# Lesson 3: Statistics

# Outline

1. Introduction to Data, Computing and Statistics

1. More about basic Statistics model (population and variables),Frequency and Distribution

# 1.1 Data, Computing and Statistics

In the vast majority of activities of human beings, we are continuously collecting (using our five senses) and analyzing **data** and making decisions (using brain) based on the analysis (often maybe unconsciously).

**Statistics** is a science and art of collecting, analyzing and drawing conclusions from data.

**Computing** provides a more general manipulation and reasoning with data. Statics is an important part of this manipulation and reasoning.

# 1.2 Human and Data

## 1.2.1 MOTIVATION:

Our **sensors** (eyes, ears, skin, nose, and tongue) constantly **collect** information/**data** from our world. These data will be sent to our **brain**, an amazing data processor, and the brain will remember (some of the data), **analyze** them, **make decisions** and act accordingly.

# 1.2 Human and Data

## 1.2.2 EXAMPLES:

**Playing a video game**

- We see the pictures (data) and hear the sound (data) from a game

- After our brain got the data, it will analyze the data, make decisions and issue actions to our hand which will control the game by keyboard, touch screen or game controllers.

# 1.2 Human and Data

**Food**

- We smell and taste the food in a restaurant
- We remember the pleasant taste
- When we are hungry, we can recall the restaurant, its location and food

**Watching a basketball game**

- We see players playing on the courts: shooting, blocking, assisting
- Our brain will remember a lot of the data obtained from our eyes.
- We can then (by analyzing the data may unconsciously) and tell other people how well a player play: scores, # of defenses and blocks etc.

# 1.3 Data and Computing

**1.3.1 MOTIVATION:** Numerous **devices** are invented for a computer to sense the world (keyboard, mouse, the game controller, camera, microphone …) to get the data about the world. Data were sent to program(s) running on the CPU of the computer, and the **program** will **analyze** the data, **make decisions** and issue actions.

**1.3.2 EXAMPLES:**

**A phone unlocks itself using your face (a phone seems to recognize your)**

- Before you use your cell phone, its camera captures your face and sends the picture (*data*) to a program (face recognition software) running on the (CPU of the) phone. The program will analyze the picture and its earlier record of your face pictures and make a decision on whether it will unlock the phone for you

# 1.3 Data and Computing

**A basketball game software (on a CPU/computer)**

- Will get data from your game controller for example shooting button is pressed
- It then will decide whether the shoot will hit or miss and then produce an animation: the ball leaving the hand, flying and then hitting or missing the basket
- The game software can certainly remember the scores of the game, the scores of each player etc.

# 1.4 Statistics

## 1.4.1 Introduction:

Statistics focuses on a population (more general than our daily word of population, it refers to any set of objects or things) and data about this population. There are two types of statistics –

- Descriptive Statistics
- Inferential Statistics

Computing provides all kinds of operations on data while statistics provide a specific way of processing the data, which happens to be very useful in our daily life.

# 1.4 Statistics

## 1.4.2 Descriptive Statistics:

Assume we know the data about a population. Since the data is huge, people are not interested in very detail of the data. **Descriptive statistics** provides ways for us to summarize the data. The summary will give us a good idea about the population.

# 1.4.2 Descriptive Statistics (example)

Consider two populations in our high school: one is 9th graders and the other is 10th graders. We are interested in the height of the students, and we have the data about each student.

Sometimes, we would like to compare the heights of 9th graders and 10th graders. A one by one comparison will not be very interesting. But a summary of the height of the population may be more meaningful.

For example, the height of all 9th graders can be summarized by the average - one number, and so can that of all 10th graders. Likely we can find that the average height of 10th graders is taller than that of 9th graders, which is a meaningful pattern. However, if we use individual data, it would be difficult to obtain such a result.

## 1.4.3 Inferential Statistics:

The population sometimes is so big that it is impossible or very difficult to get the data about this population. **Inferential statistics** provides methods to get the interesting information about the whole population without getting the data for each individual.

# 1.4.3 Inferential Statistics (example)

Consider the population of all 9th graders of the whole nation. We want to know the *average height of them*. Difficult to measure and collect the height of each student. The method of *inferential statistics* is to take a **sample** of manageable size from all 9th graders, measure the height of each student in the sample and get the average height of these students. This height will be an estimation of the *average height of **ALL** the 9th graders*.

# Outline

1. Introduction to Data, Computing and Statistics

1. Basic Statistics model: population and variables, Frequency and Distribution

# 2.1 Population and Variables

In statistics, we always start with a **problem**. To solve the problem, we need to first figure out the basic information in the problem:

- **population**: the *set* of the *main* objects that we want to know about in the problem. Each element of this population is called an **individual**.
- **statistics variable**s (of the population): are the aspects of the individuals of the population that we would like to know.

All statistics methods are based on *population* and *statistics variables*.

# 2.1 Population and Variables

**Example Problem.** We have the following information about NASCAR drivers

| Driver | Car | Home State | Races Won |
|---|---|---|---|
| **Dale Earnhardt** | #3 | North Carolina | 21 |
| **Denny Hamlin** | #11 | Florida | 63 |
| **Jeff Gordon** | #24 | California | 5 |
| **Jimmie Johnson** | #48 | California | 83 |
| **Kyle Busch** | #1 | Nevada | 212 |
| **Richard Petty** | #43 | North Carolina | 200 |

*Questions*.

