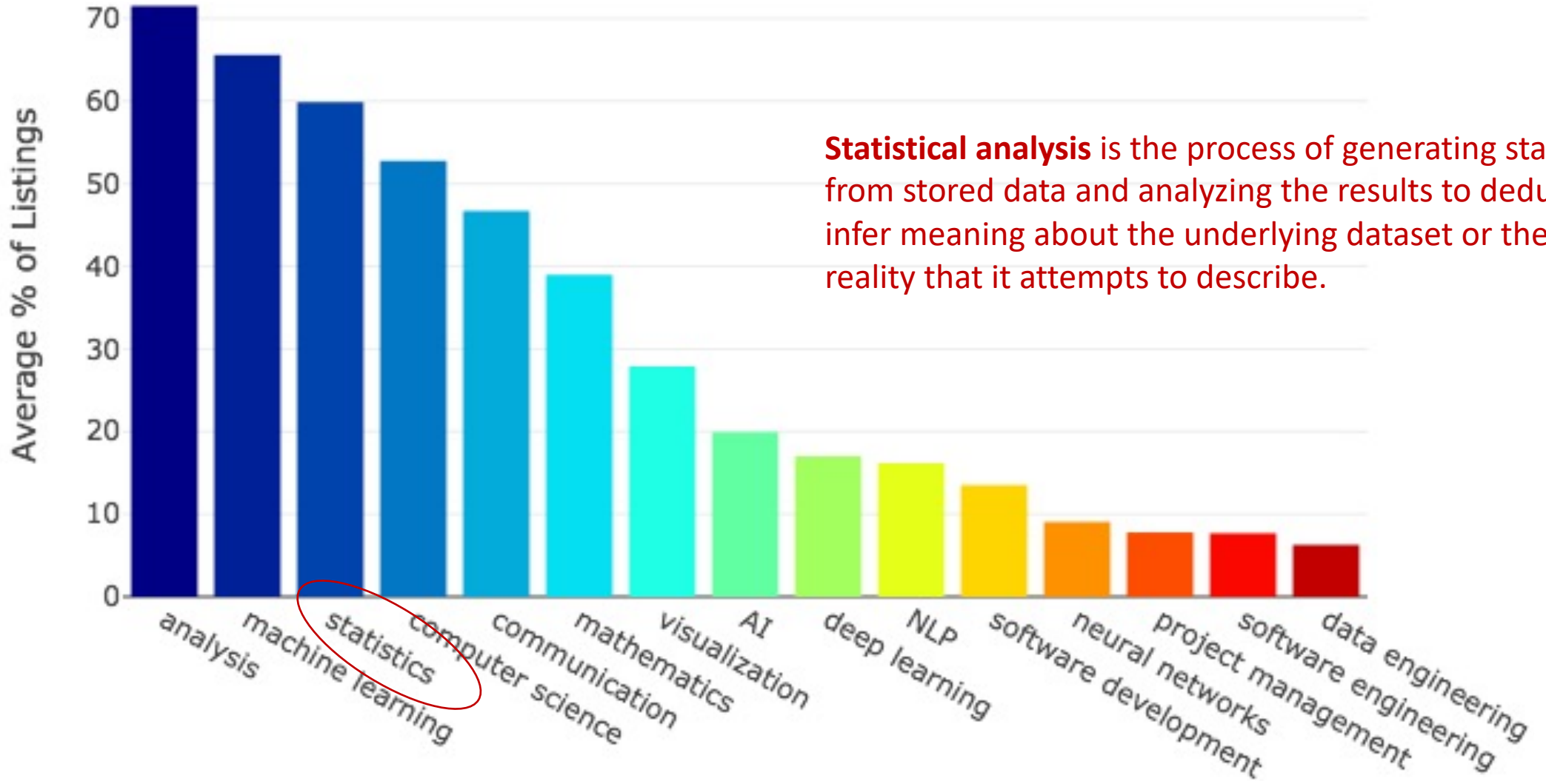
Statistics for Data Science

with Examples

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

-- Josh Wills --

General Skills in Data Scientist Job Listings



Statistical analysis is the process of generating statistics from stored data and analyzing the results to deduce or infer meaning about the underlying dataset or the reality that it attempts to describe.

https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-4a4a8db896db

DATA SCIENCE SKILL SET

BASIS 〉 CORE 〉 ENABLEMENT

TECHNICAL SKILLS

- CODING SKILLS
- HANDLING DATA
- COMPUTATIONAL TOOLS
- BASIC SOFTWARE DEVELOPMENT
- BIG DATA
- HIGH PERFORMANCE COMPUTING
- PARALLEL COMPUTING

TECH-SAVVY

ANALYTICAL SKILLS

- ADVANCED STATISTICS & INFERENCE
- MODELLING & SIMULATION
- MACHINE LEARNING
- COMPUTER SCIENCE
- ADVANCED MATH
- DATA VISUALIZATION
- EXPERIMENT DESIGN
- RESEARCH EXPERTISE

SCIENTIFIC PROBLEM SOLVING

BUSINESS SKILLS

- EVALUATION AND DEVELOPMENT OF BUSINESS CASES
- PROJECT MANAGEMENT
- BUSINESS PROCESSES
- CHANGE MANAGEMENT
- COMMUNICATION SKILLS
- LEADERSHIP SKILLS

APPETITE FOR BUSINESS PROBLEMS

40 Statistics Interview Problems and Answers for Data Scientists

1. How do you assess the statistical significance of an insight?
2. Explain what a long-tailed distribution is and provide three examples of relevant phenomena that have long tails. Why are they important in classification and regression problems?
3. What is the Central Limit Theorem? Explain it. Why is it important?
4. What is the statistical power?
5. Explain selection bias (with regard to a dataset, not variable selection). Why is it important? How can data management procedures such as missing data handling make it worse?

https://towardsdatascience.com/40-statistics-interview-problems-and-answers-for-data-scientists-6971a02b7eee

Learning Objectives:

1. To explain the importance of statistical knowledge for data scientist.
2. To review the types and steps of statistical data analysis.
3. To discuss 7 key ideas from statistics that permeate data science
    i. Samples and populations
    ii. Sample statistics
    iii. Bootstrap
    iv. Outliers
    v. Statistical models
    vi. Confounding and accounting for other factors
    vii. p-values
4. To elaborate the concepts with some R examples.
5. To describe A/B testing.

What is ‘Statistics’ ?

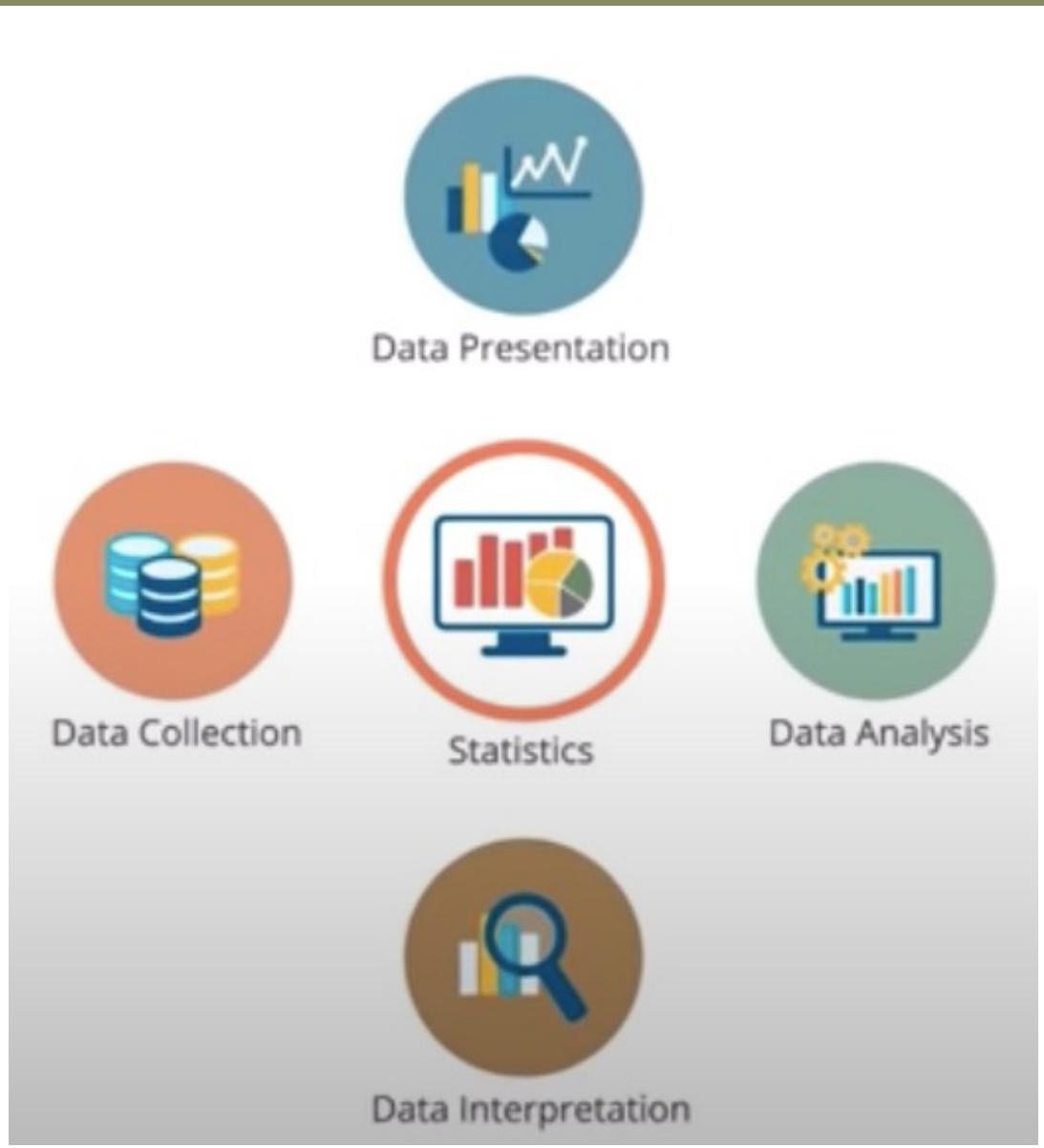
sta·tis·tics (st -t s t ks)

n.

1. *(used with a sing. verb)* The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.
2. *(used with a pl. verb)* Numerical data.

http://www.thefreedictionary.com/Statistics

- The science of learning from data.
- Statistical knowledge helps us use the proper methods to collect the data, employ the correct analyses, and effectively present the results.
- A crucial process behind how discoveries are made in science, make decisions based on data, and make predictions.
- Allows us to understand a subject much more deeply.

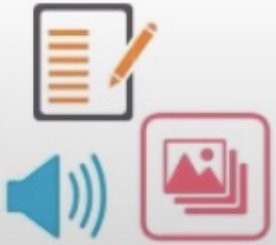


An analysis of any situation can be done in two ways:

Statistical analysis

It is the science of collecting, exploring, and presenting large amounts of data to identify patterns and trends.
It is also called quantitative analysis.

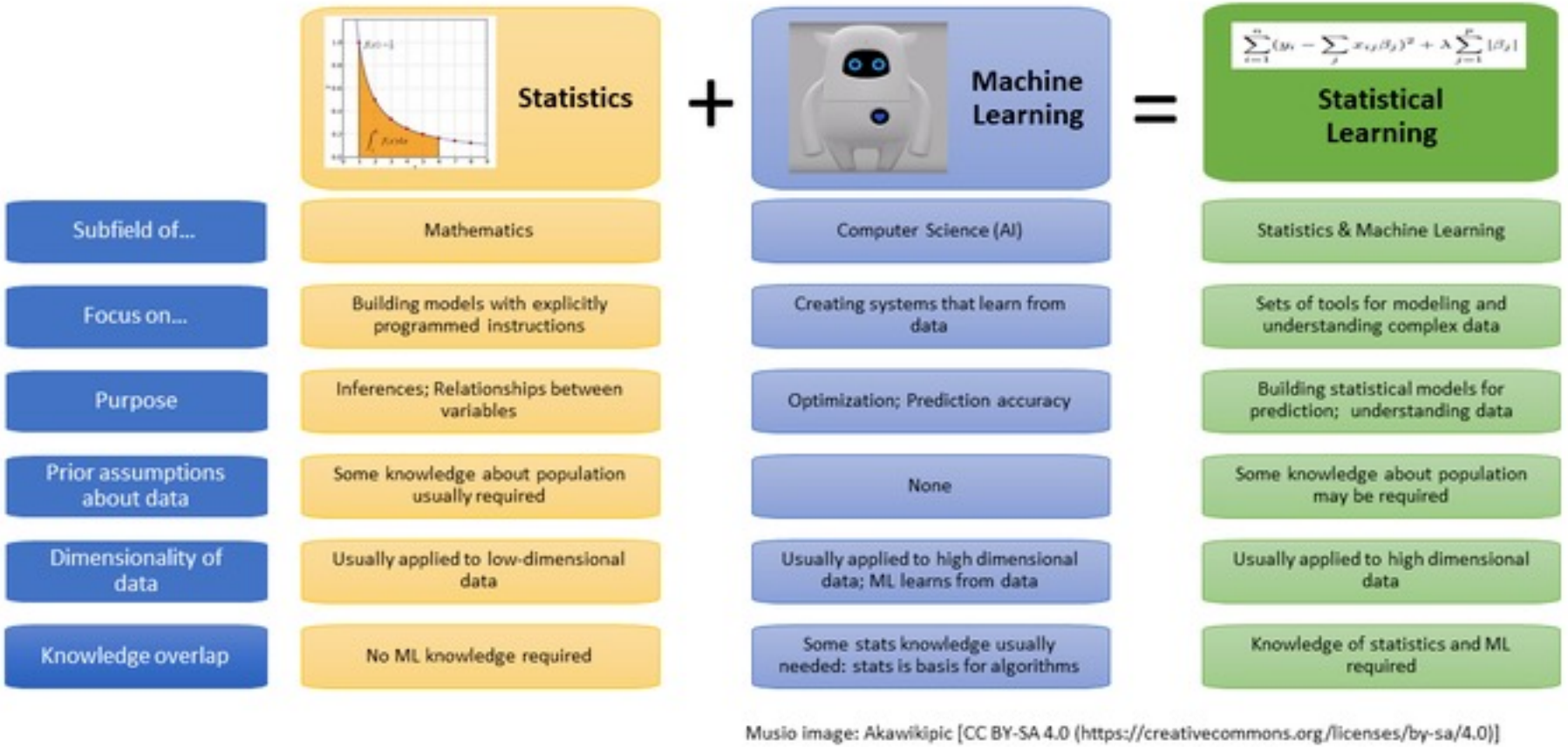
Non-statistical analysis



It provides generic information and includes text, sound, still images, and moving images.
It is also called qualitative analysis.

