

MATH 3339

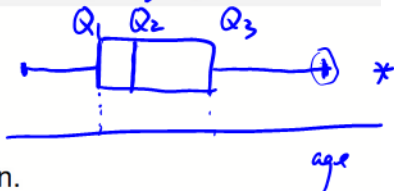
Statistics for the Sciences

sec 2.3; 2.6

Wendy Wang, Ph.D.
wwang60@central.uh.edu

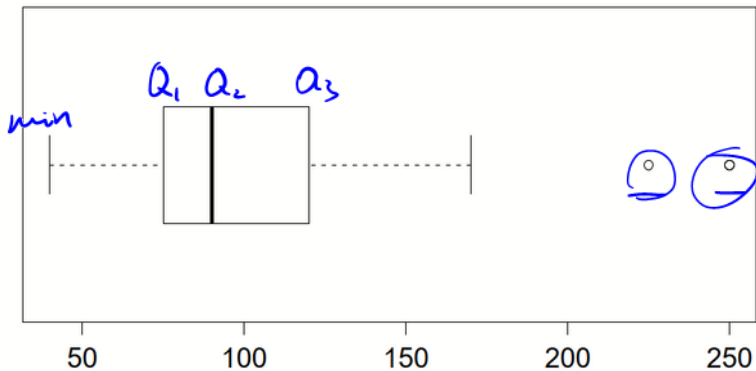
Lecture 4 - 3339

A Graph of the Five Number Summary: Boxplot



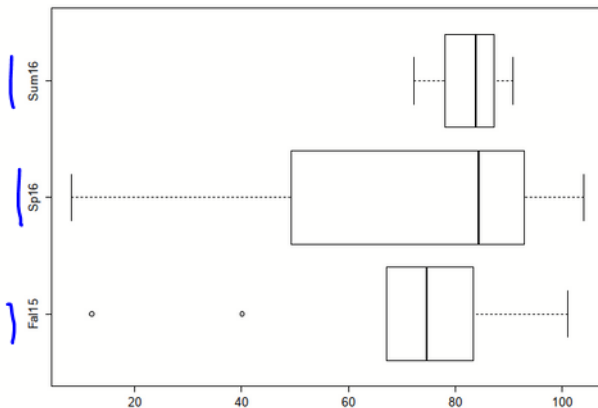
- A central box spans the quartiles.
- A line inside the box marks the median.
- Lines extend from the box out to the smallest and largest observations.
- Asterisks represents any values that are considered to be outliers.
- Boxplots are most useful for side-by-side comparison of several distributions.
- Rcode: `boxplot(dataset name$variable name)`

Boxplot of Prices



```
boxplot(shoes$Price, horizontal = T)
```

Boxplot of Course Scores by Session



```
boxplot (grades$Score~grades$Session,horizontal=TRUE)
```

Outline

- 1 Understanding Standard Deviation
- 2 Calculating The Standard Deviation
- 3 Measures of Variability
- 4 Jointly Distributed Data
- 5 Scatterplots for Jointly Distributed Variables
- 6 Covariance and Correlation

Measuring Spread: The Standard Deviation

- Measures spread by looking at how far the observations are from their mean.
- Most common numerical description for the spread of a distribution.
- A larger standard deviation implies that the values have a wider spread from the mean.
- Denoted s when used with a sample. This is the one we calculate from a list of values.
- Denoted σ when used with a population. This is the "idealized" standard deviation.
- The standard deviation has the same units of measurements as the original observations.

Definition of the Standard Deviation

The standard deviation is the average distance each observation is from the mean.

- Using this list of values from a sample: 3, 3, 3, 15, 15, 15
- The mean is 9.
- By definition, the average distance each of these values are from the mean is 6. So the standard deviation is 6.

