Statistical Foundations of Data Science

How to lie with statistics

University of the Witwatersrand

2023

Quotes

- "The first principle is that you must not fool yourself and you are the easiest person to fool."- Richard Feynman, 1965 Noble Prize Winner in Physics.

Quotes

- "The first principle is that you must not fool yourself and you are the easiest person to fool."- Richard Feynman, 1965 Noble Prize Winner in Physics.
- "My mother is the greatest statistician in the world, she can reach a conclusion with only one data point" -Anon.

Review Question

- I flip a fair coin three times. Find the probability that I get exactly two heads.

Lesson Plan

- Review Question

Lesson Plan

- Review Question
- Examples of places where statistical thinking helps

Review Question

- I flip a fair coin three times. Find the probability that I get exactly two heads.

Review Question

- I flip a fair coin three times. Find the probability that I get exactly two heads.
- Solution: There are eight ways to flip three coins. TTT, TTH, THT, THH, HTT, HTH, HHT and HHH, three of these have two heads. So $\frac{3}{8}$

Review Question

- I flip a fair coin three times. Find the probability that I get exactly two heads.
- Solution: There are eight ways to flip three coins. TTT, TTH, THT, THH, HTT, HTH, HHT and HHH, three of these have two heads. So $\frac{3}{8}$
- Solution 2: There are $\frac{1}{8}$ chances of getting 0 or 3 heads so a $\frac{3}{4}$ chance of getting one or two heads. These are symmetrical so equally likely. Hence the chance of getting exactly two heads is $\frac{3}{8}$.

Choosing a high-school

- The average Pretoria Boys' High school alumni has a net worth of over R150 million!

Choosing a high-school

- The average Pretoria Boys' High school alumni has a net worth of over R150 million!
- Pretoria Boys graduates about 300 learners a year and has been open for about 100 years so 30000 alums.

Choosing a high-school

- The average Pretoria Boys' High school alumni has a net worth of over R150 million!
- Pretoria Boys graduates about 300 learners a year and has been open for about 100 years so 30000 alums.
- Elon Musk (a Pretoria Boys alum) has a net worth of about four and a half trillion rand.

Choosing a high-school

- The average Pretoria Boys' High school alumni has a net worth of over R150 million!
- Pretoria Boys graduates about 300 learners a year and has been open for about 100 years so 30000 alums.
- Elon Musk (a Pretoria Boys alum) has a net worth of about four and a half trillion rand.
- The moral here is that averages can be misleading when we have outliers.

Gathering data

- The first discovered extra-solar planets have two characteristics. They tend to be far closer to their stars than our earth is to our sun and they tended to be far larger than the earth.

Gathering data

- The first discovered extra-solar planets have two characteristics. They tend to be far closer to their stars than our earth is to our sun and they tended to be far larger than the earth.
- Turns out these are the easiest planets to observe.

Gathering data

- The first discovered extra-solar planets have two characteristics. They tend to be far closer to their stars than our earth is to our sun and they tended to be far larger than the earth.
- Turns out these are the easiest planets to observe.
- The Literary Digest correctly predicted the U.S. election in 1920, 1924 1928 and 1932. In 1936 they pooled ten million people (with two million four hundred thousand responses) and predicted that Landon would beat FDR. Landon only won two states Maine and Vermont. The problem is that they sampled their relatively affluent readers who were skewed Republican.

Extrasolar planets and Elections

- The first discovered extra-solar planets have two characteristics. They tend to be far closer to their stars than our earth is to our sun and they tended to be far larger than the earth.

Extrasolar planets and Elections

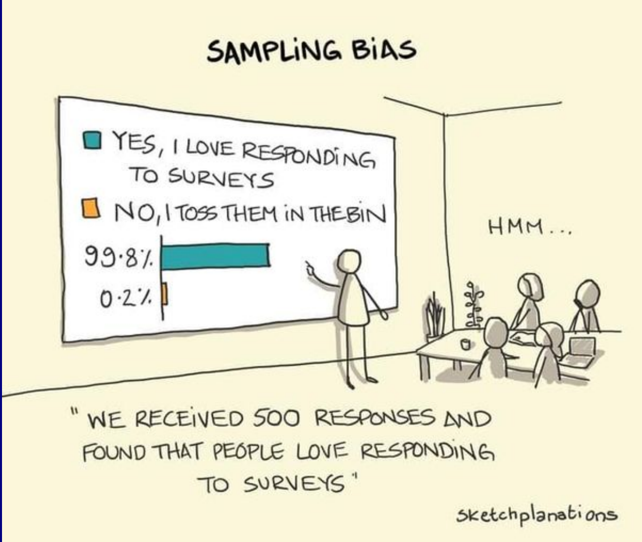
- The first discovered extra-solar planets have two characteristics. They tend to be far closer to their stars than our earth is to our sun and they tended to be far larger than the earth.
- Turns out these are the easiest planets to observe.