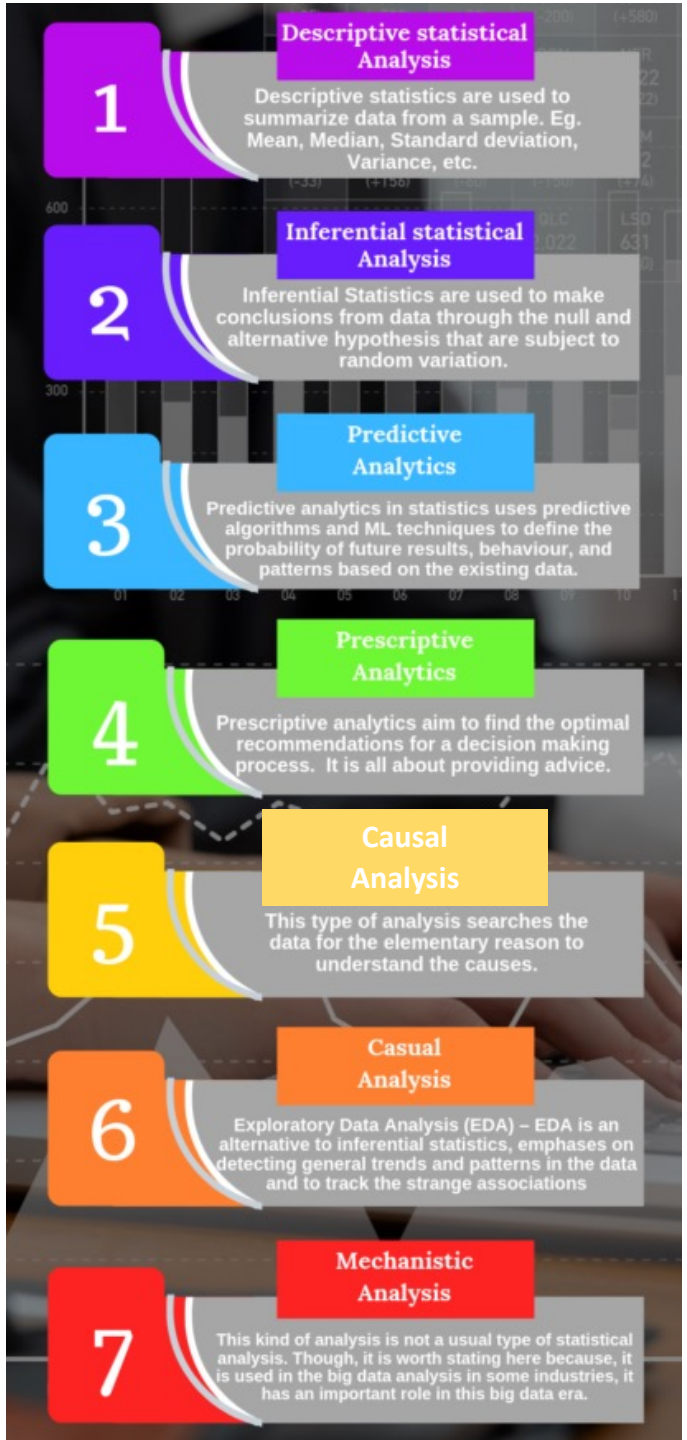
Statistical Learning

- **Statistical learning** theory is a framework for machine learning drawing from the fields of statistics and functional analysis.
- Statistical learning theory deals with the problem of finding a **predictive function** based on data.



Musio image: Akawikipic [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)]

https://www.datasciencecentral.com/profiles/blogs/machine-learning-vs-statistics-in-one-picture



Statistical data analysis can be simplified into 5 steps, as follows:

1. The primary step involves the identification of the nature of the data to be analyzed.
2. Secondly, explore the association between the data and the underlying population in the study.
3. Build a suitable model to summarize the data and proceed for further analysis.
4. Check the validity of the model and take decisions about the hypotheses.
5. Explore predictive analysis to run situations that will guide us for future actions.

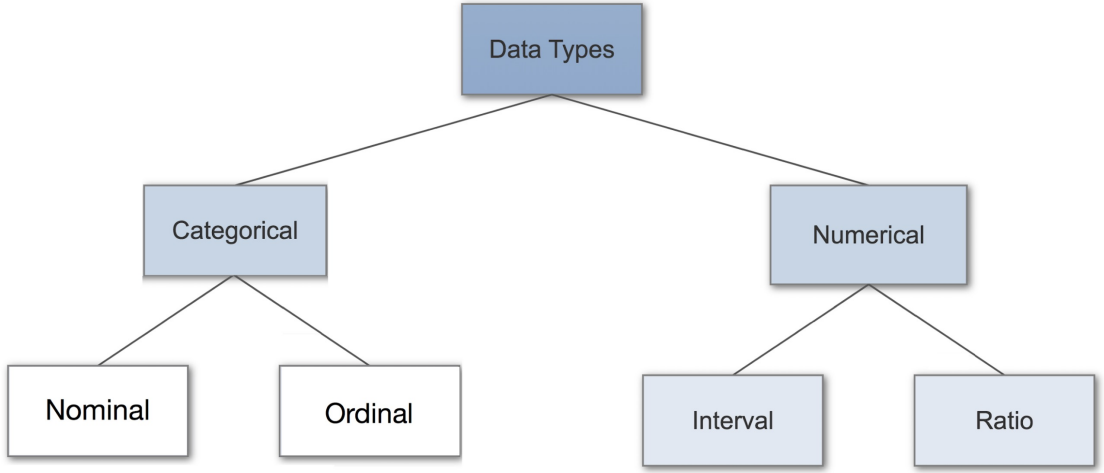
Credit: Statswork.com

Inference Versus Prediction

- There are situations where we are interested in understanding the way that **Y** is affected as **X** change.
- In this situation we wish to estimate **f**, but our goal is not necessarily to make predictions for Y.
- Here we are more interested in understanding relationship between X and Y. Now **f** cannot be treated as a black box, because we need to know its exact form.
- This is **inference**.

- In situations where a set of inputs X are readily available, but the output Y is not known, we often treat **f** as black box (not concerned with the exact form of **f**), as long as it yields accurate predictions for Y .
- This is **prediction**.

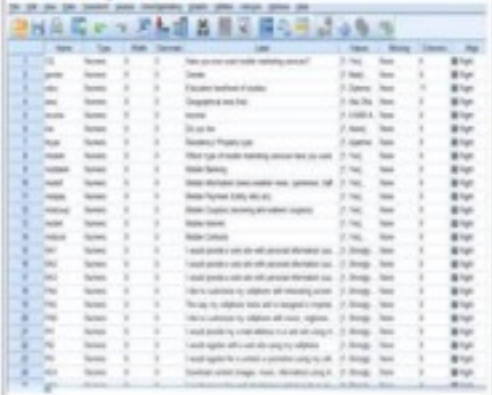
Types of Data



Nominal

These are the values/observations with no natural ordering.

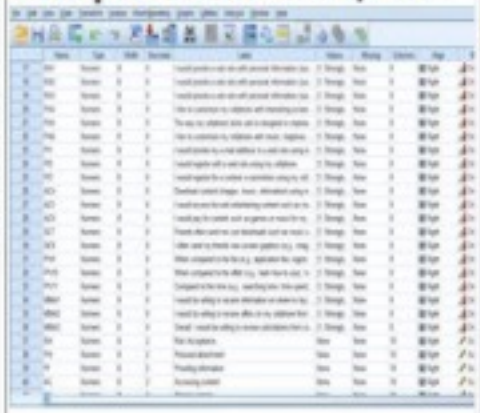
Examples: Gender, eye colour, etc.



Ordinal

These are the values or observations can be put in order or can be ranked or containing a rating scale. You can order and count these variables but it cannot be measured.

Example: Likert scale, etc.



Discrete

Values or observations that can be counted as separate and distinct. It can take only particular values. Examples: number of pens in a box; number of students in a class, etc.

From below table explains the number of birds and animals in each village

	Chaitanya Nagar	Malua	Panchgram	Bakrihawar	Gobindapur	Shibsarai
	n(%)					
Cow	11	5	5	8	9	
Buffalo	-	-	-	-	-	
Bull	1	-	-	-	-	
Goat	16	8	7	8	12	
Hen	14	11	9	7	10	
Duck	11	7	4	6	8	

Continuous

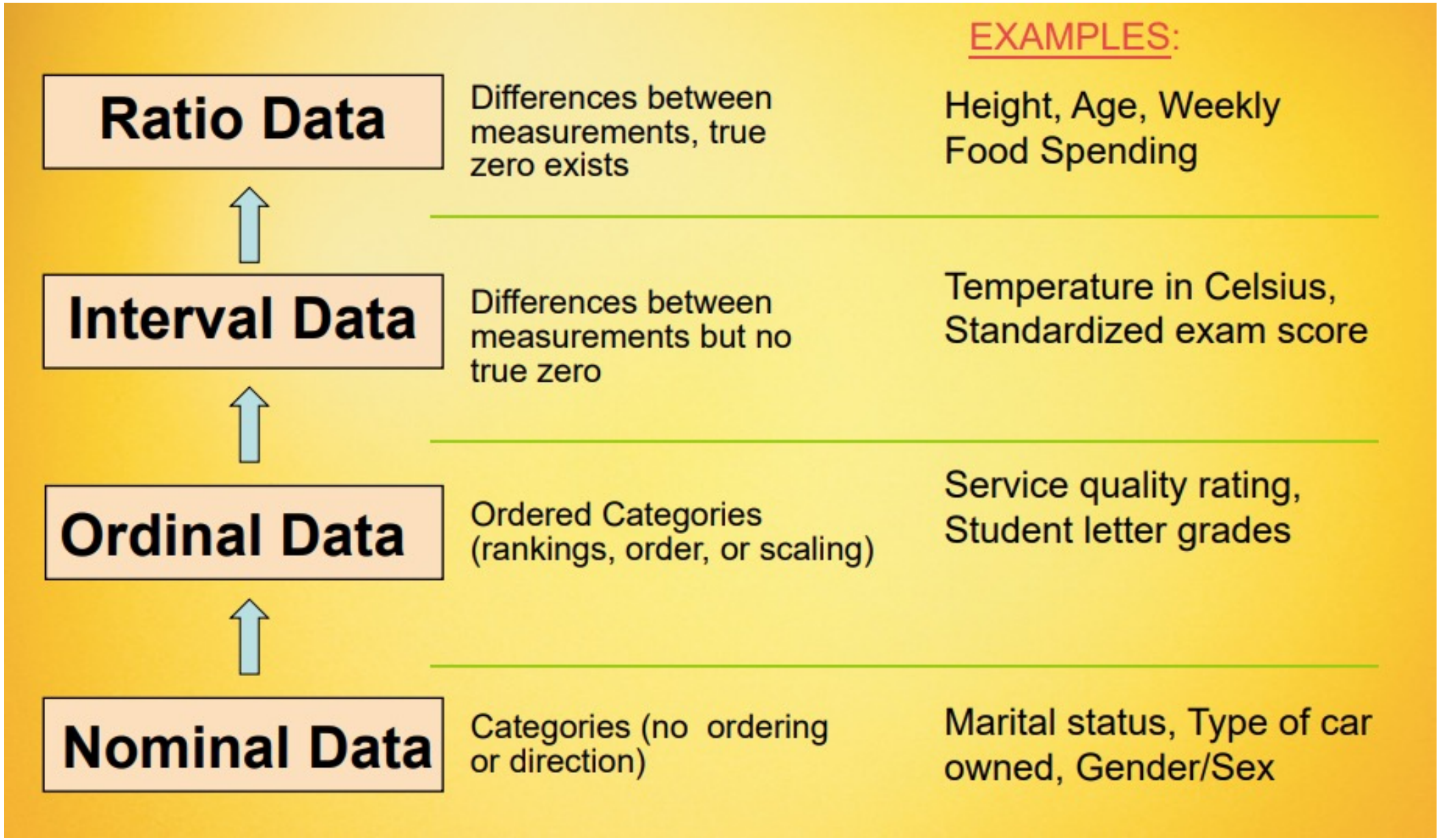
Values that can be measured are considered as continuous data. They can be further divided into two types: Finite and Infinite. And, it can take any value from minus infinity to plus infinity. Examples: height, time and temperature.