- Where does Dale come from?
- How many races does Denny win?

# 2.1 Population and Variables

**Example Problem.** We have the following information about NASCAR drivers

| Driver | Car | Home State | Races Won |
|---|---|---|---|
| **Dale Earnhardt** | #3 | North Carolina | 21 |
| **Denny Hamlin** | #11 | Florida | 63 |
| **Jeff Gordon** | #24 | California | 5 |
| **Jimmie Johnson** | #48 | California | 83 |
| **Kyle Busch** | #1 | Nevada | 212 |
| **Richard Petty** | #43 | North Carolina | 200 |

*Questions*.

- Where does Dale come from? North Carolina
- How many races does Denny win? 63

What is the **population** of the problem? (the *set* of the *main* objects that we want to know about in the problem)

# 2.1 Population and Variables

**Example Problem.** We have the following information about NASCAR drivers

| Driver | Car | Home State | Races Won |
|---|---|---|---|
| **Dale Earnhardt** | #3 | North Carolina | 21 |
| **Denny Hamlin** | #11 | Florida | 63 |
| **Jeff Gordon** | #24 | California | 5 |
| **Jimmie Johnson** | #48 | California | 83 |
| **Kyle Busch** | #1 | Nevada | 212 |
| **Richard Petty** | #43 | North Carolina | 200 |


What is **population** of the problem?

the set of drivers {Dale, Denny,Jeff,Jimmie,Kyle,Richard}

What are **statistics variables** of the problem? (A *statistics variable* is an *aspect* of individuals of the population we would like to know.)

# 2.1 Population and Variables

**Example Problem.** We have the following information about NASCAR drivers

| Driver | Car | Home State | Races Won |
|---|---|---|---|
| **Dale Earnhardt** | #3 | North Carolina | 21 |
| **Denny Hamlin** | #11 | Florida | 63 |
| **Jeff Gordon** | #24 | California | 5 |
| **Jimmie Johnson** | #48 | California | 83 |
| **Kyle Busch** | #1 | Nevada | 212 |
| **Richard Petty** | #43 | North Carolina | 200 |

What is **population** of the problem?

- *Population* of this problem: is the set of drivers {Dale, Denny,Jeff,Jimmie,Kyle,Richard}

What are **statistics variables** of the problem?

- *car*: we want to know which car a driver drives
- *racesWin*: we want to know the number of races to win.
- *homeState*: we want to know the homestate of each driver

# 2.1 Population and Variables

**Example Problem.** We have the following information about NASCAR cars

| Car | Color | Driver | Speed (mph) |
|---|---|---|---|
| **#1** | Yellow | Kyle Busch | 215 |
| **#3** | Black | Dale Earnhardt | 200 |
| **#11** | White | Denny Hamlin | 205 |
| **#24** | Blue | Jeff Gordon | 185 |
| **#43** | Blue | Richard Petty | 195 |
| **#48** | White | Jimmie Johnson | 205 |

*Questions*.

- What is the color of car #1?
- What is the speed of car #11?

# 2.1 Statistics Variables as Functions

**Problem.** Consider NASCAR drivers again

*population*:
{Dale, Denny,Jeff,Jimmie,Kyle,Richard}
*statistics variable*: *car*

| Driver | Car | Home State | Races Won |
|---|---|---|---|
| **Dale Earnhardt** | #3 | North Carolina | 21 |
| **Denny Hamlin** | #11 | Florida | 63 |
| **Jeff Gordon** | #24 | California | 5 |
| **Jimmie Johnson** | #48 | California | 83 |
| **Kyle Busch** | #1 | Nevada | 212 |
| **Richard Petty** | #43 | North Carolina | 200 |

The statistics variable *car* in fact is a function mapping drivers to cars (see the table)!

- the car of Dale is car3
- the car of Denny is car 11
- …

| Driver | Car |
|---|---|
| **Dale Earnhardt** | #3 |
| **Denny Hamlin** | #11 |
| **Jeff Gordon** | #24 |
| **Jimmie Johnson** | #48 |
| **Kyle Busch** | #1 |
| **Richard Petty** | #43 |

# 2.1 Statistics Variables as Functions

The statistics variable *car* in fact is a function mapping drivers to cars (see the table)!

| Driver | Car |
|---|---|
| Dale Earnhardt | #3 |
| Denny Hamlin | #11 |
| Jeff Gordon | #24 |
| Jimmie Johnson | #48 |
| Kyle Busch | #1 |
| Richard Petty | #43 |

**Practice**: write an R expression (using named vectors) to represent the *car* function (you can associate an *R program variable* to it).

# 2.1 Statistics Variables as Functions

The statistics variable *car* in fact is a function mapping drivers to cars (see the table)!

| Driver | Car |
|---|---|
| Dale Earnhardt | #3 |
| Denny Hamlin | #11 |
| Jeff Gordon | #24 |
| Jimmie Johnson | #48 |
| Kyle Busch | #1 |
| Richard Petty | #43 |

**Practice**: write an R expression (using named vectors) to represent the *car* function (you can associate an *R program variable* to it).

```
car <- c("Dale" ="car3", "Denny" = "car11", "Jeff" = "car24",
"Jimmie" = "car48", "Kyle" = "car1", "Richard" = "car43")
```

# 2.1 Statistics Variables as Functions

The statistics variable *car* in fact is a function mapping drivers to cars (see the table)!

| Driver | Car |
|---|---|
| Dale Earnhardt | #3 |
| Denny Hamlin | #11 |
| Jeff Gordon | #24 |
| Jimmie Johnson | #48 |
| Kyle Busch | #1 |
| Richard Petty | #43 |

For the *car* function mapping a driver to their car.

