Basic Statistics and Sampling

Jerome Dumortier

17 August 2023

Lecture Overview

Topics covered:

- Law of Large Numbers
- Central Limit Theorem

Law of Large Numbers

Measuring unemployment rate in the United States:

- Current Population Survey (CPS)
- Monthly survey among 60,000 households
- Classification: *Employed*, *Unemployed*, *Not in the labor force*

Law of large numbers:

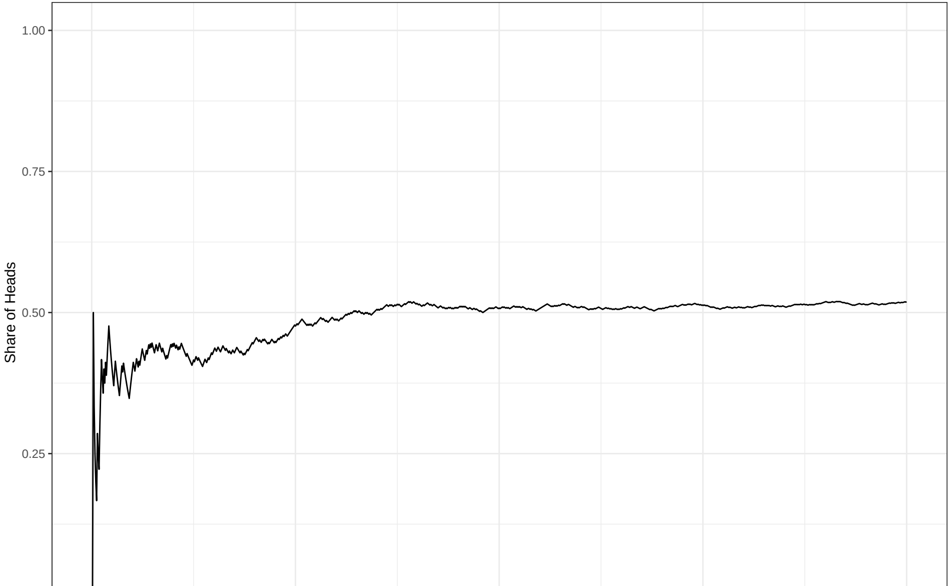
- Any feature of a distribution can be recovered from repeated sampling.

Example of flipping a coin:

- Two possible outcomes: Heads or tails
- Key condition: Independence
- Expected value of heads (or tails): $E(H) = E(T) = 0.5$

Difficulty to predict the share of heads from a single coin flip but high prediction precision from several thousand flips.

Law of Large Numbers: Flipping a Coin



Refresher: Sample versus Population

Why sampling is necessary:

- Sampling the entire population may be expensive or impossible.
- Sampling the entire population may be destructive (e.g., sampling all tires).

Random sample:

- Every item or person in the population (more specifically sample frame) has the same probability of getting selected into the sample.

Example for polling before an election:

- Every person with voting rights is in the sample frame and has the same chance of getting selected by a news agency for polling.

Estimation of the Sample Mean and the Sample Variance

Estimation of the population mean based on a sample:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Estimation of the population variance based on a sample:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

And this is important:

- In R, `var()` and `sd()` calculate the variance assuming a sample, i.e., division by $N - 1$.

Illustration: Estimating the Population Variance I

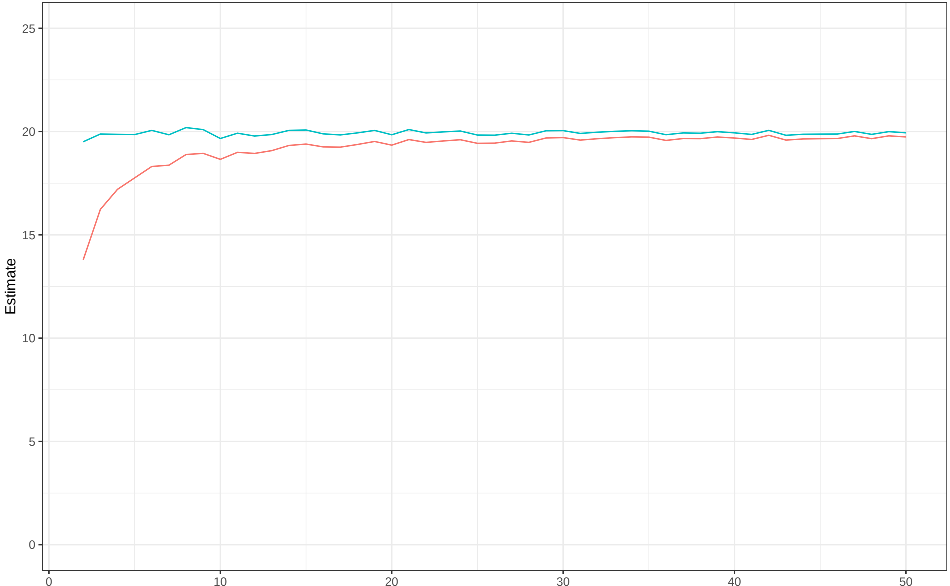
What we know about the population:

- Population size: 100,000
- Mean: $\mu = 50$
- Standard deviation: $\sigma = 20$

Sampling:

- Sample size ranging from 2 to 50
- Repeating the sampling 1000 times

Illustration: Estimating the Population Variance II



Sampling Distribution and Central Limit Theorem

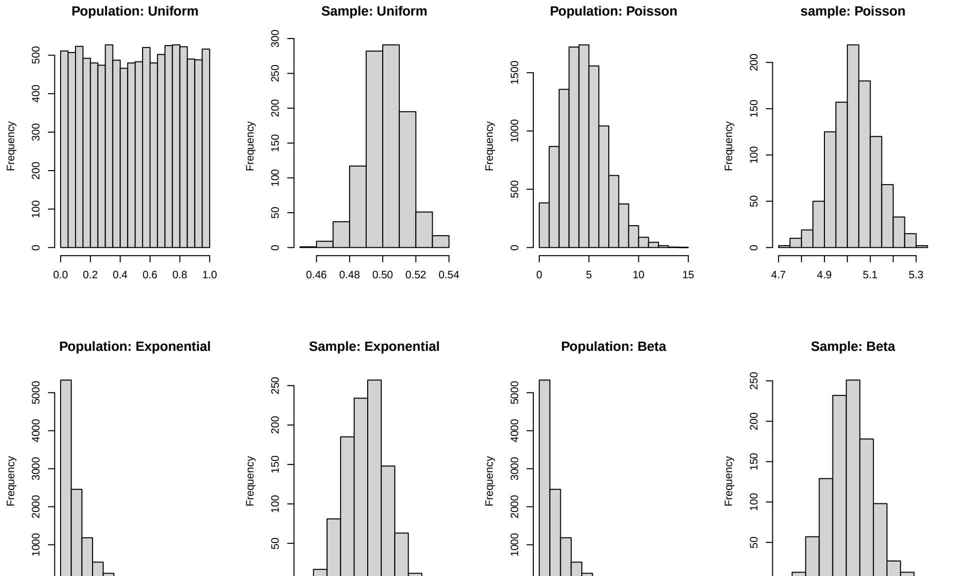
A statistic is a random variable (with its own probability distribution) based on a sample. For example, repeated polling of 1,000 people about their political preferences will result in a different outcome each time. For the sampling distribution of the mean $\bar{x}$, we have the following:

- Mean of the sampling distribution: $\mu_{\bar{X}}$
- Variance of the sampling distribution: $\sigma_{\bar{X}}^2$
- Standard deviation of the sampling distribution (commonly known as standard error): $\sigma_{\bar{X}}$

Central Limit Theorem

- Independent of the underlying distribution, as the sample size increases, the sampling distribution of the mean will follow a normal distribution.

Central Limit Theorem: Illustration



Central Limit Theorem: Implications for Estimation

The standard error of the mean is given by:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

The sample standard deviation is the statistic defined by:

$$s = \sqrt{s^2}$$

Suppose you have to predict the share of heads after flipping a coin multiple times. The variance of n coin flips is:

$$\text{Var}(n) = \frac{p \cdot (1 - p)}{n}$$

Hence: $\text{Var}(1) = 0.5$, $\text{Var}(10) = 0.025$, $\text{Var}(1000) = 0.00025$, etc.

Application: Insurance Market

Risk aversion for individuals as well as for firms.

- Why do insurance companies exist?

Example:

$$Pr(\textit{fire}) = 1/250$$

Simulation

1. Simulate the damage of n homeowners
2. Calculate the share
3. Repeat 1,000 times
4. Generate histogram

Insurance Market