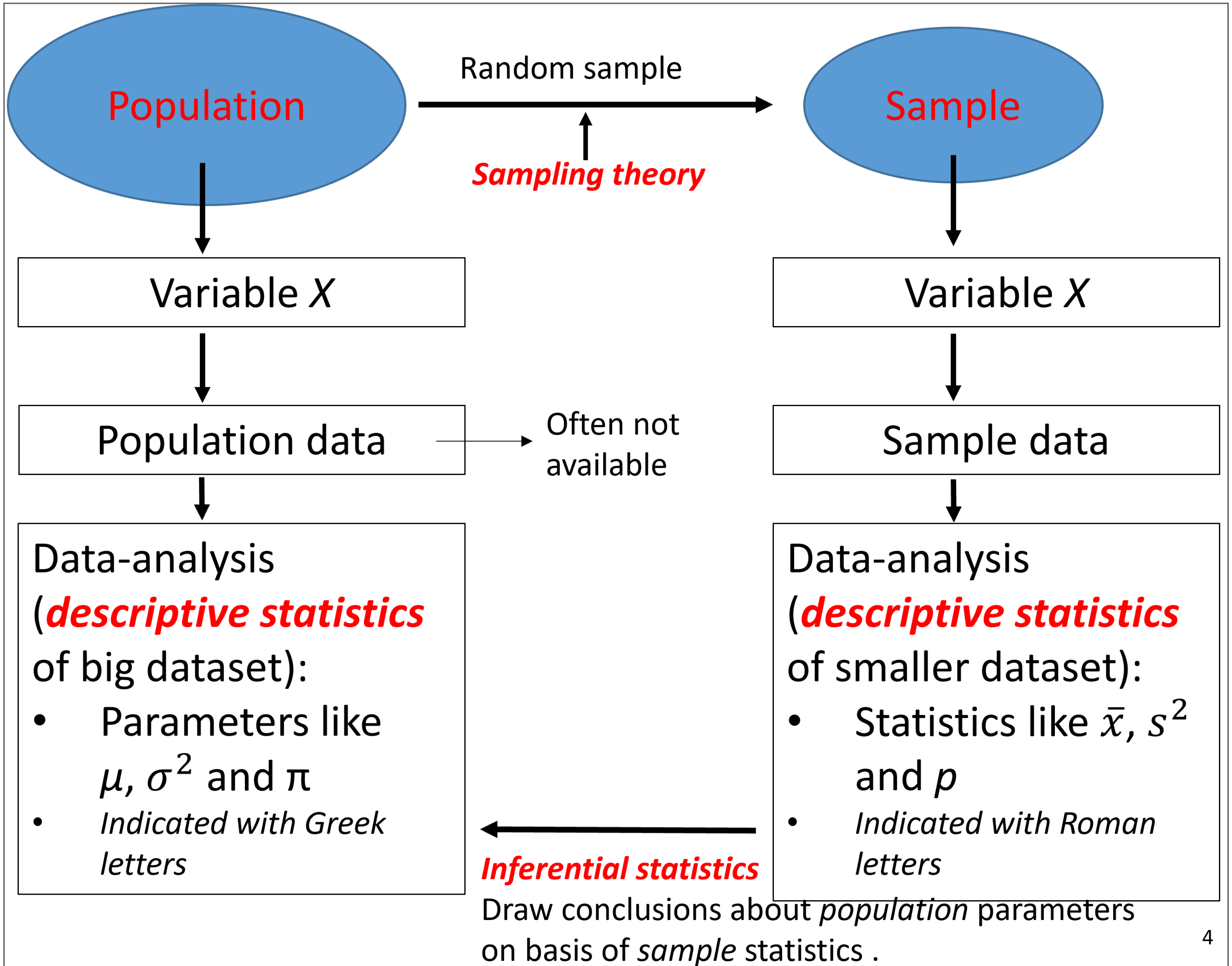
Lecture 3 - Sampling Distributions

Caspar J. van Lissa

2023-08-13

Sampling error

Sampling theory


Population
Random sample
Sample
Sampling theory
Variable X
Variable X
Population data
Often not available
Sample data
Data-analysis (*descriptive statistics* of big dataset):
• Parameters like μ, σ^2 and π
• *Indicated with Greek letters*
Data-analysis (*descriptive statistics* of smaller dataset):
• Statistics like $\bar{x}$, s^2 and *p*
• *Indicated with Roman letters*
Inferential statistics
Draw conclusions about *population* parameters on basis of *sample* statistics .