Values of the Standard Deviation

- The standard deviation is a value that is greater than or equal to zero.
- It is equal to zero only when all of the observations have the same value.
- By the definition of standard deviation determine s for the following list of values.
 - ▶ 2, 2, 2, 2 : standard deviation = 0
 - ▶ 125, 125, 125, 125, 125: standard deviation = 0

Adding or Subtracting a Value to the Observations

- Adding or subtracting the same value to all the original observations does not change the standard deviation of the list.
- Using this list of values: 3, 3, 3, 15, 15, 15 mean = 9, standard deviation = 6.
- If we add 4 to all the values: 7, 7, 7, 19, 19, 19
mean = 13, standard deviation = 6

Multiplying or Dividing a Value to the Observations

- Multiplying or dividing the same value to all the original observations will change the standard deviation by that factor.
- Using this list of values: 3, 3, 3, 15, 15, 15: mean = 9, standard deviation = 6.
- If we double all the values: 6, 6, 6, 30, 30, 30

mean = 18, standard deviation = 12 $= 2 * 6$

Population Variance and Standard Deviation

① calculate Variance

② s.d. dev = $\sqrt{\text{Variance}}$

If N is the number of values in a population with mean μ , and x_i represents each individual in the population, then the population variance is found by:

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

and the **population standard deviation** is the square root, $\sigma = \sqrt{\sigma^2}$.

Sample Variance and Standard Deviation

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad \leftarrow \text{biased sample variance}$$

Most of the time we are working with a sample instead of a population. So the **sample variance** is found by:

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

and the **sample standard deviation** is the square root, $s = \sqrt{s^2}$. Where n is the number of observations (samples), x_i is the value for the i^{th} observation and \bar{x} is the sample mean.

n : sample size

N : population size

Calculating the Standard Deviation By Hand

When calculating by hand we will calculate s .

1. Find the mean of the observations \bar{x} .
2. Calculate the difference between the observations and the mean for each observation $x_i - \bar{x}$. This is called the deviations of the observations.
3. Square the deviations for each observation $(x_i - \bar{x})^2$.
4. Add up the squared deviations together $\sum_{i=1}^n (x_i - \bar{x})^2$.
5. Divide the sum of the squared deviations by one less than the number of observations $n - 1$. This is the variance

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

Step 6: Standard Deviation

6. Find the square root of the variance. This is the **standard deviation**

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

Example: Section A

Determine the sample standard deviation of the test scores for Section A.

| Section A Scores (X_i) |
|-------------------------------|
| 65 |
| 66 |
| 67 |
| 68 |
| 71 |
| 73 |
| 74 |
| 77 |
| 77 |
| 77 |

Step 1: Calculate the Mean

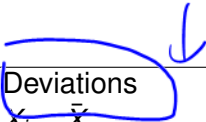
The sample mean is $\bar{x} = 71.5$.



Use Table To Calculate Standard Deviation

| Variable Score (X_i) | Deviations $X_i - \bar{X}$ | Deviations Squared $(X_i - \bar{X})^2$ |
|--------------------------|----------------------------|--|
| 65 | $65 - 71.5$ | $(65 - 71.5)^2$ |
| 66 | $66 - 71.5$ | $(66 - 71.5)^2$ |
| 67 | $67 - 71.5$ | $(67 - 71.5)^2$ |
| 68 | $68 - 71.5$ | $(68 - 71.5)^2$ |
| 71 | $71 - 71.5$ | $(71 - 71.5)^2$ |
| 73 | $73 - 71.5$ | $(73 - 71.5)^2$ |
| 74 | $74 - 71.5$ | $(74 - 71.5)^2$ |
| 77 | $77 - 71.5$ | $(77 - 71.5)^2$ |
| 77 | $77 - 71.5$ | $(77 - 71.5)^2$ |
| | sum | \sum |

Step 2: Calculate Deviations For All Values



| Variable Score (X_i) | Deviations $X_i - \bar{X}$ | Deviations Squared $(X_i - \bar{X})^2$ |
|-----------------------------|-------------------------------|---|
| 65 | $65 - 71.5 = -6.5$ | |
| 66 | $66 - 71.5 = -5.5$ | |
| 67 | $67 - 71.5 = -4.5$ | |
| 68 | $68 - 71.5 = -3.5$ | |
| 71 | $71 - 71.5 = -0.5$ | |
| 73 | $73 - 71.5 = 1.5$ | |
| 74 | $74 - 71.5 = 2.5$ | |
| 77 | $77 - 71.5 = 5.5$ | |
| 77 | $77 - 71.5 = 5.5$ | |
| 77 | $77 - 71.5 = 5.5$ | |
| | sum | |

