Applied Research Study Session 2: The World as a Data-Generating-Process

Dr. L.S. Sanna Stephan

2023

RUG Groningen

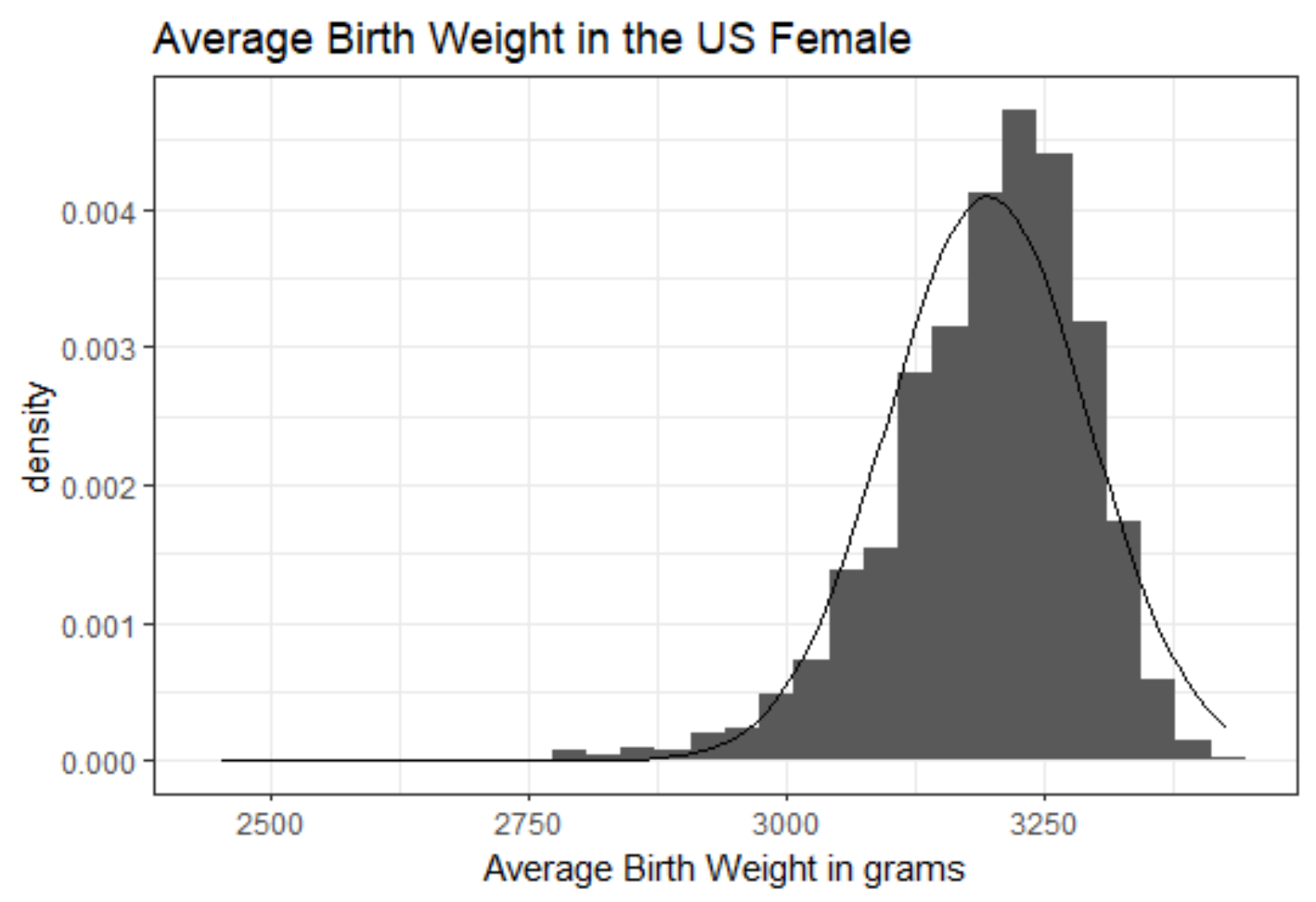
contents

We will discuss

1. examples of probability distributions in the real world,
2. how people assess the probability of an uncertain event or the value of an uncertain quantity,
3. what is p-hacking and
4. what is the Simpson's paradoxon.