- What is the *domain* of *car*?

# 2.1 Statistics Variables as Functions

The statistics variable *car* in fact is a function mapping drivers to cars (see the table)!

| Driver | Car |
|---|---|
| Dale Earnhardt | #3 |
| Denny Hamlin | #11 |
| Jeff Gordon | #24 |
| Jimmie Johnson | #48 |
| Kyle Busch | #1 |
| Richard Petty | #43 |

For the *car* function mapping a driver to their car.

- What is the *domain* of *car*?
  Drivers = {Dale, Denny, Jeff, Jimmie, Kyle, Richard}
- What is the *range* of function of *car*?

# 2.1 Statistics Variables as Functions

For the *car* function mapping a driver to their car.

- What is the *domain* of *car*?

  Drivers = {Dale, Denny,Jeff,Jimmie,Kyle,Richard}

- What is the *range* of function of *car*?

  Cars = {“car3”, “car11”,”car24”,”car48”,”car1”,”car43”}

- Write the signature of the function of *car*?

# 2.1 Statistics Variables as Functions

| Driver | Car |
|---|---|
| **Dale Earnhardt** | #3 |
| **Denny Hamlin** | #11 |
| **Jeff Gordon** | #24 |
| **Jimmie Johnson** | #48 |
| **Kyle Busch** | #1 |
| **Richard Petty** | #43 |

For the *car* function mapping a driver to their car.

- What is the *domain* of *car*?
  Drivers = {Dale, Denny, Jeff, Jimmie, Kyle, Richard}
- What is the *range* of function of *car*?
  Cars = {“car3”, “car11”,“car24”,“car48”, “car1”,“car43”}
- Write the signature of the function of *car*?
  *car*: Drivers -> Cars

# 2.1 Values of Variables

For the *car* function

- *domain* of *car* : *population* ={Dale, Denny,Jeff,Jimmie,Kyle,Richard}

- *range* of function of *car* : {“car3”, “car11”,”car24”,”car48”,”car1”,”car43”}

We usually call the values of the range of car the values of the statistics variable *car*. We also say that “car3” is a value of variable *car*.

**Definition (values of a variable)** The values of the range of a variable are also called the **values** of the **variable**.

# 2.1 Statistics Variables - Multiple Meanings

Note that we use the *statistic variable* to refer to a few different but related things

- An *attribute* of  individuals of a population
- The *function* mapping the individuals to the values of their attributes
- The *function name* of the function mapping the individuals to the values of their attributes

**Example** statistics variable *car* can refer to

- An attribute of a driver (In this case, *car* is simply a normal English word)
- The name of the function mapping drivers to their cars (A math (set theory) concept)
  *car*(dale) = ?
- The function mapping drivers to their cars. (A math (set theory) concept)

# 2.1 Statistics Variables - Multiple Meanings

**Example** statistics variable *car* can refer to

- An attribute of a driver.
- The name of the function mapping drivers to their cars *car*(dale) = ?
- The function mapping drivers to their cars (right table).

| Driver | Car |
|---|---|
| Dale Earnhardt | #3 |
| Denny Hamlin | #11 |
| Jeff Gordon | #24 |
| Jimmie Johnson | #48 |
| Kyle Busch | #1 |
| Richard Petty | #43 |

You need to associate a statistics variable to its right meaning in terms of context.

Now you may realize why we give a definition of *values of variables*.

# 2.1 Statistics Variables, Variables, R Program Variables

In contrast to *statistics variable*, the *variables* we use in algebra (math) have one meaning: referring to something.

An *R program variable* also refers to anything AND it means a memory unit in computer.

Unfortunately, people usually shorten all these into *variables*. You should be able to tell what they refer to (an algebra variable, a statistics variable or an R program variable).

# 2.2 Categorical and Quantitative Variables

Consider the NASCAR problem again.

We have

- *population*:
  {Dale, Denny,Jeff,Jimmie,Kyle,Richard}
- *statistics variable*:
  - *homeState*
  - *speed*

| Driver | Car | Home State | Races Won |
|---|---|---|---|
| **Dale Earnhardt** | #3 | North Carolina | 21 |
| **Denny Hamlin** | #11 | Florida | 63 |
| **Jeff Gordon** | #24 | California | 5 |
| **Jimmie Johnson** | #48 | California | 83 |
| **Kyle Busch** | #1 | Nevada | 212 |
| **Richard Petty** | #43 | North Carolina | 200 |

The statistics variable *speed* is called *quantitative* because the values of its range are numbers. The statistics variable *homeState* is called *categorical* because the values of its range are not numbers. These values (e.g., California, Florida) are usually called *categories*.

# 2.2 Categorical and Quantitative Variables

**Definition (Quantitative Variable. Categorical Variables)**

A statistics variable is **quantitative** if the values of its range are numbers.

A statistics variable is **categorical** if the values of its range are not numbers.

These values are usually called **categories**.