0

> 0

Step 3: Calculate Squared Deviations

| Variable Score (X_i) | Deviations $X_i - \bar{X}$ | Deviations Squared $(X_i - \bar{X})^2$ |
|--------------------------|----------------------------|--|
| 65 | $65 - 71.5 = -6.5$ | $(-6.5)^2 = 42.25$ |
| 66 | $66 - 71.5 = -5.5$ | $(-5.5)^2 = 30.25$ |
| 67 | $67 - 71.5 = -4.5$ | $(-4.5)^2 = 20.25$ |
| 68 | $68 - 71.5 = -3.5$ | $(-3.5)^2 = 12.25$ |
| 71 | $71 - 71.5 = -0.5$ | $(-0.5)^2 = 0.25$ |
| 73 | $73 - 71.5 = 1.5$ | $1.5^2 = 2.25$ |
| 74 | $74 - 71.5 = 2.5$ | $2.5^2 = 6.25$ |
| 77 | $77 - 71.5 = 5.5$ | $5.5^2 = 30.25$ |
| 77 | $77 - 71.5 = 5.5$ | $5.5^2 = 30.25$ |
| 77 | $77 - 71.5 = 5.5$ | $5.5^2 = 30.25$ |
| | sum | |

Step 4: Calculate the Sum of the Squared Deviations

| Variable Score(X_i) | Deviations $X_i - \bar{X}$ | Deviations Squared $(X_i - \bar{X})^2$ |
|-------------------------|----------------------------|--|
| 65 | $65 - 71.5 = -6.5$ | $(-6.5)^2 = 42.25$ |
| 66 | $66 - 71.5 = -5.5$ | $(-5.5)^2 = 30.25$ |
| 67 | $67 - 71.5 = -4.5$ | $(-4.5)^2 = 20.25$ |
| 68 | $68 - 71.5 = -3.5$ | $(-3.5)^2 = 12.25$ |
| 71 | $71 - 71.5 = -0.5$ | $(-0.5)^2 = 0.25$ |
| 73 | $73 - 71.5 = 1.5$ | $1.5^2 = 2.25$ |
| 74 | $74 - 71.5 = 2.5$ | $2.5^2 = 6.25$ |
| 77 | $77 - 71.5 = 5.5$ | $5.5^2 = 30.25$ |
| 77 | $77 - 71.5 = 5.5$ | $5.5^2 = 30.25$ |
| 77 | $77 - 71.5 = 5.5$ | $5.5^2 = 30.25$ |
| | sum | $\sum_{i=1}^n (X_i - \bar{X})^2 = 204.5$ |

Step 5: Calculate the Variance

$$n=10$$

$$\begin{aligned}\text{variance} &= s^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{204.5}{9} \\ &= \underline{22.7222}\end{aligned}$$

Step 6: Take the Square Root of the Variance

$$\begin{aligned}\text{standard deviation} &= s \\ &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \sqrt{22.7222} \\ &= \underline{4.77}\end{aligned}$$

Sample Standard Deviation of Section A test scores

- Sample standard deviation is $s = 4.77$.
- This implies that from the sample of the 10 students from section A the tests scores has a spread, on average, of 4.77 points from the mean of 71.50 points.

Multiply by 2, You try, Question 1

For the following dataset the mean is $\bar{x} = 4.5$, the variance is $s^2 = 3.5$ and the standard deviation is $s = 1.870829$.

