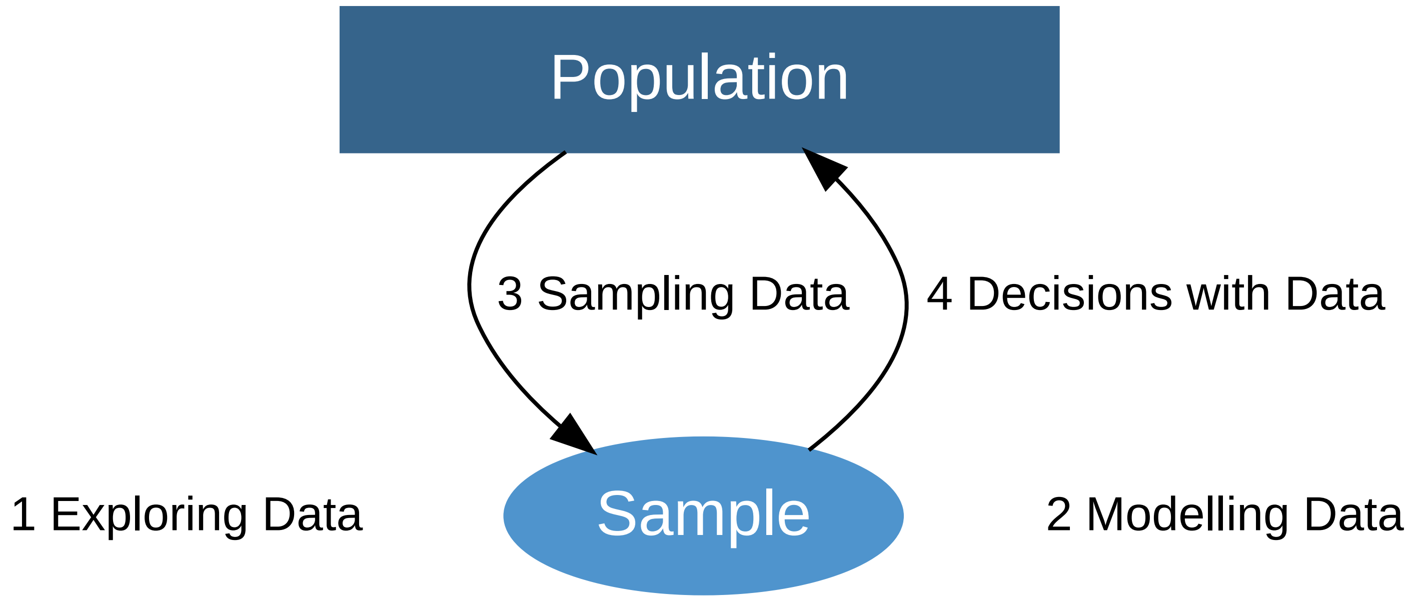
Chance

Sampling Data | Understanding Chance

© University of Sydney DATA1001/1901

Unit Overview



Module3 Sampling Data

Understanding Chance

What is chance?

Chance Variability

How can we model chance variability by a box model?

Sample Surveys

How can we model the chance variability in sample surveys?

Topic 19 Chance

Data Story | Did OJ murder his wife?

The prosecutor's fallacy

Properties of chance

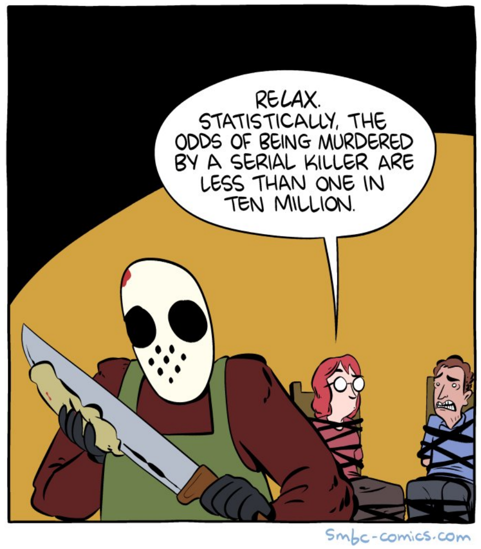
Conditional Probability

Summary

Data Story

Did OJ murder his wife?