Table shows the Height, Weight, BMI and BP of the respondents

	Mean	SD	Maximum	Minimu
Height (cm)	156.4	14.6	180.4	12
Weight (kg)	72.7	9.3	82.2	5
Body Mass Index (kg/m^2)	30.6	7.7	55.8	1
Blood Pressure (mmHg)	134.5	3.1	139.0	13

SD-Standard Deviation

Measurement Scales



Predictability of Data

Data are predictable

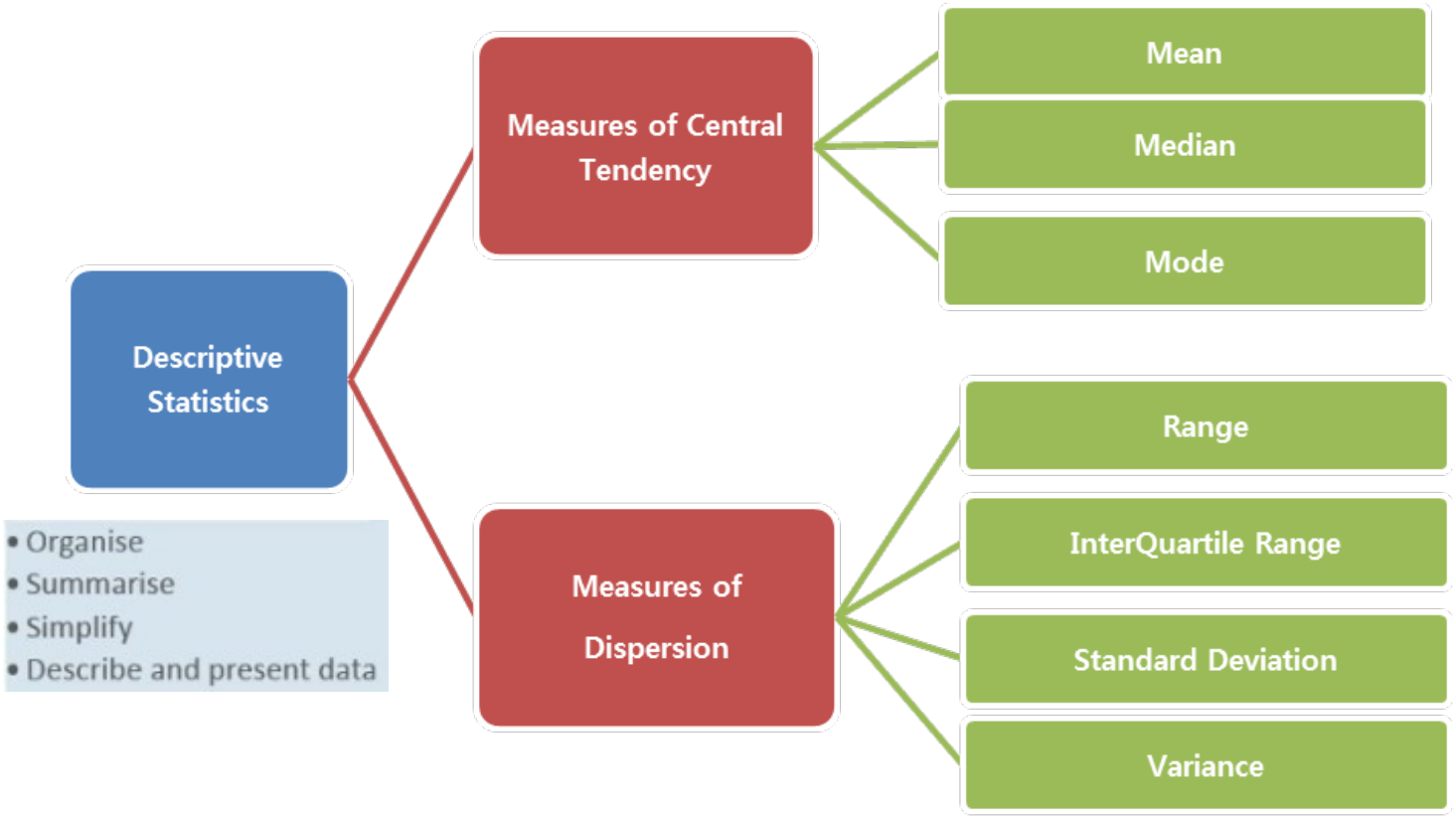
- on the basis of a set of **features** (e.g. diet or clinical measurements)
- from a set of (observed) **training data** on these features

For a set of **objects** (e.g. people).

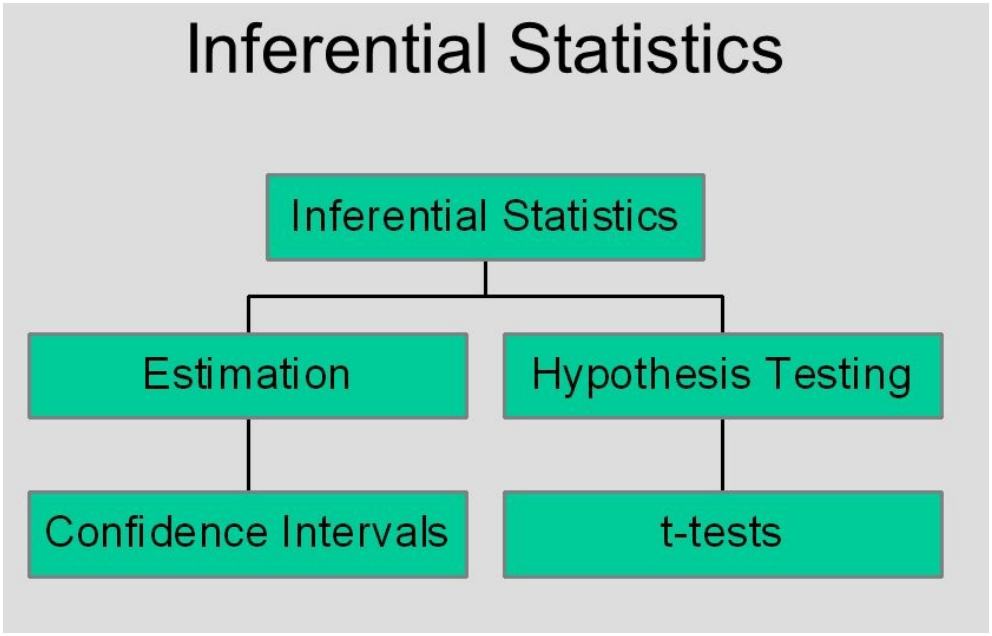
- **Inputs** for the problems are also called **predictors** or **independent variables**
- **Outputs** are also called **responses** or **dependent variables**

The prediction model is called a **learner** or **estimator** (Schätzer).

- **Supervised learning**: learn on outcomes for observed features
- **Unsupervised learning**: no feature values available



- Organise
- Summarise
- Simplify
- Describe and present data



- Generalise from samples to populations
- Hypothesis testing
- Make predictions

What type of statistic (descriptive or inferential) would you employ to answer the research questions below?
Select the appropriate category for each question.

DESCRIPTIVE STATISTICS	QUESTION	INFERENTIAL STATISTICS
	We would like to know how many university students experience high stress levels.	
	We would like to know whether female students experience higher stress levels than male students.	
	We would like to know what study strategies are used by first year university students.	
	We would like to know whether there is are a relationship between university students' study strategies and their academic results.	

Key ideas from statistics that permeate data science:

1. Samples and populations
2. Sample statistics
3. Bootstrap
4. Outliers
5. Statistical models
6. Confounding and accounting for other factors
7. p-values

Test Your Statistical Knowledge

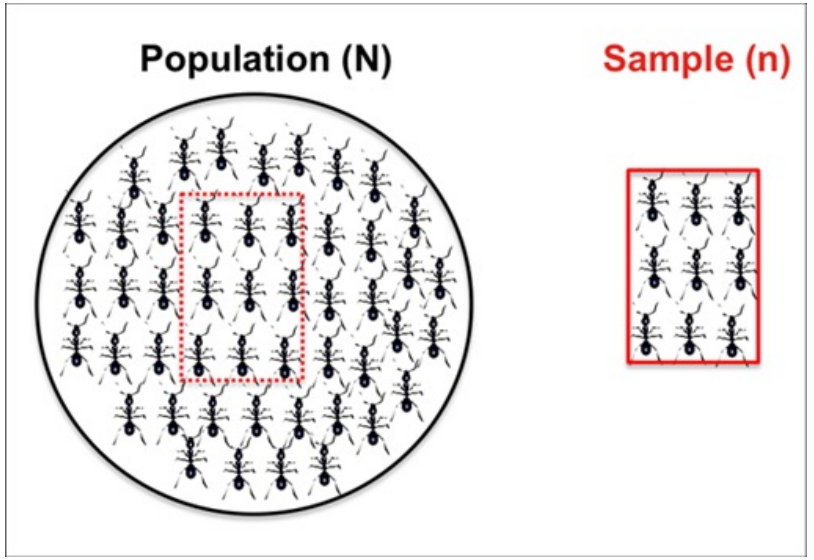
You are conducting a survey of the people in Malaysia to find out how popular the racket sports are. You randomly choose people to call, and make 1,000 phone calls to people scattered across the country.
In this study, what is the statistics term for **THE PEOPLE IN MALAYSIA**, and what is the statistics term for **THE PEOPLE YOU CALLED**?

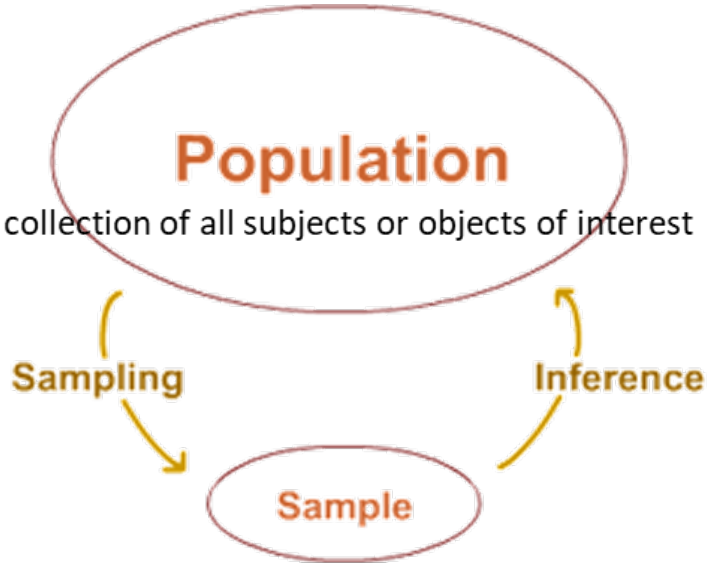
Because it would be impossible or inconvenient to call every single person in the country (the entire POPULATION) to find out how popular the racket sports are, you take a SAMPLE by only calling some of them. Provided the sample is randomly selected, and provided the sample is representative of the population as a whole, you can draw statistical conclusions from the data you obtain.

1. Samples and populations

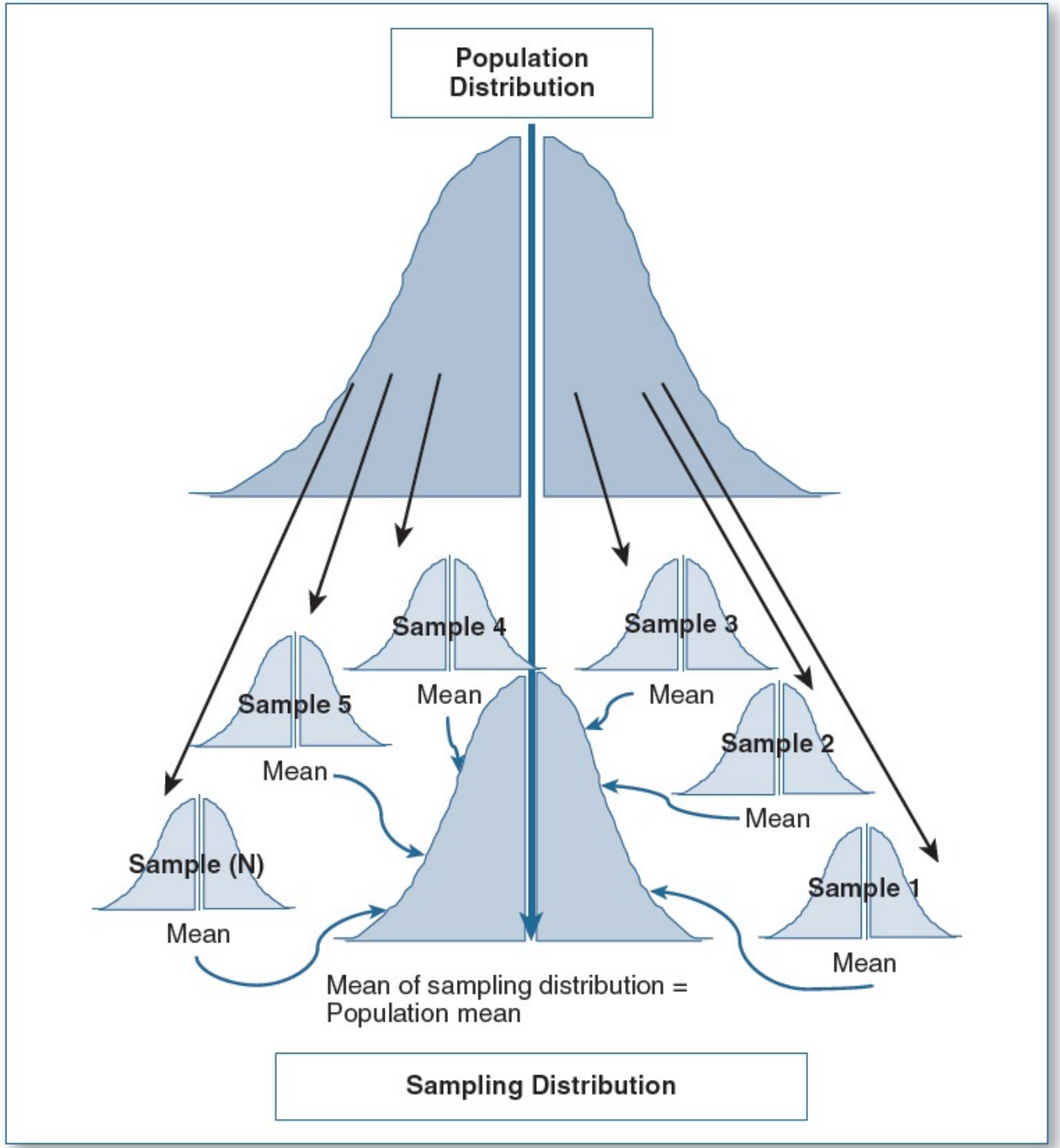
- The connection between the data we've got (the sample) and the population.
- Basic concepts of statistics –

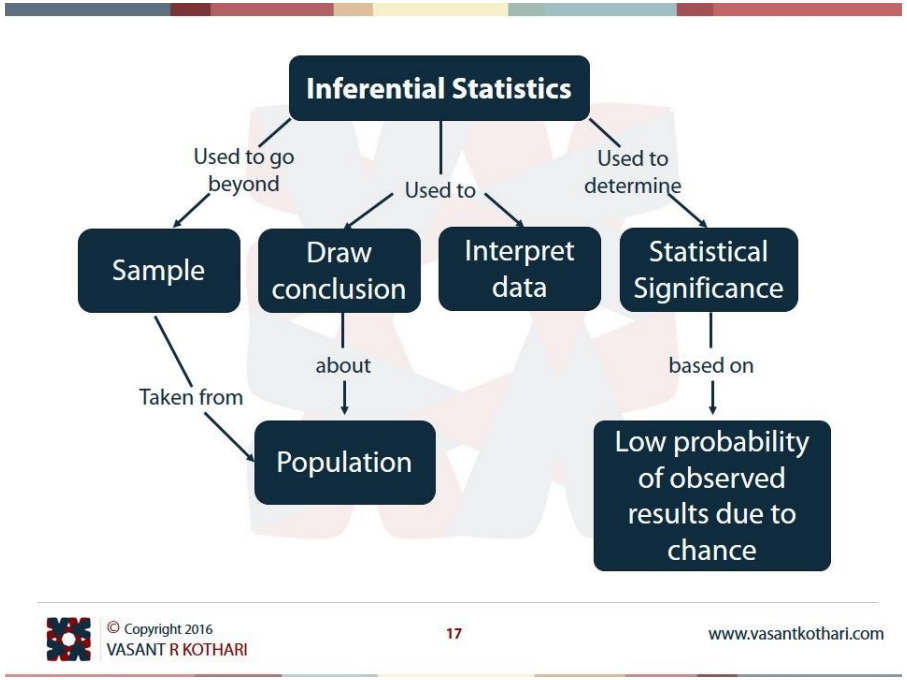
https://www.slideshare.net/tkjainbkn/basic-concepts-of-statistics-393112





subset of the population used to make inferences about the characteristics of the population


Population Distribution
Sample 4
Sample 3
Sample 5
Mean
Mean
Sample 2
Mean
Sample (N)
Mean
Sample 1
Mean
Mean of sampling distribution = Population mean
Sampling Distribution


Inferential Statistics
Used to go beyond
Used to
Used to determine
Sample
Draw conclusion
Interpret data
Statistical Significance
about
based on
Taken from
Population
Low probability of observed results due to chance
© Copyright 2016
VASANT R KOTHARI
17
www.vasantkothari.com

Sampling from Population - An Example

Suppose you were asked to help develop **a travel policy for business travelers** based in **New York City**.

