Chapter 20: chance error in sampling

Context **2**
Overview . . . 3
Population and parameter. . . . 4
Sample and statistic . . . 5

Expected value and SE for percentage **6**
Example . . . 7
When we do this 10,000 times... . . . 8
Example . . . 9
Strategy. . . . 10
Example . . . 11
Formulas for percentage . . . 12
Exercise . . . 13
Accuracy depends on sample size. . . . 14
How does accuracy depend on sample size?. . . . 15
Compare sample size 100 and 400, for 10,000 repetitions... . . . 16
Normal approximation . . . 17
Other sampling methods . . . 18

Overview

- So far we have looked at simple chance processes: flipping coins, rolling dice, playing roulette
- We will now look at a new chance process: sampling
- Outlook for the coming lectures:
  - Today: Assume we know the box/population, and study counts and percentages in samples (Ch 20)
  - Next: We do not assume that we know the box/population. From the sample, we want to make inference about the population, like in election polls (Ch 21, 23).

3 / 18

Population and parameter

- Recall: Usually we want to know a parameter (=numerical fact) about a population.
- Example: What is the percentage of Democratic voters in the next Presidential election?
  - Population: voters in next Presidential election
  - Parameter: percentage of Democratic voters
- The population parameter
  - is a fixed number
  - that we will never know exactly (because it is infeasible to look at the entire population)

4 / 18

Sample and statistic

- We only examine part of the population, a sample
- We compute a statistic from the sample
  - Percentage of Democratic voters in our sample
- The statistic
  - is a random variable. It changes if we take a different sample, due to chance error.
  - can be computed exactly for each sample.
- The parameter is what we want to know. The statistic is what we know.
- Statistic = parameter (+ bias) + chance error
- How big is the chance error?

5 / 18

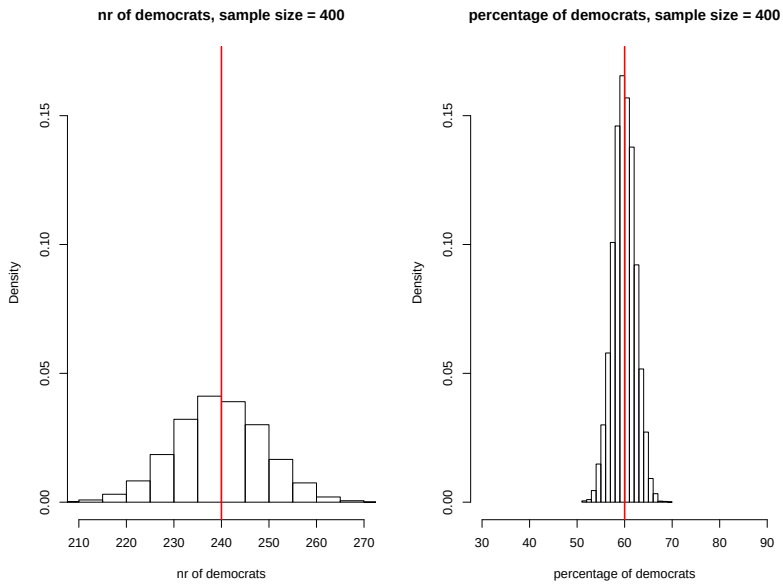
Expected value and SE for percentage

Example

- Village: 5,000 voters
  - 3,000 Democrats (60%)
  - 2,000 Republicans (40%)
- Take a simple random sample of size 400, that is,
  - drawing at random without replacement

  and compute number and percentage of Democrats. Do this 10,000 times...

When we do this 10,000 times...



Example

- The number of democrats in a sample of size 400 is around ..... (expected value for count), give or take .....(SE for count) or so. Fill in the blank.
- The percentage of democrats in a sample of size 400 is around .... (expected value for percentage), give or take .... (SE for percentage) or so. Fill in the blank.

Strategy

- Make a box model: 3000 tickets 1, and 2000 tickets 0
- We first look at the count of democrats in our sample.
- The number of democrats in a sample of size 400 is like the sum of 400 draws from this box without replacement
- We first consider the sum of the draws with replacement. Why? Because this is easier and we know how to do it.
- Average of the box: $\frac{3000\times1+2000\times0}{5000} = 0.6$
- SD of the box: shortcut formula
  (large nr - small nr) x $\sqrt{\text{fraction large nr} \times \text{fraction small nr}}$
  $= (1-0) \times \sqrt{0.6 \times 0.4} \approx 0.5$
- Expected value for sum of the draws:
  (nr of draws) $\times$ (average of the box) $= 400 \times 0.6 = 240$
- SE for sum of the draws: square root formula: $\sqrt{\text{nr of draws}} \times (\text{SD of box}) = \sqrt{400} \times 0.5 = 10$

