

Lecture 2: Introduction to Statistics

STAT 324

Ralph Trane
University of Wisconsin–Madison

Spring 2020



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Re: homework:

- Submission: Knit to html -> go to canvas -> upload .html file
- Homework 1 will be assigned later today, and will be due at 6pm next Thursday

Important point about R Markdown I forgot to mention:

- When you knit, you start from scratch!
- Please take the survey
- Learning Center
- Updated office hours on Canvas

Objectives for Today



1. Expectations
2. "Define" statistics
3. Why is statistics important? A.K.A. why you should care...
4. Broad overview of STAT 324

Before we get started...



... let's clear the air. Based on my personal experiences most of you think statistics is boring.

So I expect a lot of this...



Before we get started...



... let's clear the air. Based on my personal experiences most of you think statistics is boring.

My prayer to you: give us a chance!

Take a leap of faith...



... I promise you won't get hurt!

We will give you...

... basic knowledge of statistics and probability

... some foundational knowledge of the theory behind statistics, enabling you to learn additional methods in the future

... the ability to summarize data graphically and numerically

... the ability to identify and perform appropriate analyzes for some simple data sets

... functional skills using R

You should **NOT** expect a flow chart called "how to perform the perfect statistical analysis".

We expect that you...

... participate in lecture and discussions

... spend time on your homework and use it as a *learning experience*

... ask us questions instead of giving up

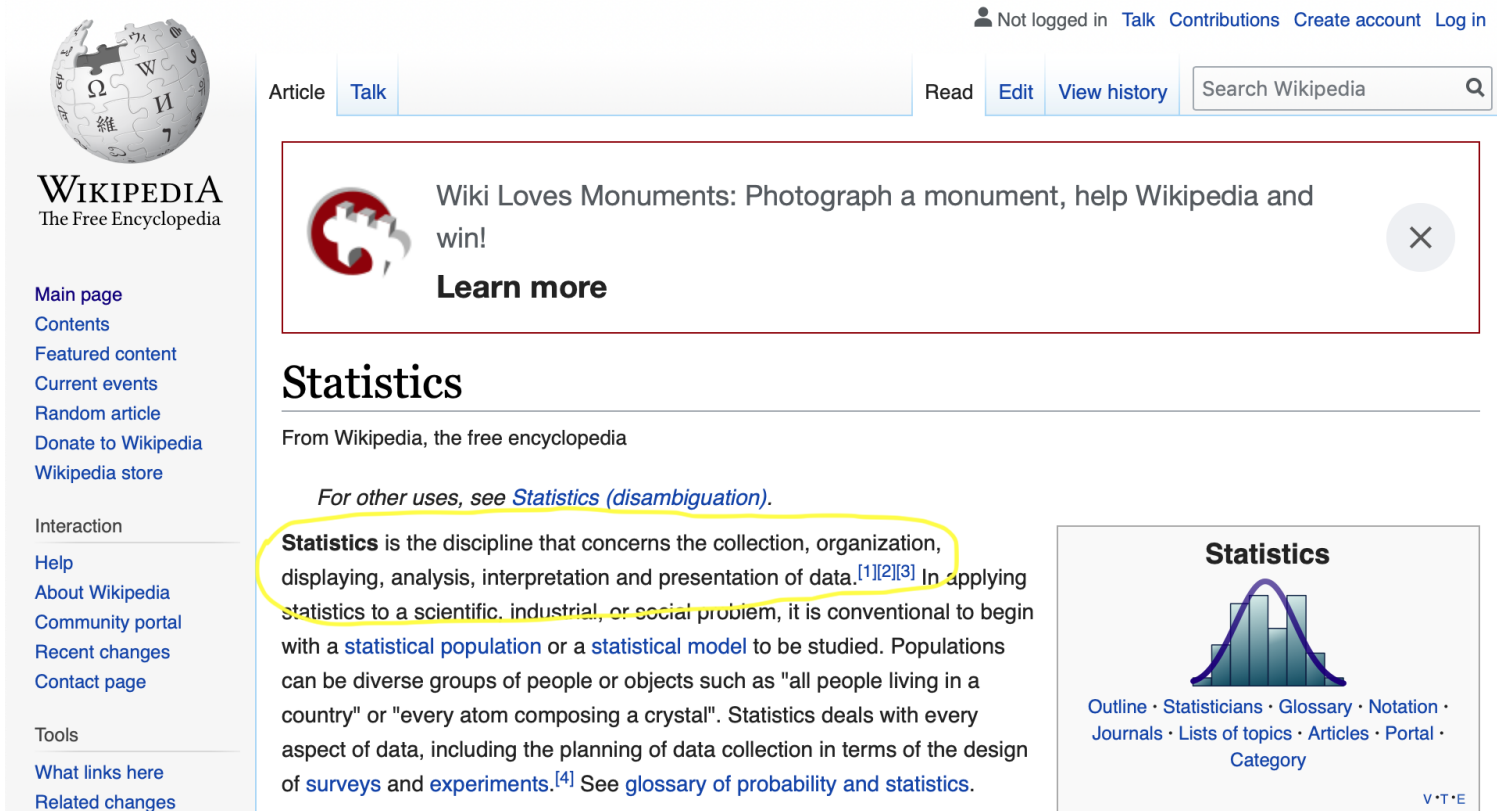
We do **NOT** expect you to be a completely self-sufficient data analyst by the end of this course.