Why study chance?



Meme

Why are many people scared of “probability”? Why is understanding “chance” important?

How Stats fools juries

OJ Simpson (1994-2017, US)

CBS News

Details of the case

- Orenthal James Simpson ("OJ") (July 1947-present) was a National Football League (NFL) player and actor.
  - In 1994, OJ was tried for the murders of his former wife Nicole Brown Simpson and her friend Ron Goldman.
  - In 1995, he was **acquitted**.
  - In 1997, a civil court awarded a $33.5 million judgment **against** Simpson for the victims' wrongful deaths.
- In 2007, OJ was arrested for armed robbery and kidnapping, and subsequently sentenced to 33 years imprisonment. In 2017, he was granted parole.

Sally Clark (1998-2003, UK)

Guardian

Details of the case

- Sally Clark (August 1964 15 March 2007) was an English solicitor.
  - In December 1996, her 1st baby son Christopher died suddenly at home.
  - In January 1998, her 2nd son Harry died in a similar way.
  - In February 1998, Sally was arrested, and in November 1999 was **convicted** of both their murders.
  - Paediatrician Professor Sir Roy Meadow testified that the chance of two children from an affluent family suffering from sudden infant death syndrome (SIDS) was 1 in 73 million.
- In January 2003, a 2nd appeal was successful in **overturning** the conviction.
  - It was discovered that the prosecution forensic pathologist Dr Alan Williams, who had examined both of the babies, had incompetently failed to disclose microbiological reports that suggested the second of her sons had died of natural causes.

People vs Collins (1968)

- On June 18 1964, Juanita Brooks was attacked in an alley near her home in Los Angeles and her purse stolen.
  - A witness John Bass reported that a woman in dark clothes and a dark blond pony tail ran from the scene and fled in a yellow car driven by a Negro man with a beard and moustache.
  - Police arrested a couple, Malcom Ricardo Collins and Janet Louise Collins, who fitted the description.
- In 1964, the defendants were found **guilty** of second degree robbery.
  - The prosecutors argued that there was a 1 in 12 million chance that a couple share these characteristics.
- In 1968, the Collins were **acquitted**.

Statistical Thinking

- What is common in all these trials?
- How can stats fool juries?

The prosecutor’s fallacy

The prosecutor’s fallacy

Prosecutor’s fallacy

The **prosecutor’s fallacy** is a mistake in statistical thinking, whereby it is assumed that the probability of a random match is equal to the probability that the defendant is innocent.

- It has used by the prosecution to argue for the guilt of a defendant during famous criminal trials.
- It can also be used by defense lawyers to argue for the innocence of their client.

Example of prosecutor’s fallacy

Facts

- Suppose there are about 5 million people living in Sydney.
- A murder occurs with DNA left on the weapon.
- A person matching the DNA is arrested.

Faulty argument

- The chance of a DNA match is 1 in 500,000 (very small).
- Hence, the chance that the arrested person is guilty is very high.

Error in thinking

- We fill out the following table:

	DNA Match	DNA doesn’t match
Guilty	1	0
Innocent	9	4,999,990

Note:

- Only 1 person is guilty and has a DNA match.
- No-one (0) is guilty and doesn’t match DNA.
- If 1 in 500,000 people matches DNA, then for a city size about 5 million, we expect 10 people to match (which is 1 guilty person and 9 innocent people).
- This leaves almost 5 million innocent people (4,999,990) that don’t match.