3, 6, 2, 7, 4, 5

Now, multiply each value by 2. What is the new mean, new variance and the new standard deviation?

a) 4.5, 14, 3.5

b) 4.5, 3.5, 1.870829

c) 9, 14, 3.7416

d) 9, 7, 3.7416

$$(sd)^2$$

$$(2sd)^2$$

$$= 2^2 sd^2$$

$$= 4 \text{ Variance}$$

Calculating Standard Deviation

- For larger data sets use a calculator or computer software.
- Each calculator is different if you cannot determine how to compute standard deviation from your calculator ask your instructor.
- For this course we will be using R as the software.
- The function for the sample standard deviation in R is

sd(data name\$variable name).

↗
for sample st. dev.
not population st. dev.

You try, Question 2

2. This is a standard deviation contest, which list of numbers have the largest standard deviation? No calculations are required.

a) ~~10, 10, 10, 10~~ 0

b) ~~20, 20, 20, 20~~ 0

c) 10, 10, 20, 20 15

d) 10, 15, 15, 20 15

Measures of Variability: Coefficient of Variation

- This is to compare the variation between two groups.
- The **coefficient of variation** (cv) is the ratio of the standard deviation to the mean.
- $CV = \frac{sd}{mean}$
- A smaller ratio will indicate less variation in the data.

CV of test scores

| | Section A | Section B |
|-----------------------------|-----------------------------|-------------------------------|
| → Sample Size | 10 | 10 |
| Sample Mean | 71.5 | 71.5 |
| → Sample Standard Deviation | 4.770 | 18.22 |
| CV | $\frac{4.77}{71.5} = 0.066$ | $\frac{18.22}{71.5} = 0.2548$ |

CV Example

The following statistics were collected on two different groups of stock prices:

A is riskier.

| | Portfolio A | Portfolio B |
|---------------------------|-------------|-------------|
| Sample size | 10 | 15 |
| Sample mean | \$52.65 | \$49.80 |
| Sample standard deviation | \$6.50 | \$2.95 |

What can be said about the variability of each portfolio?

$$CV_A = \frac{6.5}{52.65} = 0.1235$$

$$CV_B = \frac{2.95}{49.80} = 0.0592$$

Motivating examples

- In the housing market, for a larger size house, the price of the house increases.
- In a stadium, is the number of hot dogs sold related to the number of sodas sold?
- A survey wants to know if there is a relationship between age and health care cost.
- An insurance wants know if there is a relationship between color of a car and the number of accidents.

Jointly Distributed Data

- **Jointly Distributed Data** is data for two different variables and we want to know the relationship between these two variables.
- Several applications look at the relationship between two variables.
- The data can be quantitative or categorical.
- If both variables are categorical we can look at bar graphs and cross tabulations to determine if there is a relationship.
- If one variable is quantitative and another variable is categorical we can use the side-by-side box plot to look at the relationship.
- If both variables are quantitative we look at the scatter plot to determine if there is a relationship.

Two Jointly Distributed Quantitative Variables

- A **response variable** measures an outcome of a study. Sometimes called a dependent variable. Usually the y-variable.
- An **explanatory variable** explains or influences changes in a response variable. Sometimes called an independent variable or predictor or factor. Usually the x-variable

Types of variables

In the housing market for a larger size house, the price of the house increases.

- The **response variable** is the price of the house.
- The **explanatory variable** is the size of the house.

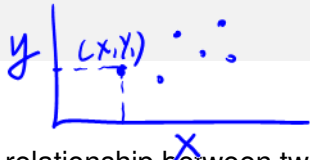
You try

y

For each of the scenarios determine the response variable.

3. A survey wants to know if there is a relationship between age and health care cost.
- a) age b) health care cost c) Not enough information
4. In a stadium, is the number of hot dogs sold related to the number of sodas sold?
- a) Number of hot dogs sold b) Number of sodas sold
- c) Not enough information

Plotting the data: Scatterplots



- The best way to initially observe the relationship between two **quantitative** variables measured on the same individuals.
- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.
- Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.
- Scatterplots can show if there is some kind of association between the two quantitative variables.
- To create a scatterplot in R **plot(explanatory, response)**

