Problem Set 4: Sampling and data collection

ADS2

Semester 1 2023/24

We expect this problem set to take around an hour to complete. But professors are sometimes wrong![citation missing] If this or future problem sets are too long, please let us know, so we can adjust and plan accordingly.

Learning Objectives

- Explain the relationship between a population and a sample
- Explain the concept of sampling bias
- Give examples of sampling biases that can occur
- Design data collection procedures that avoid sampling bias

Sampling from a population

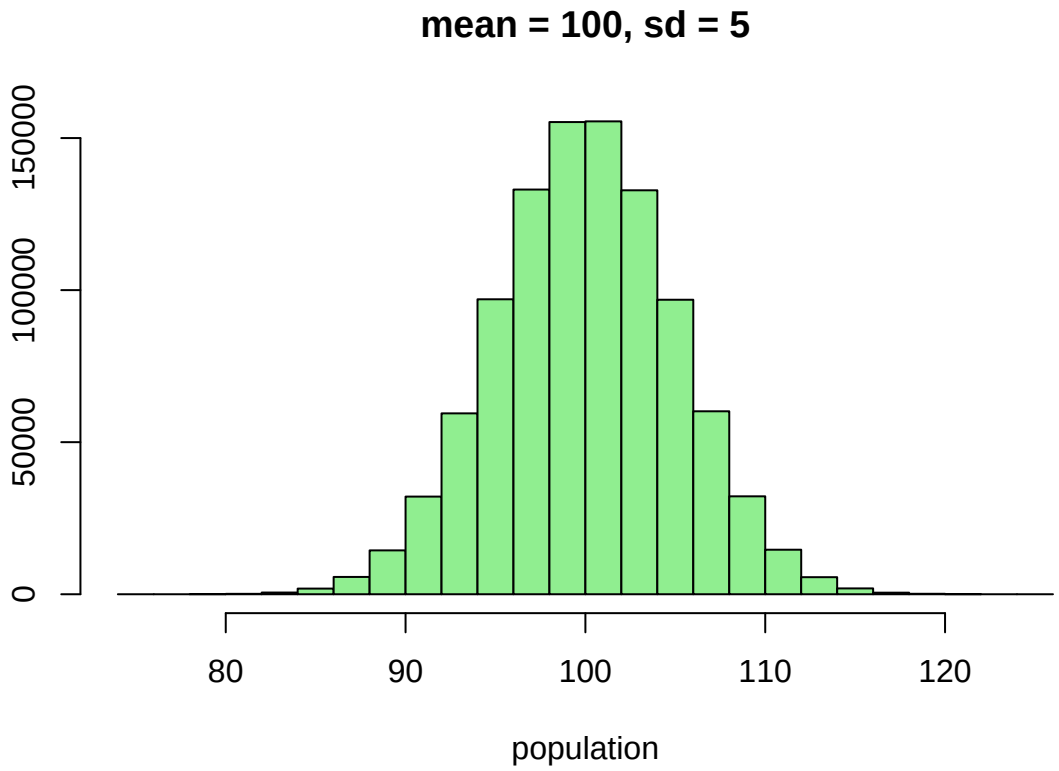
In the lecture, you saw a population of 1 million individuals, with mean 100 and standard deviation 5. This was generated and characterised using the following R code:

```
population <- rnorm(1e6,100,5)
popmean <- round(mean(population),1)
popmean
```

```
## [1] 100
```

```
popsd <- round(sd(population),1)
popsd
```

```
## [1] 5
```



- We talked a bit about sampling error, but let's take a closer look. If you take a sample of 5 individuals from this population, what would you expect the mean to be (approximately)? What would you expect the standard deviation to be (approximately)?
- Test your intuition by taking 1000 samples of size 5 and record the mean and standard deviation for each sample. Plot all the means and standard deviations (for instance, as a histogram). Take the mean of the means and the mean of the standard deviations. What do you see?
- How would you expect this to change if your samples were bigger? Make a prediction first, then test it with samples of size 100!

Breakfast and student performance

According to a BBC news article, a healthy breakfast helps pupils perform well in class [S. Messenger. Healthy breakfast 'helps pupils do well'. BBC News, 17 November 2015]

This article is based on a research paper by Littlecott et al. [H.J. Littlecott et al. (2015). Public Health Nutrition 19(9):1575-1582]

For the study, the scientists interviewed children between the ages of 9 and 11 about what they had eaten the day before, and later recorded those students' academic performance over the next months.

They found that students who ate a healthy breakfast did better in school than students who did not.

- In what way could this sample be biased?
- If there is sampling bias, do you think it affects the conclusions of the study?
- How could a study be designed to avoid sample bias?

Getting information about sensitive topics

In survey studies, it is sometimes difficult to get honest answers to questions that relate to a person's privacy, or questions about behaviours that are illegal, stigmatised or in other ways socially undesirable. For instance, if people are asked whether they have ever shoplifted, it is unlikely that everybody will answer honestly. People who have previously shoplifted may decline to participate in the survey (non-response bias), or they may participate and lie about their behaviour (social desirability bias).

One way to avoid this bias is the "unrelated question randomised response" technique [Greenberg et al. (1969), Journal of the American Statistical Association 64(326):520-539]. It works like this:

- Ask the participant to flip a coin. Only the participant can see the outcome of the coin flip.
- If the coin lands on "head", the participant is asked to answer question A: "Were you born between 1 January and 30 June?"
- If the coin lands on "tail", the participant is asked to answer question B: "Have you ever shoplifted?"

- Let's try it!
  - Throw a coin in R (*how?*)
  - Answer the question according to the instructions.
- Can you see how this design makes it easier for participants to answer truthfully?
- Imagine a study as described above. Out of 300 participants, 112 answered "yes" and 188 answered "no". Assuming this was a representative sample, what percentage of the population has a history of shoplifting?

Last update by DJ MacGregor in 2023