What is Statistics?



Statistics is... very hard to define.

From Wikipedia:



The screenshot shows the Wikipedia article for "Statistics". The page layout includes a sidebar on the left with navigation links, a top navigation bar with user status and search, and a main content area. A red-bordered banner at the top of the article area promotes "Wiki Loves Monuments". The article title "Statistics" is prominently displayed, followed by a subtitle "From Wikipedia, the free encyclopedia". The first paragraph of the article is highlighted with a yellow circle. To the right of the text is a box containing a histogram with a normal distribution curve and a list of related links.

Wikipedia logo and navigation links (Main page, Contents, Featured content, Current events, Random article, Donate to Wikipedia, Wikipedia store, Interaction, Help, About Wikipedia, Community portal, Recent changes, Contact page, Tools, What links here, Related changes).

Top navigation: Not logged in, Talk, Contributions, Create account, Log in.

Article navigation: Article, Talk, Read, Edit, View history, Search Wikipedia.

Wiki Loves Monuments banner: Photograph a monument, help Wikipedia and win! Learn more.

Statistics

From Wikipedia, the free encyclopedia

For other uses, see [Statistics \(disambiguation\)](#).

Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.^{[1][2][3]} In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a [statistical population](#) or a [statistical model](#) to be studied. Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of [surveys](#) and [experiments](#).^[4] See [glossary of probability and statistics](#).

Statistics

Outline · Statisticians · Glossary · Notation · Journals · Lists of topics · Articles · Portal · Category

V · T · E

So it seems that statistics is everything that has to do with data...?

STATISTICS IS NOT AN EXACT SCIENCE!!

To me, it is more accurately described as a *decision science*.

Very unfortunate misconception. Lead to terms as "statistically significant", arbitrary cutoffs used as THE way to determine importance, etc.

Why you should care



These days, statistics is all around us, but most have problems with even simple things, such as interpreting a probability.

Who will win the presidency?

