Basic Statistics and Sampling

Jerome Dumortier

28 October 2024

Lecture Overview

Topics covered

- Sampling
- Law of large numbers
- Central limit theorem

Sampling

Refresher: Population versus Sample

Population

- Entire group of individuals or items about which information is needed
- Characterized by unknown parameters

Sample

- Subset of the population
- Application of statistics allows inference on population characteristics
- Used in (social) science to collect data without surveying entire population
- Example from natural science: Sampling of a field for soil characteristics

Importance of sampling

- Reduction in time and resources needed to infer population characteristics
- Ability of decision-making based on sample data

Sample needs to be correctly taken which is the subject of research method classes

Sampling Methods

Probability sampling

- Simple random sampling: Equal chance of each population member to be selected
- Stratified sampling: Separation of population into subgroups with subsequent sampling from each subgroup (e.g., based on location)
- Cluster sampling: Separation of population into clusters with subsequent selection of cluster (e.g., fisheries)

Non-probability sampling (Usually biased but useful for exploratory research)

- Convenience sampling: Participants are selected based on availability
- Quota sampling: Ensures specific characteristics are represented

Comparison of Sampling Methods

Exempt Organizations: Setup

Focus on the mean income of nonprofit organizations

```
eo        = exemptorgs[c("name","income","ntee")]
eo        = na.omit(eo)
n         = 50 # Sample size
meaninc   = mean(eo$income)
```

Mean (population) income $11,126,548

Simple Random Sampling

Each individual has an equal chance of being selected

```
simplerandom    = sample(eo$income,n,replace=FALSE)
meaninc         = mean(simplerandom)
```

Mean sample income $2,481,548

Systematic Sampling

Systematic sampling involves selecting every nth individual from a list after a random start point

```
# Define a population and sample size
interval             = round(length(eo$income)/n)
start                = sample(1:interval,1)
systematicsample     = eo$income[seq(start,length(eo$income),
                                     by=interval)]
meaninc              = mean(systematicsample)
```

Mean sample income $2,790,165

Stratified Sampling by NTEE

```
nteecodes = data.frame(code=c("A","B","N"),
                       description=c("Arts, Culture, and Humanities",
                                     "Education",
                                     "Recreation and Sports"))
for(i in 1:nrow(nteecodes)){
     df                    = subset(eo,ntee==nteecodes$code[i])
     df                    = sample(df$income,n,replace=FALSE)
     nteecodes$mean[i]     = sprintf("$%s",format(mean(df),
                                                  big.mark=",",
                                                  nsmall=0))}
```

Convenience Sampling

Participants are chosen based on availability. In this example, we select the first 50 individuals from the data frame.

```
# Take the first 50 individuals as a convenience sample
conveniencesample   = head(eo$income,50)
meaninc             = mean(conveniencesample)
```

Mean sample income $3,488,097

More on Sampling

Problems with Sampling

Sampling Bias

- Selection bias: When sample selection leads to overrepresentation of particular groups
- Non-response bias: Results when certain respondents do not participate (i.e., respondents with similar characteristics)

Sampling errors vs. non-sampling errors

- Sampling error: The difference between the sample estimate and the true population parameter
- Non-sampling error: Errors not related to the sampling process such as measurement errors

Sample Size and Representativeness

Importance of sample size

- Larger samples provide more precise estimates
- Reduction in the so-called margin of error

Representativeness of a sample

- Stratified sampling often ensures diverse representation
- Randomization helps avoid selection bias
- Example: Representative sample of registered voters

Basic Calculations in Sampling

Point Estimates

- Sample mean, median, and proportions are common estimates used in public policy analysis
- Allows approximation of population parameters from sample data

Confidence Intervals

- Confidence interval (CI): Range within which a population parameter is estimated to be in

