Lecture 3: Data and Descriptive Statistics

ECON 480 - Econometrics - Fall 2018

Ryan Safner

September 5, 2018

Data Basics

Basic Statistics

Data Basics

Data Basics

- Data are information with context

Data Basics

- Data are information with context
- Data values, or observations describe information about some entity

Data Basics

- Data are information with context
- Data values, or observations describe information about some entity
- Metadata describe the process about how data is collected

- Individuals are the entities described by a set of data

- Individuals are the entities described by a set of data
    - e.g. persons, households, firms, countries

Data Basics II

- Individuals are the entities described by a set of data
  - e.g. persons, households, firms, countries
- Variables are particular characteristics about an individual

Data Basics II

- Individuals are the entities described by a set of data
  - e.g. persons, households, firms, countries
- Variables are particular characteristics about an individual
  - e.g. age, income, profits, population, GDP, marital status, type of legal institutions

Data Basics II

- Individuals are the entities described by a set of data
    - e.g. persons, households, firms, countries
- Variables are particular characteristics about an individual
    - e.g. age, income, profits, population, GDP, marital status, type of legal institutions
- Observations are the individuals described by a collection of variables

Data Basics II

- Individuals are the entities described by a set of data
  - e.g. persons, households, firms, countries
- Variables are particular characteristics about an individual
  - e.g. age, income, profits, population, GDP, marital status, type of legal institutions
- Observations are the individuals described by a collection of variables
  - e.g. for one individual, we have their age, sex, income, education, etc.

- Individuals are the entities described by a set of data
    - e.g. persons, households, firms, countries
- Variables are particular characteristics about an individual
    - e.g. age, income, profits, population, GDP, marital status, type of legal institutions
- Observations are the individuals described by a collection of variables
    - e.g. for one individual, we have their age, sex, income, education, etc.
    - individuals and observations are *not necessarily* the same:

- Individuals are the entities described by a set of data
  - e.g. persons, households, firms, countries
- Variables are particular characteristics about an individual
  - e.g. age, income, profits, population, GDP, marital status, type of legal institutions
- Observations are the individuals described by a collection of variables
  - e.g. for one individual, we have their age, sex, income, education, etc.
  - individuals and observations are *not necessarily* the same:
    - e.g. we can have separate observations on the same individual over time

- Categorial variables place an individual into one of several possible categories

Question	Categories or Responses
Do you invest in the stock market?	__ Yes __ No
What kind of advertising do you use?	__ Newspapers __ Internet __ Direct mailings
What is your class at school?	__ Freshman __ Sophomore __ Junior __ Senior
I would recommend this course to another student.	__ Strongly Disagree __ Slightly Disagree __ Slightly Agree __ Strongly Agree
How satisfied are you with this product?	__ Very Unsatisfied __ Unsatisfied __ Satisfied __ Very Satisfied

Categorical Variables

- Categorial variables place an individual into one of several possible categories
    - e.g. sex, season, political party

Question	Categories or Responses
Do you invest in the stock market?	__ Yes __ No
What kind of advertising do you use?	__ Newspapers __ Internet __ Direct mailings
What is your class at school?	__ Freshman __ Sophomore __ Junior __ Senior
I would recommend this course to another student.	__ Strongly Disagree __ Slightly Disagree __ Slightly Agree __ Strongly Agree
How satisfied are you with this product?	__ Very Unsatisfied __ Unsatisfied __ Satisfied __ Very Satisfied

Categorical Variables

- Categorial variables place an individual into one of several possible categories
  - e.g. sex, season, political party
  - may be responses to questions

Question	Categories or Responses
Do you invest in the stock market?	___ Yes ___ No
What kind of advertising do you use?	___ Newspapers ___ Internet ___ Direct mailings
What is your class at school?	___ Freshman ___ Sophomore ___ Junior ___ Senior
I would recommend this course to another student.	___ Strongly Disagree ___ Slightly Disagree ___ Slightly Agree ___ Strongly Agree
How satisfied are you with this product?	___ Very Unsatisfied ___ Unsatisfied ___ Satisfied ___ Very Satisfied

Categorical Variables

- Categorial variables place an individual into one of several possible categories
  - e.g. sex, season, political party
  - may be responses to questions
  - can be quantitative (e.g. age, zip code)

Question	Categories or Responses
Do you invest in the stock market?	___ Yes ___ No
What kind of advertising do you use?	___ Newspapers ___ Internet ___ Direct mailings
What is your class at school?	___ Freshman ___ Sophomore ___ Junior ___ Senior
I would recommend this course to another student.	___ Strongly Disagree ___ Slightly Disagree ___ Slightly Agree ___ Strongly Agree
How satisfied are you with this product?	___ Very Unsatisfied ___ Unsatisfied ___ Satisfied ___ Very Satisfied

Representing Categorical Variables

Cut	Fair	Good	Very Good	Premium	Ideal
Count	1610	4906	12082	13791	21551
Proportion	0.030	0.091	0.224	0.256	0.400

Cut characteristics of 53,940 diamonds

- A good way to represent categorical variables is with a frequency table

Representing Categorical Variables

Cut	Fair	Good	Very Good	Premium	Ideal
Count	1610	4906	12082	13791	21551
Proportion	0.030	0.091	0.224	0.256	0.400

Cut characteristics of 53,940 diamonds

- A good way to represent categorical variables is with a frequency table
- **Count**: frequency (total number) of individuals in a category

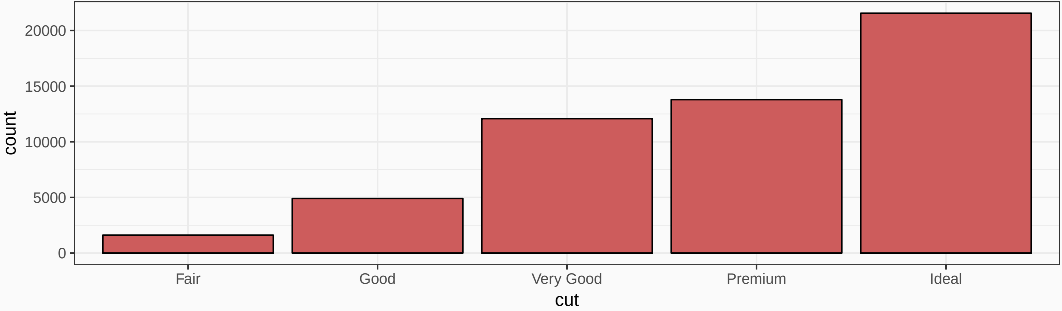
Cut	Fair	Good	Very Good	Premium	Ideal
Count	1610	4906	12082	13791	21551
Proportion	0.030	0.091	0.224	0.256	0.400

Cut characteristics of 53,940 diamonds

- A good way to represent categorical variables is with a frequency table
- **Count**: frequency (total number) of individuals in a category
- **Proportion**: *relative* frequency (percentage of all individuals) in a category

