Quantitative methods

Week #4-5

Gergely Daróczi

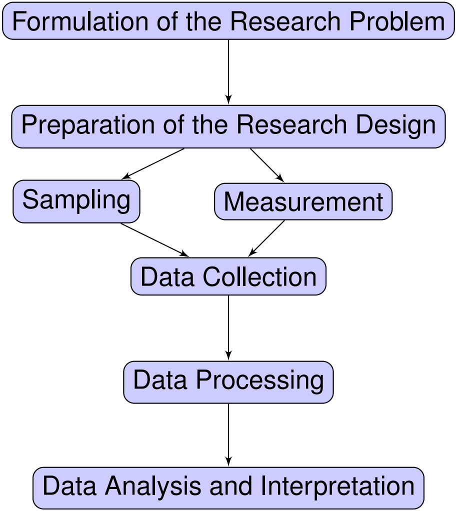
Corvinus University of Budapest, Hungary

1 March 2013

Outline

Stages of Social Research

A flowchart



Preparation of Research Design

Conceptualization and Operationalization

Preparation of Research Design

Conceptualization and Operationalization

Conceptualization:

> **Definition**
>
> *Conceptual definition is the process of formulating and clarifying concepts.*

Operationalization:

> **Definition**
>
> *Operational definition describes the research operations that will specify the value or category of a variable on each case.*

Time magazine reported in the late 1950s that

"the average Yaleman, class of 1924, makes $ 25,111 a year"

which would be equivalent to well over $ 150,000 today!

Time's estimate turns out to have been based on replies received to a sample survey questionnaire mailed to those members of the Yale class of 1924 whose addresses were known in the late 1950s by the Yale administration.

1. selection bias,
2. nonresponse bias,
3. response bias.

Sample-bias

Other historical examples

1936: the American *Literary Digest* magazine collected over two million postal surveys and predicted that the Republican candidate in the U.S. presidential election, Alf Landon, would beat the incumbent president, Franklin Roosevelt by a large margin.

Sample-bias

Other historical examples

1936: the American *Literary Digest* magazine collected over two million postal surveys and predicted that the Republican candidate in the U.S. presidential election, Alf Landon, would beat the incumbent president, Franklin Roosevelt by a large margin.

- records of registered automobile owners and telephone users,
- George Gallup: quota sampling with 50.000 respondents.

Sample-bias

Other historical examples

1936: the American *Literary Digest* magazine collected over two million postal surveys and predicted that the Republican candidate in the U.S. presidential election, Alf Landon, would beat the incumbent president, Franklin Roosevelt by a large margin.

- records of registered automobile owners and telephone users,
- George Gallup: quota sampling with 50.000 respondents.

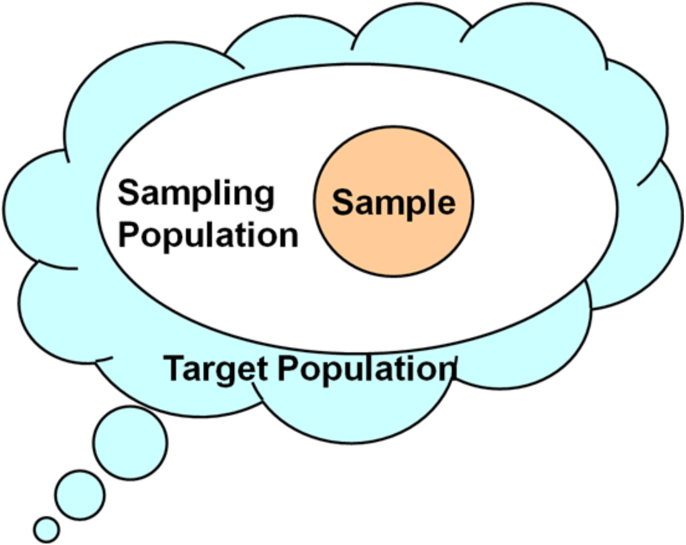
1948: *Chicago Tribune* printed the headline “DEWEY DEFEATS TRUMAN” based on a Gallup poll.

1936: the American *Literary Digest* magazine collected over two million postal surveys and predicted that the Republican candidate in the U.S. presidential election, Alf Landon, would beat the incumbent president, Franklin Roosevelt by a large margin.

- records of registered automobile owners and telephone users,
- George Gallup: quota sampling with 50.000 respondents.

1948: *Chicago Tribune* printed the headline “DEWEY DEFEATS TRUMAN” based on a Gallup poll.

- telephone interviews,
- quota matrix had changed a lot!



Sampling Population
Sample
Target Population

Sampling theory

Definition

Sampling is the process of selecting units (e.g., people, organizations) from a population of interest so that by studying the sample we may fairly generalize our results back to the population from which they were chosen.

Elements:

1. population,
2. respondents, units of analysis,
3. sampling frame,
4. sampling methods.