- Imagine that the traveler has a meeting in San Francisco (airport code SFO) at a specified time **t**.
- The **policy** to be formulated will say **how much earlier than t an acceptable flight should arrive in order to avoid being late to the meeting due to a flight delay**.

The set of **336,776** flights in 2013 in the **nycflights13** package, which gives airline delays from New York City airports in 2013.

nycflights13 package → https://github.com/hadley/nycflights13 OR https://moderndive.github.io/moderndive_labs/static/previous_versions/v0.4.0/2-getting-started.html

nycflights13 package

This package contains five datasets saved as “data frames” with information about all domestic flights departing from New York City in 2013, from either Newark Liberty International (EWR), John F. Kennedy International (JFK), or LaGuardia (LGA) airports :

- `flights` : information on all 336,776 flights
- `airlines` : translation between two letter IATA carrier codes and names (16 in total)
- `planes` : construction information about each of 3,322 planes used
- `weather` : hourly meteorological data (about 8710 observations) for each of the three NYC airports
- `airports` : airport names and locations

Dataset

Tibbles are data frames

```
1  install.packages("nycflights13")
2  library(nycflights13)
3  data(flights)
4  flights
5
```

```
flights
```

```
# A tibble: 336,776 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     1     1      517            515         2      830            819
 2  2013     1     1      533            529         4      850            830
 3  2013     1     1      542            540         2      923            850
 4  2013     1     1      544            545        -1     1004           1022
 5  2013     1     1      554            600        -6      812            837
 6  2013     1     1      554            558        -4      740            728
 7  2013     1     1      555            600        -5      913            854
 8  2013     1     1      557            600        -3      709            723
 9  2013     1     1      557            600        -3      838            846
10  2013     1     1      558            600        -2      753            745
# ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
#   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
#   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Simulate this situation by drawing a sample from the population of flights into SFO.

First set up the population.

```
install.packages("dplyr")
library(dplyr)
install.packages("nycflights13")
library(nycflights13)
SF <- flights %>%
   filter(dest == "SFO",!is.na(arr_delay))
```

Working with just a sample from the population, where sample size is n = 25

```
set.seed(101)
Sample25 <- SF %>%
  sample_n(size = 25)
```

We are interested in the **arrival delay**. Focus on the variable **arr_delay**, and compute some useful statistics.

```
require(mosaic)
favstats( ~ arr_delay, data = Sample25)

 min  Q1 median Q3 max  mean   sd  n missing
 -50 -23     -7  4 124 -2.96 35.3 25       0
```

Notice that most of delays are less than 50 minutes, but the maximum delay is 124 minutes

The **maximum delay** is 124 minutes which is about 2 hours.

So, should the travel policy be that the traveler should plan on arriving in SFO at least 2 hours ahead?

Naive policy: *book a flight that is scheduled to arrive at least 124 minutes before.*

What was the actual worst delay in 2013 if we compute on the complete set of flights?

favstats() function in the mosaic package provides a concise summary of many useful statistics.

Naive policy: book a flight that is scheduled to arrive at least 124 minutes before.

How well the “naïve policy” will work?

Let us see how well this policy works. Since in our imaginary situation we have the entire population, we can get an answer of this. In particular, we can derive how often the travelers from NYC will be late for a meeting in San Francisco, when the naive policy is used.

```
SF %>%
  mutate(is.late = arr_delay > 124) %>%
  summarize(prop.late = mean(is.late))
```

```
## # A tibble: 1 x 1
##    prop.late
##        <dbl>
## 1 0.02937827
```

This seems reasonable, if being late or missing a meeting with 2.9% of chance is acceptable.

What was the actual worst delay in **2013** if we compute on the complete set of flights?

Working with the **population**, where the size is N = **336,776**

```
favstats( ~ arr_delay, data = SF)

 min  Q1 median Q3  max mean   sd     n missing
 -86 -23     -8 12 1007 2.67 47.7 13173       0
```

The **maximum delay** is 1007 minutes which is about 17 hours.

Notice that *the results from the sample are **different** from the results for the population*.

This suggest that to avoid missing a meeting, one should travel the day before the meeting. Ok, but it is

- costly in terms of lodging, meals etc.
- No guarantee that there will never be a delay of more than 17 hours.

Sampling Error

- The sampling error reflects the fact that the result we get from our sample is not going to be exactly equal to the result we would have got if we had been able to measure the entire population. And each possible sample we could take would give a different result.

A **practical travel policy** will trade off small probabilities of being late against the savings in cost and traveler's time. For instance, you might judge it acceptable to be late just **2%** of the time—a **98%** chance of being on time.

The **98th percentile** of the arrival delays in our data sample:

```
qdata( ~ arr_delay, p = 0.98, data = Sample25)

      p quantile
   0.98    87.52
```

The calculation is easy, but how good is the answer?

A delay of **88 minutes**.

Suppose we used the 90-minute travel policy. How well would that have worked in achieving our intention to be late for meetings only 2% of the time? With the population data in hand, it's easy to answer this question.

```
tally( ~ arr_delay < 90, data = SF, format = "proportion")

arr_delay < 90
   TRUE   FALSE
0.9514 0.0486
```

The 90-minute policy would miss its mark **5%** of the time, much worse than we intended. To correctly hit the mark 2% of the time, we will want to increase the policy from 90 minutes to what value?

With the **population**, calculate the **98th percentile** of the **arrival delays**:

```
qdata( ~ arr_delay, p = 0.98, data = SF)