Extrasolar planets and Elections

- The first discovered extra-solar planets have two characteristics. They tend to be far closer to their stars than our earth is to our sun and they tended to be far larger than the earth.
- Turns out these are the easiest planets to observe.
- The Literary Digest correctly predicted the U.S. election in 1920, 1924 1928 and 1932. In 1936 they pooled ten million people (with two million four hundred thousand responses) and predicted that Landon would beat FDR. Landon only won two states Maine and Vermont. The problem is that they sampled their relatively affluent readers who were skewed Republican.

Selection Bias

Selection Bias



Fighter Planes

- In world war two the Allies had bomber planes that sometimes got slot at.

Fighter Planes

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?

Fighter Planes

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?
- Wald's great insight was to see that his data didn't include the places where the planes that didn't return were hit.

Fighter Planes

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?
- Wald's great insight was to see that his data didn't include the places where the planes that didn't return were hit.
- You need to put the armour were the bullets aren't!

Fighter Planes

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?
- Wald's great insight was to see that his data didn't include the places where the planes that didn't return were hit.
- You need to put the armour were the bullets aren't!
- As it happens this means over the fuel system and engine were a shot will bring them down.

Other Survivor bias examples

- In world war two the Allies had bomber planes that sometimes got slot at.

Other Survivor bias examples

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?

Other Survivor bias examples

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?
- Wald's great insight was to see that his data didn't include the places where the planes that didn't return were hit.

Other Survivor bias examples

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?
- Wald's great insight was to see that his data didn't include the places where the planes that didn't return were hit.
- You need to put the armour were the bullets aren't!

Other Survivor bias examples

- In world war two the Allies had bomber planes that sometimes got slot at.
- So they counted up the bullet holes on planes and asked Abraham Wald how much armour to put on the vunerable regions?
- Wald's great insight was to see that his data didn't include the places where the planes that didn't return were hit.
- You need to put the armour were the bullets aren't!
- As it happens this means over the fuel system and engine were a shot will bring them down.

The birthday problem

- There are 23 people enrolled in this course. What is the probability that two of you share a birthday?

The birthday problem

- There are 23 people enrolled in this course. What is the probability that two of you share a birthday?
- Let's make some simplifying assumptions. No February 29ths and all other days are equally likely.

The birthday problem

- There are 23 people enrolled in this course. What is the probability that two of you share a birthday?
- Let's make some simplifying assumptions. No February 29ths and all other days are equally likely.
- Well let's compute the probability that no two have the same birthday and let's start with fewer people

The birthday problem

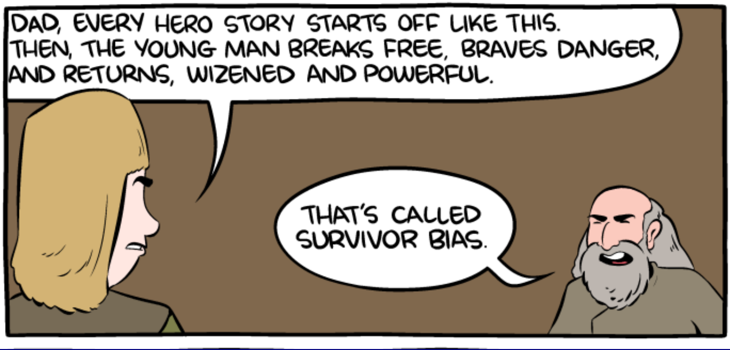
- There are 23 people enrolled in this course. What is the probability that two of you share a birthday?
- Let's make some simplifying assumptions. No February 29ths and all other days are equally likely.
- Well let's compute the probability that no two have the same birthday and let's start with fewer people
- One person: Probability 1 doesn't have the same birthday as himself.

The birthday problem

- There are 23 people enrolled in this course. What is the probability that two of you share a birthday?
- Let's make some simplifying assumptions. No February 29ths and all other days are equally likely.
- Well let's compute the probability that no two have the same birthday and let's start with fewer people
- One person: Probability 1 doesn't have the same birthday as himself.
- Two people: Probability $1(1-\frac{1}{365})$ don't share a birthday.

Survivor bias

Survivor bias



Survivor bias

Survivor bias

Survivor bias

Survivor bias

The birthday problem

- Three people: Probability $1(1 - \frac{1}{365})(1 - \frac{2}{365})$ don't share a birthday.