Representing Categorical Variables II

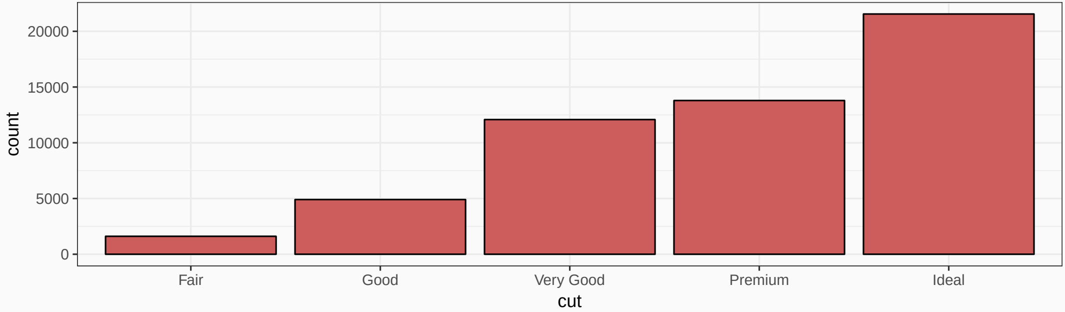
```
ggplot(diamonds, aes(x=cut))+
  geom_bar(fill="indianred", color="black")
```



- Charts and graphs are *always* better ways to visualize data

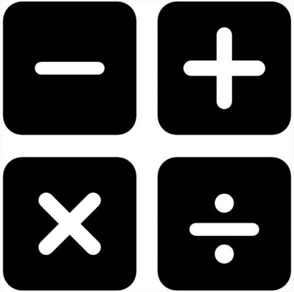
Representing Categorical Variables II

```
ggplot(diamonds, aes(x=cut))+
  geom_bar(fill="indianred", color="black")
```



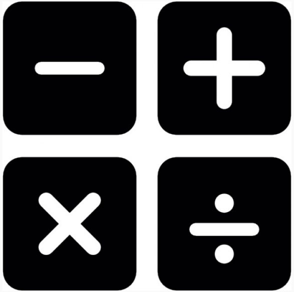
- Charts and graphs are *always* better ways to visualize data
- A bar chart represents categories as bars, with lengths proportional to the count or relative frequency fo each category

- Quantitative variables take on numerical values of equal units that describe an individual



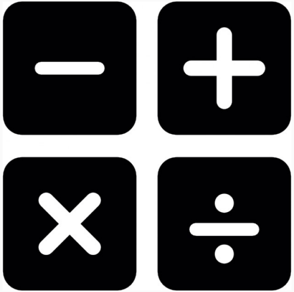
Quantitative Variables

- Quantitative variables take on numerical values of equal units that describe an individual
  - Units: points, dollars, inches



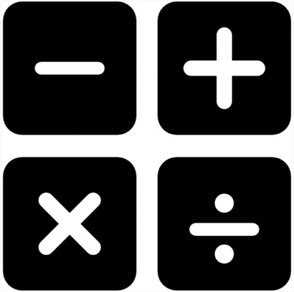
Quantitative Variables

- Quantitative variables take on numerical values of equal units that describe an individual
  - Units: points, dollars, inches
  - Context: GPA, prices, height



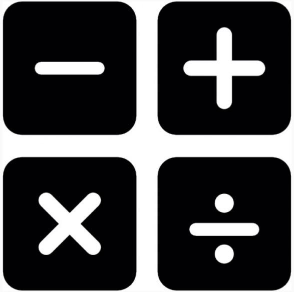
Quantitative Variables

- Quantitative variables take on numerical values of equal units that describe an individual
  - Units: points, dollars, inches
  - Context: GPA, prices, height
- We mathematically manipulate quantitative variables *only* (even if categorical variables are numbers!)



Quantitative Variables

- Quantitative variables take on numerical values of equal units that describe an individual
  - Units: points, dollars, inches
  - Context: GPA, prices, height
- We mathematically manipulate quantitative variables *only* (even if categorical variables are numbers!)
  - e.g. sum, average, standard deviation



Context!

- How variables are classified depends on the *purpose* of collecting and using the data

Context!

- How variables are classified depends on the *purpose* of collecting and using the data

Example

- *Age*, measured in years (quantitative) vs. categories of child, adult, senior, etc.

Example
What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own

Example
What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own
2. The type of car you drive

Categorical or Quantitative?

Example

What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own
2. The type of car you drive
3. Where you go on vacation

HOOD COLLEGE

Example

What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own
2. The type of car you drive
3. Where you go on vacation
4. The amount of money spent on a Super Bowl ad

Example

What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own
2. The type of car you drive
3. Where you go on vacation
4. The amount of money spent on a Super Bowl ad
5. Customer ratings

Example
What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own
2. The type of car you drive
3. Where you go on vacation
4. The amount of money spent on a Super Bowl ad
5. Customer ratings
6. The date a purchase was made

Categorical or Quantitative?

Example

What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own
2. The type of car you drive
3. Where you go on vacation
4. The amount of money spent on a Super Bowl ad
5. Customer ratings
6. The date a purchase was made
7. Transaction ID

Example
What kind of data (categorical or quantitative) does each variable describe ?

1. The number of pairs of shoes you own
2. The type of car you drive
3. Where you go on vacation
4. The amount of money spent on a Super Bowl ad
5. Customer ratings
6. The date a purchase was made
7. Transaction ID
8. Number of correct answers on an exam

Discrete Data

- Discrete data are finite, with a countable number of alternatives

Discrete Data

- Discrete data are finite, with a countable number of alternatives
  - Categorical: e.g. letter grades A, B, C, D, F

- Discrete data are finite, with a countable number of alternatives
  - Categorical: e.g. letter grades A, B, C, D, F
  - Quantitative: integers, e.g. SAT Score, number of children

Continuous Data

- Continuous data are infinitely divisible, with an uncountable number of alternatives

Continuous Data

- Continuous data are infinitely divisible, with an uncountable number of alternatives
  - e.g. weights, temperature, GPA

Continuous Data

- Continuous data are infinitely divisible, with an uncountable number of alternatives
  - e.g. weights, temperature, GPA
- Many discrete variables may be treated as if they are continuous

- Continuous data are infinitely divisible, with an uncountable number of alternatives
    - e.g. weights, temperature, GPA
- Many discrete variables may be treated as if they are continuous
    - e.g. SAT scores, wages

Example
What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds

Example

What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds
2. Price in dollars

Example

What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds
2. Price in dollars
3. Grade (Letter)

Example

What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds
2. Price in dollars
3. Grade (Letter)
4. Temperature

Example

What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds
2. Price in dollars
3. Grade (Letter)
4. Temperature
5. Amazon Star Rating

Example

What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds
2. Price in dollars
3. Grade (Letter)
4. Temperature
5. Amazon Star Rating
6. Number of customers

Discrete or Continous?