Law of Large Numbers

Overview

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*

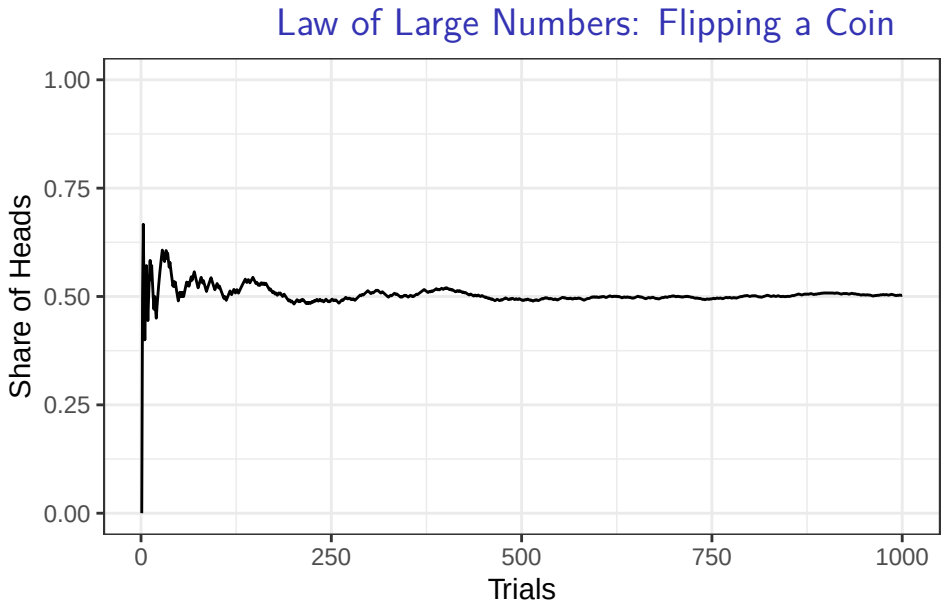
Law of large numbers:

- 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
Share of Heads
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
Trials

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.

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

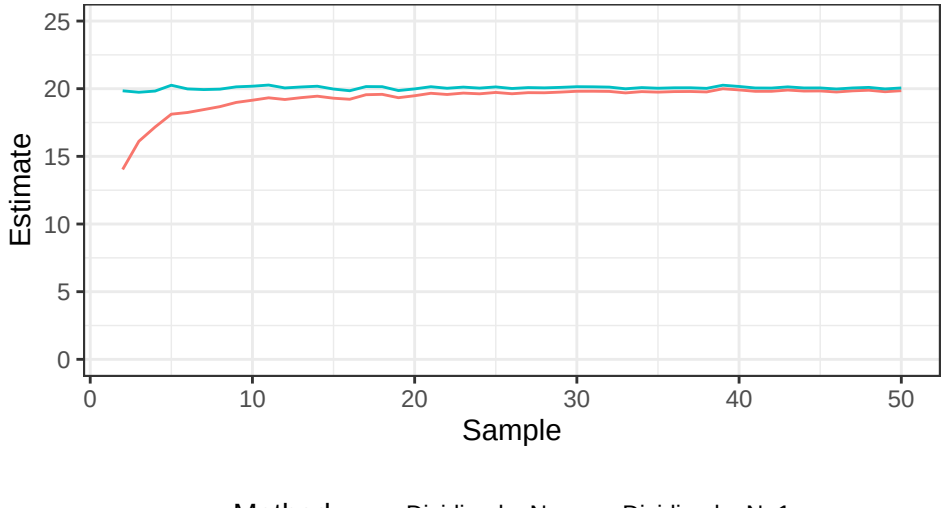
What we know about the population:

