

Biostats Lecture 1: Introduction

Public Health 783

Ralph Trane
University of Wisconsin–Madison

Fall 2019



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Practicalities

Office hours: Wednesdays 11am till noon in PHS 666 (WARF 6th floor)



Practicalities

Office hours: Wednesdays 11am till noon in PHS 666 (WARF 6th floor)

Lecture notes will be hosted [here](#)

Updated as we go with some material available before lectures, some after

Lecture slides will be available on Canvas *after* lectures

Expectations

You can expect that I will ...

- ... be brutally honest

- ... do everything I can to keep things interesting

- ... fight for you!

In return, I expect that you will...

- ... help me out by participating in lectures (i.e. ask me all your question, answer all my questions)

- ... show up prepared when I ask that of you

- ... give me a chance!

Before we get started...



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



Before we get started...



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

Give me a shot: I will try to make this different than most statistics classes.

Take a leap of faith...



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

Objectives for Today



1. "Define" statistics
2. Better understanding of what statistics is/isn't
3. Why is statistics important?
4. Broad overview of statistics in 783

What is Biostatistics?



Statistics applied to biology related questions

- Average Age of Death in 1842 Great Britain

	Manchester	Rutland
Professionals	38	52
Tradesmen	20	41
Laborers	17	38

- Are there genes associated with an increased risk of cancer?
- What are common factors of increased risk of cardiovascular disease?
- Does exposure to green spaces improve mental health?

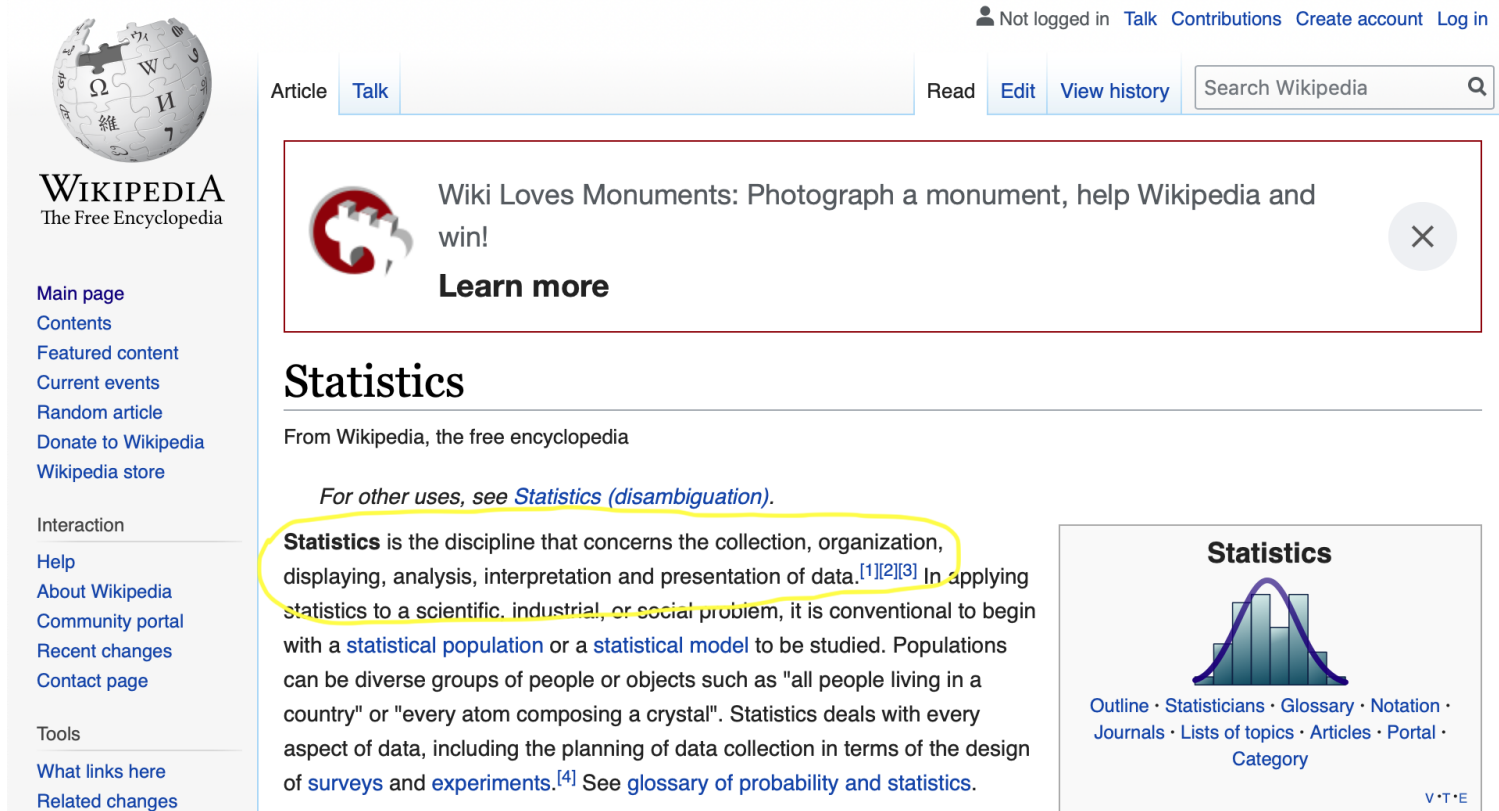
So the question really is: what is statistics?