# 2.3.1 Frequency (Motivation)

**Problem.** On a test, Aaron got 90, Bill got 95, Cecilia got 90, Dina got 88, and Eric got 90.

**Questions**. How many students got 90 on the test? How many got 95?

We would call the number of students getting 90 the *frequency* of 90, or the *distribution* of students to grade 90. In statistics, these numbers are very interesting and important and thus have special names.

# 2.3.2 Frequency - Formal Definition

**Definition (frequency)**

The **frequency**, also called **absolute frequency**, of a value *v* of a variable is the number of individuals that the variable maps to *v*.

**Example.** In earlier example, for the variable *grade*, the frequency of value 90 of *grade* is 3, i.e., the number of students (i.e., individuals) who get 90.

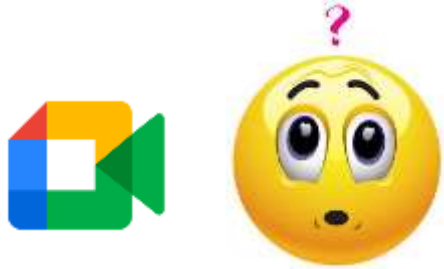
The frequency of a value can be understood as how frequently the individuals (as a whole) take this value for their attribute (variable).

# 2.3.2 Frequency - Set Builder Notation

**Definition (frequency)**

The **frequency**, also called **absolute frequency**, of a value $v$ of a variable is the number of individuals that the variable maps to $v$.

By the definition above, represent the frequency of $v$ of a variable $x$ using set builder notation?



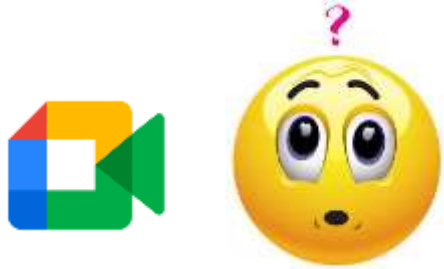
# 2.3.2 Frequency - Set Builder Notation

**Definition (frequency)**

The **frequency**, also called **absolute frequency**, of a value $v$ of a variable is the number of individuals that the variable maps to $v$.

By the definition above, represent the frequency of $v$ of a variable $x$ using set builder notation?

- We first get all students whose score is $v$:



# 2.3.2 Frequency - Set Builder Notation

**Definition (frequency)**

The **frequency**, also called **absolute frequency**, of a value *v* of a variable is the number of individuals that the variable maps to *v*.

By the definition above, represent the frequency of *v* of a variable *x* using set builder notation?

- We first get all students whose score is *v*:
  $\{individual: x(individual) = v\}$
- We then apply cardinality function to this set to get the frequency of *v*:

# 2.3.2 Frequency - Set Builder Notation

**Definition (frequency)**

The **frequency**, also called **absolute frequency**, of a value *v* of a variable is the number of individuals that the variable maps to *v*.

By the definition above, represent the frequency of *v* of a variable *x* using set builder notation?

- We first get all students whose score is *v*:
  $\{individual: x(individual) = v\}$ The set consisting of every *individual* such that its value understand statistics variable *x* is *v*.
- We then apply cardinality function to this set to get the frequency of *v*:
  $|\{individual: x(individual) = v\}|$

# 2.3 Frequency (Example)

| Driver | Home State |
|---|---|
| **Dale Earnhardt** | North Carolina |
| **Denny Hamlin** | Florida |
| **Jeff Gordon** | California |
| **Jimmie Johnson** | California |
| **Kyle Busch** | Nevada |
| **Richard Petty** | North Carolina |

Write an R expression (using named vectors) to represent the *Home State* function (you can associate an *R program variable* to it).

# 2.3 Frequency (Example)

| Driver | Home State |
|---|---|
| **Dale Earnhardt** | North Carolina |
| **Denny Hamlin** | Florida |
| **Jeff Gordon** | California |
| **Jimmie Johnson** | California |
| **Kyle Busch** | Nevada |
| **Richard Petty** | North Carolina |

Write an R expression (using named vectors) to represent the *Home State* function (you can associate an *R program variable* to it).

homeState <- c("Dale Earnhardt" = "North Carolina", "Denny Hamlin" = "Florida", "Jeff Gordon" = "California", "Jimmie Johnson" = "California", "Kyle Busch" = "Nevada", "Richard Petty" = "North Carolina")

Write an R expression(s) to find frequency of North Carolina (NS) of variable homeState?

# 2.3 Frequency (Example)

| Driver | Home State |
|---|---|
| **Dale Earnhardt** | North Carolina |
| **Denny Hamlin** | Florida |
| **Jeff Gordon** | California |
| **Jimmie Johnson** | California |
| **Kyle Busch** | Nevada |
| **Richard Petty** | North Carolina |

Write an R expression for the frequency of North Carolina (NC) of variable homeState?

Step1: the set (as a vector) of drivers from North Carolina:

# 2.3 Frequency (Example)

| Driver | Home State |
|---|---|
| **Dale Earnhardt** | North Carolina |
| **Denny Hamlin** | Florida |
| **Jeff Gordon** | California |
| **Jimmie Johnson** | California |
| **Kyle Busch** | Nevada |
| **Richard Petty** | North Carolina |

Calculate frequency of North Carolina (NC) of variable homeState?