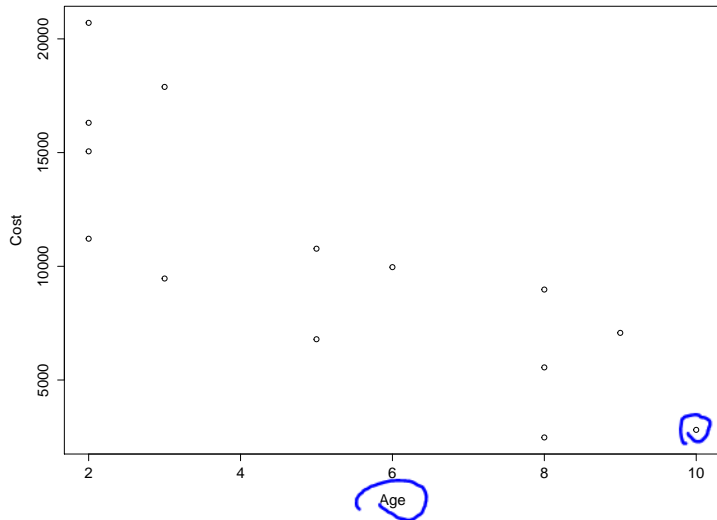
plot(x, y)

Example

We want to look at the relationship between the estimated cost of the car and age of the vehicle.

| Vehicle | Estimated Cost | Age |
|--------------------|----------------|-----------|
| Honda Insight | \$5,555 | 8 |
| Toyota Prius | \$17,888 | 3 |
| Toyota Prius | \$9,963 | 6 |
| Toyota Echo | \$6,793 | 5 |
| Honda Civic Hybrid | \$10,774 | 5 |
| Honda Civic Hybrid | \$16,310 | 2 |
| Chevrolet Prizm | \$2,475 | 8 |
| Mazda Protege | <u>\$2,808</u> | <u>10</u> |
| Toyota Corolla | \$7,073 | 9 |
| Acura Integra | \$8,978 | 8 |
| Scion xB | \$11,213 | 2 |
| Scion xA | \$9,463 | 3 |
| Mazda3 | \$15,055 | 2 |
| Mini Cooper | \$20,705 | 2 |

Scatterplot



Creating a Scatterplot in R

plot(x,y)

> plot(carsreg\$Age, carsreg\$Cost)

x y

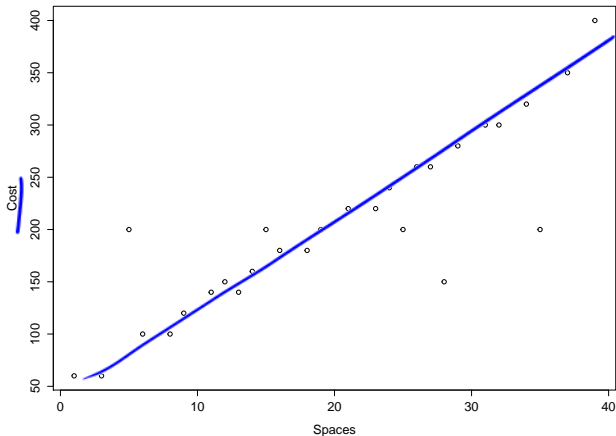
Interpreting scatterplots



- Look for **Direction**: A pattern that runs from the upper left to the lower right is said to have a negative direction. A trend that is running the other way has a positive direction.
- Look for **Form**: If there is a straight line relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form. This is linear form.
- Look for **Strength**: How much scatter the plot has. Do the points appear to follow a single stream? This is a very strong association. Or does the swarm of points seem to form a vague cloud through which we can barely discern any trend or pattern? This is a weak association. Look for the unexpected.

