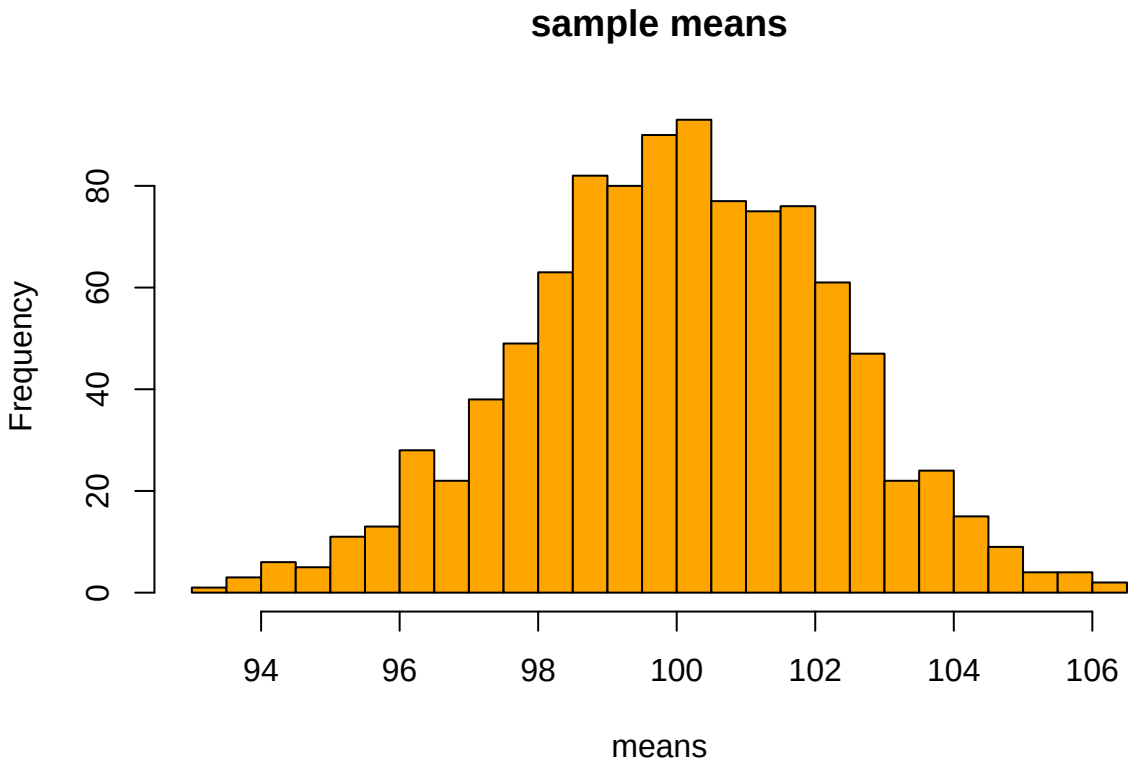
Problem Set 4 Notes: Sampling and data collection