Example
What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds
2. Price in dollars
3. Grade (Letter)
4. Temperature
5. Amazon Star Rating
6. Number of customers
7. Transaction ID

Example
What kind of data (discrete or continuous) does each variable describe ?

1. Weight in pounds
2. Price in dollars
3. Grade (Letter)
4. Temperature
5. Amazon Star Rating
6. Number of customers
7. Transaction ID
8. Number of correct answers on an exam

```
##   ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- The most common data structure we use is a spreadsheet

```
##   ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- The most common data structure we use is a spreadsheet
    - Note: *R* calls this a `data frame`

```
##   ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- The most common data structure we use is a spreadsheet
    - Note: *R* calls this a `data frame`
- A **row** contains data about all variables for a single individual

```
##   ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- The most common data structure we use is a spreadsheet
    - Note: *R* calls this a `data frame`
- A **row** contains data about all variables for a single individual
- A **column** contains data about a single variable across all individuals

```
##   ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- The most common data structure we use is a spreadsheet
    - Note: *R* calls this a `data frame`
- A **row** contains data about all variables for a single individual
- A **column** contains data about a single variable across all individuals
- It is good practice to have an **ID variable** to count and keep track of each observation

Some Notation

- It is common to some notation like the following:

Example

- It is common to some notation like the following:
- Let $\{x_1, x_2, \cdots, x_n\}$ be a simple data series

Example

- It is common to some notation like the following:
- Let $\{x_1, x_2, \cdots, x_n\}$ be a simple data series
    - n individual observations

Example

- It is common to some notation like the following:
- Let $\{x_1, x_2, \cdots, x_n\}$ be a simple data series
    - n individual observations
    - x_i is the value of the i^{th} observation for $i = 1, 2, \cdots, n$

Example

HOOD COLLEGE

- It is common to some notation like the following:
- Let $\{x_1, x_2, \cdots, x_n\}$ be a simple data series
    - n individual observations
    - x_i is the value of the i^{th} observation for $i = 1, 2, \cdots, n$

Example

- Let x represent the score on a homework assignment:

$$75, 100, 92, 87, 79, 0, 95$$

Some Notation

- It is common to some notation like the following:
- Let $\{x_1, x_2, \cdots, x_n\}$ be a simple data series
    - n individual observations
    - x_i is the value of the i^{th} observation for $i = 1, 2, \cdots, n$

Example

- Let x represent the score on a homework assignment:

$$75, 100, 92, 87, 79, 0, 95$$

- What is n?

Some Notation

- It is common to some notation like the following:
- Let $\{x_1, x_2, \cdots, x_n\}$ be a simple data series
  - n individual observations
  - x_i is the value of the i^{th} observation for $i = 1, 2, \cdots, n$

Example

- Let x represent the score on a homework assignment:

$$75, 100, 92, 87, 79, 0, 95$$

- What is n?
- What is x_1?

Some Notation

- It is common to some notation like the following:
- Let $\{x_1, x_2, \cdots, x_n\}$ be a simple data series
  - n individual observations
  - x_i is the value of the i^{th} observation for $i = 1, 2, \cdots, n$

Example

- Let x represent the score on a homework assignment:

$$75, 100, 92, 87, 79, 0, 95$$

- What is n?
- What is x_1?
- What is x_6?

Datasets: Cross-Section

```
##    ID     Name Age    Sex Income
## 1  1     John  23   Male  41000
## 2  2    Emile  18   Male  52600
## 3  3  Natalya  28 Female  48000
## 4  4  Lakisha  31 Female  60200
## 5  5    Cheng  36   Male  81900
```

- Cross-sectional data: observations of individuals at a given point in time

Datasets: Cross-Section

```
##   ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- Cross-sectional data: observations of individuals at a given point in time
    - Each observation is a unique individual

Datasets: Cross-Section

```
##   ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- Cross-sectional data: observations of individuals at a given point in time
    - Each observation is a unique individual
    - Simplest and most common data

Datasets: Cross-Section

```
##    ID    Name Age    Sex Income
## 1  1    John  23   Male  41000
## 2  2   Emile  18   Male  52600
## 3  3 Natalya  28 Female  48000
## 4  4 Lakisha  31 Female  60200
## 5  5   Cheng  36   Male  81900
```

- Cross-sectional data: observations of individuals at a given point in time
    - Each observation is a unique individual
    - Simplest and most common data
    - A “snapshot” to compare differences across individuals

```
##   Year  GDP Unemployment CPI
## 1 1950  8.2         0.06 100
## 2 1960  9.9         0.04 118
## 3 1970 10.2         0.08 130
## 4 1980 12.4         0.08 190
## 5 1985 13.6         0.06 196
```

- Time-series data: observations of the same individuals over time

```
##   Year  GDP Unemployment CPI
## 1 1950  8.2         0.06 100
## 2 1960  9.9         0.04 118
## 3 1970 10.2         0.08 130
## 4 1980 12.4         0.08 190
## 5 1985 13.6         0.06 196
```

- Time-series data: observations of the same individuals over time
  - Each observation is an individual-year

```
##   Year  GDP Unemployment CPI
## 1 1950  8.2         0.06 100
## 2 1960  9.9         0.04 118
## 3 1970 10.2         0.08 130
## 4 1980 12.4         0.08 190
## 5 1985 13.6         0.06 196
```

- Time-series data: observations of the same individuals over time
    - Each observation is an individual-year
    - Often used for macroeconomics, finance, and forecasting

```
##   Year  GDP Unemployment CPI
## 1 1950  8.2         0.06 100
## 2 1960  9.9         0.04 118
## 3 1970 10.2         0.08 130
## 4 1980 12.4         0.08 190
## 5 1985 13.6         0.06 196
```

- Time-series data: observations of the same individuals over time
    - Each observation is an individual-year
    - Often used for macroeconomics, finance, and forecasting
    - Unique challenges for time series

Datasets: Time Series

```
##   Year  GDP Unemployment CPI
## 1 1950  8.2         0.06 100
## 2 1960  9.9         0.04 118
## 3 1970 10.2         0.08 130
## 4 1980 12.4         0.08 190
## 5 1985 13.6         0.06 196
```

- Time-series data: observations of the same individuals over time
  - Each observation is an individual-year
  - Often used for macroeconomics, finance, and forecasting
  - Unique challenges for time series
  - A “moving picture” to see how individuals change over time

```
##                City Year Murders Population Unemployment Police
## 1     Philadelphia 1986       5      3.700          8.7    440
## 2     Philadelphia 1990       8      4.200          7.2    471
## 3 Washington D.C. 1986       2      0.250          5.4     75
## 4 Washington D.C. 1990      10      0.275          5.5     85
## 5         New York 1986       3      6.400          9.6    102
```

- Panel dataset, or longitudinal dataset: a time-series for *each* cross-sectional entity

Datasets: Panel Data