     p quantile
0.98   153.00
```

It should have been about **150 minutes**, not 90.
But in many important real–world settings, we do not have access to the population data. We have only our **sample**.

How can we use our sample to judge whether the result we get from the sample is going to be good enough to meet the 98% goal?
And if it's not good enough, how large should a sample be to give a result that is likely to be good enough?

This is where the concepts and methods from statistics come in.

2. Sample Statistics

- Characteristics of the sample such as the sample mean, the sample variance, and the sample proportion are called **sample statistics**.
- It can be used to provide **estimates** of the corresponding **population parameters**.
  - e.g. % of voters in an opinion poll in Malaysia who think the Government is doing a good job to control inflation.
  - Null and alternative hypotheses are statements about population parameters.
- More examples on next slide.

Heights of women, aged 25 - 29

Suppose you measured the heights of a random sample of 100 women aged 25-29 years and the calculated

sample mean = 165 cm
sample standard deviation s = 5 cm

What can you conclude about the heights of all women in this population aged 25-29 years?

Supposing any bias is negligibly small, the population mean is approximately 165 cm. But how close is the approximation? How might the estimate have varied if a different random sample had been selected?

In **parameter estimation**, we assume that the distribution of the variable we are interested in is adequately described by a distribution with one or more (unknown) parameters. We attempt to estimate the population parameter using the sample data.

To emphasize the difference between sample and population, the parameters we wish to estimate are called **population parameters**.

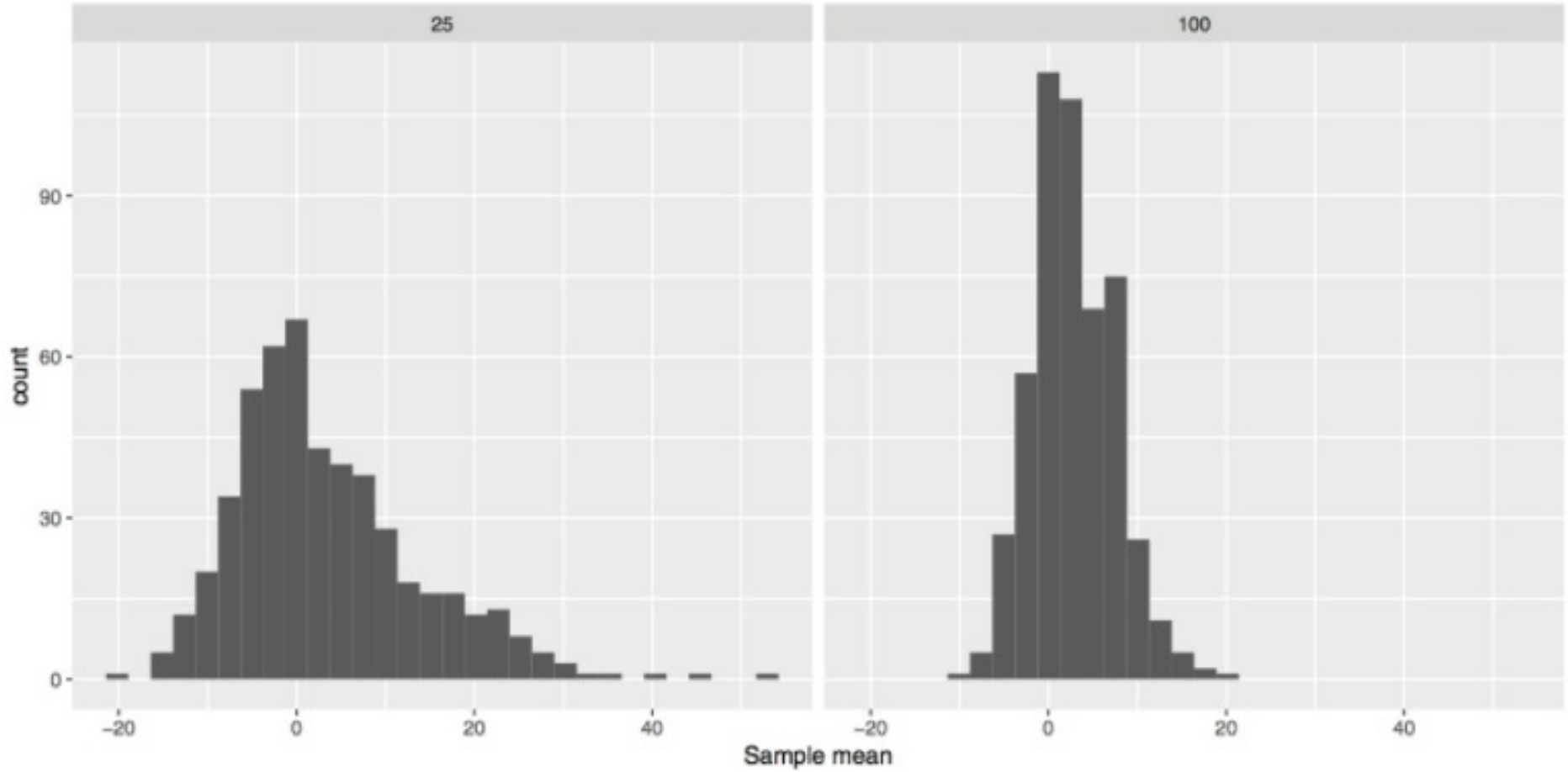
Estimates of the population parameters obtained from a sample are called **sample statistics** (or sample estimates).

Sample statistics

How to define **reliability**?

- The **sample size** is the **number of cases in the sample**, usually denoted with n. In the above, the sample size is n = 25.
- The **sampling distribution** is the collection of the sample statistic from all of the trials. We carried out 500 trials here, but the exact number of trials is not important so long as it is large.
- The **shape** of the sampling distribution is worth noting. Here it is a little skewed to the right.
- The **standard error** is the **standard deviation of the sampling distribution**. It describes the width of the sampling distribution. For the trials calculating the sample mean in samples with n = 25, the standard error is 10.3 minutes.

We had access to the population data and so we could find the sampling distribution by repeatedly sampling from the population.



Comparing the two sampling distributions, one with **n = 25** and the other with **n = 100** shows some patterns that are generally true for statistics such as the mean:

- Both sampling distributions are centered at the same value.

- A **larger sample size** produces a **standard error that is smaller**. That is, a **larger sample size is more reliable** than a smaller sample size. You can see that the standard deviation for n = 100 is one-half that for n = 25. As a rule, the standard error of a √ sampling distribution scales as $1/\sqrt{n}$.
- For **large sample sizes n**, the shape of the sampling distribution tends to **bell-shaped**. In a bit of archaic terminology, this shape is often called the **normal distribution**. Indeed, the distribution arises very frequently in statistics, but there is nothing abnormal about any other distribution shape.

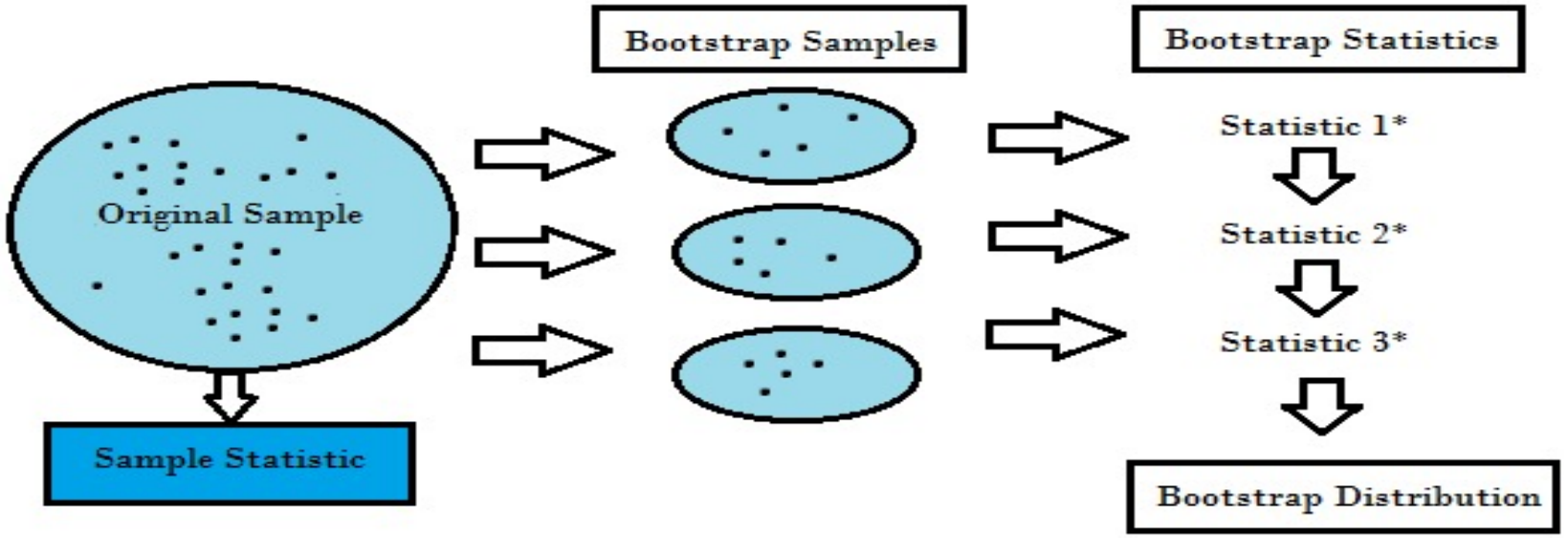
The **reliability of a sample statistic** is typically measured by:

- the **mean** of the statistic (mean of the sampling distribution). Better if it is closer to the truth (the population mean) .
- the **standard error** of the statistic (standard of the sampling distribution). Better if it is small.

http://pitt.edu/~jub69/material/Lec15_SamplingDist.html

3. Bootstrap

A bootstrap sample is a smaller sample that is "bootstrapped" from a larger sample. **Bootstrapping** is a **type of resampling** where large numbers of smaller samples of the same size are repeatedly drawn, with **replacement**, from a single original sample.

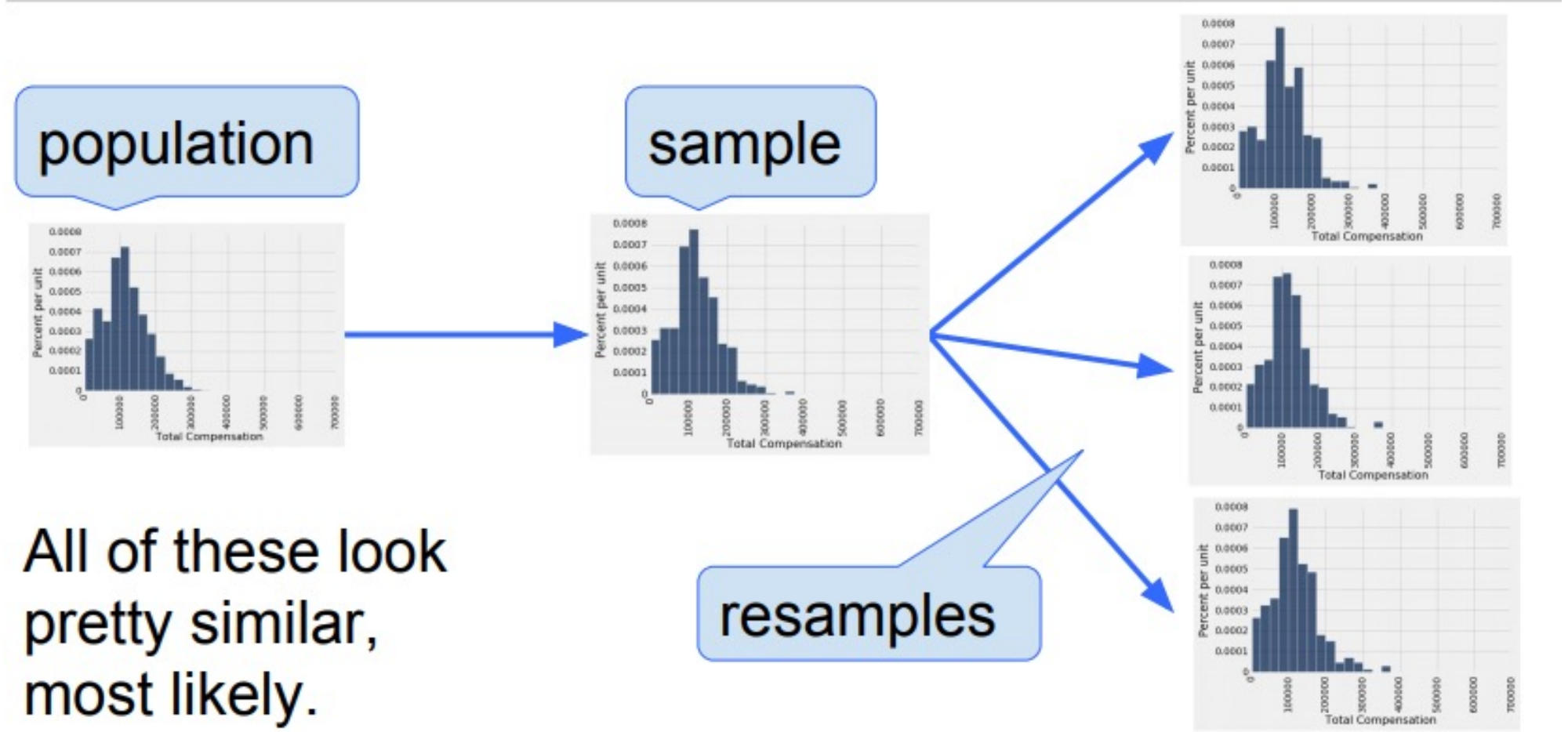


Bootstrap sampling

https://youtu.be/tTZybQTE0dw

Bootstrapping is loosely based on the **law of large numbers**, which states that if you **sample over and over again**, your **data should approximate the true population data**.

Bootstrapping is a way of replicating a sample so that you get a sample that is similar but most likely not exactly the same as the original sample.



Law of Large Numbers

- It is a theorem that **describes the result of performing the same experiment a large number of times**.
- This theorem forms the basis of **frequency-style thinking**.
- It says that the sample mean, the sample variance and the sample standard deviation converge to what they are trying to estimate.
- Example: roll a dice, expected value is 3.5. For a large number of experiments, the average converges to 3.5.

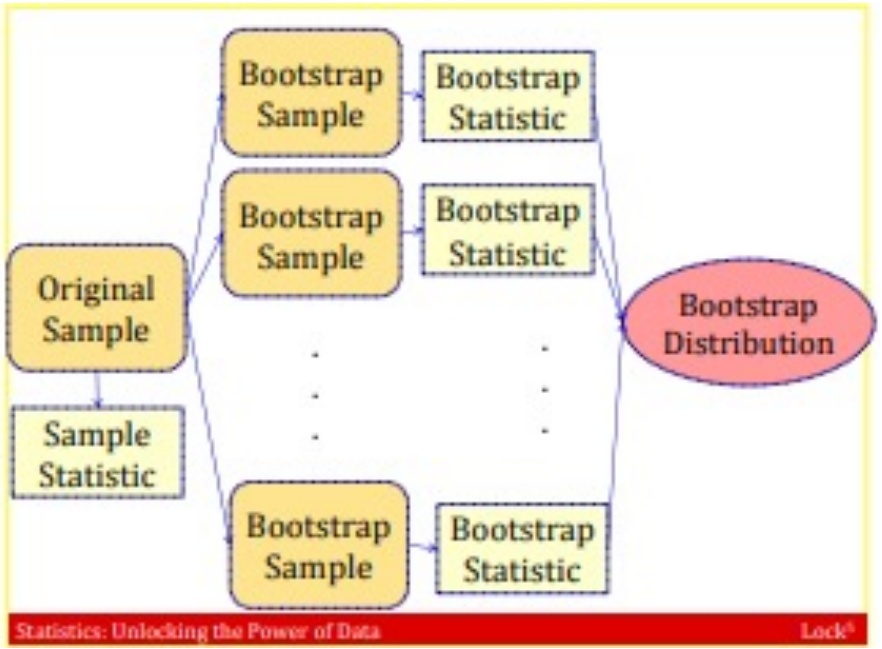
Bootstrapping is random sampling from the sample with replacement.

Law of large numbers

Find likelihood of "Heads"

P(Heads) = 0.5

Number of flips	Relative frequency	Percent
10	0.4 to 0.6	66
100	0.49 to 0.51	92
1,000	0.499 to 0.501	97
10,000	0.4999 to 0.5001	99



Bootstrap Sample

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21, 22

a) Yes
b) No

22 is not a value from the original sample

Statistics: Unlocking the Power of Data

Bootstrap Sample

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 19, 20, 21

a) Yes
b) No

Bootstrap samples must be the same size as the original sample

Statistics: Unlocking the Power of Data

Bootstrap Sample

Your original sample has data values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 18, 19, 20, 21

Same size, could be gotten by sampling with replacement

a) Yes
b) No

Statistics: Unlocking the Power of Data

Bootstrap Sample

You have a sample of size $n = 50$. You sample with replacement 1000 times to get 1000 bootstrap samples.

What is the sample size of each bootstrap sample?

Bootstrap samples are the same size as the original sample

(a) 50
(b) 1000

Statistics: Unlocking the Power of Data

Bootstrap Distribution

You have a sample of size $n = 50$. You sample with replacement 1000 times to get 1000 bootstrap samples.

How many bootstrap statistics will you have?

(a) 50
(b) 1000

One bootstrap statistic for each bootstrap sample

Statistics: Unlocking the Power of Data

Credit to source from:
https://www2.stat.duke.edu/courses/Fall12/sta101.002/Sec3-34.pdf

Bootstrap

We only have **one sample** and **not the entire population**.

- The **bootstrap** is a statistical method that allows us to approximate the sampling distribution even without access to the population.
- The logical leap involved in the bootstrap is to think of our sample itself as if it were the population.
- Just as in the previous examples we drew many samples from the population, now we will **draw many new samples from our original sample**. This process is called **resampling**: drawing a new sample from an existing sample.
- When sampling from a **population**, we would of course make sure **not to duplicate** any of the cases. When **resampling**, however, we do allow such **duplication**.

No duplication

```
Small <- sample_n(SF, size = 3, replace = FALSE)
```

```
# A tibble: 3  7
   year month   day dep_time sched_dep_time dep_delay arr_time
  <int> <int> <int>    <int>          <int>     <dbl>    <int>
1  2013     4    27     1653           1700        -7     1952
2  2013     5    14     1810           1800        10     2104
3  2013     5    16     1729           1732        -3     2133
```

Resampling from `Small` is done by setting the `replace` argument to `TRUE`, which allows the sample to include duplicates.

```
Small %>% sample_n(size = 3, replace = TRUE)
```

```
# A tibble: 3  7
   year month   day dep_time sched_dep_time dep_delay arr_time
  <int> <int> <int>    <int>          <int>     <dbl>    <int>
1  2013     5    16     1729           1732        -3     2133
2  2013     5    16     1729           1732        -3     2133
3  2013     5    16     1729           1732        -3     2133
```

Using bootstrapping to find the reliability of the mean arrival time calculated on a sample of size 200.

```
n <- 200
Orig_sample <- SF %>% sample_n(size = n, replace = FALSE)
```

Now, with the original sample in hand, we can draw a resample and calculate the mean arrival delay.

```
mean( ~ arr_delay,
      data = sample_n(Orig_sample, size = n, replace = TRUE))

[1] -2.2
```

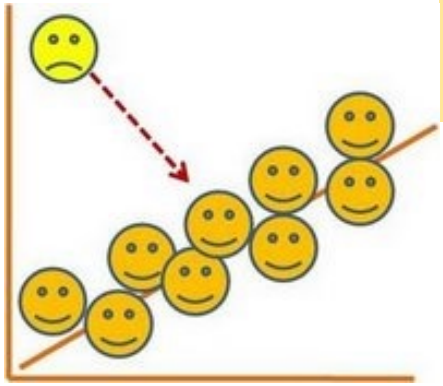
By repeating this process many times, we'll be able to see how much variation there is from sample to sample:

```
Bootstrap_trials <- do(500) * mean( ~ arr_delay,
  data = sample_n(Orig_sample, size = n, replace = TRUE))
favstats( ~ mean, data = Bootstrap_trials)

   min    Q1 median     Q3  max  mean   sd   n missing
 -9.04 -3.98  -2.25 -0.564 4.57 -2.28 2.37 500       0
```

We can compare this to a (hypothetical) sample of size $n = 1,000$ from the original SF flights.

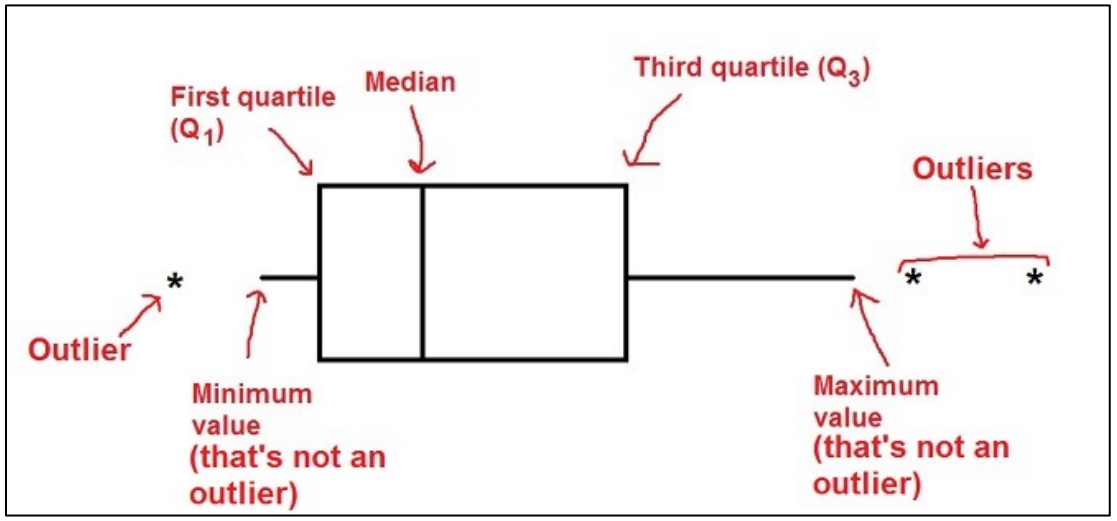
```
Trials_200 <- do(500) *
  mean( ~ arr_delay, data = sample_n(SF, size = n, replace = FALSE))
favstats( ~ mean, data = Trials_200)