What is Statistics?



Statistics is... very hard to define.

From Wikipedia:



The screenshot shows the Wikipedia article for "Statistics". At the top, there's a navigation bar with "Article" and "Talk" tabs, and a search bar. Below the navigation bar, there's a banner for "Wiki Loves Monuments". The main heading is "Statistics". Below the heading, it says "From Wikipedia, the free encyclopedia". The first sentence of the article is "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."

The text "Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.^{[1][2][3]}" is highlighted with a yellow circle.

On the right side of the article, there is a box titled "Statistics" containing a histogram with a normal distribution curve overlaid. Below the histogram, there are links: "Outline · Statisticians · Glossary · Notation · Journals · Lists of topics · Articles · Portal · Category".

On the left side of the article, there is a sidebar with the Wikipedia logo and a list of 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", and "Related changes".

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

STATISTICS IS NOT AN EXACT SCIENCE!!

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

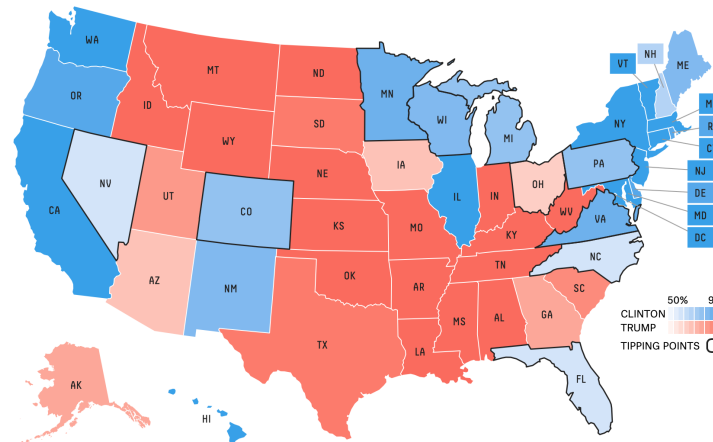


These days, statistics is all around us, but most have problems with even simple things, such as interpreting a probability.

Who will win the presidency?