Kish (1995) posited four basic problems of sampling frames:

1. **Missing elements:** Some members of the population are not included in the frame.
2. **Foreign elements:** The non-members of the population are included in the frame.
3. **Duplicate entries:** A member of the population is surveyed more than once.
4. **Groups or clusters**: The frame lists clusters instead of individuals.

Sampling theory

A not so well choosen sampling frame

We started a small research company and someone proposed to use the public phonebook to build samples:

1. based on public phonebook: only those are on the list who holds a phone,
2. only those with *public* phone number,
3. mobile numbers are not called for surveying (expensive),
4. repeated calls to the same number are forbidden,
5. only those are reached, who are willing to asnwer to our questions on the line.

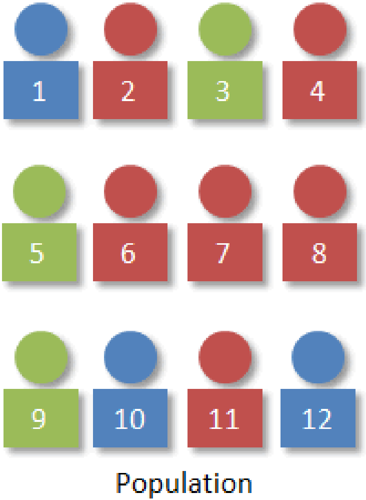
Probability sampling:

1. Simple Random Sampling,
2. Stratified Random Sampling,
3. Systematic Random Sampling,
4. Cluster (Area) Random Sampling,
5. Multi-Stage Sampling.

A subset of the population.

Simple Random Sampling

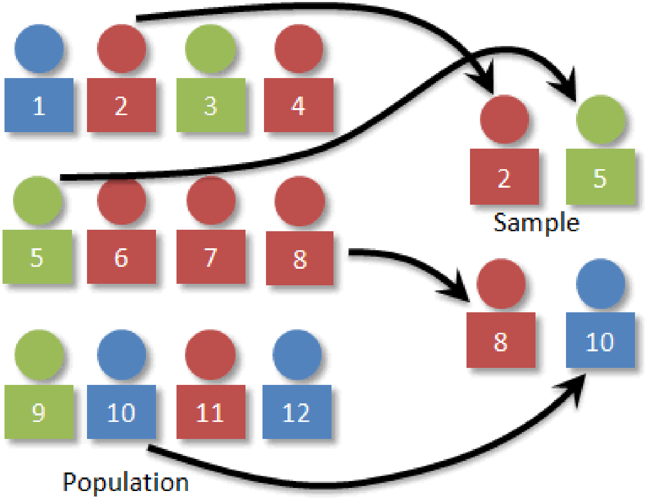
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Simple Random Sampling

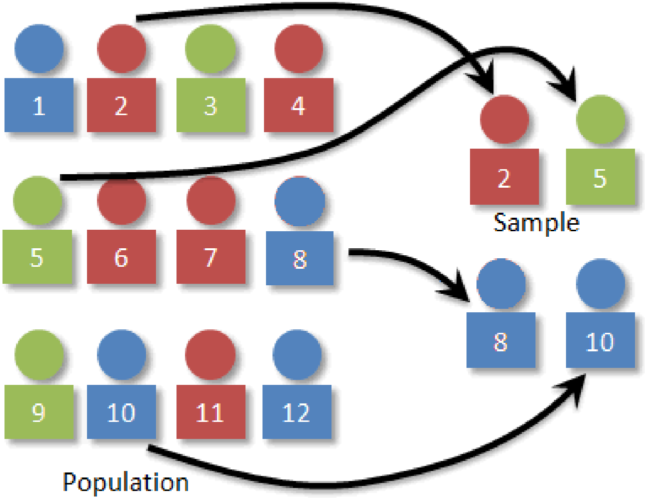
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Simple Random Sampling

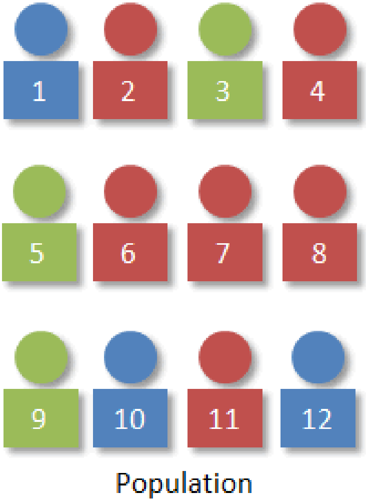
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Systematic Random Sampling