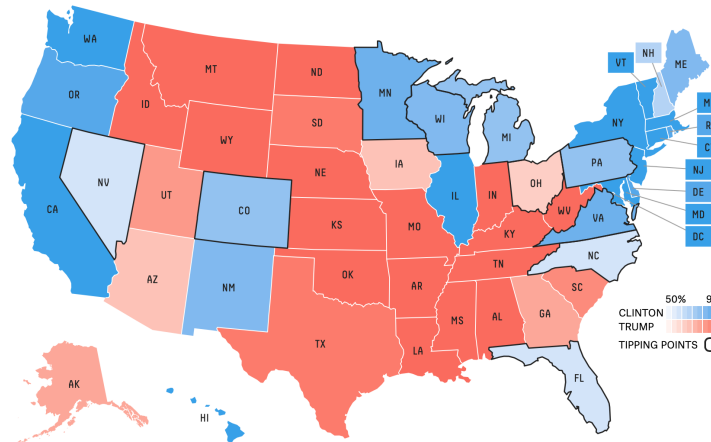


Chance of winning



Hillary Clinton
71.4%

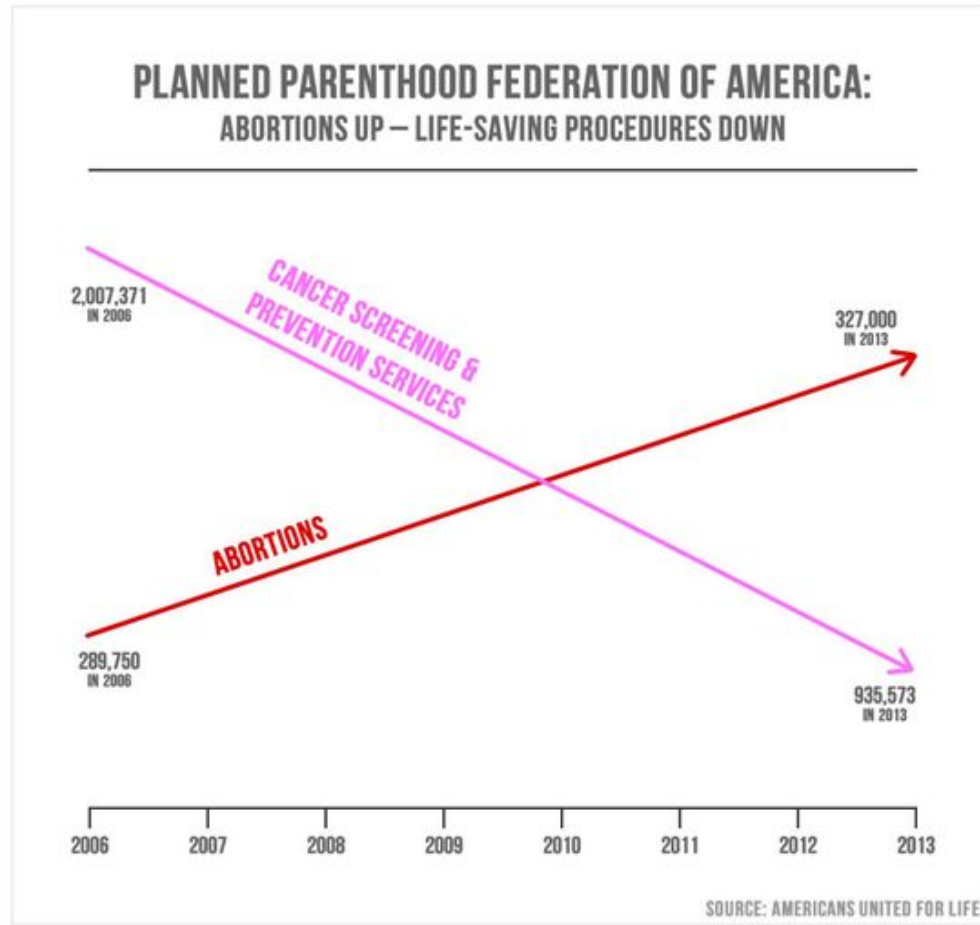
Donald Trump
28.6%



Why you should care



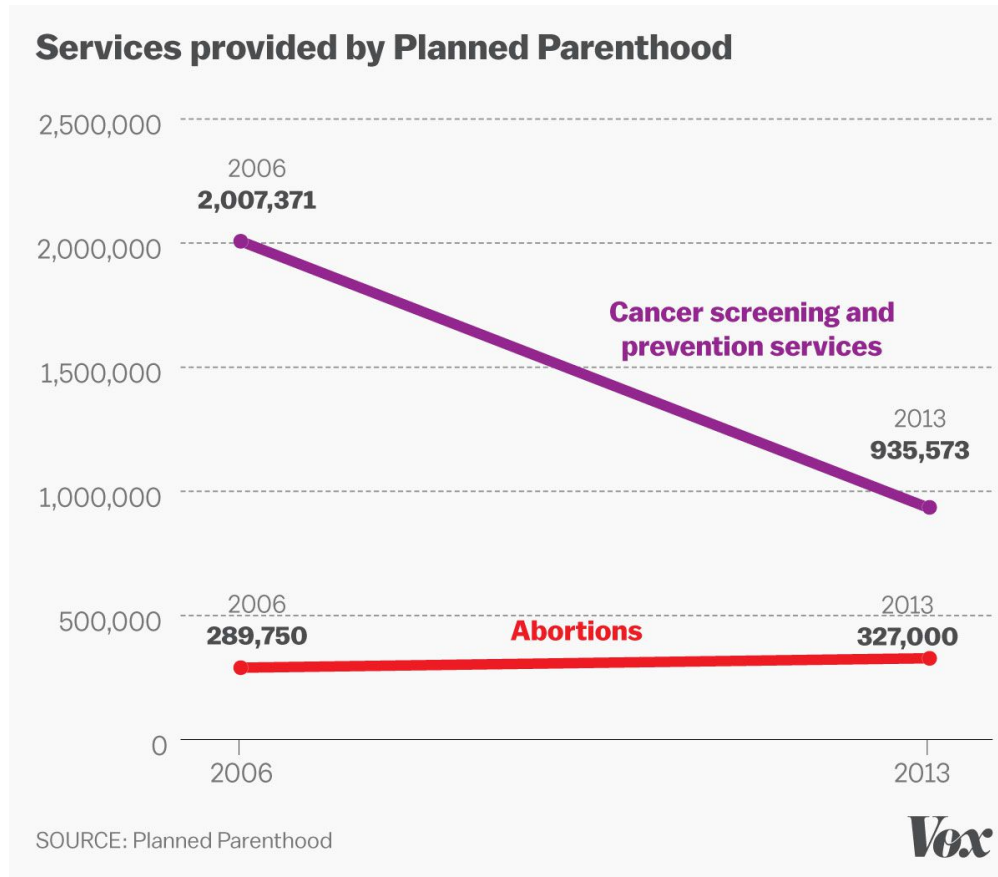
Too often studies are misrepresented.



Why you should care



Too often studies are misrepresented.



People get away with bad practices (on purpose or accidentally):

OPEN ACCESS Freely available online

 PLOS one

How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data

Daniele Fanelli*

INNOGEN and ISSTI-Institute for the Study of Science, Technology & Innovation, The University of Edinburgh, Edinburgh, United Kingdom

Abstract

The frequency with which scientists fabricate and falsify data, or commit other forms of scientific misconduct is a matter of controversy. Many surveys have asked scientists directly whether they have committed or know of a colleague who committed research misconduct, but their results appeared difficult to compare and synthesize. This is the first meta-analysis of these surveys. To standardize outcomes, the number of respondents who recalled at least one incident of misconduct was calculated for each question, and the analysis was limited to behaviours that distort scientific knowledge: fabrication, falsification, "cooking" of data, etc... Survey questions on plagiarism and other forms of professional misconduct were excluded. The final sample consisted of 21 surveys that were included in the systematic review, and 18 in the meta-analysis. A pooled weighted average of 1.97% (N = 7, 95%CI: 0.86–4.45) of scientists admitted to have fabricated, falsified or modified data or results at least once – a serious form of misconduct by any standard – and up to 33.7% admitted other questionable research practices. In surveys asking about the behaviour of colleagues, admission rates were 14.12% (N = 12, 95% CI: 9.91–19.72) for falsification, and up to 72% for other questionable research practices. Meta-regression showed that self reports surveys, surveys using the words "falsification" or "fabrication", and mailed surveys yielded lower percentages of misconduct. When these factors were controlled for, misconduct was reported more frequently by medical/ pharmacological researchers than others. Considering that these surveys ask sensitive questions and have other limitations, it appears likely that this is a conservative estimate of the true prevalence of scientific misconduct.

Full

paper (from 2009) can be found [here](#)

My point is: whether you'll be producers or consumers of statistics, it's important to have a basic understanding of what's going on.

Generally speaking, there are two kinds of statistics:

- descriptive
- inferential

We will mostly consider inferential (this is by far the more difficult and more fun of the two!)

Descriptive Statistics

Reduce data to a few key take-aways

This is what people usually think about when they hear the words "statistics"

- means, medians, quantiles, etc.
- variances, standard deviations, etc.

