010 - Sampling Distributions

EPIB 607

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on September 23, 2021

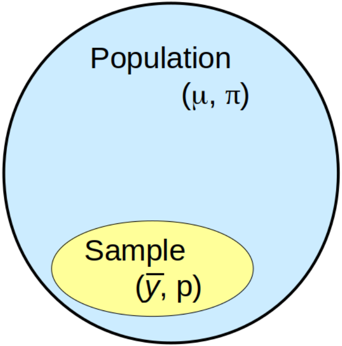
Samples

Sampling Distributions

Exercise 1 Results

Parameters and Statistics

- **Paramter**: An unknown numerical constant pertaining to a population/universe, or in a statistical model.
  - μ: population mean π: population proportion
- **Statistic**: A numerical quantity calculated from a sample. The empirical counterpart of the parameter, used to *estimate* it.
  - $\bar{y}$: sample mean p: sample proportion



Examples

Proportions:

- Proportion of Earth’s surface covered by water
- Proportion who saw a medical doctor last year
- Proportion of Québécois who don’t have a family doctor

Examples

Proportions:

- Proportion of Earth’s surface covered by water
- Proportion who saw a medical doctor last year
- Proportion of Québécois who don’t have a family doctor

Means:

- Mean depth in *n* randomly selected ocean locations
- Mean household size in *n* randomly selected households.
- Median number of persons under-5 in a sample of *n* households

Samples must be random

- The validity of inference will depend on the way that the sample was collected. If a sample was collected badly, no amount of statistical sophistication can rescue the study.

Samples must be random

- The validity of inference will depend on the way that the sample was collected. If a sample was collected badly, no amount of statistical sophistication can rescue the study.

- Samples should be **random**. That is, there should be no systematic set of characteristics that is related to the scientific question of interest that causes some people to be more likely to be sampled than others. The simplest type of randomization selects members from the population with equal probability (a uniform distribution).

Samples must be random

- The validity of inference will depend on the way that the sample was collected. If a sample was collected badly, no amount of statistical sophistication can rescue the study.

- Samples should be **random**. That is, there should be no systematic set of characteristics that is related to the scientific question of interest that causes some people to be more likely to be sampled than others. The simplest type of randomization selects members from the population with equal probability (a uniform distribution).

- When conducting a study, it is always better to seek statistical advice sooner rather than later. Get a statistician involved at the *planning* stage of the study... by the analysis stage, it may be too late!

Samples must be random - No cheating!

Do not cheat by

Samples must be random - No cheating!

Do not cheat by

- Taking 5 people from the same household to estimate
  - proportion of Québécois who don't have a family doctor
  - who saw a medical doctor last year
  - average rent

Samples must be random - No cheating!

Do not cheat by

- Taking 5 people from the <u>same</u> household to estimate
  - proportion of Québécois who don't have a family doctor
  - who saw a medical doctor last year
  - average rent

- Sampling the depth of the ocean <u>only around Montreal</u> to estimate
  - proportion of Earth's surface covered by water

Collecting data takes effort

In general

- The larger the sample → the more accurate the estimate (if sampling is done correctly)

Collecting data takes effort

In general

- The larger the sample → the more accurate the estimate (if sampling is done correctly)

CAVEAT

- Collecting more data takes effort and money!
- We will also soon discover the curse of the $\sqrt{n}$

Collecting data takes effort

In general

- The larger the sample → the more accurate the estimate (if sampling is done correctly)

Collecting data takes effort

In general

- The larger the sample → the more accurate the estimate (if sampling is done correctly)

CAVEAT

- Collecting more data takes effort and money!
- We will also soon discover the curse of the $\sqrt{n}$

Samples

Sampling Distributions

Exercise 1 Results

Sampling Distributions

- Given a sample of n observations from a population, we will be calculating estimates of the population mean, proportion, standard deviation, and various other population characteristics (parameters)

- Prior to obtaining data, there is uncertainty as to which of all possible samples will occur

- Because of this, estimates such as $\bar{y}$ (the sample mean) will vary from one sample to another