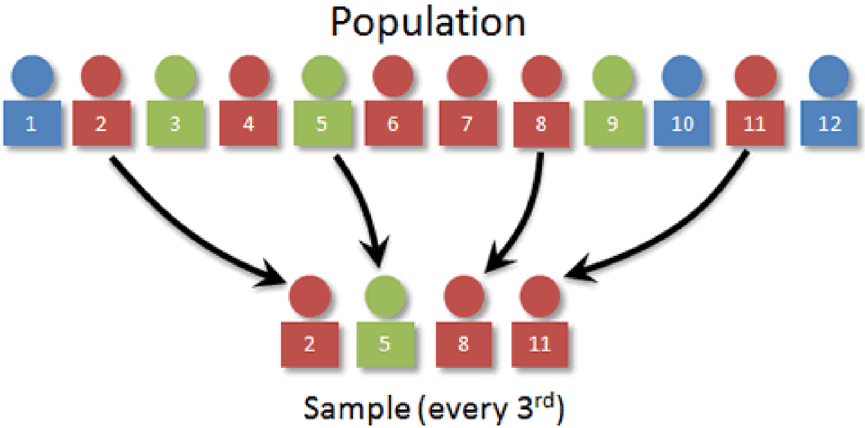
Drawing a sample



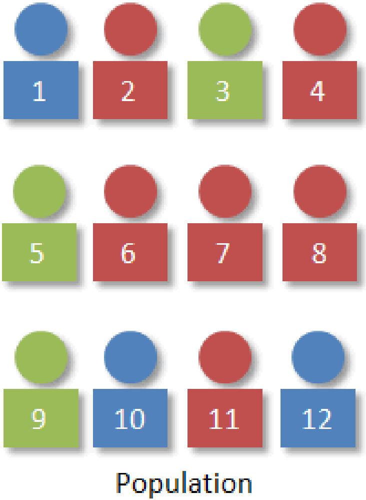
Source: Dan Kerlner, Elgin Community College

Systematic Random Sampling

Drawing a sample



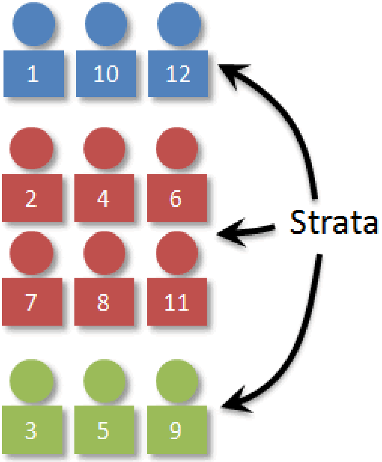
Source: Dan Kerlner, Elgin Community College



Source: Dan Kerlner, Elgin Community College

Stratified Sampling

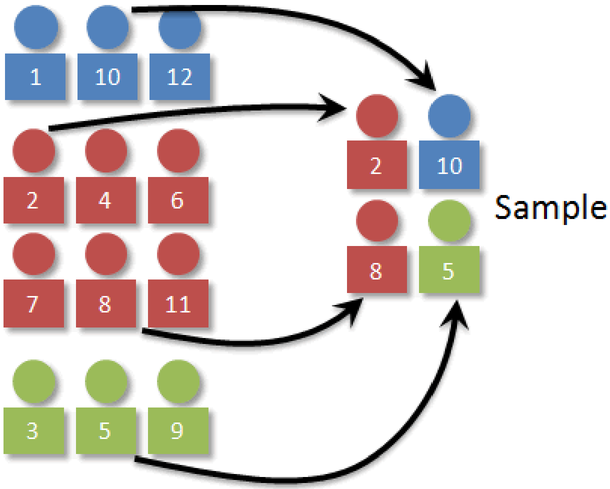
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Stratified Sampling

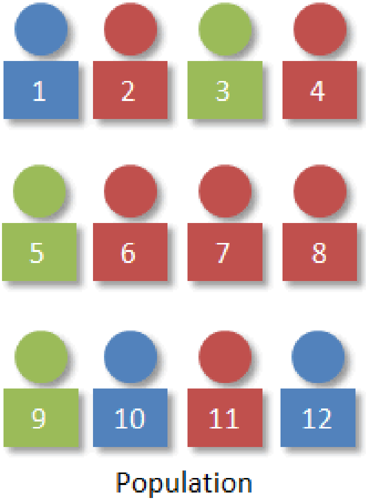
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Systematic+Stratified Random Sampling

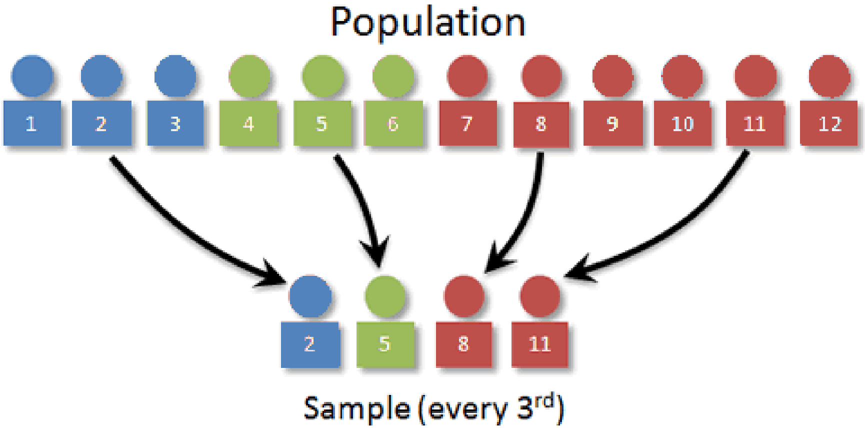
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Systematic+Stratified Random Sampling