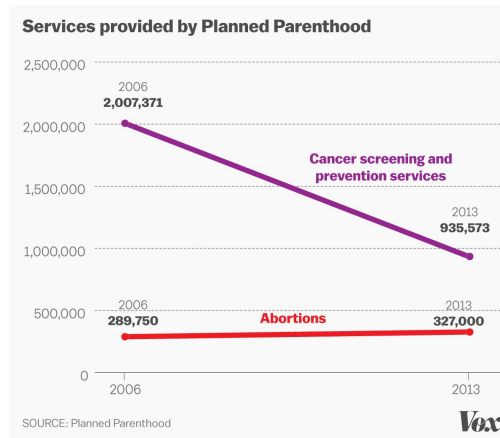
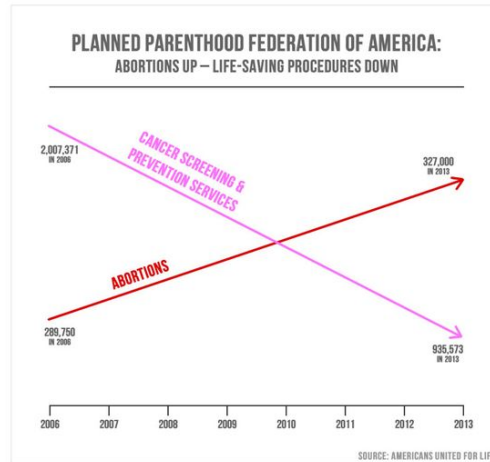
Chance of winning



Why you should care



Too often studies are misrepresented.



Why you should care



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

OPEN ACCESS Freely available online



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](#)

Why you should care



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.

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
- However, it is **NOT** clear what "higher" means

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!



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

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

Unfortunately, this is basically impossible!



What do we do instead? 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 that 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 of incidence rates.

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

(Inspired by the "Average Age of Death (1842)" example)

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

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?

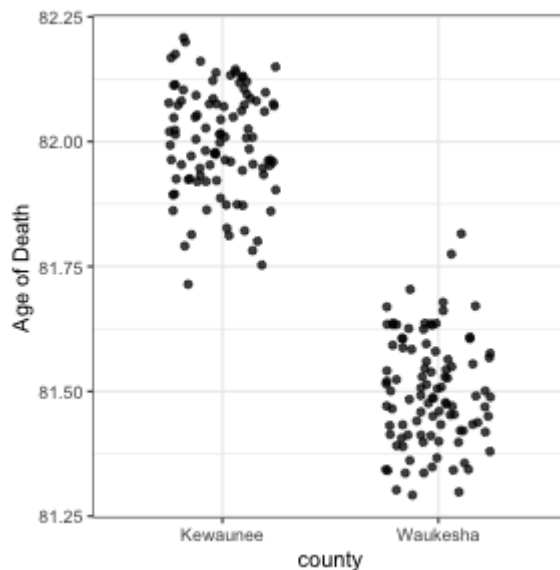
Biostatistics in PUBLHLTH 783



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

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

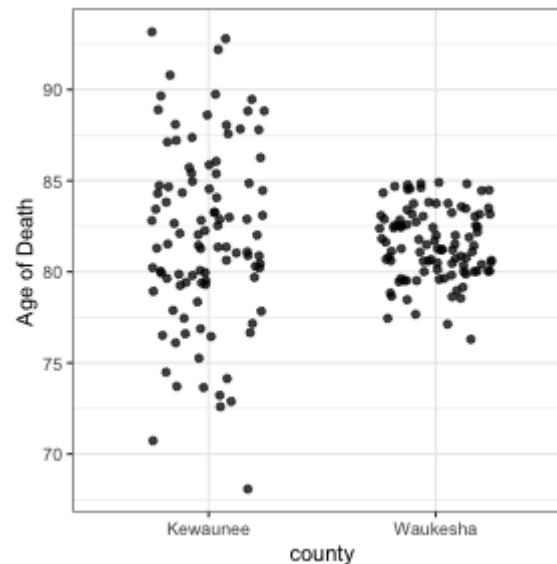
county	n	Average Age of Death
Kewaunee	100	82.0
Waukesha	100	81.5



Would you prefer Kewaunee? I definitely would...

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...



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



Components

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

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!

Work Flow:

- Want to test hypothesis
- Take sample from population
- Calculate quantity of interest
- Ask: *IF* the hypothesis is true, how likely are we to see what we saw
 - If unlikely, evidence **against** the hypothesis
 - If likely, evidence **for** the hypothesis
- Answer: get an idea of how much the observed quantity is expected to vary, then use probability theory

How do we get an idea of how much this quantity would vary with repetition?