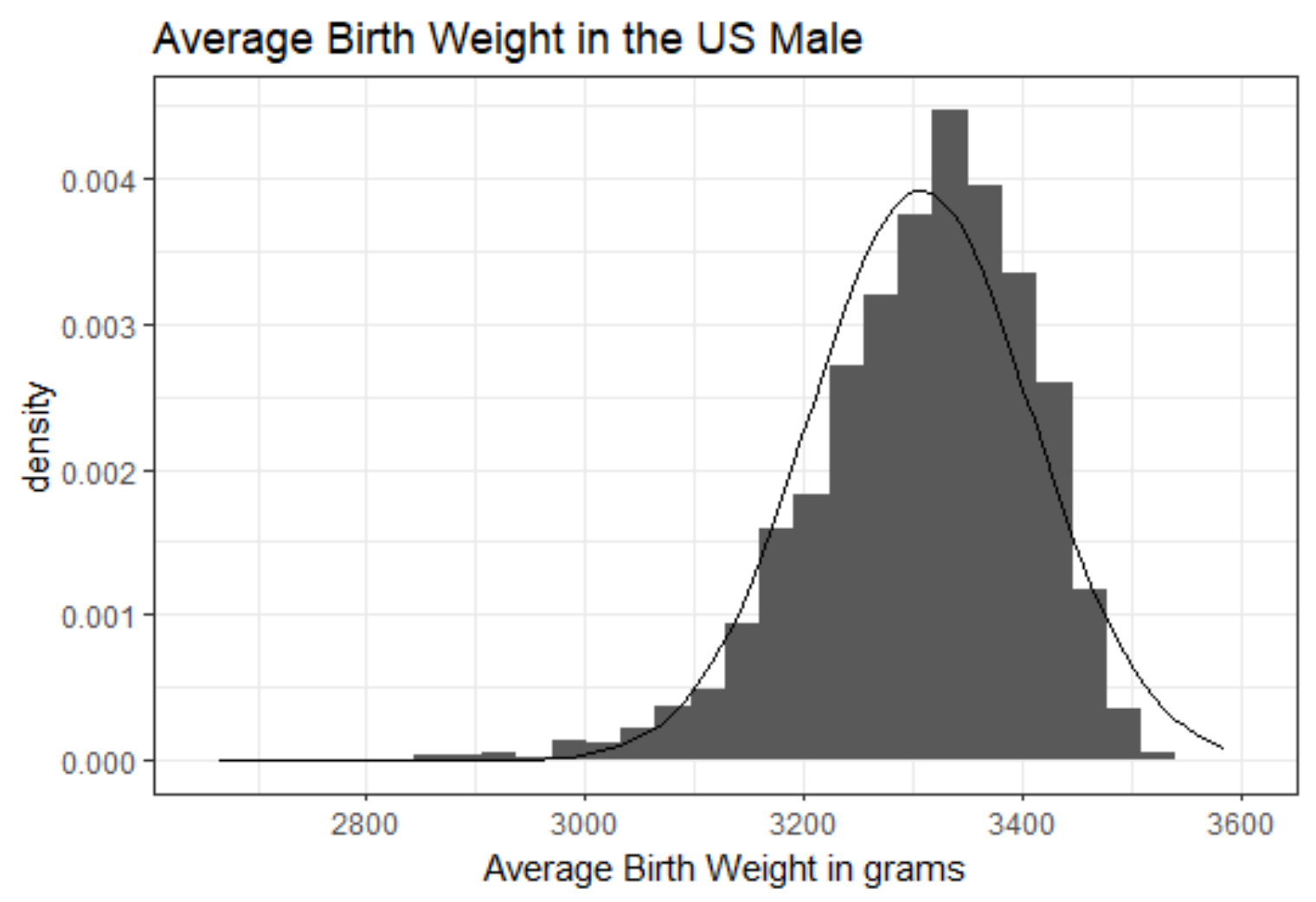
Probability Distributions in the Real World