Hence:

- The chance that DNA matches, given innocent person = tiny!

$$\mathrm{P}(\text{DNA Match}|\text{Innocent}) = \frac{9}{4,999,999}$$

- But the chance that the person is innocent, given a DNA match = 0.9 = high!

$$\mathrm{P}(\text{Innocent}|\text{DNA Match}) = \frac{9}{10}$$

- Note $\mathrm{P}(\text{DNA Match}|\text{Innocent}) \neq P(\text{Innocent}|\text{DNA Match})$
- So for any person with DNA match, we can't say $\mathrm{P}(\text{Guilty}|\text{DNA Match})$ is high.

$$\mathrm{P}(\text{Guilty}|\text{DNA Match}) = 1 - \mathrm{P}(\text{Innocent}|\text{DNA Match}) \neq 1 - \mathrm{P}(\text{DNA Match}|\text{Innocent}) = 1 - \frac{9}{4,999,999}$$

Mistake in the OJ case

- **Fact**: The crime scene blood matched Simpson's blood, with characteristics shared by 1 in 400 people.
- **Mistake (confusion about conditional probability)**: The prosecution presented evidence that Simpson had been violent toward his wife, while the defense argued that there was only 1 woman murdered for every 2500 women who were subjected to spousal abuse, and that any history of Simpson being violent toward his wife was irrelevant to the trial. It was ignored that Simpson's wife had not only been subjected to domestic violence, but subjected to domestic violence and murdered.
    - **Defence lawyer:** $\mathrm{P}(\text{murder}|\text{domestic violence}) = \frac{1}{2500}$
    - **Prosecution:** $\mathrm{P}(\text{murder}|\text{death}) = \frac{8}{9}$

Article

Mistakes in the Sally Clark case

- **Mistake1 (dependent probabiity)**: The chance of 2 SIDS deaths in an affluent family was claimed to be 1 in 73 million (tiny). This figure was improperly derived by ignoring the association between events. So The Royal Statistical Society in the UK issued a public statement pointing out the statistical invalidity of this number.
- **Mistake2 (Prosecutor's Fallacy)**: The chance Sally Clark was guilty was said to be very high.

$$\mathrm{P}(\text{Guilty}|2\text{ deaths}) = 1 - \mathrm{P}(\text{Innocent}|2\text{ deaths}) \neq 1 - \mathrm{P}(2\text{ deaths}|\text{Innocent}) = 1 - \mathrm{P}(2\text{ SIDS}) = 1 - \frac{1}{73M}$$

- **Mistake3 (ignoring evidence)**: The evidence of natural causes for the second death was ignored.

Mistakes in the People vs Collins case

- **Mistake1 (dependent probability)**: The chance of finding a bearded black man accompanied by a ponytailed blonde in a yellow car was claimed to be 1 in 12 million (tiny). Again, this probability is tiny because the association between events is ignored.

- **Mistake2 (Prosecutor's Fallacy)**: The chance that the arrested couple were guilty given a match was claimed to be very high.

Properties of Chance

What is chance?

Chance (frequentist definition)

Chance (or probability) is the percentage of time a certain event is expected to happen, if the same process is repeated long-term.

Basic properties of Chance

1. Chances are between 0% (impossible) and 100% (certain).

$$\mathrm{P}(\text{Impossible event}) = 0$$

$$\mathrm{P}(\text{Certain event}) = 1$$

2. The chance of something equals 100% minus its opposite (**complement**).

$$\mathrm{P}(\text{Event}) = 1 - \mathrm{P}(\text{Complement event})$$

3. Drawing at **random** means that a collection of objects have the same chance of being picked.

Statistical Thinking

Rates of SIDS vary in developed countries from one in a 1000 to one in 10 000.

- How were these chances determined?
- What is the chance of a baby not having SIDS?
- What other factors might need to be considered?

Conditional probability

Conditional probability

Conditional probability

Conditional probability is the chance that a certain event occurs, given another event has occurred.

$$\mathrm{P}(\text{Event 1} \mid \text{Event 2})$$

Multiplication Principle

The probability that 2 events occur is the chance of the 1st event multiplied by the chance of 2nd event, given the 1st has occurred.

$$\mathrm{P}(\text{Event 1 and Event 2 occur}) = \mathrm{P}(\text{Event 1}) \times \mathrm{P}(\text{Event 2} \mid \text{Event 1})$$

Statistical Thinking

Assume a rate of 1 in 8543 SIDS deaths per live births in the UK (assuming no risk factors). We want to work out the chance that 2 babies in a family die of SIDS? Is this true?

$$\mathrm{P}(2 \text{ SIDS deaths}) \stackrel{?}{=} \mathrm{P}(1 \text{ SIDS death}) \times \mathrm{P}(1 \text{ SIDS death}) = 1 \text{ in } 73 \text{ million}$$

Article

Example

In a pack of 52 cards there are 4 aces. Assume we draw without replacement.

