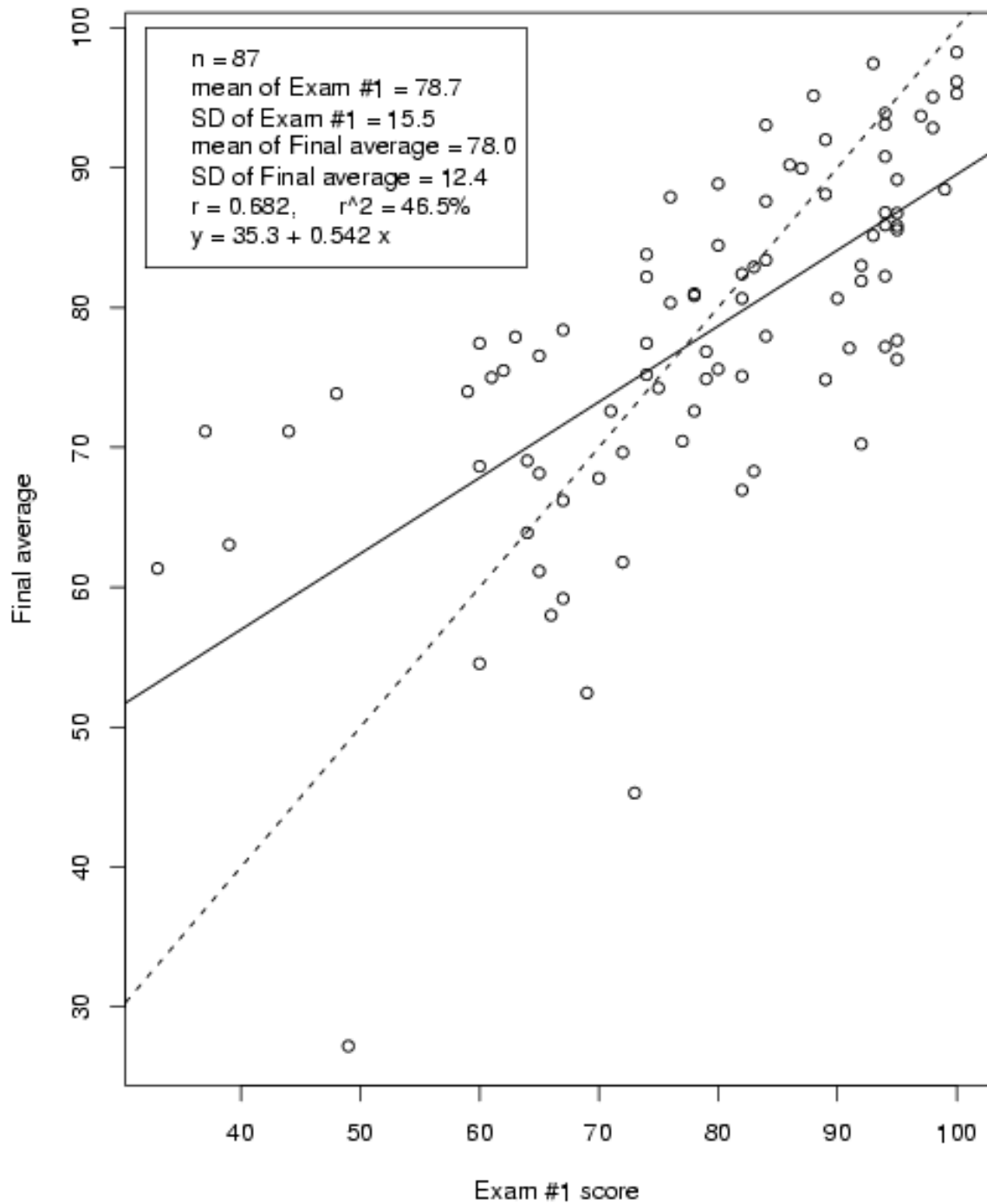
## 3.6 The Covariance and the Coefficient of Correlation