Examples of probability distributions in the real world



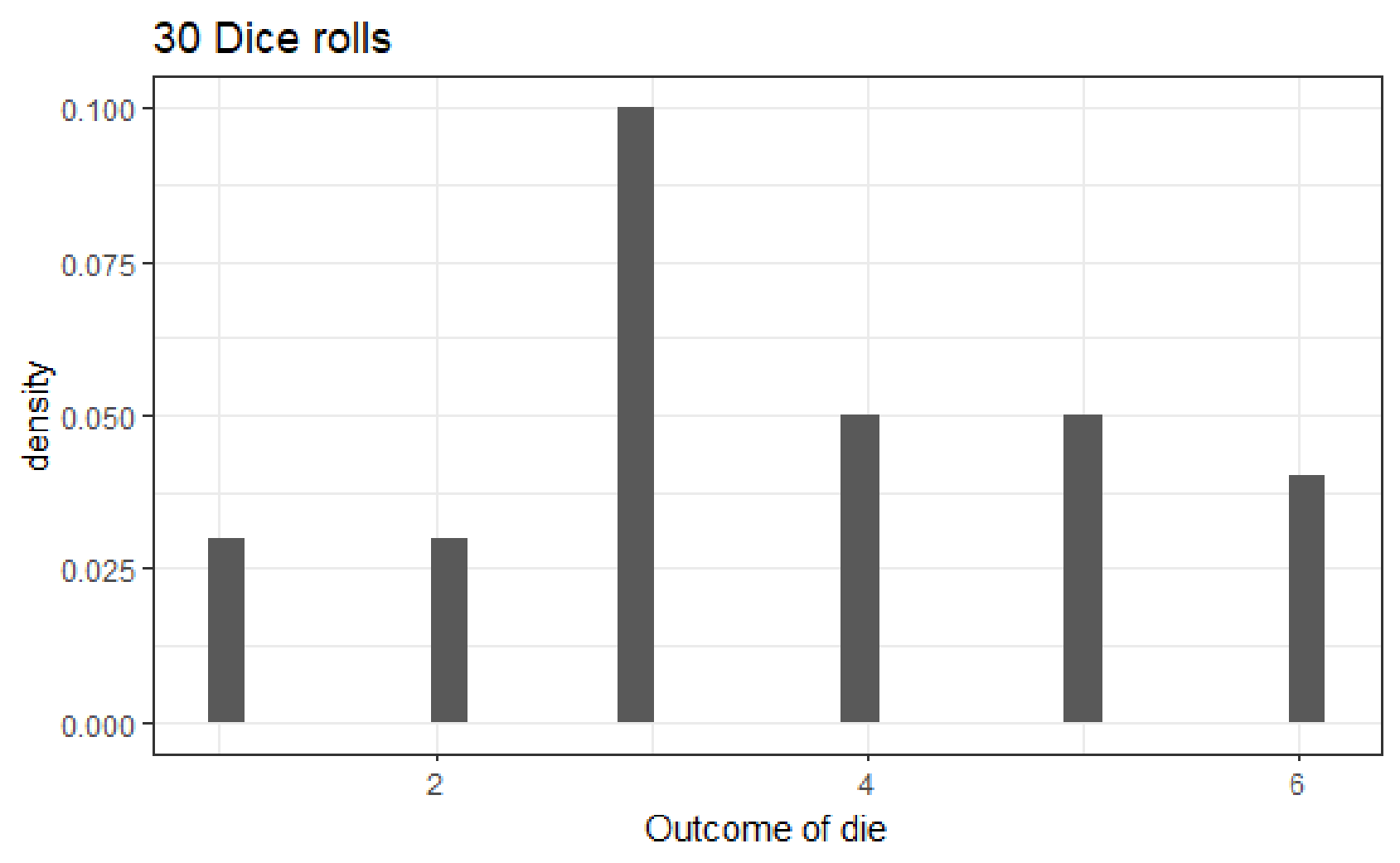
Data Source: CDC WONDER

Examples of probability distributions in the real world

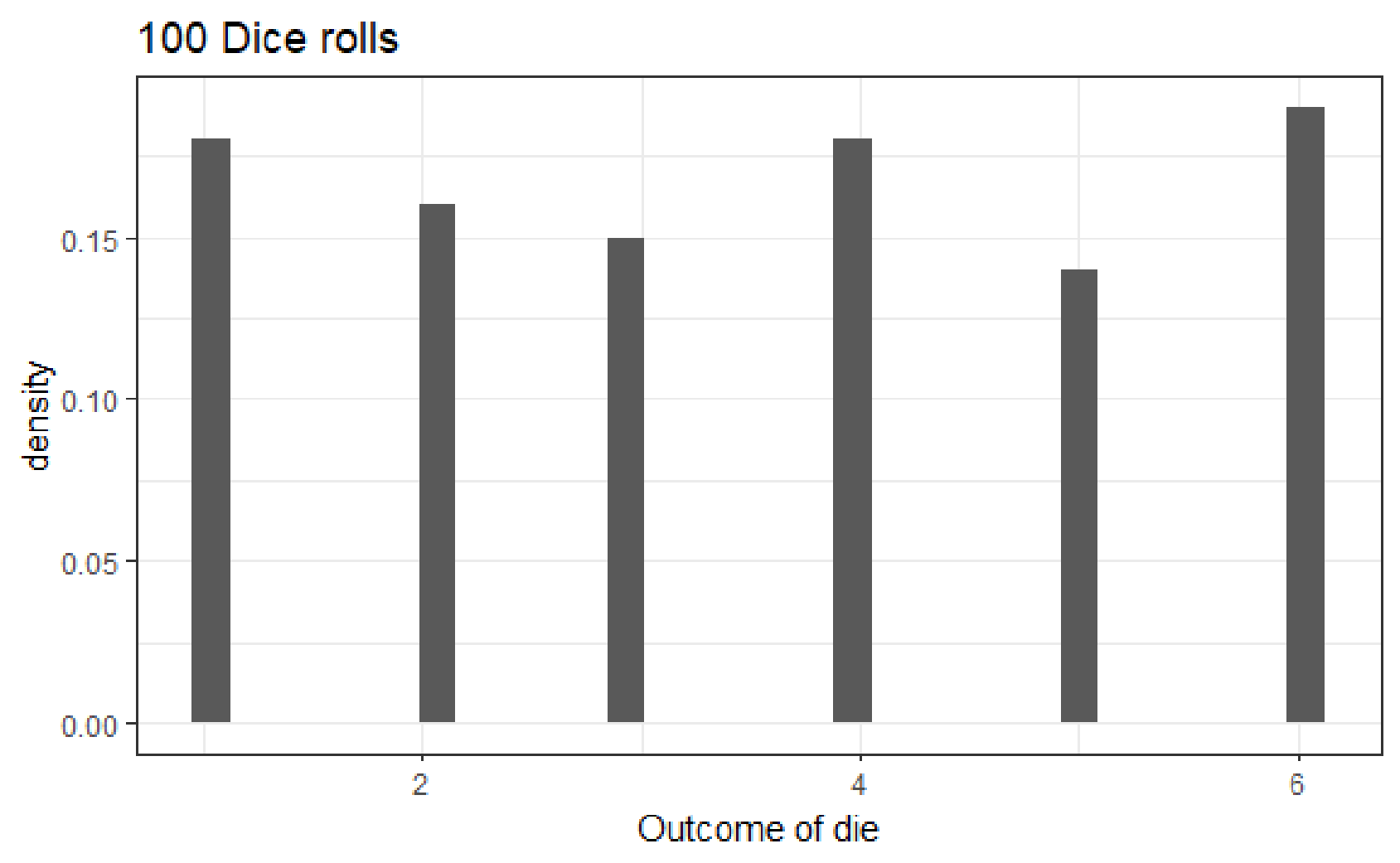