- What is the chance of drawing 2 aces?

P(2 ♠A) = P(1st ♠A) $\times$ P(2nd ♠A|1st ♠A) = $\frac{4}{52} \times \frac{3}{51} = 0.0045$

- What is the chance of drawing 2 kings?

Note: Drawing without replacement often results in dependent events, as the probability of each event depends on the outcome of the previous event. So it is not simply a product of each unconditional probability.

Independence & dependence

Independence

- 2 events are **independent** if the chance of the 2nd given the 1st is the same as the 2nd, ie P(2nd event|1st event) = P(2nd event) independent of the outcome of 1st event

- Drawing randomly with **replacement** ensures independence.

- If 2 events are independent, then the chance of both occuring is the product of their unconditional probabilities. [special case of multiplication principle]

$$\text{P}(\text{Event 1 \& Event 2 occur}) = \text{P}(\text{Event 1})\text{P}(\text{Event 2})$$

Dependence

- 2 events are **dependent** if the chance of the 2nd given the 1st is not the same as the chance of the 2nd, that is,

$$P(\text{2nd event given 1st event}) = P(\text{2nd event}|\text{1st event}) \neq P(\text{2nd event}),$$

as it depends on the result of the 1st event.

- Drawing **without replacement** ensures dependence.

Statistical Thinking

In the Sally Clark case, are the deaths of the 2 babies dependent or independent?

Independent:

$$\mathrm{P}(2\ \mathrm{SIDS}) \stackrel{?}{=} \mathrm{P}(\mathrm{SIDS}) \times \mathrm{P}(\mathrm{SIDS}) = \left(\frac{1}{8543}\right)^2 = 1 \text{ in } 73 \text{ million}$$

Dependent:

$$\mathrm{P}(2\ \mathrm{SIDS}) \stackrel{?}{=} \mathrm{P}(1\mathrm{st}\ \mathrm{SIDS}) \times \mathrm{P}(2\mathrm{nd}\ \mathrm{SIDS} \mid 1\mathrm{st}\ \mathrm{SIDS})$$

Note: $\mathrm{P}(2\mathrm{nd}\ \mathrm{SIDS} \mid 1\mathrm{st}\ \mathrm{SIDS})$ tends to be higher because of possible association between deaths.

Statistical Thinking

In the Collins case, why is this working not correct?

Characteristic	Chance
Black man with beard	1 in 10
Man with mustache	1 in 4
White woman with pony tail	1 in 10
White woman with blond hair	1 in 3
Yellow motor car	1 in 10
Interracial couple in car	1 in 1,000

Working: The probability of matching all characteristics is the product of all the individual chances = 1 in 12 million.

Example

A coin is tossed twice. If the coin lands heads on the 2nd toss, you win \$1.

- If the 1st coin is a head, what is the chance of winning \$1?
- If the 1st coin is a tail, what is the chance of winning \$1?
- Are the tosses independent?
- What is the chance of winning \$1?

Answers:

1. Event HH out of {HH,HT}; P(2nd H| 1st H) = $\frac{1}{2}$. 2. Event TH out of {TH,TT}; P(2nd H| 1st T) = $\frac{1}{2}$. 3. Yes. 4. $\frac{1}{2}$ regardless of the first event.

Summary

Key Words

prosecutor’s fallacy, chance, subjective, frequentist, theoretical, complement, random, conditional probability, multiplication principle, independence, dependence,

Further Thinking

Keith Devlin

Calling Bullshit 4.3: P Values and the Prosecutor’s Fallacy

The Prosecutor's Fallacy