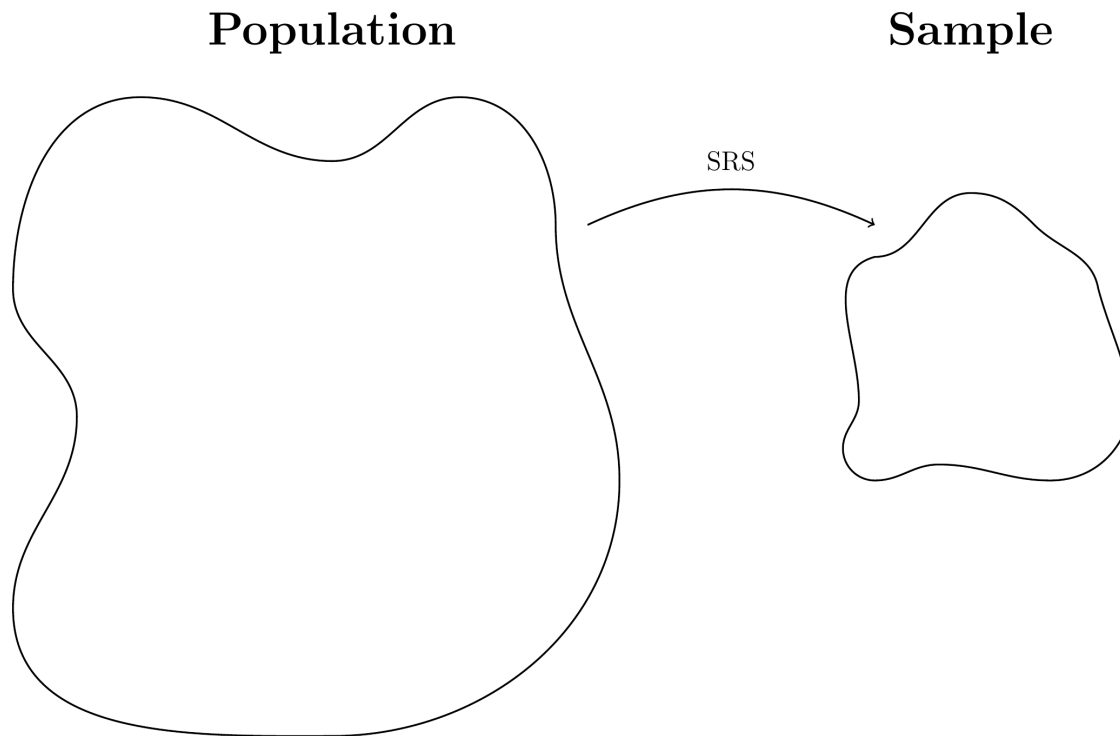
This includes things like sports statistics, census information.

Descriptive Statistics can take the form of graphical and/or numerical summaries of the data.

Inferential Statistics

Where descriptive statistics aim at *describing* the data at hand...

... inferential statistics aim at *inferring* information about a *population* based on the data at hand (the *sample*)



Inferential Statistics

In the beginning, there is a hypothesis: "Cigarette smoking causes lung cancer".

- Very vague, hard to verify or dismiss

More specific: "People who smoke cigarettes have a higher incidence of lung cancer over a 10-year period than people who do not smoke cigarettes."

- Clear what it means to "cause lung cancer" -- higher incidence

Inferential Statistics

Say we wanted to answer this question. Only one way to obtain the "Truth™".
It's a simple three-step procedure:

1. Ask the people in your population of interest if they have ever smoked, and if they have developed lung cancer.
2. Calculate incidence rates
3. See which is larger

Rejoice in newfound knowledge!



Inferential Statistics

Say we wanted to answer this question. Only one way to obtain the "Truth™".
It's a simple three-step procedure:

1. Ask the people in your population of interest if they have ever smoked, and if they have developed lung cancer.
2. Calculate incidence rates
3. See which is larger

Unfortunately, this is basically impossible!



Inferential Statistics

What do we do instead? Inferential Statistics!

1. Define the population(s) you're interested in, and specify the feature you'll be looking at
 - populations are people who smoke, and people who do not
 - feature of interest would be incidence rate
2. Get a representative sample from the population
 - preferably sample by random from the two populations
3. From the sample, calculate quantities (i.e. "things") that can help you say things about the "truth"
 - when interested in the incidence rate, simply calculate the incidence rates in the samples

The thing is, the samples won't mirror the populations *exactly* -- take a new sample, get new estimates.

The question is then: is the difference due to "differences in the truths", or is it simply "differences due to random samples"?