Step1: the set (as a vector) of drivers from North Carolina:

ncDrivers <- names(which(homeState == “North Carolina”))

Step2: the number of elements in the set

# 2.3 Frequency (Example)

| Driver | Home State |
|---|---|
| **Dale Earnhardt** | North Carolina |
| **Denny Hamlin** | Florida |
| **Jeff Gordon** | California |
| **Jimmie Johnson** | California |
| **Kyle Busch** | Nevada |
| **Richard Petty** | North Carolina |

Calculate frequency of North Carolina (NC) of variable homeState?

Step1: the set (as a vector) of drivers from North Carolina:

ncDrivers <- names(which(homeState == “North Carolina”))

Step2: the number of elements in the set

length(ncDrivers)

# 2.4 Distribution (Motivation)

Recall statistics variable *type* of roller coasters:

| Roller coaster | Type |
|---|---|
| Wildfire | Wood |
| Skyline | Steel |
| Goliath | Wood |
| Helix | Steel |
| Banshee | Steel |
| Black Hole | Steel |

We know the frequency of each value of type:

- the frequency of wood is 2,
- the frequency of steel is 4.

Here we have a function mapping a value to its frequency! People called this function *distribution* of the variable.

Write an R expression to represent the distribution of variable *type*?

# 2.4 Distribution (Motivation)

Recall statistics variable *type* of roller coasters:

| Roller coaster | Type |
|---|---|
| Wildfire | Wood |
| Skyline | Steel |
| Goliath | Wood |
| Helix | Steel |
| Banshee | Steel |
| Black Hole | Steel |

We know the frequency of each value of type:

- the frequency of wood is 2,
- the frequency of steel is 4.

Here we have a function mapping a value to its frequency! People called this function *distribution* of the variable.

Write an R expression to represent the distribution of variable *type*?

```
disType <-c( "Wood"=2, "Steel"=4)
```

# 2.4 Distribution - Formal Definition

**Definition (distribution of a variable):**
The **distribution** of a variable is the function that maps each value of the variable to its frequency.

In statistics, most advanced concepts are based on the distribution of a variable.

# 2.5 Visualization of Distribution

People can visualize the distribution of a variable using

- **Bar graph** (for categorical variables)
- **Dotplot** (for categorical or quantitative variables)

# 2.5 Visualization of Distribution - Bar Graph

Function **`barplot`**(*dist*), where *dist* is an R expression whose value is a named vector, draws the **bar graph** of the name vector. Note *dist* has to be a distribution of a statistics variable.

**Practice**. Type the following R expressions in the editor, NOT the console, of repl.it.

```
distribution <- c("wood"= 2,"steel" = 4)
barplot(distribution)
```

- Then click the `Run` button above the editor
- Then in the left-most pane, click the `Rplots.pdf` (which holds the drawing produced by the barplot function in your R program)

# 2.5 R Program and Its Execution

**Definition (R Program)** A *sequence* of R expressions is called an **R program**.

To **execute** an R program is to evaluate the R expressions from top to bottom.

```
distribution <- c("wood" = 2, "steel" = 4)
barplot(distribution)
```

Meaning of **running or executing an R program** using example above:

- Evaluate the first expression. Its value is a named vector. (NOTE: no print is done here! In R console, after we hit enter, we tell the console to evaluate the expression AND print the value to the console)
- Evaluation of function barplot(...) causes the print of the graph you saw. Note this is called **side effect** of the function. It is **NOT** the output (recall output of a function in set theory and math) of the function. The *side effect of a function* is sth created by the function that affects the world outside the function.

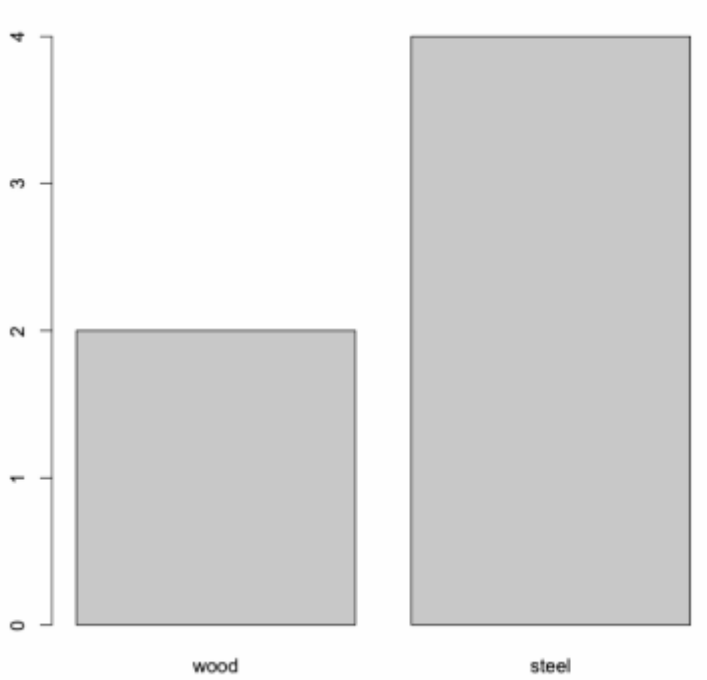
# 2.5 Visualization of Distribution - Bar Graph

A **bar graph** of the distribution of a categorical variable:

- x-axis - the *values* of the variable
- y-axis - the frequencies of the values of the variable
- There is a bar at each value, and its height is the frequency of the value.