   min    Q1 median   Q3  max mean   sd   n missing
 -5.64 0.241   2.29 4.51 13.3 2.47 3.11 500       0
```

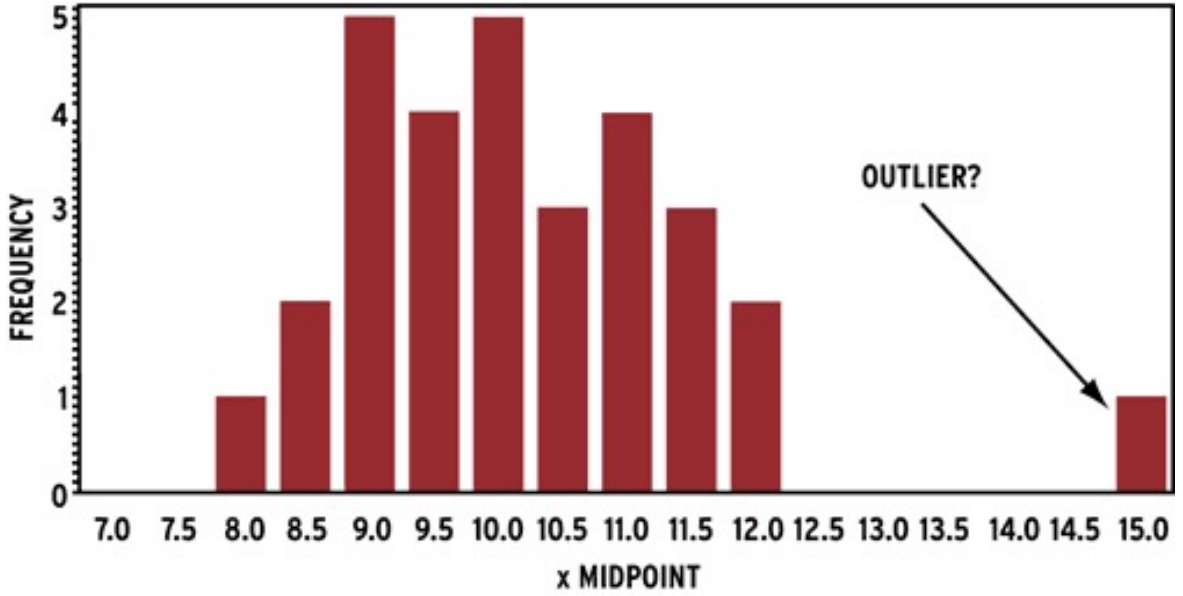
- What's remarkable here is that the **standard error** calculated in this way, 2.4 minutes, is a **reasonable approximation** to the standard error of the sampling population calculated in the previous section (3.1 minutes).
- The distribution of values in the **bootstrap trials** is called the **bootstrap distribution**. It's not exactly the same as the sampling distribution, but for moderate to large sample sizes it has been **proven to approximate** those aspects of the sampling distribution that we care most about, such as the standard error.



4. Outliers

An observation that lies an abnormal distance from other values in a random sample from a population.





The most common ways to treat outlier values –

1) To change the value and bring in within a range.
2) To just remove the value.

Outliers

One place where more data is helpful is in identifying unusual or extreme events: **outliers**.

Suppose we consider any flight delayed by seven hours (420 minutes) or more as an extreme event.

```
SF %>%
  filter(arr_delay >= 420) %>%
  select(month, day, dep_delay, arr_delay, carrier)

# A tibble: 7  5
  month   day dep_delay arr_delay carrier
  <int> <int>     <dbl>     <dbl>   <chr>
1    12     7       374       422      UA
2     7     6       589       561      DL
3     7     7       629       676      VX
4     7     7       653       632      VX
5     7    10       453       445      B6
6     7    10       432       433      VX
7     9    20      1014      1007      AA
```

Arrive extra early in July and to avoid VX flight.

Reminder on Outliers

- Outliers can often tell us interesting things.
- How they should be handled depends on their cause.
- Outliers due to data irregularities or errors should be fixed.
- Other outliers may yield important insights.
- Outliers should never be dropped unless there is a clear rationale. If outliers are dropped this should be clearly reported.

5. Statistical Model

A **statistical model** is a special class of mathematical model.

- What distinguishes a statistical model from other mathematical models is that a statistical model is **non-deterministic**. Thus, in a statistical model specified via mathematical equations, some of the variables do not have specific values, but instead have probability distributions; i.e. some of the variables are **stochastic** (something that was randomly determined).

There are three purposes for a statistical model,

- Predictions
- Extraction of information
- Description of stochastic structures

Model: A representation for a purpose. Blueprints, dolls, model of airplanes.

- **Mathematical model**: a description of a system using mathematical concepts and language.
  - can take many forms, including dynamical systems, statistical models, differential equations, or game theoretic models.
- **Statistical model**: A class of mathematical model founded on data.
  - embodies a set of assumptions concerning the generation of some **sample** data, and similar data from a larger **population**.
  - What distinguishes a statistical model from other mathematical models is that a statistical model is **non-deterministic**.

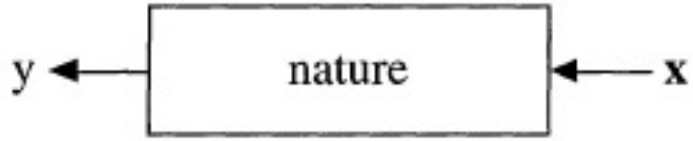
A system in which the output cannot be predicted because there are multiple possible outcomes for each input.

Two purposes for a statistical model:

- Extraction of information
- Predictions

Two purposes for a Statistical Model:

1. To **extract** some **information** about how nature is associating the response variables to the input variables.
2. To be able to **predict** what the responses are going to be to future input variables

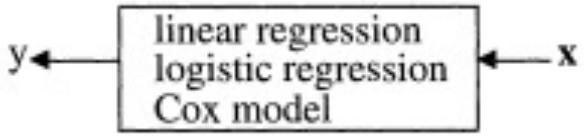


The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

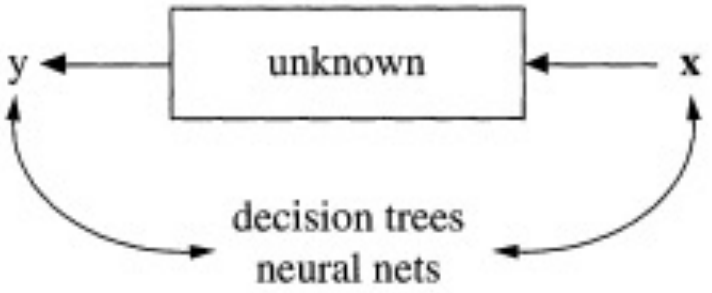
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on $\mathbf{x}$ to predict the responses $\mathbf{y}$. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Source: Statistical Modeling: The Two Cultures- by Leo Breiman

Statistical models: Explaining variation.

Statistical modeling provides a way to relate variables to one another. What impact, if any, does scheduled time of departure have on expected flight delay?

We first begin by considering time of day. In the `nycflights13` package, the `flights` data frame has a variable (`hour`) that specifies the *scheduled* hour of departure.

```
tally( ~ hour, data = SF)

hour
   5    6    7    8    9   10   11   12   13   14   15   16   17   18   19
  55  663 1696  987  429 1744  413  504  476  528  946  897 1491 1091  731
  20   21
 465   57
```

We see that many flights are scheduled in the early to mid-morning and from the late afternoon to early evening. None are scheduled before 5 am or after 10 pm.

Let’s examine how the arrival delay depends on the hour.

<u>Can do in 2 ways</u>:

First using standard box-and-whiskers to show the distribution of arrival delays;

Second with a kind of statistical model over the course of the day.

```
SF %>%
  ggplot(aes(x = hour, y = arr_delay)) +
  geom_boxplot(alpha = 0.1, aes(group = hour)) + geom_smooth(method = "lm") +
  xlab("Scheduled hour of departure") + ylab("Arrival delay (minutes)") +
  coord_cartesian(ylim = c(-30, 120))
```

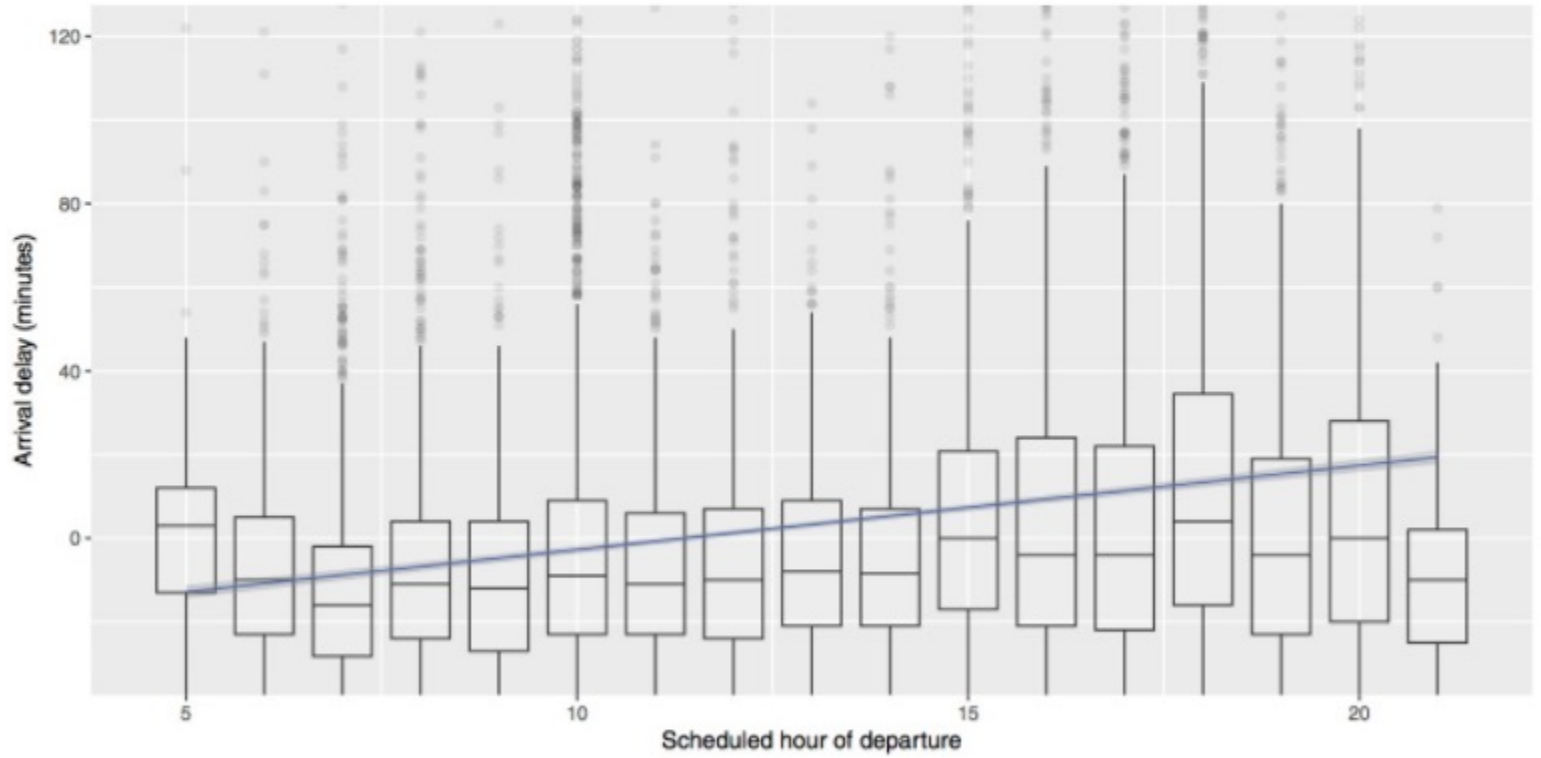


Figure 7.3: Association of flight arrival delays with scheduled departure time for flights to San Francisco from New York airports in 2013.

```
mod1 <- lm(arr_delay ~ hour, data = SF)
msummary(mod1)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.9327     1.2328   -18.6   <2e-16 ***
hour          2.0149     0.0915    22.0   <2e-16 ***