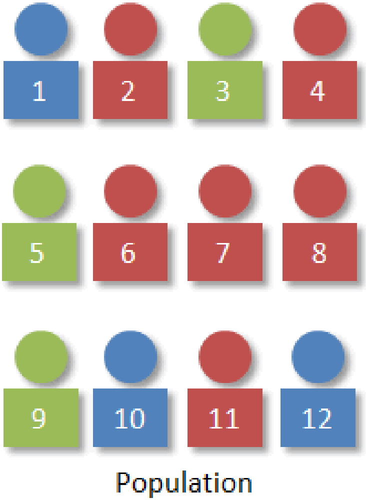
Drawing a sample



Source: Dan Kerlner, Elgin Community College

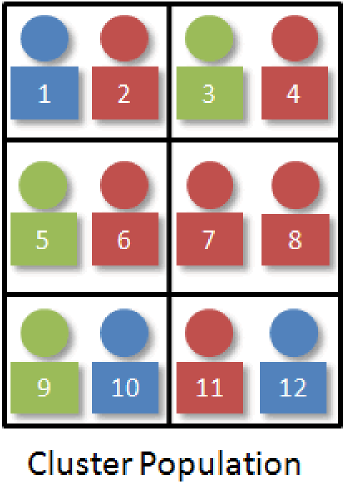
Multi-Stage Sampling

Drawing a sample

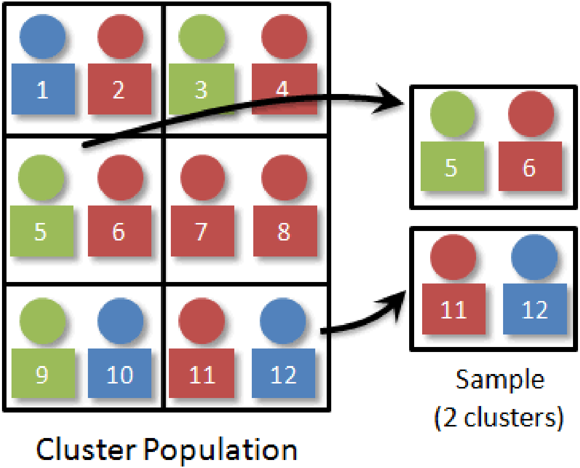


Source: Dan Kerlner, Elgin Community College

Multi-Stage Sampling
Drawing a sample



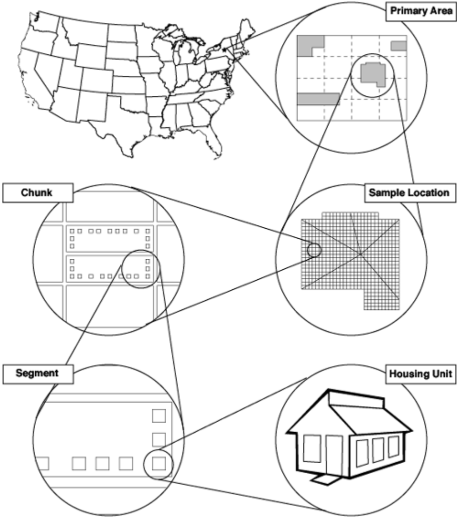
Source: Dan Kerlner, Elgin Community College



Source: Dan Kerlner, Elgin Community College

Cluster Sampling

Drawing a sample



Nonprobability sampling:

1. Accidental, Haphazard or Convenience Sampling,
2. Purposive Sampling:
    1. Modal Instance Sampling,
    2. Expert Sampling,
    3. Quota Sampling:
        1. Proportional Quota Sampling,
        2. Nonproportional Quota Sampling.
    4. Heterogeneity Sampling,
    5. Snowball Sampling.