Data Source: CDC WONDER

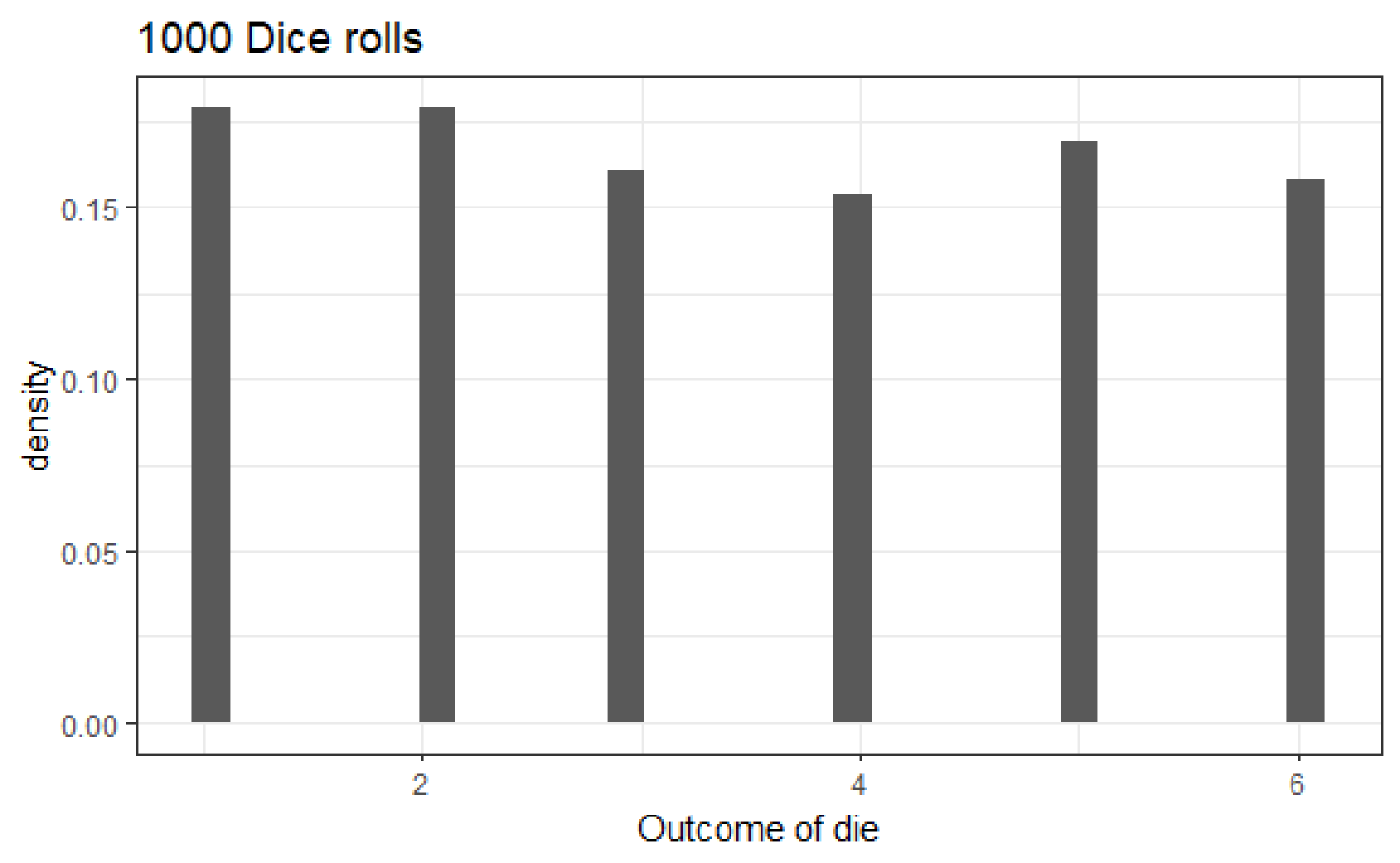
Examples of probability distributions in the real world



Examples of probability distributions in the real world



Examples of probability distributions in the real world



Examples of probability distributions in the real world

Example of the binomial distribution in the real world

If 10% of the orders get returned and you received 80 orders this week.