```
##                City Year Murders Population Unemployment Police
## 1    Philadelphia 1986       5      3.700          8.7    440
## 2    Philadelphia 1990       8      4.200          7.2    471
## 3 Washington D.C. 1986       2      0.250          5.4     75
## 4 Washington D.C. 1990      10      0.275          5.5     85
## 5        New York 1986       3      6.400          9.6    102
```

- Panel dataset, or longitudinal dataset: a time-series for *each* cross-sectional entity
  - Must be the *same* cross-sectional entities over time

```
##                City Year Murders Population Unemployment Police
## 1     Philadelphia 1986       5      3.700          8.7    440
## 2     Philadelphia 1990       8      4.200          7.2    471
## 3 Washington D.C. 1986       2      0.250          5.4     75
## 4 Washington D.C. 1990      10      0.275          5.5     85
## 5         New York 1986       3      6.400          9.6    102
```

- Panel dataset, or longitudinal dataset: a time-series for *each* cross-sectional entity
    - Must be the *same* cross-sectional entities over time
    - More common today for serious researchers

```
##                City Year Murders Population Unemployment Police
## 1     Philadelphia 1986       5      3.700          8.7    440
## 2     Philadelphia 1990       8      4.200          7.2    471
## 3 Washington D.C. 1986       2      0.250          5.4     75
## 4 Washington D.C. 1990      10      0.275          5.5     85
## 5         New York 1986       3      6.400          9.6    102
```

- Panel dataset, or longitudinal dataset: a time-series for *each* cross-sectional entity
  - Must be the *same* cross-sectional entities over time
  - More common today for serious researchers
  - Unique challenges for panel data