- 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



Central Limit Theorem

Sampling Distribution and Central Limit Theorem

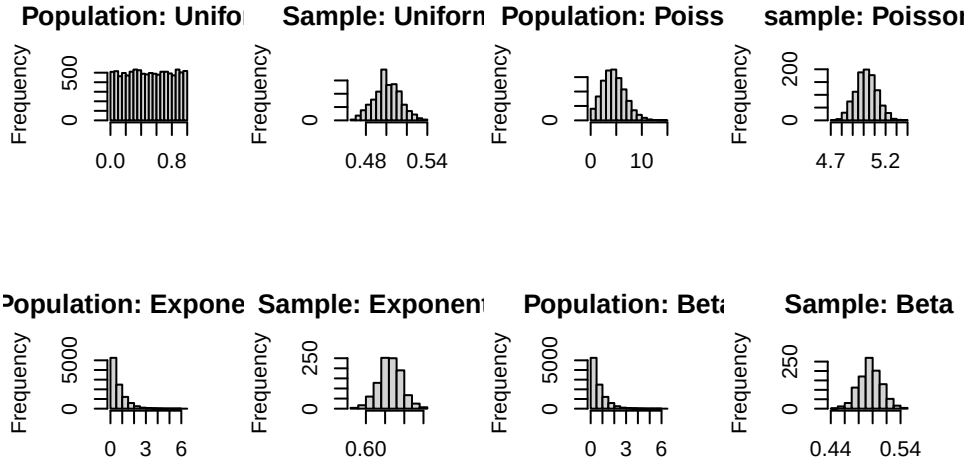
A statistic is a random variable (with its own probability distribution) based on a sample. For example, repeated polling of 1,000 people about their political preferences will result in a different outcome each time. For the sampling distribution of the mean $\bar{x}$, we have the following:

- Mean of the sampling distribution: $\mu_{\bar{X}}$
- Variance of the sampling distribution: $\sigma^2_{\bar{X}}$
- 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.

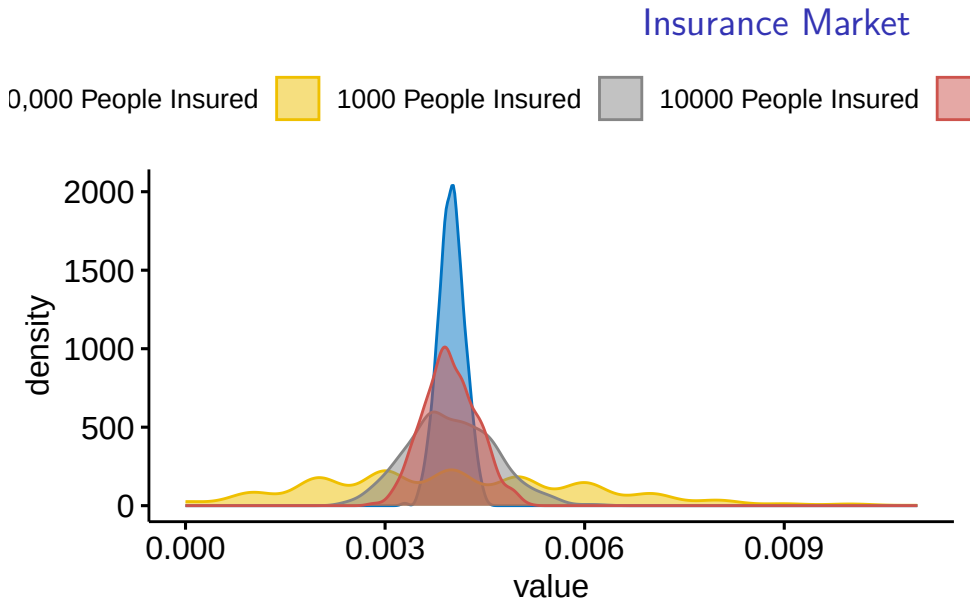
- Why do insurance companies exist?

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
0,000 People Insured
1000 People Insured
10000 People Insured
2000
1500
1000
500
0
density
0.000
0.003
0.006
0.009
value

Sampling Variance

Sampling Variance is the variability in a sample statistic (e.g., sample mean) across different samples drawn from the same population.

- Reflection of the spread of sample estimates around the population parameter
- Lower variance indicates a more stable estimate, while higher variance suggests more fluctuation between samples.
- Key in determining the margin of error and confidence intervals

Factors Influencing Sampling Variance

Sample Size and Sampling Variance

- **Larger samples** generally have lower sampling variance, resulting in more precise estimates.
- **Smaller samples** tend to have higher variance, making estimates less reliable.

Population Variability - If the population itself has high variability, samples will also exhibit higher sampling variance. - Lower population variability translates to lower sampling variance, even with smaller samples.

A larger sample size reduces sampling variance, leading to smaller confidence intervals This reduction follows the Law of Large Numbers, which states that as sample size increases, the sample mean approaches the population mean.