Computation

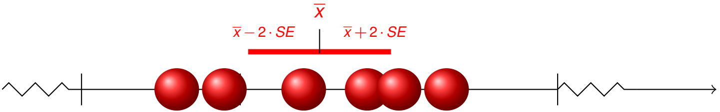
Required formulas

For Simple Random Sampling:

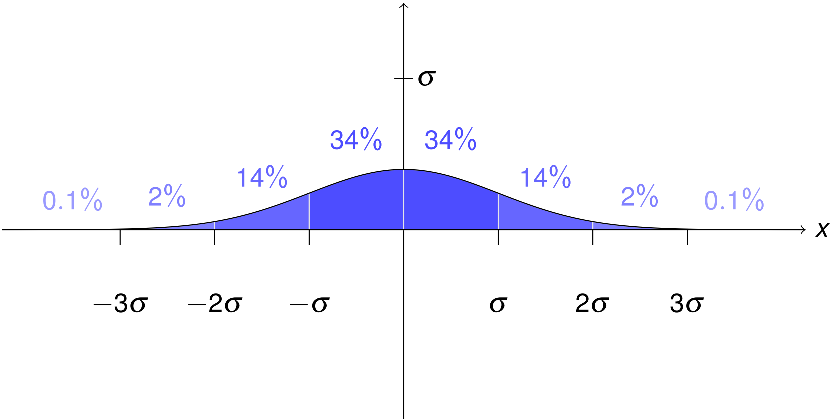
- mean: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- standard deviation: $\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$
- standard error: $SE = \frac{\sigma}{\sqrt{n}} \cdot FPC$
- Finite Population Correction: if sampling fraction is large (>5%)

$$FPC = \sqrt{1 - \frac{n}{N}}$$

- confidence interval: $\overline{x} \pm z \cdot SE$, where $z = 1,96$
- confidence interval: $[\overline{x} - 2 \cdot SE; \overline{x} + 2 \cdot SE]$



$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-x^2}{2\sigma^2}\right)$$

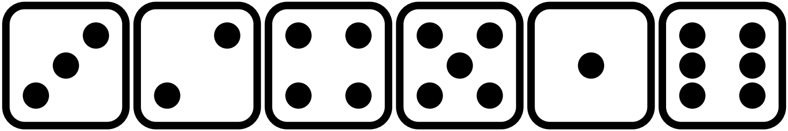


standard normal distribution: $\overline{x} = 0, \sigma = 1$

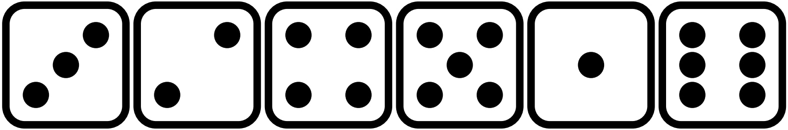
Game rules

Roll the dice!
If the result is even, the player wins the rolled value in dollars.
If the result is odd, the playes pays 2 dollars to the bank.

After rolling the below values, what would you think about the expected value of the game?



Would you continue playing?



$$X = \{-2, 2, 4, -2, -2, 6\}$$

$$\bar{x} = \frac{-2+2+4+2+2+6}{6} = \frac{6}{6} = \frac{1}{1} = 1$$

$$\sigma = \sqrt{\frac{(-2-1)^2+(2-1)^2+(4-1)^1+(-2-1)^1+(-2-1)^2+(6-1)^2}{5}} =$$

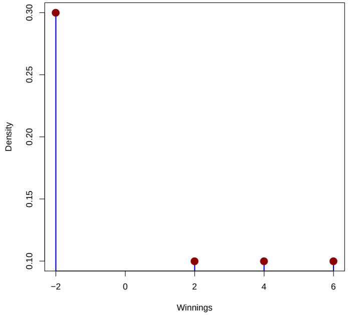
$$= \sqrt{\frac{9+1+9+9+9+25}{5}} = \sqrt{\frac{62}{5}} = \sqrt{12.4} = 3.521363$$

$$SE = \frac{3.521363}{\sqrt{6}} = \frac{3.521363}{2.44949} = 1.437591$$

The expected value can vary between -1.87 and 3.87 at 95% CI.

Good luck!

Forget about the experiment and try to determine the **real** expected value of the game!



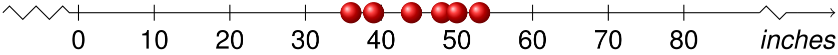
What is wrong with the above plot?

Computation

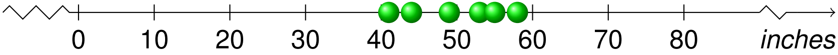
Comparison of samples

The height, in inches, of six trees at a nursery are shown at the specificed dates.
Find the mean, standard deviation and standard error of the heights!
Is there a significant difference between the means of samples?

1. **2013 January 1**: 36 48 50 44 53 39



2. **2013 March 1**: 41 53 55 49 58 44



The height, in inches, of six trees at a nursery are shown at the specificed dates.
Find the mean, standard deviation and standard error of the heights!
Is there a significant difference between the means of samples?