```
##               City Year Murders Population Unemployment Police
## 1    Philadelphia 1986       5      3.700          8.7    440
## 2    Philadelphia 1990       8      4.200          7.2    471
## 3 Washington D.C. 1986       2      0.250          5.4     75
## 4 Washington D.C. 1990      10      0.275          5.5     85
## 5        New York 1986       3      6.400          9.6    102
```

- Panel dataset, or longitudinal dataset: a time-series for *each* cross-sectional entity
    - Must be the *same* cross-sectional entities over time
    - More common today for serious researchers
    - Unique challenges for panel data
    - A combination of "snapshot" comparisons and differences over time

Basic Statistics

- Variable shave a distribution of different individual values (and how frequently they take these values)

- Variable shave a distribution of different individual values (and how frequently they take these values)
- We want to *visualize* and *analyze* distributions to search for meaningful patterns using **statistics**

Two Types of Statistics

- Two main categories or uses of statistics:

Two Types of Statistics

- Two main categories or uses of statistics:
  1. Descriptive Statistics: describes or summarizes the properties of a sample

- Two main categories or uses of statistics:
  1. Descriptive Statistics: describes or summarizes the properties of a sample
     1.2 Inferential Statistics: uses a sample in order to infer properties about a larger population

- A common way to present a variable's distribution is a histogram

Histograms

- A common way to present a variable's distribution is a histogram
  - The quantitative analog to the bar graph for a categorical variable

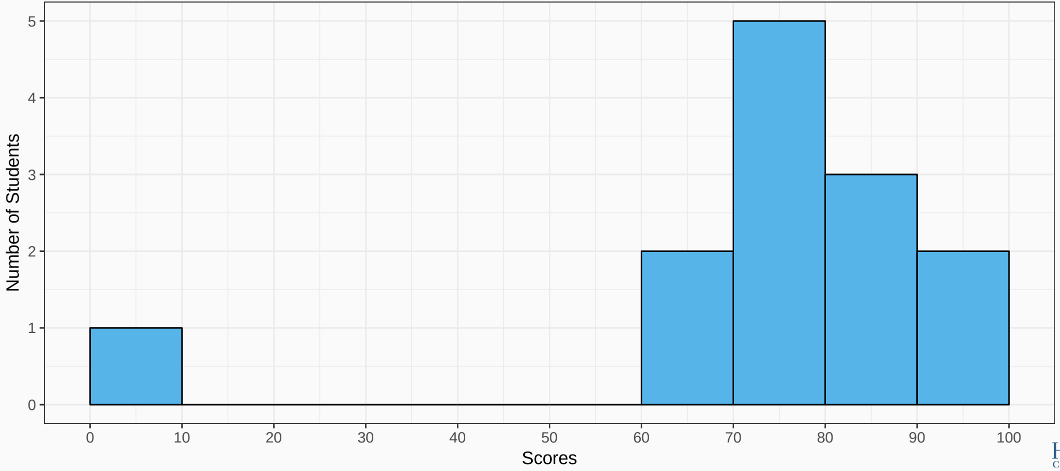
Histograms

- A common way to present a variable's distribution is a histogram
    - The quantitative analog to the bar graph for a categorical variable
- We divide up the data values into **bins** of a certain size, and count the number of values falling within each bin, representing them visually as bars

Example

A class of 13 students takes a quiz (out of 100 points) with the following results:

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$


5
4
3
2
1
0
Number of Students
0
10
20
30
40
50
60
70
80
90
100
Scores

- We are often interested in the *shape* or *pattern* of a distribution, particularly:

- We are often interested in the *shape* or *pattern* of a distribution, particularly:
    - Measures of central tendency

- We are often interested in the *shape* or *pattern* of a distribution, particularly:
    - Measures of central tendency
    - Measures of dispersion

- We are often interested in the *shape* or *pattern* of a distribution, particularly:
    - Measures of central tendency
    - Measures of dispersion
    - Shape of distribution

- The mode of a variable is simply its most frequent value

Example
$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$

- The mode of a variable is simply its most frequent value
- A variable can have multiple modes

Example
{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95}

- The mode of a variable is simply its most frequent value
- A variable can have multiple modes

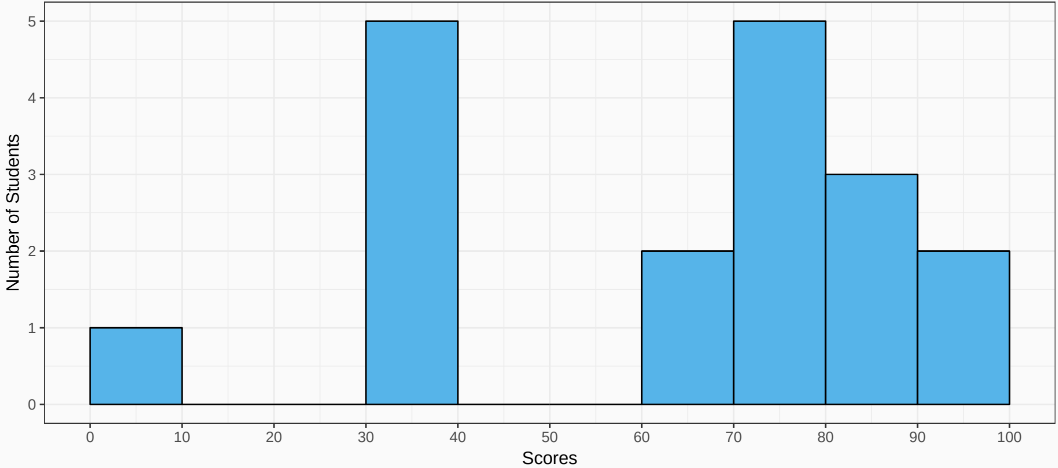
Example