Recall distribution

```
("wood" = 2, "steel" = 4)
```




Bar graph of the distribution of the variable *type*

# 2.5 Visualization of Distribution - dotplot

Dotplot is often used for quantitative (statistics) variables.

Function **dotchart**(*e*), where *e* is an R expression whose value is a *numerical* vector, draws the **dotplot** of the vector.

To draw a dotplot of the distribution of a quantitative variable, dotchart(...) function allows us to use statistics variable as direct input (i.e., we don't need to get the distribution of the variable as we did for drawing bar graph).

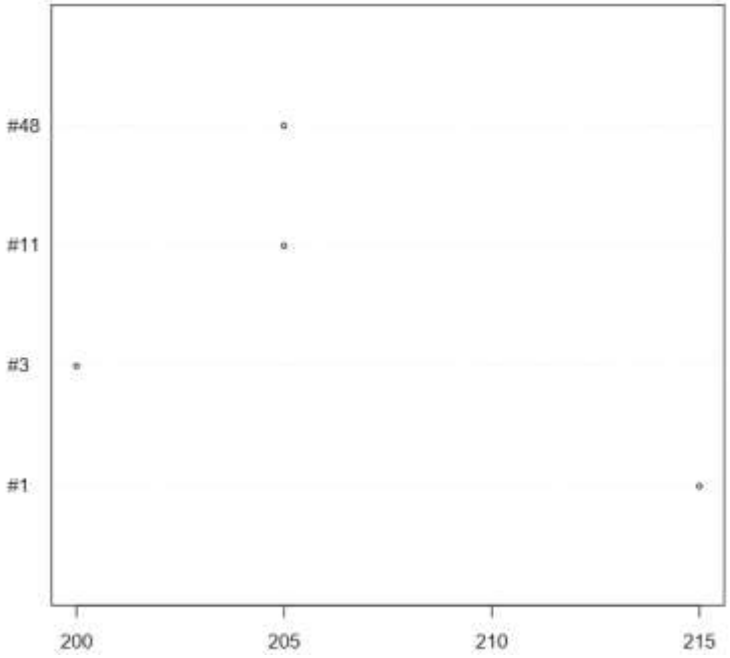
**Practice**. Recall the statistics variable *speed* of the population of cars. Type the following R program into your repl.it editor, run it and see the drawing.