1. **2011 March 22**: 36 48 50 44 53 39

$$X = \{36, 48, 50, 44, 53, 39\}$$

$$\overline{x} = \frac{36+48+50+44+53+39}{6} = \frac{270}{6} = 45$$

$$\sigma = \sqrt{\frac{(36-45)^2+(48-45)^2+(50-45)^2+(44-45)^2+(54-45)^2+(39-45)^2}{5}} =$$

$$= \sqrt{\frac{81+9+25+1+64+36}{5}} = \sqrt{\frac{216}{5}} = \sqrt{43.2} = 6.57$$

$$SE = \frac{6.57}{\sqrt{6}} = \frac{6.57}{2.44} = 2.68$$

The expected value can vary between 40.5 and 49.5 at 95% CI.

The height, in inches, of six trees at a nursery are shown at the specificed dates.

Find the mean, standard deviation and standard error of the heights!

Is there a significant difference between the means of samples?

1. **2011 April 1**: 41 53 55 49 58 44

$$X = \{41, 53, 55, 49, 58, 44\}$$

$$\overline{x} = \frac{41+53+55+49+58+44}{6} = \frac{300}{6} = 50$$

$$\sigma = \sqrt{\frac{(41-50)^2+(53-50)^2+(55-50)^2+(49-50)^2+(58-50)^2+(44-50)^2}{5}} =$$

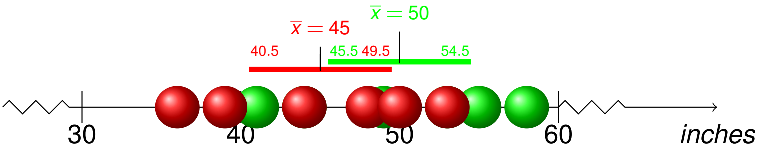
$$= \sqrt{\frac{81+9+25+1+64+36}{5}} = \sqrt{\frac{216}{5}} = \sqrt{43.2} = 6.57$$

$$SE = \frac{6.57}{\sqrt{6}} = \frac{6.57}{2.44} = 2.68$$

The expected value can vary between 45.5 and 54.5 at 95% CI.

The height, in inches, of six trees at a nursery are shown at the specificed dates.
Find the mean, standard deviation and standard error of the heights!
Is there a significant difference between the means of samples?

1. **2013 January 1**: 36 48 50 44 53 39
2. **2013 March 1**: 41 53 55 49 58 44



Computation

Standard error in finite population

We have seen in the dice example, that the standard error (1.437591) could be relatively high compared to the mean (1).

If we would check the exact same values (-2, 2, 4, -2, -2, 6) denoting the temperature measured from Monday to Saturday, then would you think that the average temperature at the audited week cannot be estimated more precisely than the earlier computed confidence interval (-1.87 – 3.87)? You have only one missing data!

$$SE = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$$

Is there any difference between computing the standard error in Hungary or in the United States?

Computation

Standard error in finite population

$$X = \{-2, 2, 4, -2, -2, 6\}$$

$$\bar{x} = \frac{-2+2+4+2+2+6}{6} = \frac{6}{6} = \frac{1}{1} = 1$$

$$\sigma = \sqrt{\frac{(-2-1)^2+(2-1)^2+(4-1)^1+(-2-1)^1+(-2-1)^2+(6-1)^2}{5}} =$$

$$= \sqrt{\frac{9+1+9+9+9+25}{5}} = \sqrt{\frac{62}{5}} = \sqrt{12.4} = 3.521363$$

$$SE = \frac{3.521363}{\sqrt{6}} \cdot FPC = \frac{3.521363}{2.44949} \cdot FPC = 1.437591 \cdot FPC$$

$$FPC = \sqrt{1-\frac{n}{N}} = \sqrt{1-\frac{6}{7}} = 0.377$$

$$SE = 0.54$$

The expected value can vary between 0.46 and 1.54 at 95% CI (opposed to: 1.87, 3.87).

„The gas prices dramatically increased in 2011 in Hungary. We asked drivers about how much they would pay for one litre of gasoline. The results showed that there are some drivers who would even pay more then 450 forints for a litre, others do not tend to refill at the prices of 400."

Forensis Autóklub (November of 2011)

„How much would you pay for one litre of gas?”

410, 420, 420, 430, 500, 450, 400, 425, 460

„How much would you pay for one litre of gas?"

410, 420, 420, 430, 500, 450, 400, 425, 460

Descriptive statistics:

- **mean**: $\overline{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **median**: 425
- **mode**: 420
- **minimum**: 400
- **maximum**: 500
- **range**: 100

Computation

Exercise

„How much would you pay for one litre of gas?"

410, 420, 420, 430, 500, 450, 400, 425, 460

- **mean**: $\overline{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **standard deviation**: $S^* = 30.619$
- **standard error**: $SE = \frac{30{,}619}{\sqrt{9}} = \frac{30{,}619}{3} = 10,206$
- **confidence interval**: $435 \pm 2 \cdot 10,206 = [414,59; 455,41]$

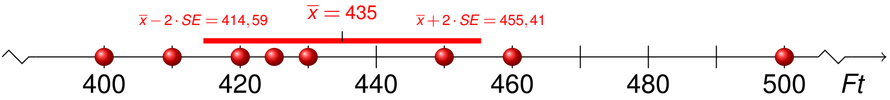
Computation

Exercise

„How much would you pay for one litre of gas?"