For example:
$\mathbf{P}(returns = 6) = \binom{80}{6} 0.1^6 \cdot 0.9^{74} = 0.12354$
$\mathbf{P}(returns < 6) = \mathbf{P}(returns = 0) + \mathbf{P}(returns = 1) + ... + \mathbf{P}(returns = 5) = 0.17692$

How Do People Assess the Probability of an Uncertain Event or the Value of an Uncertain Quantity?

Quiz

- Two hospitals: large (45 babies per day) and small (15 babies per day).
- Population: 50 % of babies boys.
- Daily average varies.
- During one year, record days on which percentage of boys exceeds 60

  Which hospital do you think recorded more such days?

a) The larger hospital

b) The smaller hospital

c) About the same (that is, within 5 percent of each other)

Quiz

- Survey families with six children.
- In 72 families the exact order of boys and girls was G B G B B G.

  In, on average, how many families will the exact order of births was B B B B B B?

Quiz

- Sample word from English text.

  It is more likely that...

a) the word starts with a "k" or

b) the word has a "k" in third position ?

Heuristics

- Judge uncertain event $\Rightarrow$ use **heuristics** instead of probability theory.
- Heuristic: rule that is **easy** and **roughly right in many cases**.
- But: common heuristics are known $\Rightarrow$ predicatble and systematic error in peoples' forecasting.

Represtativeness heuristic

Representativeness heuristic

The subjective probability of an event, is determined by the degree to which it:

(i) is similar in essential characteristics to its parent population

(ii) reflects salient features of the process by which it is generated

Representativeness heuristic

Insensitivity to prior probability outcomes

- Experiment: some participants were shown brief (useless) personality descriptions.
- Task: How likely is it that the person is an engeneer/ a lawyer?
- Group 1: "Person drawn from 70 engineers and 30 lawyers".
- Group 2: "Person drawn from 30 engineers and 70 lawyers"

Useless information

"Dick is a 30 year old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues."

Representativeness heuristic

- No information: correct judgement on average.
- Useless information: both groups deviate from information on **actual** number of lawyers and engeneers.
- Both: 50%.
- Prior probabilities effectively ignored.

Representativeness heuristic

Insensitivity to sample size

Hospital question: most subject choose c) (same probability for large and small hospital).

But this is incorrect!!

Representativeness heuristic

- More data $\Rightarrow$ less deviation from average!
- 60 % is a deviation.
- More likely in **small** hospital.

  People's judgement is incentive to sample size.

Representativeness heuristic

Misconceptions of chance

- People think sequences generated by a random process must represent the essential characteristics of that process to be likely to occur.
- True for **average properties** but **not for individual events**.
- Quiz 2: actually, any **particular sequence is equally likely**!
- $P(GBGBBG) = 0.5^6 = P(BBBBBB)$
- BUT: if sequence is long enough, then relative frequency of boys 50%

Availability Heuristic

Availability heuristic

- Probability of an event is evaluated by availability (ease with which examples come to mind).
- Quiz 3: easier to think of words that start with k than words in which k is third letter.
- People choose a) (start with k)
- Reality: typical text contains twice as many words in which a "k" is in third position than words that start with "k".

Availability heuristic

Biases of imaginability

- 10 people need to form committee with r members.
- $2 \leq r \leq 8$
- For which r do they have the most options for the committee?

Availability heuristic

- Smaller committees easier to imagine.
- Larger committees are far less distinct, harder to picture.
- small committees appear more numerous than larger committees.
- Reality: $\binom{10}{r}$ which reaches a maximum of 252 for $r = 5$

what you have learned

- You have seen that real world data follow certain probability distributions
- People do not follow the principles of probability theory. Instead they use heuristics, that lead to predictable errors in certain situations

P-hacking

P-hacking

- Statistically significant results are more likely to be published.
- P-hacking involves manipulating data analysis to achieve statistically significant results.
- Common tactic: manipulating the sample.

Cherry-Picking

- Selectively choosing data points that support your hypothesis.
- Ignoring data that contradicts the desired outcome.
- Can lead to overestimation of effects and false positives.

Data Trimming

- Removing outliers or extreme values to influence the results.
- May lead to a biased representation of the population.
- Can distort the true effect size and significance.

Subgroup Analysis

- Splitting the sample into subgroups to find statistically significant effects.
- Increases the chance of finding significant results by chance.