Sampling Distributions

- The behavior of such estimates in many samples of equal size is described by what are called **sampling distributions**

Sampling Distributions

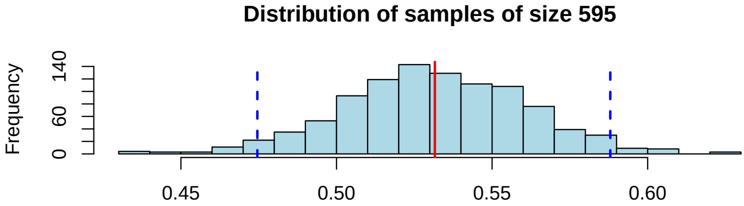
- The behavior of such estimates in many samples of equal size is described by what are called **sampling distributions**

- DVB definition: If we could see all the statistics (means, proportions, ect.) from all possible samples (Chapter 18, page 432)

Sampling distribution of correlations[1]

Lets create a pseudo population from the 595 observations by sampling **with replacement**, and calculate the correlation. Lets repeat this process 1000 times:

```
library(oibiostat); data("famuss"); B <- 1000; N <- 595
R <- replicate(B, {
  dplyr::sample_n(famuss, size = N, replace = TRUE) %>%
  dplyr::summarize(r = cor(height, weight)) %>%
  dplyr::pull(r)
})
```



[1] from 004-exploring-data-2

Why are sampling distributions important?

- Modeling how sample statistics vary from sample to sample is one of the most powerful ideas we'll see in this course.

- A sampling distribution *model* for how a sample statistics varies from sample to sample allows us to quantify that variation and to talk about how likely it is that we'd observe a sample statistic in any particular interval.

- Thus, they are used in confidence intervals for parameters. Specific sampling distributions (based on a null value for the parameter) are also used in statistical tests of hypotheses.

Exercise 1: How Deep is the Ocean?

- We will get a sense of what a sampling distribution is in Exercise 1

Exercise 1: How Deep is the Ocean?

- We will get a sense of what a sampling distribution is in Exercise 1

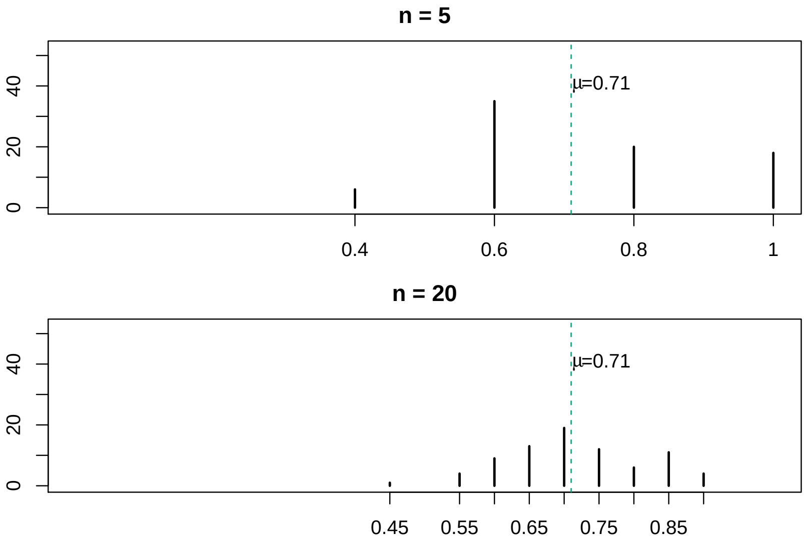
- **CAVEAT**: This is a luxury using a toy example. In actual studies, we only get one shot!