410, 420, 420, 430, 500, 450, 400, 425, 460

- **mean**: $\overline{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **standard deviation**: $S^{*} = 30.619$
- **standard error**: $SE = \frac{30{,}619}{\sqrt{9}} = \frac{30{,}619}{3} = 10,206$
- **confidence interval**: $435 \pm 2 \cdot 10,206 = [414,59; 455,41]$



Standard error and sampling

Examples

A módszertan haszna. EP választások 2009: „Hajszálpontos mérés”

	Nézőpont		Tárki	Medián	*NRC*	
	BSZ	BSZP	BSZP	??	??	**eredmény**
Fidesz	54%	**66%**	**70%**	60%	50%	**56,4%**
MSZP	12%	14%	17%	21%	**26%**	**17,4%**
Jobbik	6%	7%	4%	7%	**13%**	**14,8%**
MDF	5%	6%	**1%**	4%	4%	**5,3%**
SZDSZ	3%	4%	3%	4%	3%	**2,2%**

Standard error and sampling

Examples

A módszertan haszna. EP választások 2009: „Hajszálpontos mérés”

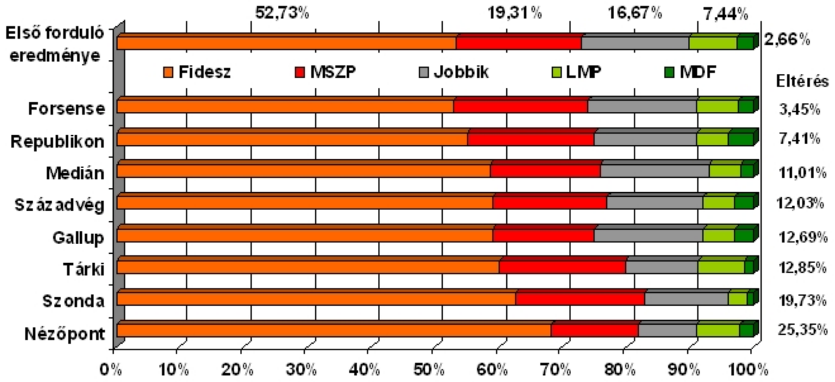
	Nézőpont		Tárki	Medián	*NRC*	
	BSZ	BSZP	BSZP	??	??	**eredmény**
Fidesz	54%	**66%**	**70%**	60%	50%	**56,4%**
MSZP	12%	14%	17%	21%	**26%**	**17,4%**
Jobbik	6%	7%	4%	7%	**13%**	**14,8%**
MDF	5%	6%	**1%**	4%	4%	**5,3%**
SZDSZ	3%	4%	3%	4%	3%	**2,2%**

	Nézőpont	TÁRKI	Medián	*NRC*
Kutatás ideje	V. 20-22.	V. 7-20	V. 22-26.	n.a.
Módszer	Telefonos lekérdezés	Személyes lekérdezés (?)	Személyes lekérdezés	Online kérdőív
Megkérdezettek száma	1000	1000	1200	1000

Source: lectures of Dr. Bartus Tamás

Standard error and sampling

Examples



Source: spss.hu

Bernoulli distribution:

- p chance for 1, q $(= 1 - p)$ chance for 0 value
- **mean**: p
- **median**: –
- **mode**: $\begin{cases} 0 & \text{if } q > p \\ 0,1 & \text{ha } q = p \\ 1 & \text{if } q < p \end{cases}$
- **standard deviation**: $\sqrt{p(1-p)}$
- **variance**: $p(1-p)$
- **standard error**: $SE = \frac{s^*}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \approx \frac{s^*}{\sqrt{n}} \approx \frac{\sqrt{p(1-p)}}{\sqrt{n}}$
- **confidence interval**: $\overline{x} \pm z \cdot SE$, where $z = 1,96$

Bernoulli distribution:

- assume the maximum of standard error,
- standard error is affected by standard deviation and sample size,
- higher sample size lowers standard error,
- higher standard deviation results in higher standard error.

Which p value would result in the maximum of standard deviation?

$$S^* = \sqrt{p(1-p)}$$

Bernoulli distribution:

- assume the maximum of standard error,
- standard error is affected by standard deviation and sample size,
- higher sample size lowers standard error,
- higher standard deviation results in higher standard error.

Which p value would result in the maximum of standard deviation?