Data Transformation

- Applying various transformations until a desired result is achieved.
- Logarithmic, exponential, or power transformations.
- Can distort the interpretation of the effect's practical significance.

Data Exclusion

- Excluding certain data points post-analysis to achieve significance.
- Should define exclusion criteria before analysis to avoid bias.
- Post hoc exclusions can lead to misleading conclusions.

Consequences

- P-hacking leads to untrue and misleading conclusions.
- P-hacking undermines scientific integrity and reproducibility.

Drawing the Wrong Conclusion due to Simpson's Paradox

Introduction to Simpson's Paradox

- Statistical phenomena: trend is present in entire sample, but not present or reversed sub-samples.
- Understanding Simpson's Paradox is crucial to avoid misleading conclusions.
- First we look at a graphical example.

Graphical Example

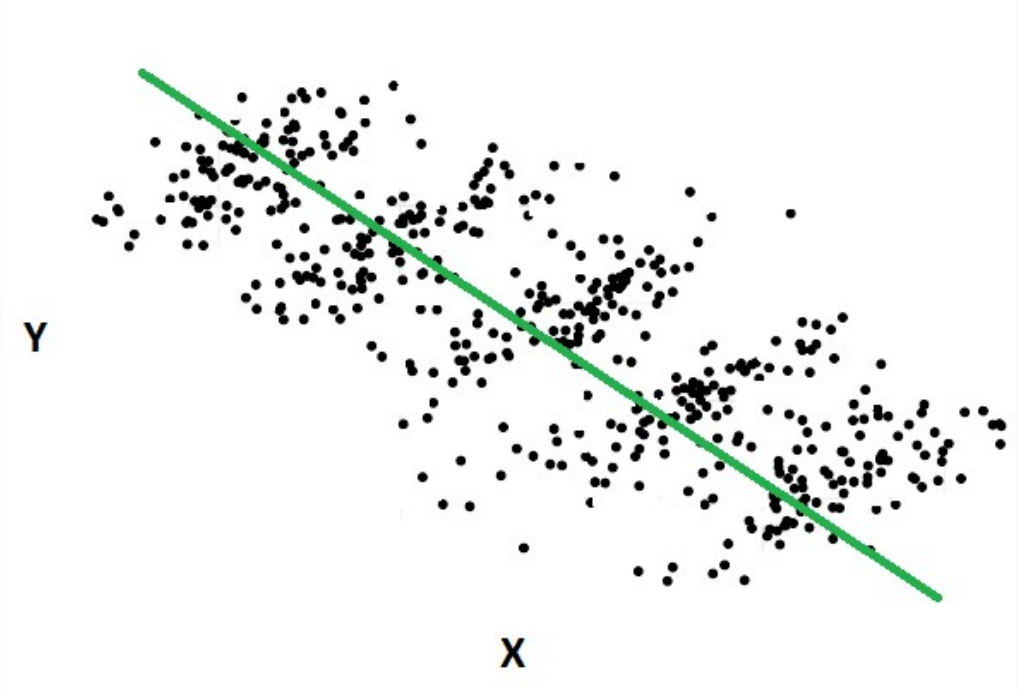


Figure 1: One group

Entire sample: negative correlation between the X and Y variable.