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
dotchart(speed)
```

# 2.5 Visualization of Distribution - dotplot

A **dotplot** of the distribution of a categorical or quantitative variable:

- x-axis - the values of the variable
- y-axis - the names of the cars
- There is stack of dots for each value and the number of the dots in the stack is the frequency of this value




Dotplot of the distribution of the variable *speed*

# 2.6 Relative Frequency (Motivation)

Consider two populations *P1 = {Aaron , Bill, Cecilia}* and *P2 = {Alice, Bob, Claudia, Dina, Steve}*.

Both populations have a statistics variable *favoriteSport* with values *football* and *basketball*. The frequency of football is 2 for P1 (i.e., 2 students have football as their favorite sport), and 3 for P2.

Question: which population has "more" students like football?

- P2 has more students than P2 like football
- We do realize that the P1 has only 3 students while the P2 has 5. Overall, the proportion of students of the P1 like football is higher than that of P2.

We note that the second option above provides a way to make information more comparable across different populations. The proportion there is called *relative frequency*. For example, the relative frequency of football for P1 is (*frequency of football* / *population size*), i.e., 2/3 = 0.67. That of the 2nd population is:

# 2.6 Relative Frequency (Motivation)

Consider two populations *P1 = {Aaron , Bill, Cecilia}* and *P2 = {Alice', Bob', Claudia', Dina', Steve'}*.

Both populations have a statistics variable *favoriteSport* with values *football* and *basketball*. The frequency of football is 2 for P1 (i.e., 2 students have football as their favorite sport), and 3 for P2.

Question: which population has "more" students like football?

- P2 has more students than P1 like football
- We do realize that the P1 has only 3 students while the P2 has 5. Overall, the proportion of students of P1 like football is higher than that of P2.

We note that the second option above provides a way to make information more comparable across different populations. The proportion there is called *relative frequency*. For example, the relative frequency of football for P1 is (*frequency of football* / *population size*), i.e., 2/3 = 0.67. That of the 2nd population is: 3/5 = 0.6.

# 2.6 Relative Frequency - Formal Definition

**Definition (relative frequency)**

The **relative frequency** of a value *v* of a variable is the proportion of individuals that the variable maps to *v*, or

the **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Example.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

What is the *relative frequency* of value 205 of statistics variable *speed*?

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Example.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

What is the *relative frequency* of value 205 of statistics variable *speed*?

By **definition** of relative frequency,

- Frequency of 205: individuals have 205: {"#11", "#48"} whose cardinality is 2, i.e., frequency of 205 is 2.
- Population size: the population is all names of the statistics variable, i.e., {"#1", "#3", "#11", "48"}
- Relative frequency of 205 is (frequency of 205) / (population size), i.e., 2/4 = 0.5

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**

Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Using R console, write an R program to get the *relative frequency* of value 205 of variable *speed*.

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**

Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Using R console, write an R program to get the *relative frequency* of value 205 of variable *speed*.

Read definition carefully, to get *relative frequency*, what other information do we need?

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Using R console, write an R program to get the *relative frequency* of value 205 of variable *speed*.

Read definition carefully, to get *relative frequency*, what other information do we need?

Step 1: the *frequency* of value 205 of *speed*

Step 2: the population *size*: (it is the number of named indexes, i.e., the individuals)

Step 3: the relative frequency of value 205 of variable speed is *frequency*/*size*

# Review of Representing a Set Using R

Assume we have `grade <- c("Aaron" = 10, "Bill" = 9, "Cecilia" = 10, "Dina" = 11)`

Recall how we write R expressions to represent the set {*x*: *grade(x) = 10*}

- Get named logical vector indicating who is 10th grader
  by R expression: `is10grade <- grade == 10`

is10grade = (TRUE, FALSE, TRUE, FALSE)

*Aaron* *Bill* *Cecilia* *Dina*

- Get the numerical indices of *is10grade* where logical value is true
  by R expression: `tenGrade <- which(is10grade)`

tenGrade = ( 1, 3 )

*Aaron* *Cecilia*

- The set {*x: grade(x) = 10*} is the named indexes {Aaron, Cecilia}
  which can be obtained by R expression:
  `tenGraders <- names(tenGrade)`

tenGraders = (Aaron, Cecilia)
note: R function names does not give us the set, but a vector (for set).

- Although tenGraders is not a set, but it reflects the set well.
  For example, to get the cardinality of the set we simply use R expression
  `length(tenGraders)`

# Assignment

# Population and Variables

**Example Problem.** We have the following information about NASCAR cars

| Car | Color | Driver | Speed (mph) |
|---|---|---|---|
| **#1** | Yellow | Kyle Busch | 215 |
| **#3** | Black | Dale Earnhardt | 200 |
| **#11** | White | Denny Hamlin | 205 |
| **#24** | Blue | Jeff Gordon | 185 |
| **#43** | Blue | Richard Petty | 195 |
| **#48** | White | Jimmie Johnson | 205 |

*Questions.*

- What is the color of car #1?
- What is the speed of car #11?

What is the **population** of the problem?

What are **statistics variables** of the problem?

# Relative Frequency

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Using R console, write an R program to get the *relative frequency* of value 205 of variable *speed*.

Step 1: the frequency of value 205 of *speed*

# Relative Frequency

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Write an R program to get the *relative frequency* of value 205 of variable *speed*.

Step 2: the population size: (it is the number of named indexes, i.e., the individuals)

# Relative Frequency

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**

Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Write an R program to get the *relative frequency* of value 205 of variable *speed*.

Step 3: the relative frequency of value 205 of variable speed is

# Review Assignment

# Population and Variables

**Example Problem.** We have the following information about NASCAR cars

| Car | Color | Driver | Speed (mph) |
|---|---|---|---|
| **#1** | Yellow | Kyle Busch | 215 |
| **#3** | Black | Dale Earnhardt | 200 |
| **#11** | White | Denny Hamlin | 205 |
| **#24** | Blue | Jeff Gordon | 185 |
| **#43** | Blue | Richard Petty | 195 |
| **#48** | White | Jimmie Johnson | 205 |

*Questions.*

- What is the color of car #1?
  - Yellow
- What is the speed of car #11?
  - 205

What is the **population** of the problem?

What are **statistics variables** of the problem?

# Population and Variables

**Example Problem.** We have the following information about NASCAR cars

| Car | Color | Driver | Speed (mph) |
|---|---|---|---|
| **#1** | Yellow | Kyle Busch | 215 |
| **#3** | Black | Dale Earnhardt | 200 |
| **#11** | White | Denny Hamlin | 205 |
| **#24** | Blue | Jeff Gordon | 185 |
| **#43** | Blue | Richard Petty | 195 |
| **#48** | White | Jimmie Johnson | 205 |

*Questions.*

- What is the color of car #1?
  - Yellow
- What is the speed of car #11?
  - 205

What is the **population** of the problem?

- {car1, car3, car11,car24,car43,car48}

What are **statistics variables** of the problem?

# Population and Variables

**Example Problem.** We have the following information about NASCAR cars

| Car | Color | Driver | Speed (mph) |
|---|---|---|---|
| **#1** | Yellow | Kyle Busch | 215 |
| **#3** | Black | Dale Earnhardt | 200 |
| **#11** | White | Denny Hamlin | 205 |
| **#24** | Blue | Jeff Gordon | 185 |
| **#43** | Blue | Richard Petty | 195 |
| **#48** | White | Jimmie Johnson | 205 |

*Questions.*

- What is the color of car #1?
  - Yellow
- What is the speed of car #11?
  - 205

What is the **population** of the problem?

- {car1, car3, car11,car24,car43,car48}

What are **statistics variables** of the problem?

- *color*: we want to know what color of the car is
- *speed*: we want to know the speed of cars
- Driver:we want to know the name of each driver

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Using R console, write an R program to get the *relative frequency* of value 205 of variable *speed*.

Step 1: the frequency of value 205 of *speed*

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Write an R program to get the *relative frequency* of value 205 of variable *speed*.

Step 1: the frequency of value 205 of *speed*

```
frequency <- length(names(which(speed == 205)))
```

Step 2: the population size: (it is the number of named indexes, i.e., the individuals)

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Write an R program to get the *relative frequency* of value 205 of variable *speed*.

Step 1: the frequency of value 205