Example 2

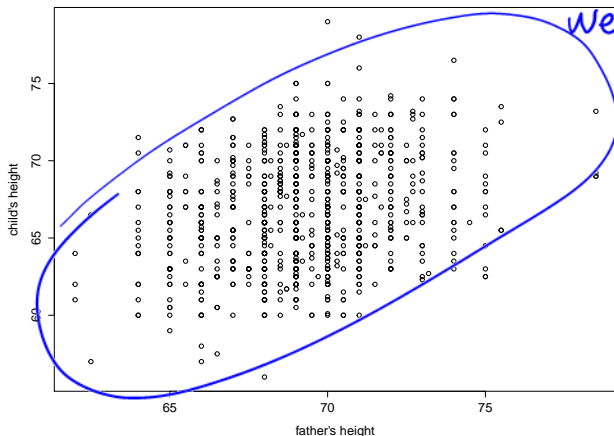
Suppose we want to know if there is an association between the number of spaces a property is from GO and the cost of the property in a monopoly game.



position
linear.
very strong

Example 3

Suppose we want to know if there is an association between the height of the father and their child.



positive
cloud,
weak

Covariance and Correlation

If we have two quantitative variables from a sample, then their covariance is calculated by:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

In R: $\text{cov}(x, y)$ $\text{COV}(X, X) = \text{Var}(X)$

```
> cov(carsreg$Age, carsreg$Cost)
[1] -13254.98
```

However, the strength of the relationship cannot be measured by the covariance, thus we divide each deviation in the sum by the standard deviation of that variable. The result is called the **correlation**.

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\text{sd}(x) * \text{sd}(y)}$$

In R: $\text{cor}(x, y)$

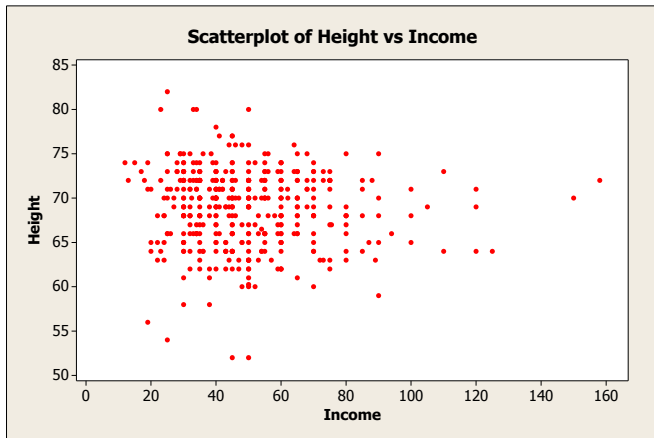
Facts About Correlation: r

- The correlation coefficient is a value that measures the **direction** and **strength** of a **straight-line** relationship between two quantitative variables.
- Correlation makes no distinction between explanatory and response variables.
- Correlation requires that both variables be quantitative.
- Because r uses the standardize values of the observations, r does not change when we change the units of measurement of x , y or both.
- Positive r indicates positive association between the variables, and negative r indicates negative association.

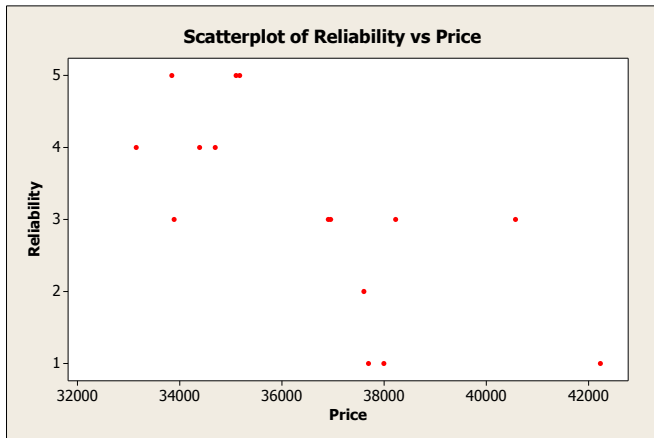
Facts About Correlation: r

- The r is always a number between -1 and 1 .
 - ▶ Values close to 0 indicate a very weak linear relationship.
 - ▶ If r is close to -1 the association is a very strong negative linear relationship.
 - ▶ If r is close to 1 the association is a very strong positive association.
- Correlation measures the strength of only a linear relationship.
- r is strongly affected by a few outlying observations.

$$r = -0.078$$



$$r = -0.707$$



$$r = 0.925$$