ADS2

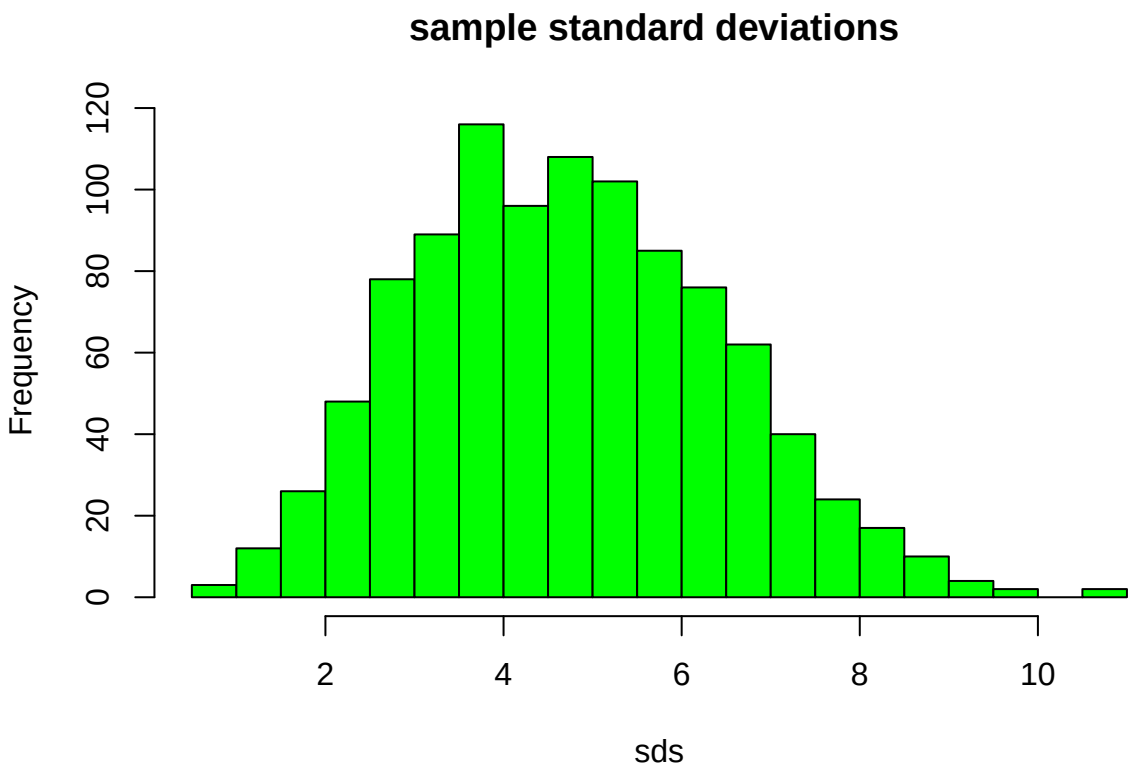
Semester 1 2023/24

Sampling from a population

Here is what we got for samples of size 5 (when sampling 1000 times):




```
## Mean of sample means:   100.0143
## Population mean:   100.0064
```




```
## Mean of sample standard deviations:   4.729067
```

```
## Population sd:   4.997899
```

This looks like the estimate of the mean from our sample is pretty accurate, but that our samples are underestimating the true (population) standard deviation. Is this consistent with what you found? Is it different if the sample is bigger? (We are going to talk more about this in the next lecture!)

Breakfast and student performance

There are several ways in which sampling bias could come into this study. Here are a few that we thought about:

- Having a healthy breakfast may not in itself increase student performance. Instead, it may be a "symptom" of other factors that make a student more likely to succeed in school (for instance, having parents or carers who look after them well, being healthy, coming from a household with enough money, having an organised morning routine, . . .) By just asking about eating, the researchers may miss the real underlying cause.
- Asking students what they ate in a particular 24 hour period may not be representative of their typical or long-term eating behaviours.
- Because students at that age know what kind of food is healthy and that a healthy breakfast is a good thing, they may not necessarily tell the truth about their own eating behaviours (social desirability bias).
- By asking pupils in school, the authors may miss children who are ill or who are educated at home.

You may have thought about yet other possible sources of bias. Not all possible sources of bias are equally important - for instance, the proportion of children educated at home may be very small, so missing them in the study would not affect the overall outcomes.

Depending on the source of bias, there are some alternative study designs that could be useful. For instance, an unrelated question randomised response method (see below) could encourage more students to answer the question truthfully.

Another approach would be not to ask students, but to give everybody a healthy breakfast for a period of time and assess how their academic performance changes. This would help distinguish between the effects of breakfast itself and other factors in the students' home life.

Getting information about sensitive topics

You can solve this problem approximately by running a simulation in R, similar to what you did last week. Or you could do the maths. Here, we are mathsing it out, because it is an opportunity to review conditional probabilities and probability trees.