```
frequency <- length(names(which(speed == 205)))
```

Step 2: the population size: (it is the number of named indexes, i.e., the individuals)

```
length(speed)
```

Step 3: the relative frequency of value 205 of variable speed is

# 2.6 Relative Frequency - Formal Definition

The **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

**Practice.**
Assume we have the *speed* variable:

```
speed <- c( "#1" = 215, "#3" = 200, "#11" = 205, "#48" = 205)
```

Write an R program to get the *relative frequency* of value 205 of variable *speed*.

Step 1: the frequency of value 205

```
frequency <- length(names(which(speed == 205)))
```

Step 2: the population size: (it is the number of named indexes, i.e., the individuals)

```
length(speed)
```

Step 3: the relative frequency of value 205 of variable speed is

```
frequency / length(speed)
```

# Some New Materials

# 2.7 Relative Frequency - Given Distribution

Recall: the **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

Also we mentioned that statistics concepts are based on distribution (but not the variable content as function). Consider the example of types of roller coaster. The distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| Wood | 2 |
| Steel | 4 |

Now assume we only know the distribution of *type* but don't know the content of the variable. Now by the definition of relative frequency, can you find the relative frequency of value wood? Write an expression for relative frequency:

# 2.7 Relative Frequency - Given Distribution

Recall: the **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

Also we mentioned that statistics concepts are based on distribution (but not the variable content as function). Consider the example of types of roller coaster. The distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| Wood | 2 |
| Steel | 4 |

Now assume we only know the distribution of *type* but don't know the content of the variable. Now by the definition of relative frequency, can you find the relative frequency of value *wood*? Write an expression for relative frequency:

- By definition, we need frequency of *wood*:
- By definition, we need to know population size: recall frequency of a value is the number of individuals taking that value: _?__ individuals take *wood* value, _?__ individuals take *Steel* value, and one individual can take only one value. Hence, population size is the sum of all frequencies: ___?__.

# 2.7 Relative Frequency - Given Distribution

Recall: the **relative frequency** of a value *v* of a variable is the frequency of *v* divided by population size of the variable.

Also we mentioned that statistics concepts are based on distribution (but not the variable content as function). Consider the example of types of roller coaster. The distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| Wood | 2 |
| Steel | 4 |

Now assume we only know the distribution of *type* but don't know the content of the variable. Now by the definition of relative frequency, can you find the relative frequency of value *wood*? Write an expression for relative frequency:

- By definition, we need frequency of *wood*: 2
- By definition, we need to know population size: recall frequency of a value is the number of individuals taking that value: two individuals take *wood* value, four individuals take *Steel* value, and one individual can take only one value. Hence, population size is the sum of all frequencies: 2+4 = 6.

# 2.7 Relative Frequency - Given Distribution

**Practice.** Consider the distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| **Wood** | 2 |
| **Steel** | 4 |

Write an R program to find the relative frequency of value *Steel*:

# 2.7 Relative Frequency - Given Distribution

**Practice.** Consider the distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| **Wood** | 2 |
| **Steel** | 4 |

Write an R program to find the relative frequency of value *Steel*:

- Write R expression to represent the distribution as a named vector and associate it to a R program variable:

# 2.7 Relative Frequency - Given Distribution

**Practice.** Consider the distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| **Wood** | 2 |
| **Steel** | 4 |

Write an R program to find the relative frequency of value *Steel*:

- Write R expression to represent the distribution as a named vector and associate it to a R program variable: `distribution <- c("Wood" = 2, "Steel" = 4)`
- Write an R expression to represent the frequency of *Steel*:

# 2.7 Relative Frequency - Given Distribution

**Practice.** Consider the distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| **Wood** | 2 |
| **Steel** | 4 |

Write an R program to find the relative frequency of value *Steel*:

- Write R expression to represent the distribution as a named vector and associate it to a R program variable: `distribution <- c("Wood" = 2, "Steel" = 4)`
- Write an R expression to represent the frequency of *Steel*: `freq <- distribution["Steel"]`
- Write an R expression to represent the population size:

# 2.7 Relative Frequency - Given Distribution

**Practice.** Consider the distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| Wood | 2 |
| Steel | 4 |

Write an R program to find the relative frequency of value *Steel*:

- Write R expression to represent the distribution as a named vector and associate it to a R program variable: `distribution <- c("Wood" = 2, "Steel" = 4)`
- Write an R expression to represent the frequency of *Steel*: `freq <- distribution["Steel"]`
- Write an R expression to represent the population size: `size <- sum(dis`
- Write an R expression to find the relative frequency of *Steel*:

# 2.7 Relative Frequency - Given Distribution

**Practice.** Consider the distribution of the *type* variable of roller coasters is as shown in the table.

| types | frequency |
|---|---|
| **Wood** | 2 |
| **Steel** | 4 |

Write an R program to find the relative frequency of value *Steel*:

- Write R expression to represent the distribution as a named vector and associate it to a R program variable: `distribution <- c("Wood" = 2, "Steel" = 4)`
- Write an R expression to represent the frequency of *Steel*: `freq <- distribution["Steel"]`
- Write an R expression to represent the population size: `size <- sum(distribution)`
- Write an R expression to find the relative frequency of *Steel*: `freq/size`