The birthday problem

- Three people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})$ don't share a birthday.
- Four people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})(1-\frac{3}{365})$ don't share a birthday.

The birthday problem

- Three people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})$ don't share a birthday.
- Four people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})(1-\frac{3}{365})$ don't share a birthday.
- Twenty three people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})...(1-\frac{22}{365})$ don't share a birthday.

The birthday problem

- Three people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})$ don't share a birthday.
- Four people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})(1-\frac{3}{365})$ don't share a birthday.
- Twenty three people: Probability $1(1-\frac{1}{365})(1-\frac{2}{365})...(1-\frac{22}{365})$ don't share a birthday.
- Don't share a birthday with probabilty: $1-1(1-\frac{1}{365})(1-\frac{2}{365})...(1-\frac{22}{365}) = 1-\prod_{i=0}^{22}(1-\frac{i}{365}) = 0.507297$

How many people live in your house

- We want to know how many people live in the average house so we collect a sample of 100 people. 50 say they live alone. The other 50 say they live in a house of 5 people. We conclude that the average house has 3 people in it, and that the split is even between houses with 1 person and houses with 5 people.

How many people live in your house

- We want to know how many people live in the average house so we collect a sample of 100 people. 50 say they live alone. The other 50 say they live in a house of 5 people. We conclude that the average house has 3 people in it, and that the split is even between houses with 1 person and houses with 5 people.
- That's wrong. **People are evenly split between living alone and five-person houses here. But it takes only a fifth as many five-person-houses to house the half of the population that lives together as it does one person houses that so that the mean house contains** $1 \times \frac{5}{6} + 5 \times \frac{1}{6} = 1.666$ **people.**

How many people live in your house

- We want to know how many people live in the average house so we collect a sample of 100 people. 50 say they live alone. The other 50 say they live in a house of 5 people. We conclude that the average house has 3 people in it, and that the split is even between houses with 1 person and houses with 5 people.
- That's wrong. **People are evenly split between living alone and five-person houses here. But it takes only a fifth as many five-person-houses to house the half of the population that lives together as it does one person houses that so that the mean house contains** $1 \times \frac{5}{6} + 5 \times \frac{1}{6} = 1.666$ **people.**
- You probably go to the dentist at a more crowded than average time. Most people do. That's what makes those times crowded.

How many people live in your house (size biassing)

- More generally let's say the the proportion of people living in a house of k people is p_k and the proportion of houses with k people is given by h_k

How many people live in your house (size biassing)

- More generally let's say the the proportion of people living in a house of k people is p_k and the proportion of houses with k people is given by h_k
- Then $h_k = \frac{\frac{p_k}{k}}{\sum_{i=1}^{\infty} \frac{p_i}{i}}$

How many people live in your house (size biassing)

- More generally let's say the the proportion of people living in a house of k people is p_k and the proportion of houses with k people is given by h_k
- Then $h_k = \frac{\frac{p_k}{k}}{\sum_{i=1}^{\infty} \frac{p_i}{i}}$
- Then $p_k = \frac{k h_k}{\sum_{i=1}^{\infty} i h_i}$

Monty Hall

- Monty Hall hosts a game show.

Monty Hall

- Monty Hall hosts a game show.
- He does this every week with the same rules so we can be pretty sure he's playing fairly.

Monty Hall

- Monty Hall hosts a game show.
- He does this every week with the same rules so we can be pretty sure he's playing fairly.
- He has three closed doors one of which has a prize (car) and the other two have booby-prizes (typically goats).

Monty Hall

- Monty Hall hosts a game show.
- He does this every week with the same rules so we can be pretty sure he's playing fairly.
- He has three closed doors one of which has a prize (car) and the other two have booby-prizes (typically goats).
- Monty let's the player choose a door. He then opens one of the other doors to reveal a goat.

Monty Hall

- Monty Hall hosts a game show.
- He does this every week with the same rules so we can be pretty sure he's playing fairly.
- He has three closed doors one of which has a prize (car) and the other two have booby-prizes (typically goats).
- Monty let's the player choose a door. He then opens one of the other doors to reveal a goat.
- He offers the player a chance to switch. Should the player?

Monty Hall

- Monty Hall hosts a game show.
- He does this every week with the same rules so we can be pretty sure he's playing fairly.
- He has three closed doors one of which has a prize (car) and the other two have booby-prizes (typically goats).
- Monty let's the player choose a door. He then opens one of the other doors to reveal a goat.
- He offers the player a chance to switch. Should the player?
- Yes, given certain knowns (which we have). That Monty knows where the car, that he always offers the chance to switch and that when the player chooses the car he chooses the door with the goat uniformly.