Samples

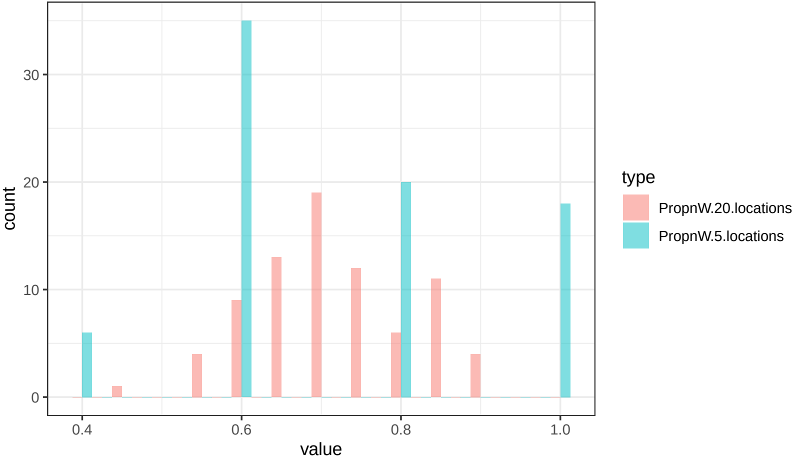
Sampling Distributions

Exercise 1 Results

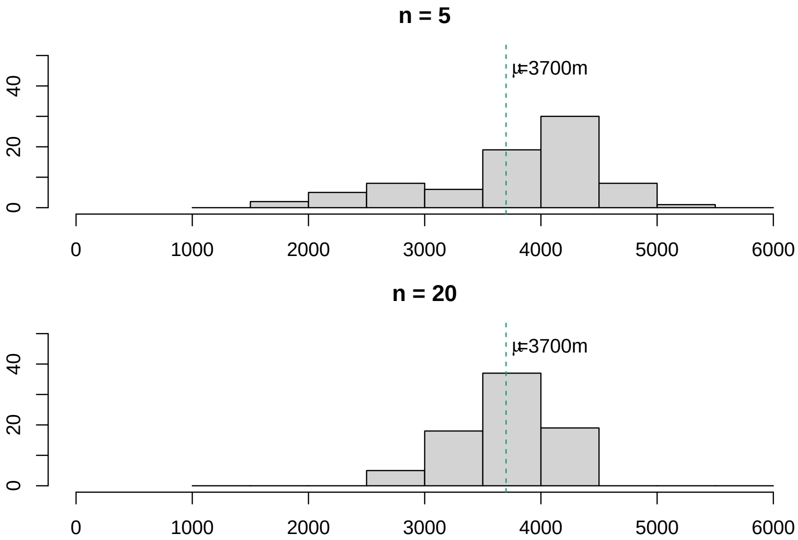
Sampling distribution: proportion covered by water



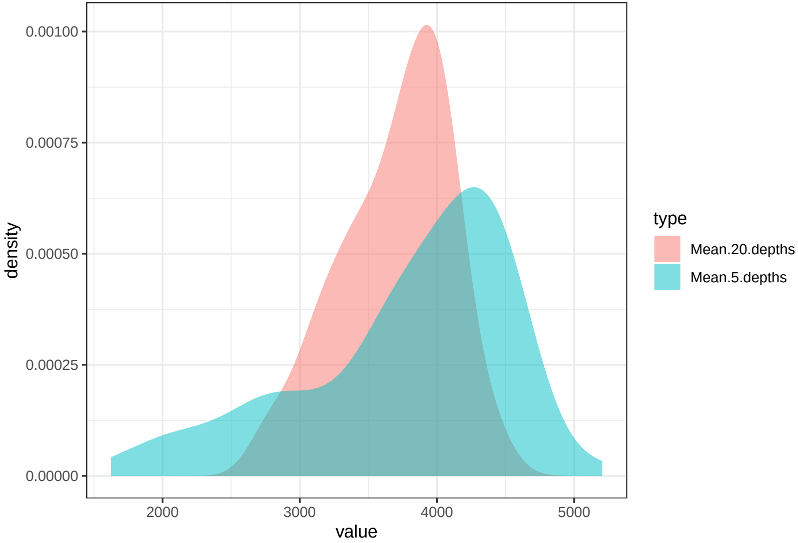
Sampling distribution: proportion covered by water



Sampling distribution: mean depth of the ocean



Sampling distribution: mean depth of the ocean


density
value
type
🟥 Mean.20.depths
🟦 Mean.5.depths