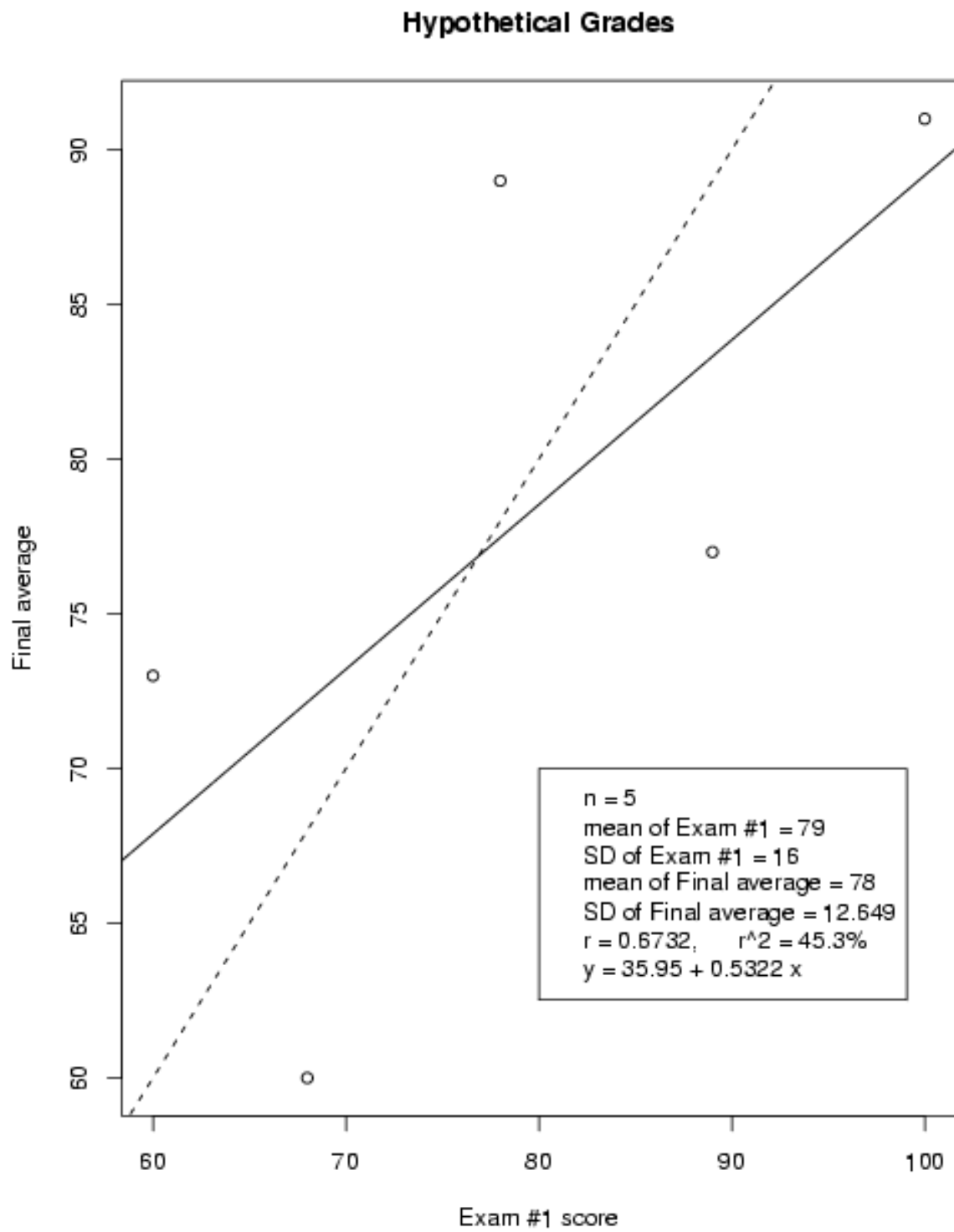
**Correlation** is a numerical measure of the **linear** association between two numerical variables.





## Real Grades







Calculation of **(Pearson's) correlation,  $r$** , for  $n$  pairs of data  $(x, y)$ .

$$r = \frac{\frac{1}{n-1} \sum (x - \bar{x})(y - \bar{y})}{s_x s_y}.$$

The textbook gives the formula

$$r = \frac{\sum z_x z_y}{n - 1},$$

where  $z_x = (x - \bar{x})/s_x$  and  $z_y = (y - \bar{y})/s_y$ .

**Example:** Determine the correlation for the following data:

| Exam #1 score | Final score |
|---------------|-------------|
| $x$           | $y$         |
| 68            | 60          |
| 100           | 91          |
| 89            | 77          |
| 78            | 89          |
| 60            | 73          |

**Remarks:**

- (a) Is  $r$  random or fixed?
- (b) What are the units on  $r$ ?
- (c) What are the possible values of  $r$ ?
- (d)  $r = 1$  implies what type of correlation?
- (e)  $r = -1$  implies what type of correlation?
- (f) Is selection of  $x$  and  $y$  relevant when calculating  $r$ ?

- (g)  $r$  makes sense for linear associations only.
- (h) A linear transformation on the data does not affect  $|r|$ .
- (i) As the number of  $(x, y)$  data pairs becomes huge,  $r$  “converges” to the **population** correlation.