Estimating population parameters

- Say we want to estimate population mean μ
- Our best guess of the population mean is the sample mean M (aka $\bar{X}$)
- But the sample mean M is not a perfect estimate of μ
- The (unknown) difference between M and μ is called *sampling error*

Sampling distribution demo

Distribution:

Normal

n:

30

Samples:

200

☐ Add

Draw

Sampling distribution

Hypothetically, imagine that

- We draw many (k) samples from the population
- Estimate μ in each sample, so $M_1, M_2, \ldots M_k$
- We could plot a distribution of the observed Ms and call it the *sampling distribution*

Central Limit Theorem

Central limit theorem

- As the number of samples increases, the sampling distribution approaches a normal distribution, $\bar{X} \sim N(\mu, SE_{\mu})$
  - The samples must be **large enough** (typically > 30)
- The mean of $M_{1...k}$ converges to the population value μ
- This happens **regardless of the distribution of the data** (not normal? no problem)

The standard error

The sampling distribution is $\sim N(\mu, SE_\mu)$

- SE_μ is its standard deviation
  - To avoid confusion with the SD of the *data*, we call it *standard error*, or SE
- SE gives us the average sampling error
- Think of this as a measure of uncertainty of M as an estimate of μ
  - “When we estimate μ using M, how wrong are we on average?”
  - If SE_μ is very small, our guesses of μ are very accurate

Properties of the standard error

$SE_M = \frac{\sigma}{\sqrt{n}}$

- SE decreases as the sample size increases (more precise estimates of μ)
  - Imagine the sample size becomes as large as the entire population
  - The sample mean M will be a perfect estimate of μ
  - So the SE goes to zero
- SE increases as the population SD increases (less precise estimates of μ)
  - Imagine everyone in our sample has the same value
  - Again, the sample mean M will be a perfect estimate of μ
  - So the SE goes to zero

The rationale for inference

- I used the mean as an example
- This applies to all other statistics, not just means
- The key lessons are:
    - Sample statistics can be used to estimate population parameters
    - Those sample statistics have a hypothetical distribution that we could observe if we took very many samples
    - The standard distribution of that hypothetical sampling distribution is called the standard error, and it is a measure of uncertainty about our estimate
    - We can use that standard error for statistical tests
- Basically, any statistic has a standard error; you learned to manually calculate the one for the mean. For other statistics the same logic applies, but the formulas may differ.
- You will use statistical software to calculate the standard errors for other statistics

Thought experiment

- There are two elevators
- One has a 6-person limit, the other a 12-person limit
- Both elevators get stuck if the *average* weight exceeds 95 kg
- Which of the two elevators would likely get stuck more often?

Thought experiment 2

- The “best schools” (highest average score on standardized tests) are often small schools
- Does that mean small schools are better?

One remaining problem

In practice we typically have only one sample so we can't *calculate* SE_μ

- **Solution**: We *estimate* the SE_M from the single sample

$$SE_\mu = \frac{\sigma}{\sqrt{n}}$$
$$SE_M = \frac{SD}{\sqrt{n}}$$

Working with standard errors

Use the normal distribution!

Last lecture we calculated probabilities using the normal distribution

- Previous lecture: calculations about the population distribution and distribution of data in one sample
- Today: calculations about the sampling distribution!
- Thanks to Central limit theorem, we can make inferences about likely values of population parameters using only sample statistics
- Use what you know about the normal distribution

Disambiguation

Today, we have talked about 3 types of normal distributions (remember interactive demo):

1. Population distribution of $X \sim N(\mu, \sigma)$
   - Typically unknown
2. Distribution of data in one sample, $X \sim N(M, SD)$
   - Typically observed
3. Sampling distribution of the means of many hypothetical samples from the population: $\bar{X} \sim N(\mu, SE_{\mu})$
   - We know its theoretical properties, we estimate its parameters from the sample

Confidence intervals

Confidence interval: window of uncertainty around estimate

- SE is a measure of uncertainty of M as an estimate of μ
  - If SE_μ is very small, our guesses of μ are very accurate
- Use this to express our confidence in M as an estimate of μ
- Remember 95% of a normal distribution is between +/- 2SD
- So $M + / - 2 * SE_m$ gives us boundaries corresponding to 95% probability
- We can never be sure that **this** confidence interval contains the population value
- But 95% of confidence intervals ought to include the population value

Confidence interval 2

M =

100

SD =

15

n =

75

%:

95

Z-scores

- Previous lecture: population & sample distribution
  - IQ is normally distributed with mean 100 and SD of 15
  - What is the probability that the IQ of a randomly chosen person exceeds 115?
  - $Z = \frac{X-\mu}{\sigma} = \frac{115-100}{15} = 1$
  - $P(Z > 1) = .025$
- Today: sampling distribution of the mean
  - What is the probability that the mean of a random sample of 9 persons exceeds 115?
  - $SE_m = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$
  - $Z = \frac{X-\mu}{SE_m} = \frac{115-100}{5} = 3$
  - $P(Z > 3) = .001$

Calculating Z-scores

Weekly fruit consumption is distributed $\sim N(\mu = 10.5, \sigma = 6.4)$

- What is the probability that the mean fruit consumption of 16 randomly chosen people is less than 7.78?
- $SE_\mu = \frac{\sigma}{\sqrt{n}} = \frac{6.4}{\sqrt{16}} = 1.6$
- $Z = \frac{X-\mu}{SE_\mu} = \frac{7.78-10.5}{1.6} = -1.7$

From Z to X

A coffee roaster uses a machine to fill 1000 bags with coffee

- The machine's accuracy is $\sigma = 10$
- For how many grams should they set the machine to ensure that at most 1 bag contains less than 250g?
- $SE_M = \frac{10}{\sqrt{1000}} = 0.32$
- $Z(P > .001) = 2.33$
- $250 + 2.33 * 0.32 = 250.75$