Residual standard error: 46.8 on 13171 degrees of freedom
Multiple R-squared:  0.0355,Adjusted R-squared:  0.0354
F-statistic:  484 on 1 and 13171 DF,  p-value: <2e-16
```

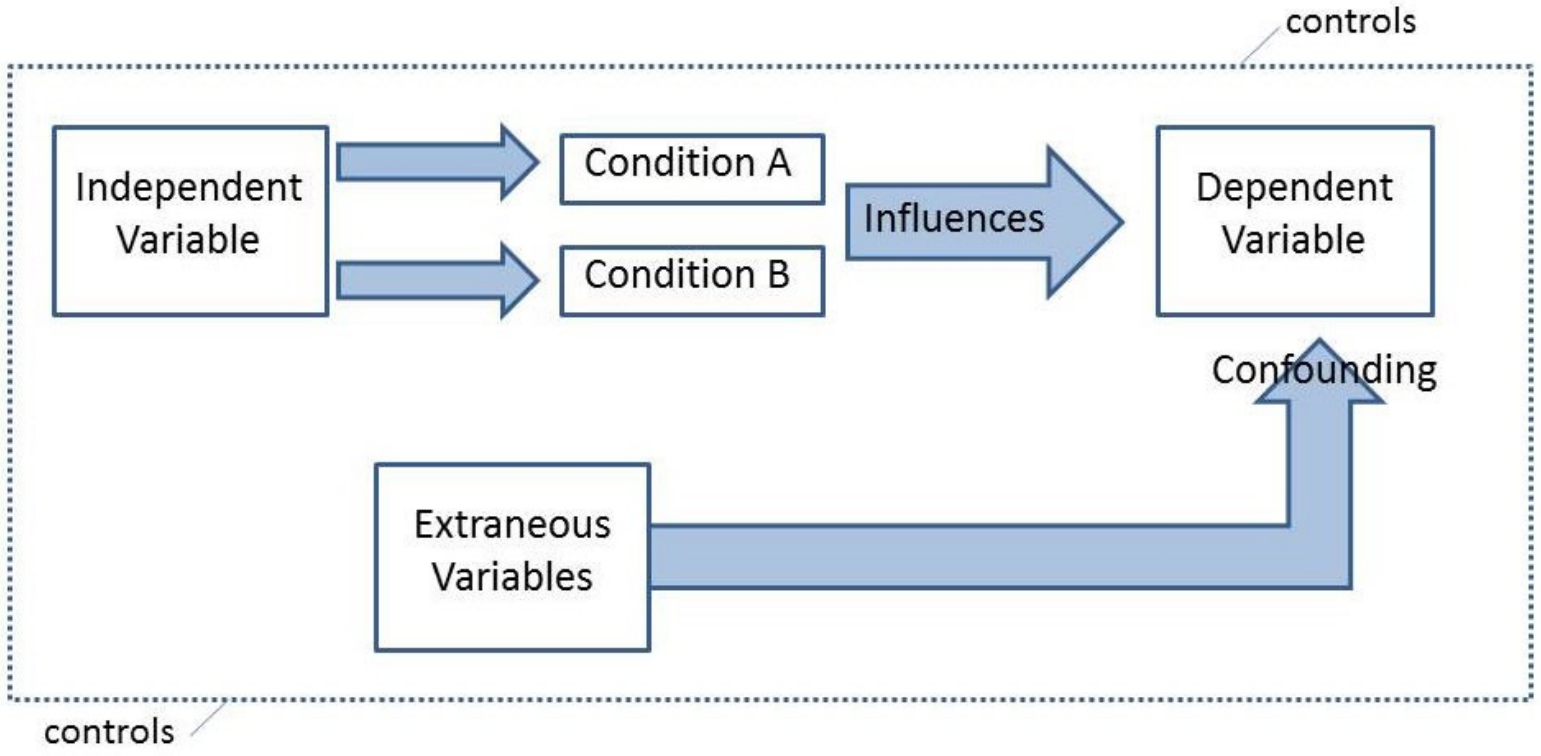
lm() to construct what are called **linear models**.
Linear models describe how the mean of the response variable varies with the explanatory variables. They are the most widely used statistical modeling technique.

The number under the "Estimate" for hour indicates that the arrival delay increases by about 2 minutes per hour. Over the 15 hours of flights, this leads to a 30-minute increase in arrival delay for flights at the end of the day.

The **msummary()** function also calculates the standard error: 0.09 minutes per hour. Or, stated as a 95% confidence interval, this model indicates that arrival delay increases by 2.0 ± 0.18 minutes per hour. The rightmost column gives the p-value, a way of translating the estimate and standard error onto a scale from zero to one.

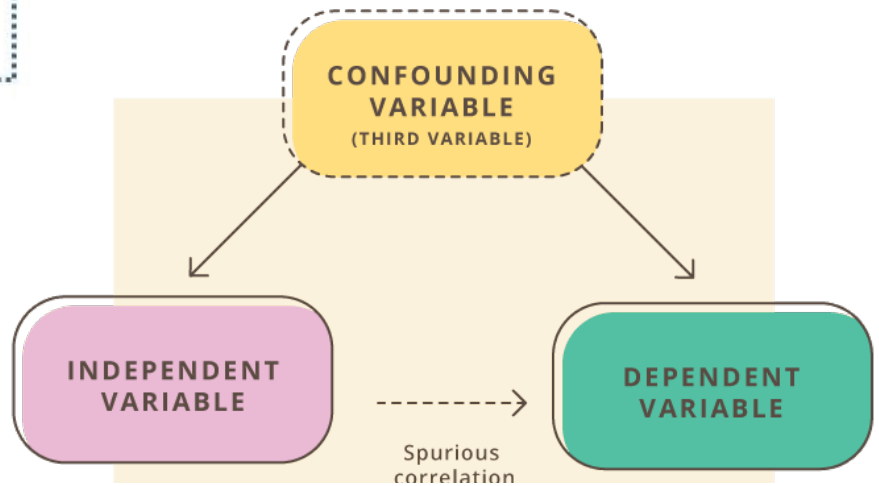
By convention, **p-values below 0.05** provide a kind of certificate testifying that random, accidental patterns would be unlikely to generate an estimate as large as that observed.

6. Confounding and Accounting for Other Factors



"Correlation does not imply causation"

While the statement is certainly true, there are times when correlations do imply causal relationships

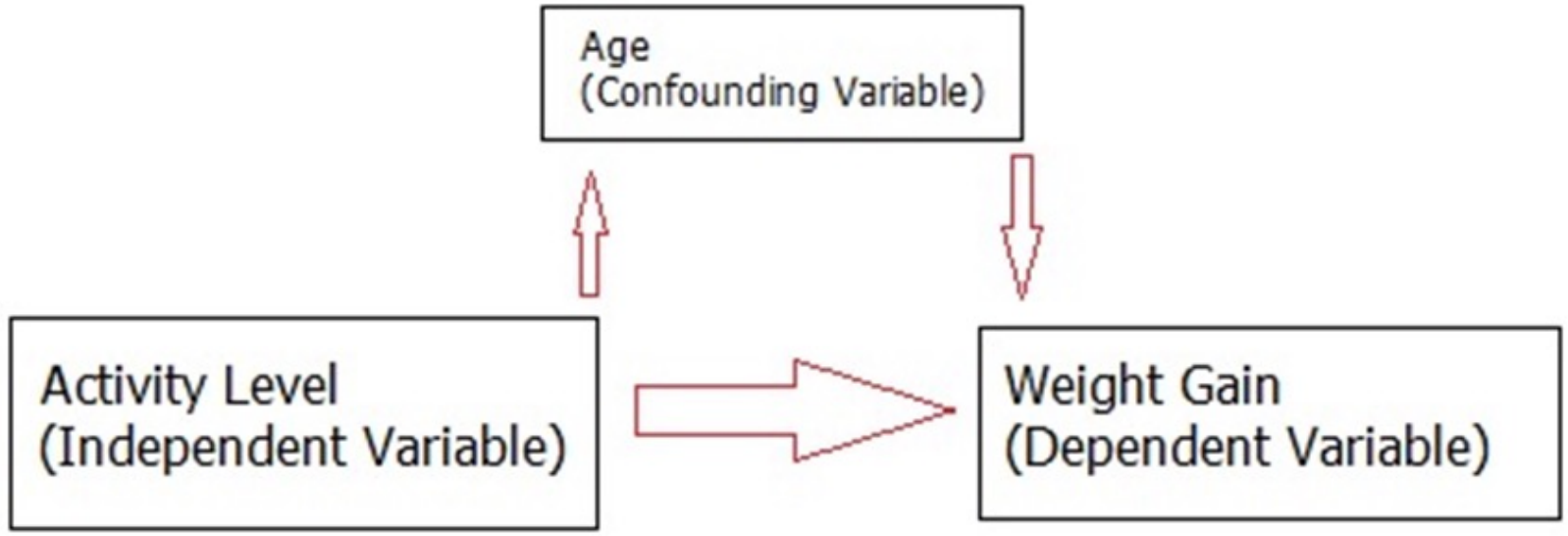


Confounding variables - These are **extraneous variables** in a statistical model that correlate directly or inversely with both the dependent and the independent variable. The estimate fails to account for the confounding factor.

aka third variables

Confounding Variables

If you are researching whether lack of exercise leads to weight gain, lack of exercise is your **independent variable** and weight gain is your **dependent variable**. **Confounding variables** are any other variable that also has an effect on your dependent variable.

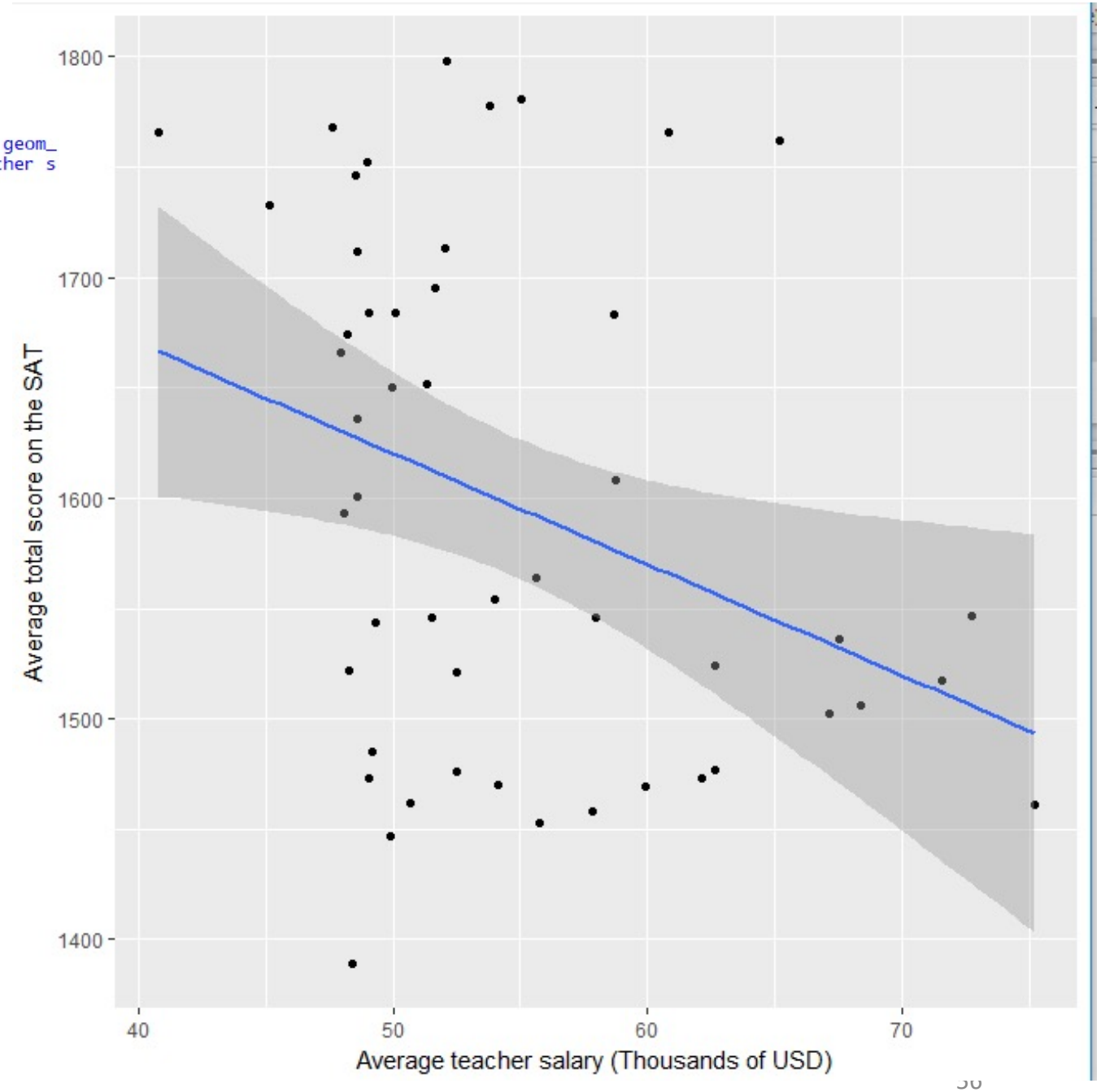


- Data scientists spend most of their time working with **observational data**.
- When seeking to **find meaning** from such data, it is important to be on the lookout for potential confounding factors that could distort observed associations.

Let's consider an example using data on average teacher salaries and average total SAT scores for the 50 United States. The SAT (Scholastic Aptitude Test) is a high-stakes exam used for entry into college. Are higher teacher salaries associated with better outcomes on the test at the state level?

```
library(mdsr)
SAT_2010 <- mutate(SAT_2010, Salary = salary/1000)
SAT_plot <- ggplot(data = SAT_2010, aes(x = Salary, y = total)) +
  geom_point() + geom_smooth(method = "lm") +
  ylab("Average total score on the SAT") +
  xlab("Average teacher salary (thousands of USD)")
SAT_plot
```

```
> SAT_2010 <- mutate(SAT_2010, Salary = salary/1000)
> SAT_plot <- ggplot(data =SAT_2010, aes (x= Salary, y = total)) +
+ geom_point() + geom_smooth(method ="lm")
> SAT_plot <- ggplot(data =SAT_2010, aes (x= Salary, y = total)) + geom_point() + geom_
smooth(method ="lm") + ylab("Average total score on the SAT") + xlab("Average teacher s
alary (Thousands of USD)")
> SAT_plot
>
```



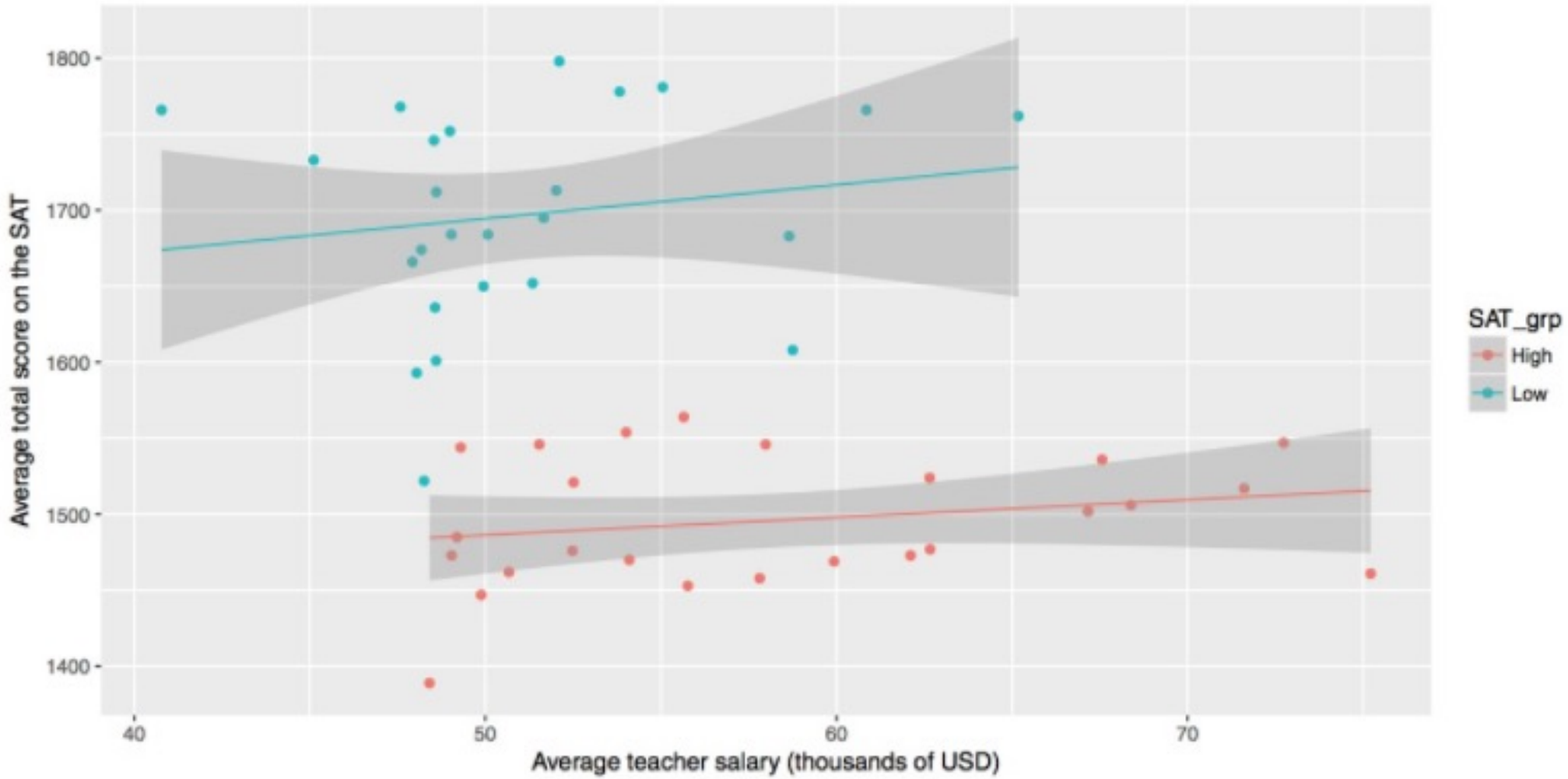
```
SAT_mod1 <- lm(total ~ Salary, data = SAT_2010)
msummary(SAT_mod1)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1871.10     113.14   16.54   <2e-16 ***
Salary         -5.02       2.05   -2.45    0.018 *

Residual standard error: 111 on 48 degrees of freedom
Multiple R-squared:  0.111,Adjusted R-squared:  0.0927
F-statistic: 6.01 on 1 and 48 DF,  p-value: 0.0179
```

```
> favstats(~ sat_pct, data = SAT_2010)
 min Q1 median Q3 max  mean       sd  n missing
   3  6     27 68  93 38.52 31.99122 50       0
```

Lurking in the background is another important factor!
% of students who take the SAT in each state varies dramatically (from **3** – **93**% in 2010).
Create a variable (SAT grp) that divides the states into two groups.