{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95}

Central Tendency: Mode

- The mode of a variable is simply its most frequent value
- A variable can have multiple modes

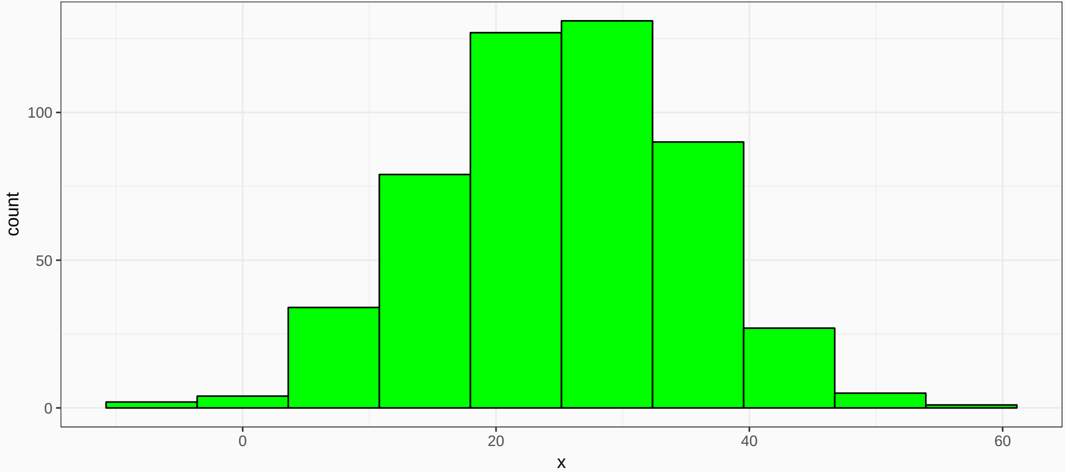
Example
{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95}

Multi-Modal Distributions



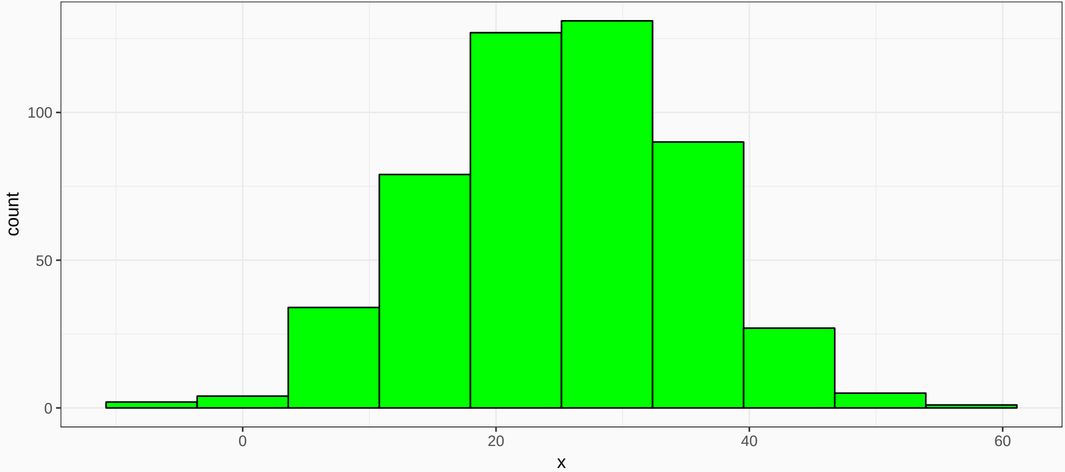
- Looking at a histogram, the modes are often the “peaks” of the distribution - May be **unimodal**, **bimodal**, **trimodal**, etc

Symmetry

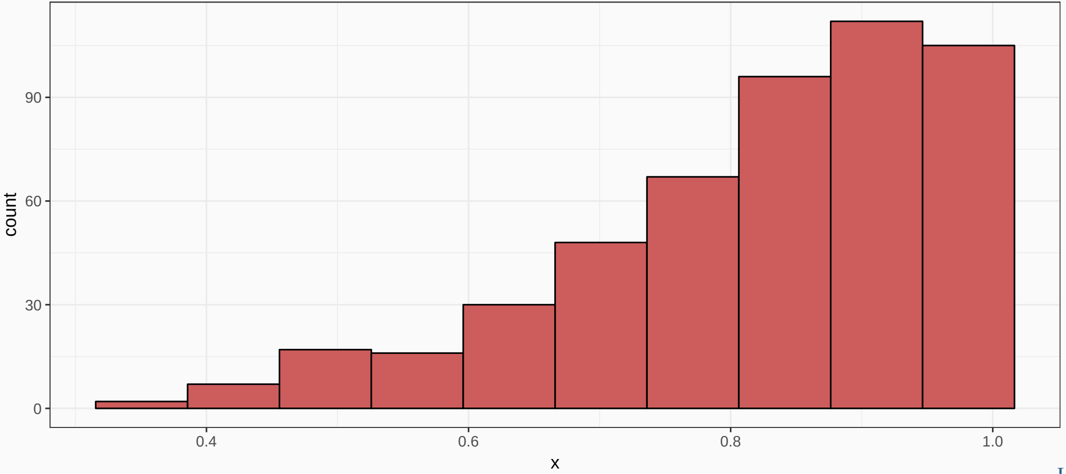


- A distribution is **symmetric** if it looks roughly the same on either side of the “center”

Symmetry

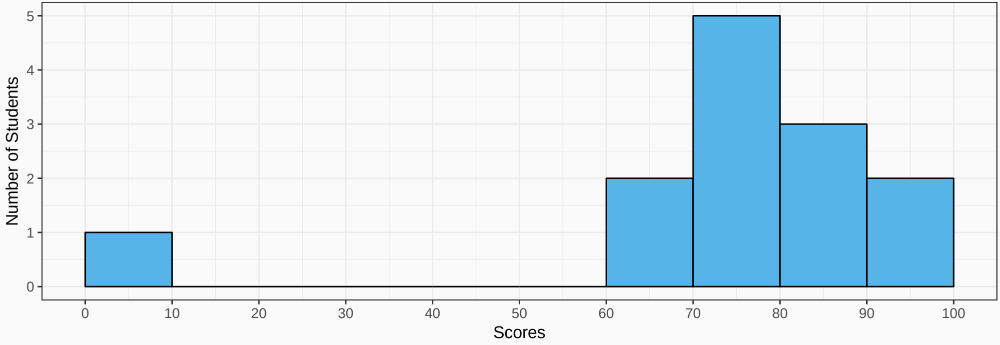


- A distribution is **symmetric** if it looks roughly the same on either side of the "center"
- The thinner ends (far left and far right) are called the **tails** of a distribution



- If one tail stretches farther than the other, distribution is **skewed** in the direction of the longer tail

- An extreme value that does not appear part of the general pattern of a distribution is an outlier



Outliers

- Outliers can strongly affect descriptive statistics about a dataset

Outliers

- Outliers can strongly affect descriptive statistics about a dataset
- Outliers can be the most informative part of the data

Outliers

- Outliers can strongly affect descriptive statistics about a dataset
- Outliers can be the most informative part of the data
- Outliers could be the result of errors

Outliers

- Outliers can strongly affect descriptive statistics about a dataset
- Outliers can be the most informative part of the data
- Outliers could be the result of errors
- Outliers should always be discussed in presentations about data

Arithmetic Mean (Population)

- The natural measure of the center of a *population*'s distribution is its "average" or arithmetic mean (μ)

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Arithmetic Mean (Population)

- The natural measure of the center of a *population*'s distribution is its "average" or arithmetic mean (μ)

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- For N values of variable x, "mu" is the sum of all individual x values $\left(x_i\right)$ from 1 to N, divided by the N number of values

Arithmetic Mean (Sample)

- When we have a *sample*, we compute the sample mean ($\bar{x}$)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

-For n values of variable x, "x-bar" is the sum of all individual x values (x_i) from 1 to n, divided by the n number of values

Sample Mean Example

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

$$\begin{aligned}
\bar{x} &= \frac{1}{13}(0 + 62 + 66 + 71 + 71 + 74 + 76 + 79 + 83 + 86 + 88 + 93 + 95) \\
&= \frac{944}{13} \\
&= 72.61
\end{aligned}$$

- Note the mean need not be an actual value of the data!

$$\{62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

- If we drop the outlier (0)

$$\bar{x} = \frac{1}{12}(62 + 66 + 71 + 71 + 74 + 76 + 79 + 83 + 86 + 88 + 93 + 95) = \frac{944}{12}$$
$$= 78.67$$

Median

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

- The median is the midpoint of the distribution

Median

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

- The median is the midpoint of the distribution