Example: Average age of death in Wisconsin

Where should you live

1. Populations of interest: people living in Wisconsin broken down by county. Feature: mean age of death
2. Use public records to get age of death for a good sample of people
3. What would be a good quantity to look at in the sample? Probably the average age of death

Example: Average age of death in Wisconsin

Results according to [this report](#):

County	Life expectancy
Kewaunee	82.0
Ozaukee	81.8
Pierce	81.6
Waukesha	81.5
Taylor	81.5
Milwaukee	77.6
Washburn	76.7
Ashland	77.5
Sawyer	77.1
Menominee	72.5

Would you prefer Kewaunee over Waukesha?

Example: Average age of death in Wisconsin

The question is: do we *really* think there's a difference?

Let's pretend the results of the actual data looked like this:

county	n	Average Age of Death
Kewaunee	100	82.0
Waukesha	100	81.5

Would you prefer Kewaunee?

I definitely would...

Example: Average age of death in Wisconsin

What if the actual data look more like this:

county	n	Average Age of Death
Kewaunee	100	82.0
Waukesha	100	81.5

Would you prefer Kewaunee?

I'm not sure...

Example: Average age of death in Wisconsin

The main question: when is a difference "big enough"? How do we make the answer less subjective?

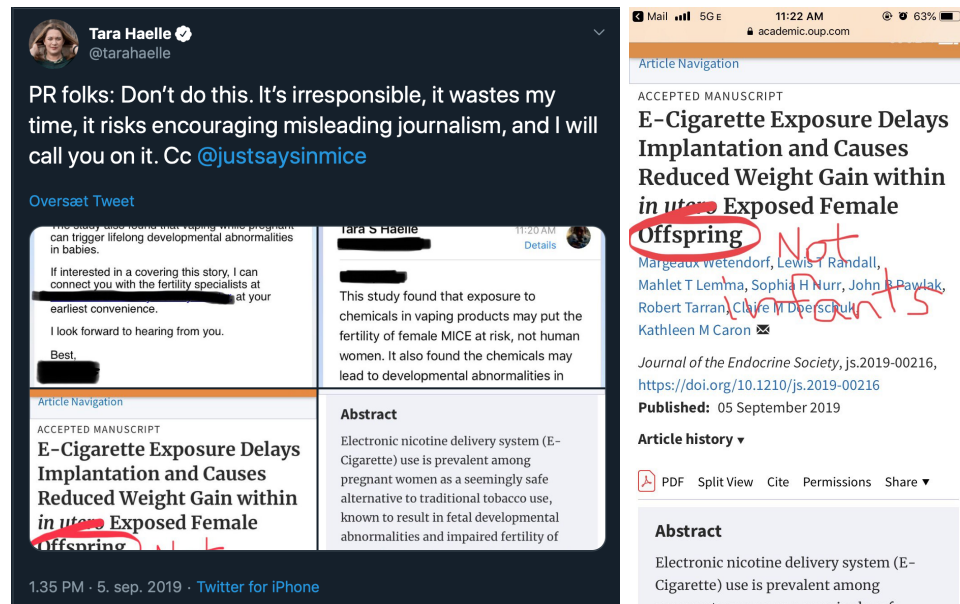
Main Parts

1. Descriptive Statistics
2. Probability
3. Inference

1. Descriptive Statistics

Q: Why is it important to describe your sample?

A: Can only draw conclusions about population that looks like your sample



The image shows a tweet from Tara Haelle (@tarahaelle) and a research article. The tweet discusses a study on vaping and pregnancy, mentioning "in utero Exposed Female Offspring" and "Not Infants". The research article, titled "E-Cigarette Exposure Delays Implantation and Causes Reduced Weight Gain within in utero Exposed Female Offspring", is from the Journal of the Endocrine Society. The article title and the tweet text are circled in red, with "Not Infants" written in red next to the article title.

Moral of the story: know your population!

2. Probability

Describes what happens when getting a sample from a population.

Probability Theory is a branch of mathematics that plays a crucial role in statistics

This is what enables us to describe the variability of sampling

3. Inference

The art of extrapolating from a sample to the population.

SUPER HARD!

To make it easier, we make assumptions.

This also means that if our assumptions are off, everything is off. Therefore, important to state **AND** check your assumptions!