Graphical Example

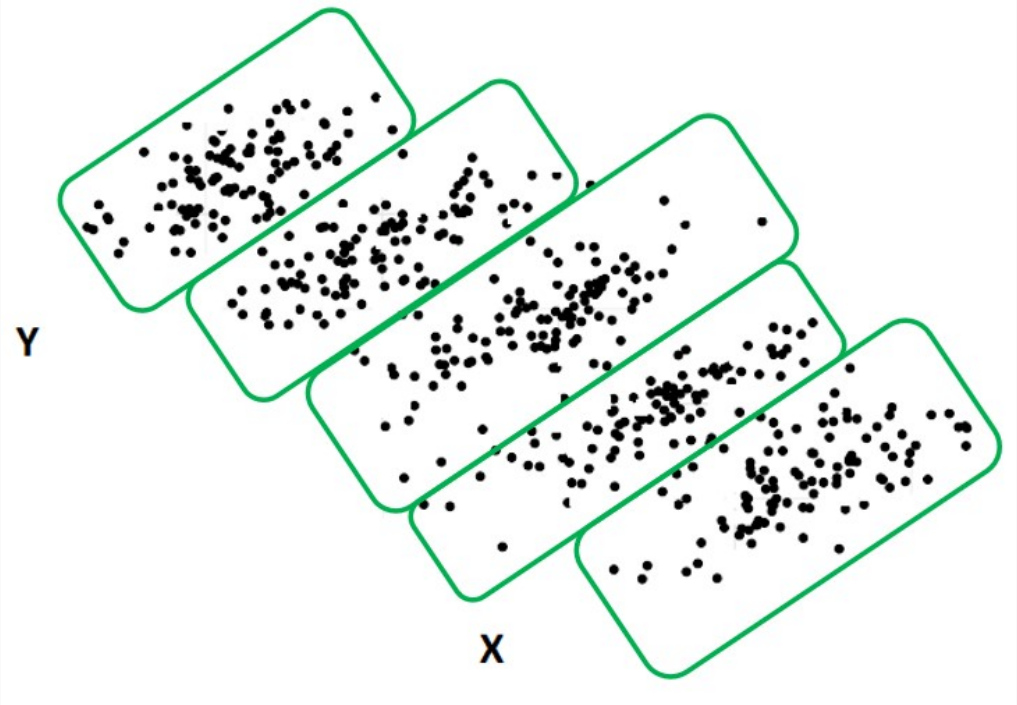


Figure 2: Subgroups

Sub-sample: positive relationship between X and Y.

Why Does Simpson's Paradox Occur?

- Simpson's Paradox arises due to the influence of confounding variables.
- Confounders affect the relationship between two variables and need to be considered for accurate interpretation.
- Confounders not taken into account when data aggregated.

Explaining the Simpson's Paradox with an Example

- UC Berkeley's admissions data.
- Initial data seems to show a preference for men in admissions.

Men	Women
45%	30%

Table 1: Overall acceptance rates

Explaining the Simpson's Paradox with an Example

- However, there is more to the story thanks to Simpson's Paradox!
- Aggregating the data from all departments removes departmental differences from the analysis.
- Some departments have much lower acceptance rates than others, making them more selective.

Explaining the Simpson's Paradox with an Example

- The following two factors create the misleading, unbalanced acceptance rates in the previous table:
  - Women tended to apply for the harder departments, lowering their overall acceptance rate.
  - Men were inclined to apply for the easier departments, boosting their rates.

Explaining the Simpson's Paradox with an Example

- To determine whether the selection process favors men, we need to assess the data at the departmental level and compare acceptance rates within each department.
- This method holds each department's acceptance rate constant, allowing for valid comparisons.

Explaining the Simpson's Paradox with an Example

Department	Men	Women
1	62%	82%
2	63%	68%
3	37%	34%
4	33%	35%
5	28%	24%
6	6%	7%

Figure 3: Department Admission Rates

- Comparing the rates within departments paints a different picture. Women have a slight advantage over men in most departments.
- The subgroup analysis accounts for the confounding variable of the varying admission rates.

Simpson's Paradox: Conclusion

- To avoid this type of confusion, researchers must carefully consider the level of data aggregation and carefully examine the data for potential confounding variables that could influence the results.
- By doing this, they can ensure that their study results accurately reflect the underlying trends and patterns in the data.

References

1. Kahneman, Daniel, and Amos Tversky. "Subjective Probability: A Judgment of Representativeness." Cognitive Psychology, vol. 3, no. 3, July 1972, pp. 430–454.
2. Tversky, Amos, and Daniel Kahneman. "Availability: A Heuristic for Judging Frequency and Probability." Cognitive Psychology, vol. 5, no. 2, Sept. 1973, pp. 207–232
3. Tversky, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases." Science, vol. 185, no. 4157, 27 Sept. 1974, pp. 1124–1131
4. Frost, Jim. "Simpsons Paradox Explained." Statistics by Jim