## Linear Regression

Examining the relationship between variables is called **regression analysis**.

Examining the **linear** relationship between **two** variables is called **simple linear regression**.

Two purposes of regression analysis:

1. explain
2. predict

Typically,

$x$  is the **explanatory** variable.

$y$  is the **response** variable.

Goal is to fit a reasonable line through the scatter plot.

The unique line which minimizes the sum of squares of the vertical distances is called the **least squares line** or **fitted regression line**.

The equation of the **least squares** line can be written

$$\hat{y} = a + b x.$$

The **slope** of the least squares line can be shown to be

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{r s_y}{s_x}$$

The **intercept** of the least squares line may be computed by noting that the least squares line goes through the point  $(\bar{x}, \bar{y})$ .

**Example:** (FIVE PAIRS OF GRADES) Return to the data for the grades of the five hypothetical students of (exam #1 score, final score): (68, 60), (100, 91), (89, 77), (78, 89), and (60, 73). Fit the regression line.

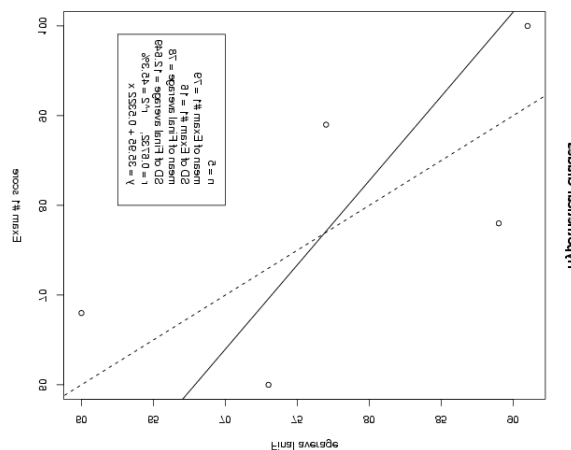
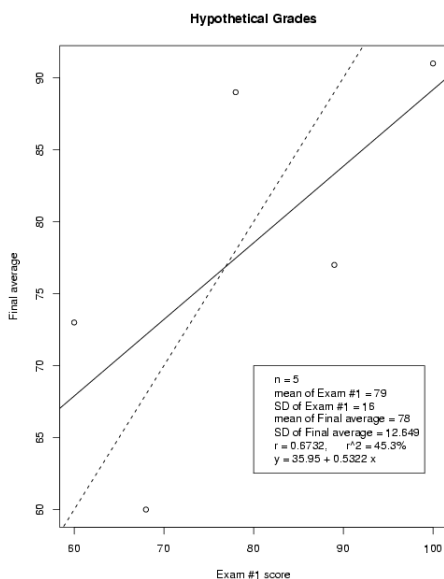
Predict a new value of  $y$  if  $x$  is 85.

Estimate the mean of  $y$  if  $x$  is 85.

Predict a new value of  $y$  if  $x$  is 40.

□

**Remark:** The least squares line of  $y$  on  $x$  differs from the least squares line of  $x$  on  $y$ .



## Assessing the Fit of a Line: r-Squared

**Definition:** The **proportional reduction in error**, denoted by  $r^2$ , gives the proportion of variation in  $y$  which can be explained by  $x$ , when the data are linear.

**Example:** (FIVE PAIRS OF GRADES) Return to the data for the grades of the five hypothetical students of (exam #1 score, final score): (68, 60), (100, 91), (89, 77), (78, 89), and (60, 73). Determine the *proportional reduction in error*.

**Example:** Under the *lofty* assumption that the final score is based upon 5 equally weighted INDEPENDENT exams with a common variance, then  $r^2$  (at least for the entire data set of 87 students) should be about what number?

□

What are the possible values of  $r^2$ ?

## Cautions in Analyzing Association

Extrapolation is dangerous.

Recall: Correlation does not imply causation.

**Example:** Consider the two variables “weight of **older** brother at age 5” and “weight of **younger** brother at age 5.”

A **lurking variable** is a third variable which confuses the relationship between the two variables of interest.

In general, association does not imply causation.

**Example:** Suppose in a large survey on alcohol consumption and lung cancer, it is determined that people who consume *a lot of alcohol* have a significantly *higher* rate of *lung cancer* than people who consume *little or no alcohol*.

Is it reasonable to conclude that heavy alcohol consumption **causes** lung cancer?

What might be a *lurking variable*?

□

**Read pp. 138–139, Appendix E3: Using Microsoft Excel for Descriptive Statistics.**