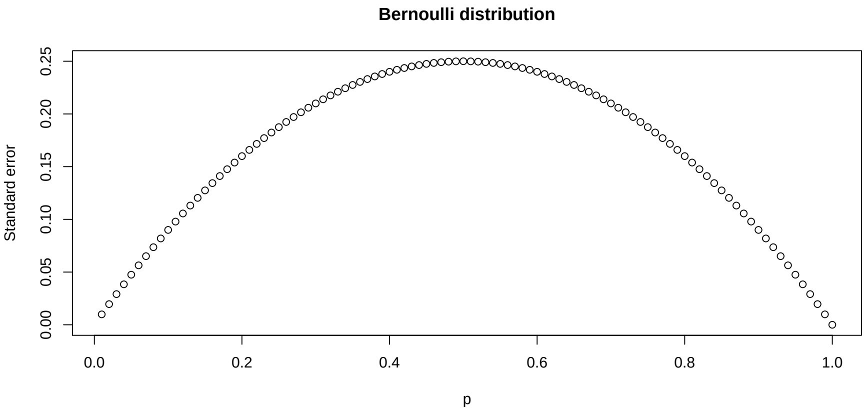
$$S^* = \sqrt{p(1-p)}$$

$$p = 0.5$$

$$VAR(x) = 0.5 \cdot (1 - 0.5) = 0.5^2 = 0.25$$

Standard error with dichotome variables

Being a pessimist



$$\textbf{standard error}: SE = \frac{S^*}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \approx \frac{S^*}{\sqrt{n}} \approx \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Compute the sample size to measure the support for a party with the precision of 2 percent!

Standard error with dichotome variables

Determining sample size

Compute the sample size to measure the support for a party with the precision of 2 percent!

- 2 percent $=> SE = 1$,
- maximum of variance: $50 \cdot (100 - 50) = 2500$
- $SE = \frac{S^*}{\sqrt{n}}$

$$\Downarrow$$

- $1 = \frac{\sqrt{2500}}{\sqrt{n}}$

$$\Downarrow$$

- $1 \cdot \sqrt{n} = \sqrt{2500}$
- $n = 2500$

Determining sample size

Example

Compute the sample size to measure the time spent in front of television among Hungarian citizents! Let us choose a precision of 5 minutes.

Determining sample size

Example

Compute the sample size to measure the time spent in front of television among Hungarian citizents! Let us choose a precision of 5 minutes.

- 5 mins $=> SE = 2.5$,
- estimated standard deviation: 10
- $SE = \frac{S^*}{\sqrt{n}}$

$$\Downarrow$$

- $2,5 = \frac{10}{\sqrt{n}}$

$$\Downarrow$$

- $2,5 \cdot \sqrt{n} = 10$
- $\sqrt{n} = 4$
- $n = 16$

Compute the sample size to measure the time spent in front of television among Hungarian citizents! Let us choose a precision of 1 minutes.

Determining sample size

Example

Compute the sample size to measure the time spent in front of television among Hungarian citizents! Let us choose a precision of 1 minutes.

- 1 mins $=>$ $SE = 0.5$,
- estimated deviation: 10
- $SE = \frac{S^*}{\sqrt{n}}$

$$\Downarrow$$

- $0,5 = \frac{10}{\sqrt{n}}$

$$\Downarrow$$

- $0,5 \cdot \sqrt{n} = 10$
- $\sqrt{n} = 20$
- $n = 400$

We asked 4 student about the number of cats at home:

	Rockers	Rappers
Girls	9	7
Boys	3	1

Imagine, what would be the results if the sample was choosen randomly and if it was stratified?

Choosing samples of n=2:

1. SRS: 6 possible samples: (1,7) (1,9) (3,7) (3,9) (1,3) (7,9)
   $\overline{x} = \frac{4+5+5+6+2+8}{6} = 5, S^{*} = \frac{1+0+0+1+9+9}{6} = 3.33$
2. Strat. Sampling: 4 possible samples: (1,7) (1,9) (3,7) (3,9)
   $\overline{x} = \frac{4+5+5+6}{4} = 5, S^{*} = \frac{1+0+0+1}{4} = 0.5$
3. Strat. Sampling: 4 possible samples: (1,3) (1,9) (3,9) (3,7)
   $\overline{x} = \frac{2+5+6+5}{4} = 4.5, S^{*} = \frac{2.5^2+0.5^2+1.5^2+0.5^2}{4} = 2.25$

Singleton, R. A. Jr. and Bruce C. Straits (1999): Approaches to Social Research. Third Edition. Oxford University Press: New York/Oxford.

Questions:

1. What is reliability? *How do the main rules concerning the order of survey questions improve the reliability and validity of survey data?* (pp. 113-117, 292-296)
2. What is meant by probability sampling? How do stratification and multistage cluster sampling affect sampling errors? Why? (pp. 141-142, 145-156)
3. What are the main types of non-probability sampling? Explain why these types do not meet the criteria of probability samples. (pp. 157-169)
4. What factors affect the desired sample size? (pp. 163-169)

It was a pleasure!

Gergely Daróczi
daroczi.gergely@btk.ppke.hu