- 50% to the left of the median, 50% to the right of the median

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

- The median is the midpoint of the distribution
- 50% to the left of the median, 50% to the right of the median
- Arrange values in numerical order

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

- The median is the midpoint of the distribution
- 50% to the left of the median, 50% to the right of the median
- Arrange values in numerical order
    - For odd n: median is middle observation

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

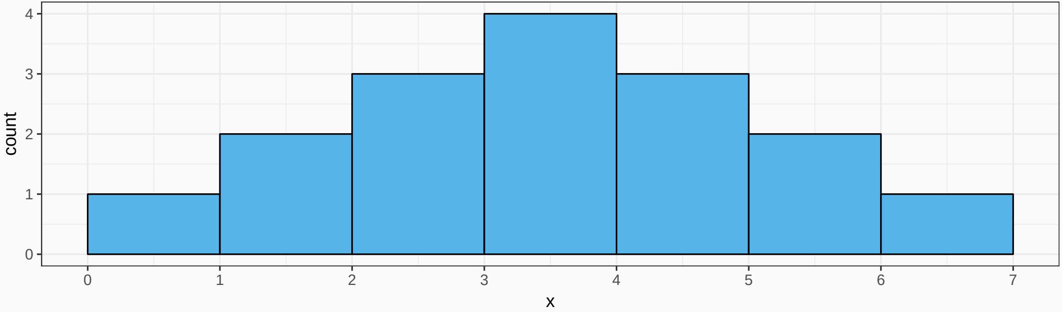
- The median is the midpoint of the distribution
- 50% to the left of the median, 50% to the right of the median
- Arrange values in numerical order
    - For odd n: median is middle observation
    - For even n: median is average of two middle observations

$$\{0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

- The median is *robust* to outliers (if 0 changes to 62)

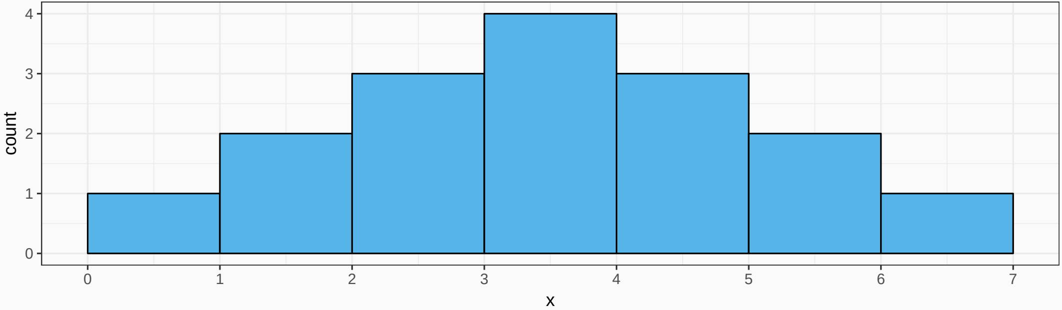
$$\{62, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95\}$$

$$\{1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7\}$$



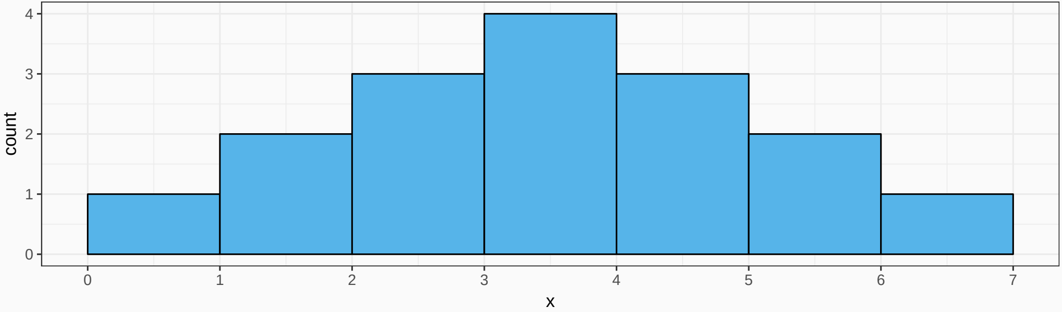
- Mean: $\frac{64}{16} = 4$

$$\{1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7\}$$



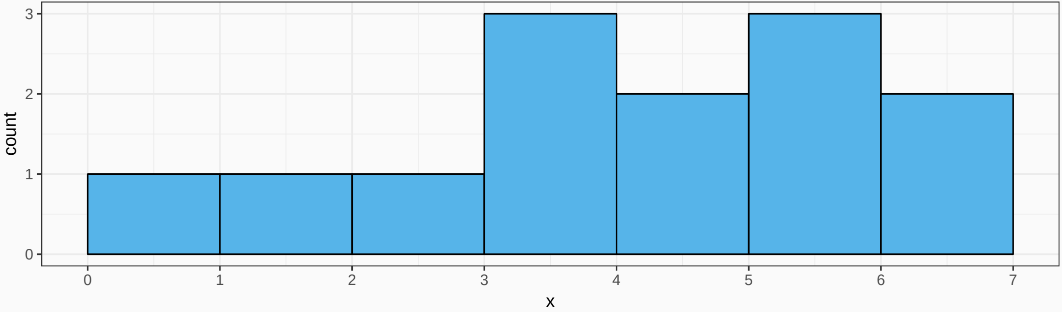
- Mean: $\frac{64}{16} = 4$
- Median: 4

$$\{1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7\}$$



- Mean: $\frac{64}{16} = 4$
- Median: 4
- For a symmetric distribution, mean=median

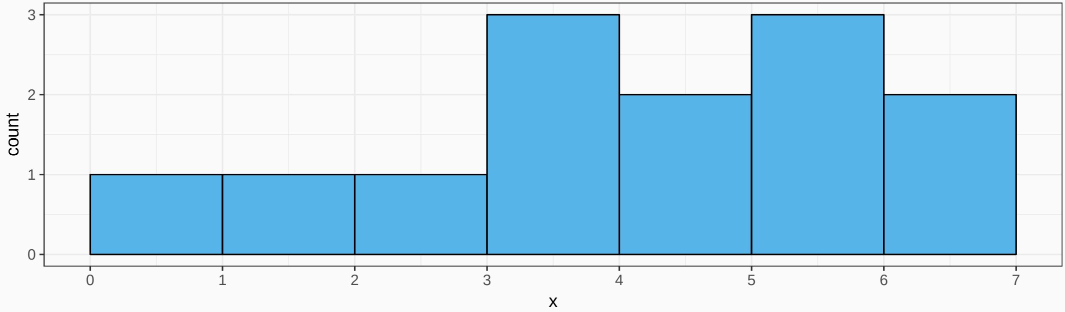
$$\{1, 2, 3, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7\}$$



- Mean: $\frac{60}{13} = 4.6$

Mean, Median, Symmetry, Skew II

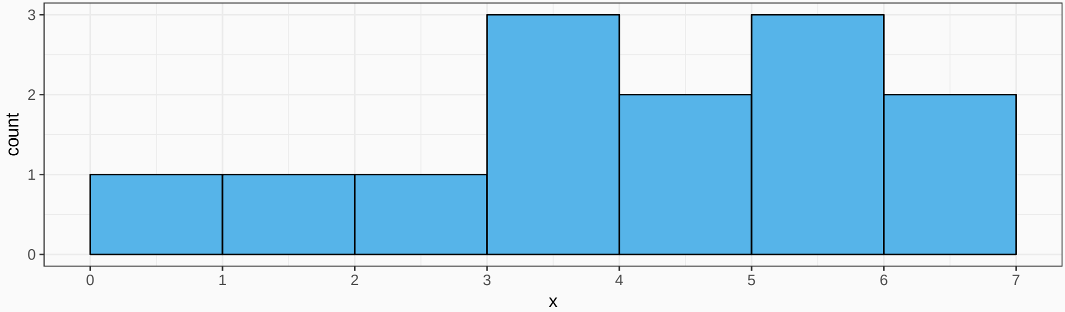
$$\{1, 2, 3, 4, 4, 4, 5, 5, 6, 6, 6, 7, 7\}$$



- Mean: $\frac{60}{13} = 4.6$
- Median: 5

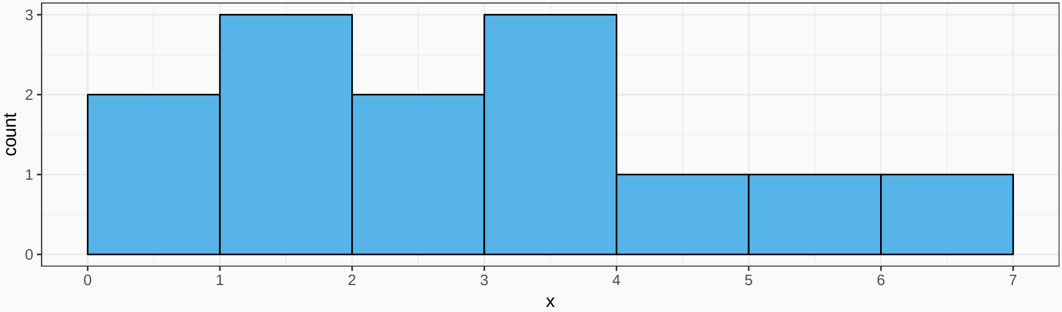
Mean, Median, Symmetry, Skew II

$$\{1, 2, 3, 4, 4, 4, \textcolor{magenta}{5}, 5, 6, 6, 6, 7, 7\}$$



- Mean: $\frac{60}{13} = 4.6$
- Median: 5
- For a left-skewed distribution, mean$<$median

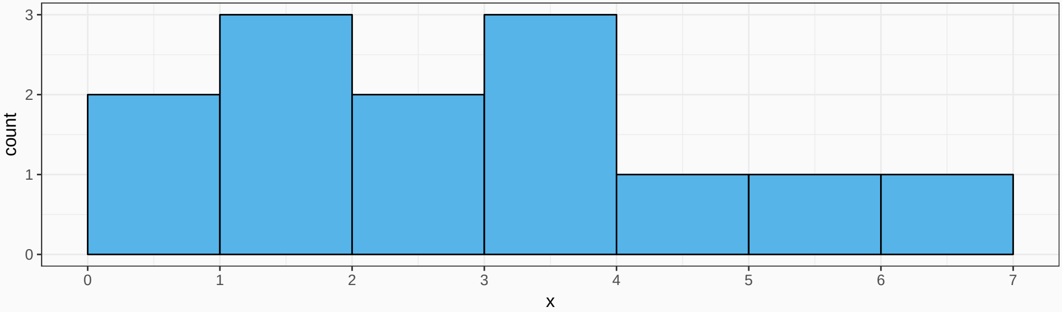
$$\{1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 5, 6, 7\}$$



- Mean: $\frac{44}{13} = 3.4$

Mean, Median, Symmetry, Skew III

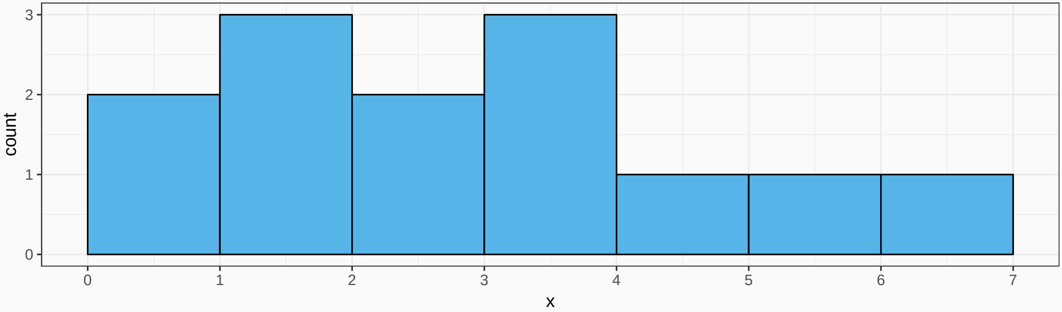
$$\{1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 5, 6, 7\}$$



- Mean: $\frac{44}{13} = 3.4$
- Median: 3

Mean, Median, Symmetry, Skew III

$$\{1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 5, 6, 7\}$$



- Mean: $\frac{44}{13} = 3.4$
- Median: 3
- For a right-skewed distribution, mean$>$median

- The more *variation* in the data, the less helpful a measure of central tendency will tell us

Measures of Spread

- The more *variation* in the data, the less helpful a measure of central tendency will tell us
- Beyond just the center, we also want to measure the spread

- The more *variation* in the data, the less helpful a measure of central tendency will tell us
- Beyond just the center, we also want to measure the spread
- Simplest metric is range=max-min

Once we know the values of the quartiles, we can construct a five-number summary of a distribution, including: 1. Minimum 2. Q_1 (25%) 3. Median (50%) 4. Q_3 (75%) 5. Maximum

Five Number Summary II