```
> SAT_2010 <- SAT_2010 %>%
+ mutate (SAT_grp= ifelse(sat_pct <=27, "Low", "High"))
> tally(~ SAT_grp, data =SAT_2010)
SAT_grp
High  Low
  25   25
```

```
SAT_plot %+% SAT_2010 + aes(color = SAT_grp)
```



Scatter plot, stratified by the % of students taking the SAT in each state.

Stratification

- Stratification (simply means grouping) can control the confounding variable sat_grp.
- For each group given by the values of sat_grp, average teacher salary is positively associated with average SAT score.
- When we collapse over this variable, average teacher salary is negatively associated with average SAT score. (sat_grp or sat_pct is confounding here.)
- This form of confounding is called **Simpson's paradox**.

```
coef(lm(total ~ Salary,
        data = filter(SAT_2010n, sat_grp == "Low")))
```

```
## (Intercept)      Salary
## 1557.658858    2.613381
```

```
coef(lm(total ~ Salary,
        data = filter(SAT_2010n, sat_grp == "High")))
```

```
## (Intercept)      Salary
## 1405.048718    1.515035
```

1. Among states with a low percentage taking the SAT, teacher salaries and SAT scores are positively associated.
2. Among states with a high percentage taking the SAT, teacher salaries and SAT scores are positively associated.
3. Among all states, salaries and SAT scores are negatively associated.

Multiple regression is another way of controlling confounding variables.

```
SAT_mod2 <- lm(total ~ Salary + sat_pct, data = SAT_2010)
msummary(SAT_mod2)

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1589.007     58.471    27.2   <2e-16 ***
Salary         2.637      1.149     2.3    0.026 *
sat_pct       -3.553      0.278   -12.8   <2e-16 ***

Residual standard error: 53.2 on 47 degrees of freedom
Multiple R-squared:  0.801,Adjusted R-squared:  0.792
F-statistic: 94.5 on 2 and 47 DF,  p-value: <2e-16
```

The slope for Salary is positive and statistically significant when we control for sat_pct.

Strategies to Reduce Confounding

1. **Randomization** (aim is random distribution of confounders between study groups)
2. **Matching** (of individuals or groups, aim for equal distribution of confounders)
3. **Stratification** (confounders are distributed evenly within each stratum)
4. **Adjustment** (usually distorted by choice of standard)
5. **Multivariate analysis** (only works if you can identify and measure the confounders)

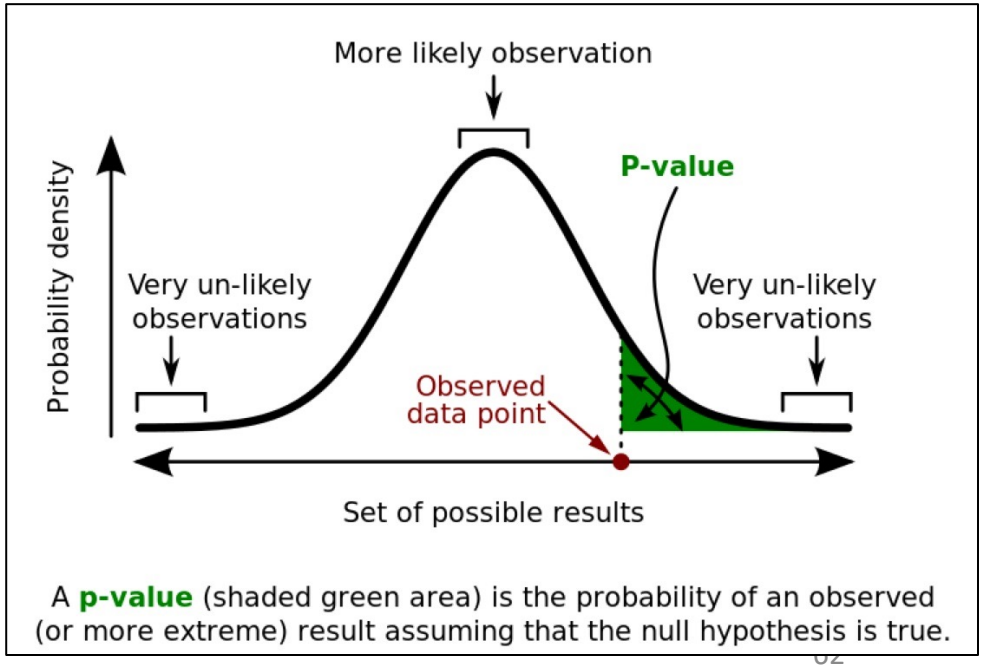
How do you know how much confidence to put in the outcome of a hypothesis test?

The statistician's criterion is the **statistical significance** of the test, or the likelihood of obtaining a given result by chance.

There are **several ways** to refer to the significance level of a test, and it is important to be familiar with them.

- ✓ The finding is significant at the 0.05 level.
- ✓ The confidence level is 95 percent.
- ✓ The Type I error rate is 0.05.
- ✓ The alpha level is 0.05.
- ✓ The p-value is 0.05.

The **smaller** the **significance level *p***, the more stringent the test and the greater the likelihood that the **conclusion is correct**.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

7. p-Values or Probability Values

- The **P value** is used all over statistics, from t-tests to regression analysis.
- P-value is used to determine the significance of results after a hypothesis test in statistics.
- P-value helps the readers to draw conclusions and is always between 0 and 1.
  - P- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.
  - P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.
  - P-value=0.05is the marginal value indicating it is possible to go either way.
- A result is called “statistically significant” whenever the p-value is less or equal to the significance level.

Understanding the p-value https://youtu.be/eyknGvncKLw

American Statistical Association endorsed a statement on p-values - **Six principles**:

i. P-values can indicate how incompatible the data are with a specified statistical model.

ii. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

iii. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

iv. Proper inference requires full reporting and transparency.

v. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

vi. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

What is A/B testing?

- **A/B testing** (also known as **split testing** or **bucket testing**) is basically statistical hypothesis testing, or, in other words, statistical inference.
- It is a method of **comparing two versions of a webpage or app** against each other to determine which one performs better.
- AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.
- The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of an interest.

What is A/B testing?

- Two-sample hypothesis testing
- Randomized experiments with two variants: A and B
- A: control; B: variation
- User-experience design: identify changes to web pages that increase clicks on a banner.
- Current website: control; NULL hypothesis
- New version: variation; alternative hypothesis

See also → https://www.analyticsvidhya.com/blog/2020/10/ab-testing-data-science/

Imagine that you are CEO of Amazon, and trying to work out whether rearranging your website into a new format affects conversion rate (i.e. the proportion of visitors to Amazon who become customers):

One approach would be to run both versions to selected customers and make a judgement based on these numbers alone:

	Layout A	Layout B
Visitors	122	118
Customers	22	25
Conversion %	18.0%	21.2%

In this case we would conclude that Layout B is superior to Layout A.

However, such a simple approach suffers from **two possible errors** that statistics students will be very familiar with:

Type I error — or falsely concluding that your intervention was successful (which here might be falsely concluding that layout B is better than Layout A). Also known as a **false positive** result.

Type II error — falsely concluding that your intervention was not successful. Also known as a **false negative** result.

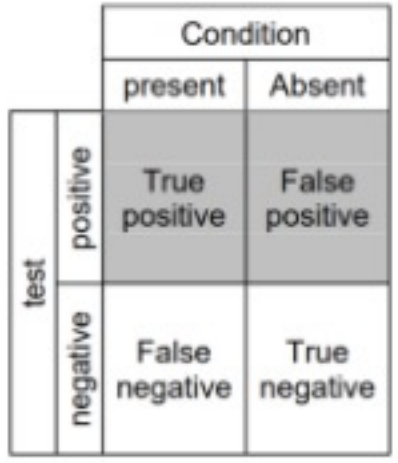
An **A/B test** - enable to accurately quantify the effect size and errors, and so calculate the probability type I or type II error.

https://towardsdatascience.com/data-science-you-need-to-know-a-b-testing-f2f12aff619a

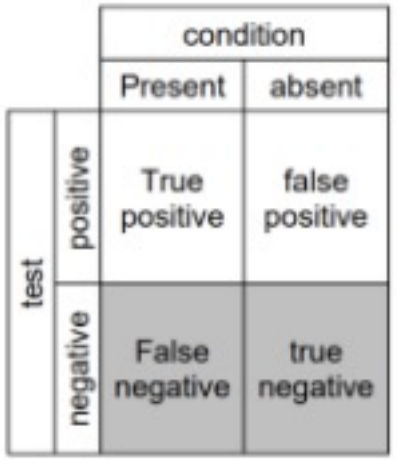
https://www.tutorialspoint.com/ab_testing/index.htm

Accuracy

- ✓ **True positive:** detects the condition when the condition is present.
- ✓ **True negative:** does not detect the condition when the condition is not present.
- ✓ **False-positive:** detects the condition when the condition is absent.
- ✓ **False-negative:** does not detect the condition when the condition is present.
- ✓ **Sensitivity:** also known as **recall**; measures the ability of a test to detect the condition when the condition is present; sensitivity = TP/(TP+FN)
- ✓ **Specificity:** measures the ability of a test to correctly exclude the condition when the condition is absent; specificity = TN/(TN+FP)



Predictive value positive



Predictive value negative

https://towardsdatascience.com/basic-statistics-you-need-to-know-for-data-science-1fdd290f59b5

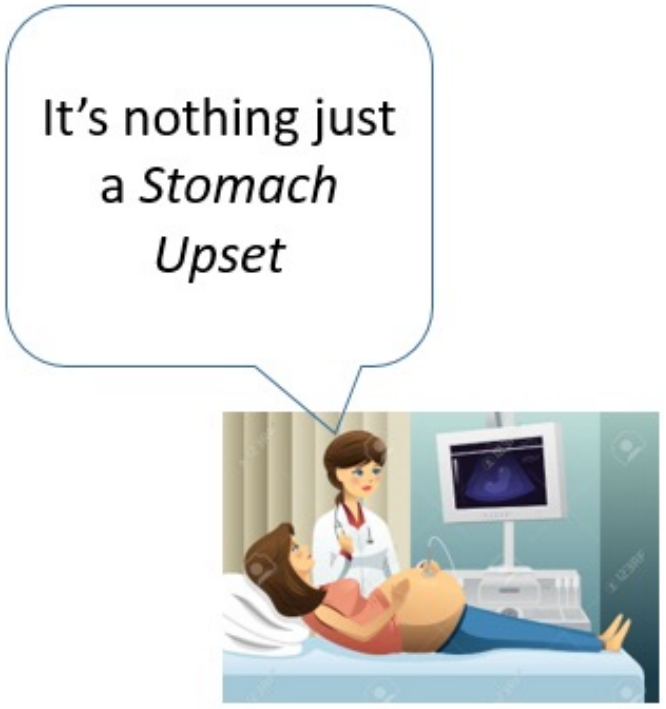
Decision you made			
		Do <u>not reject</u> the null hypothesis	**<u>Reject</u> the null hypothesis**
True condition in the population	The null hypothesis should **<u>not</u>** be **<u>rejected</u>**	**You are correct** in not rejecting the null. **(True negative)**  Probability of correctly not rejecting the null hypothesis = **1-α** (equivalent to **confidence level**)	**You made a mistake!** **Type I error**: Rejected the null hypothesis, when you should not. **(False positive)**  The risk of making Type I error = **α** (equivalent to **significance level**)
	The null hypothesis should be **<u>rejected</u>**	**You made a mistake!** **Type II error**: Research hypothesis is true but you decide to stick with the null hypothesis. **(False negative)**  The risk of making Type II error = **β**	**You are correct** in rejecting the null hypothesis and accepting the research hypothesis. **(True positive)**  Probability of getting correct = **1- β** (This is also called **statistical power**)

http://methods.sagepub.com/book/research-methods-statistics-for-public-nonprofit-administrators-practical-guide/i1218.xml

False Positive

False Positives are the cases where you wrongly classified a non-event as an event a.k.a **Type I error**.

False Negative



False Negatives are the cases where you wrongly classify events as non-events, a.k.a **Type II error**.

More Resources

```
> swirl::install_course("Statistical Inference")

| Please choose a lesson, or type 0 to return to course menu.

 1: Introduction              2: Probability1            3: Probability2
 4: ConditionalProbability    5: Expectations            6: Variance
 7: CommonDistros             8: Asymptotics             9: T Confidence Intervals
10: Hypothesis Testing       11: P Values               12: Power
13: Multiple Testing         14: Resampling
```

https://github.com/swirldev/swirl_courses

“Uncertainty is the only certainty”

Learn More

Statistics For Data Science | Data Science Tutorial | Simplilearn

https://www.youtube.com/watch?v=Lv0xcdeXaGU&feature=youtu.be

Inferential Statistics (Coursera) - https://bit.ly/2Kneaw9