Here is a tree of what happens in this questionnaire:

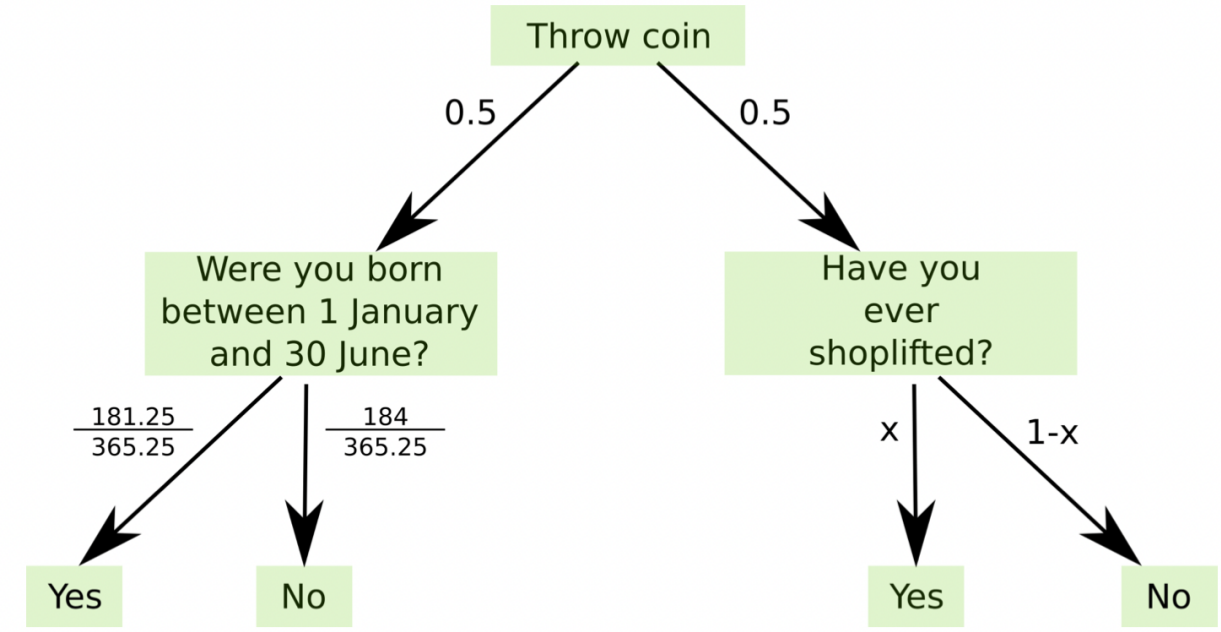


Figure 1: Probability tree for unrelated question randomised response example

What we want to know is of course x, the probability that the person has shoplifted (giving us the proportion of people who have shoplifted).

We get probabilities by multiplying along the branches of a tree. For instance, the probability that someone threw "Head" and answered "Yes" is $0.5 \times \frac{181.25}{365.25}$.

The total number of "Yes" answers is the total probability of "Yes" answers times the number of participants

$$112 = 300 \times \left(0.5 \times \frac{181.25}{365.25} + 0.5 \times x\right)$$

Re-arranging a bit gives:

$$\frac{2 \times 112}{300} - \frac{181.25}{365.25} = x$$

```
x = 2*112/300-181.25/365.25
x
```

```
## [1] 0.2504312
```

This would suggest that around a quarter of participants have shoplifted.

(Just to check our numbers are correct, test whether number of "No" answers is correct)

```
nos <- 300*(0.5*184/365.25 + 0.5*(1-x))
nos
```

```
## [1] 188
```

Note that even though we can do the math here, this is not an exact answer, because we don't know how many times exactly the coin landed on Heads. All we can do is state our best guess based on a 50:50 probability.

Last update by DJ MacGregor in 2023