```r
summary(quizzes$scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   71.00   76.00   72.62   86.00   95.00
```

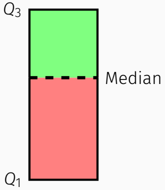
- Graphical way to visualize five number summary is a boxplot

Boxplots

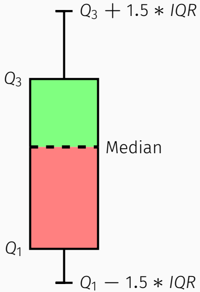
- Graphical way to visualize five number summary is a boxplot

- Graphical way to visualize five number summary is a boxplot
  - The length of the box is the IQR (Q1-Q3)

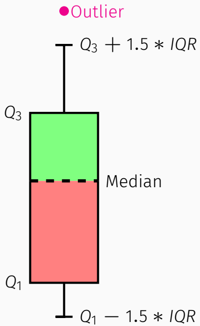
- Graphical way to visualize five number summary is a boxplot
    - The length of the box is the IQR (Q1-Q3)
    - The line within the box is the median



- Graphical way to visualize five number summary is a boxplot
  - The length of the box is the IQR (Q1-Q3)
  - The line within the box is the median
  - The "whiskers" identify data within $1.5 \times IQR$



- Graphical way to visualize five number summary is a boxplot
  - The length of the box is the IQR (Q1-Q3)
  - The line within the box is the median
  - The "whiskers" identify data within $1.5 \times IQR$
  - Points beyond the whiskers are outliers



Quiz 1: {0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95}

Quiz 2: {50, 62, 72, 73, 79, 81, 82, 82, 86, 90, 94, 98, 99}

Boxplots II

Quiz 1: {0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95}

Quiz 2: {50, 62, 72, 73, 79, 81, 82, 82, 86, 90, 94, 98, 99}

Quiz 1

Min	Q_1	Median	Q_3	Max
0	71	76	86	95

Boxplots II

Quiz 1: {0, 62, 66, 71, 71, 74, 76, 79, 83, 86, 88, 93, 95}

Quiz 2: {50, 62, 72, 73, 79, 81, 82, 82, 86, 90, 94, 98, 99}

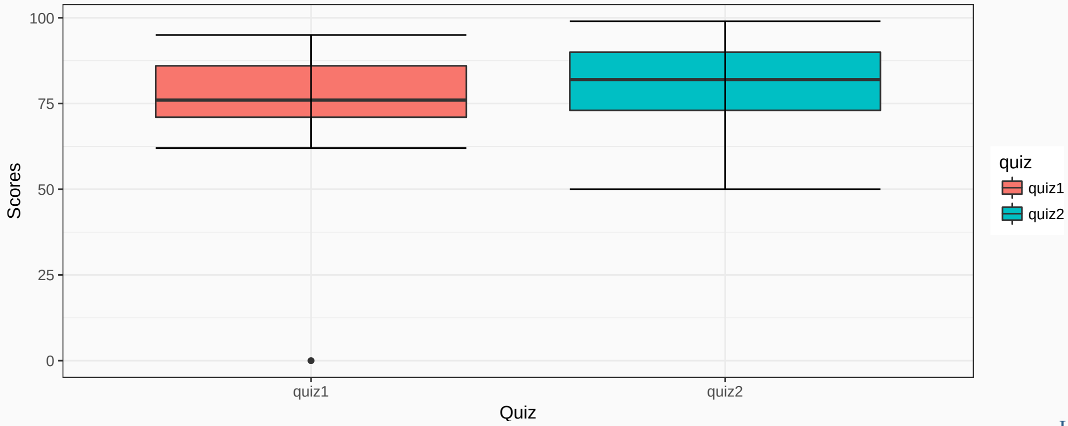
Quiz 1

Min	Q_1	Median	Q_3	Max
0	71	76	86	95

Quiz 2

Min	Q_1	Median	Q_3	Max
50	73	82	90	99

Boxplots III



Deviations

- Each observation deviates from the mean of the data:

$$deviation = x_i - \mu$$

- Each observation deviates from the mean of the data:

$$deviation = x_i - \mu$$

- There are as many deviations as there are data points (n)

- Each observation deviates from the mean of the data:

$$deviation = x_i - \mu$$

- There are as many deviations as there are data points (n)
- We can measure the *average* or standard deviation from the mean

- The population variance (σ^2) of a *population* distribution measures the average of the *squared* deviations from the population mean

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

Variance

- The population variance (σ^2) of a *population* distribution measures the average of the *squared* deviations from the population mean

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

- Why do we square deviations?

-Square root the variance to get the population standard deviation (σ), the average deviation from the mean (in x units)

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

- The sample variance (s^2) of a *sample* distribution measures the average of the *squared* deviations from the sample mean

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

-Why divide by $n - 1$?

Sample Standard Deviation

- Square root the variance to get the sample standard deviation (s), the average deviation from the mean (in x units)

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

Example
Calculate the sample standard deviation for the following series:

$$\{2, 4, 6, 8, 10\}$$

Population Parameters

- Population Size: N
- Mean: μ
- Variance:
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$
- Standard Deviation:
$$\sigma = \sqrt{\sigma^2}$$

Sample Statistics

- Sample Size: n
- Mean: $\bar{x}$
- Variance:
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
- Standard Deviation: $s = \sqrt{s^2}$