Example

- The number of democrats in a sample of size 400 is around 240, give or take 10 or so.
- Convert to percentages:
  - 240 out of 400 $= \frac{240}{400} \times 100\% = 60\%$
  - 10 out of 400 $= \frac{10}{400} \times 100\% = 2.5\%$
- So the percentage of democrats in a sample of size 400 is around 60%, give or take 2.5% or so

Formulas for percentage

$$\text{Expected value for percentage} = \text{percentage of the box}$$

$$\text{SE for percentage} = \frac{\text{SE for sum}}{\text{nr of draws}} \times 100\%$$

- These formulas are exact when we draw with replacement
- If the sample is small compared to the population (say, less then 1/10th), then the formulas are good approximations for drawing without replacement. Why?
- The book gives a correction factor for switching between drawing with and without replacement. You should know that a correction factor exists, but you don't need to use it.

Exercise

- Ch 20, Exercise set E, Problem 1
- A public opinion poll uses a simple random sample of size 1500, drawn from a population of 25,000
- Another poll uses a simple random sample of size 1500 from a town with a population of 250,000
- The polls are trying to estimate the percentage of voters who favor single-payer health insurance
- Choose one:
  - The 1st poll is quite a bit more accurate than the 2nd
  - The 2nd poll is quite a bit more accurate than the 1st
  - There is not much difference in the accuracy of the two polls

Accuracy depends on sample size

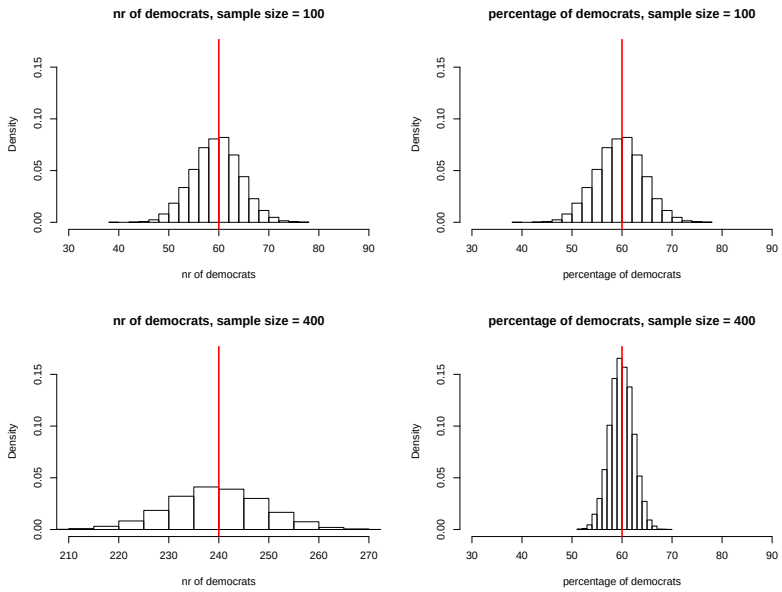
$$\begin{aligned}
\text{SE for percentage} &= \frac{\text{SE for sum}}{\text{nr of draws}} \times 100\% \\
&= \frac{\sqrt{\text{nr of draws}} \times (\text{SD of box})}{\text{nr of draws}} \times 100\% \\
&= \frac{\text{SD of box}}{\sqrt{\text{nr of draws}}} \times 100\%
\end{aligned}$$

- The SE for a percentage depends on the number of draws = sample size, and not on the population size. (This is true if the sample is small relative to the population, so that we don't have to worry about the correction factor)

How does accuracy depend on sample size?

- What happens to the SE for a percentage if we multiply the sample size by a factor of 4? Then the SE is divided by a factor $\sqrt{4} = 2$.
- So if the sample size is 4 times as large, our estimate becomes twice as precise.
- With a larger sample size we can estimate percentages more accurately.

Compare sample size 100 and 400, for 10,000 repetitions...



16 / 18

Normal approximation

- We can use the normal approximation as before:
  - Determine whether the question is about counts or percentages
  - Use new average = expected value for count/percentage
  - Use new SD = SE for count/percentage
- Example: what is the chance to get more than 65% democrats in a sample of size 400 from the village we looked at before?
  - This question is about percentages
  - New average: expected value for percentage = 60%
  - New SD: SE for percentage = 2.5%
  - See overhead. Answer: the chance to get more than 65% democrats in a sample of size 400 is about 2.5%.

17 / 18

Other sampling methods

- Our formulas work for simple random samples
- The Gallup Poll and Iraq study were (multistage) cluster samples. For such samples we need other, more complicated formulas. But statisticians can still figure out the chance error.
- For non-probability sampling methods, it is very hard/impossible to determine the chance error.

18 / 18