Quantitative methods

Week #7

Gergely Daróczi

Corvinus University of Budapest, Hungary

23 March 2012

Outline

1. Sample-bias
2. Sampling theory
3. Probability sampling
   - Simple Random Sampling
   - Stratified Sampling
   - Systematic Random Sampling
   - Multi-Stage Sampling
4. Nonprobability sampling
5. Computation
   - Required formulas
   - Standard error
   - A basic example
   - Comparison of samples
   - Standard error in finite population

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!

Sample-bias

Cause of errors

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.

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.

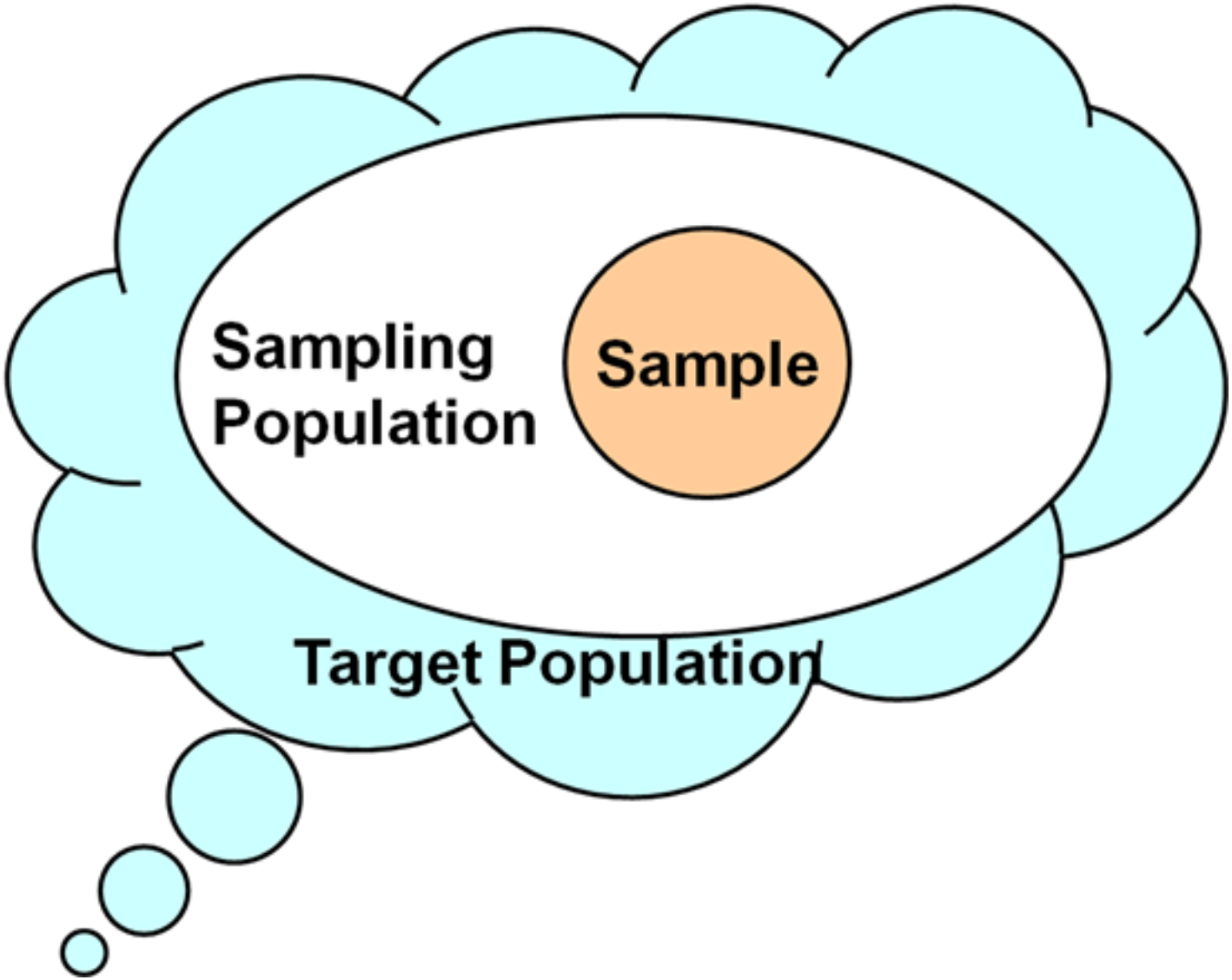
- records of registered automobile owners and telephone users,
- George Gallup: quota sampling with 50.000 respondents.

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

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

Sampling theory

Elements



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.

Sampling methods - Probability sampling

A short summary

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.

Sampling methods - Nonprobability sampling

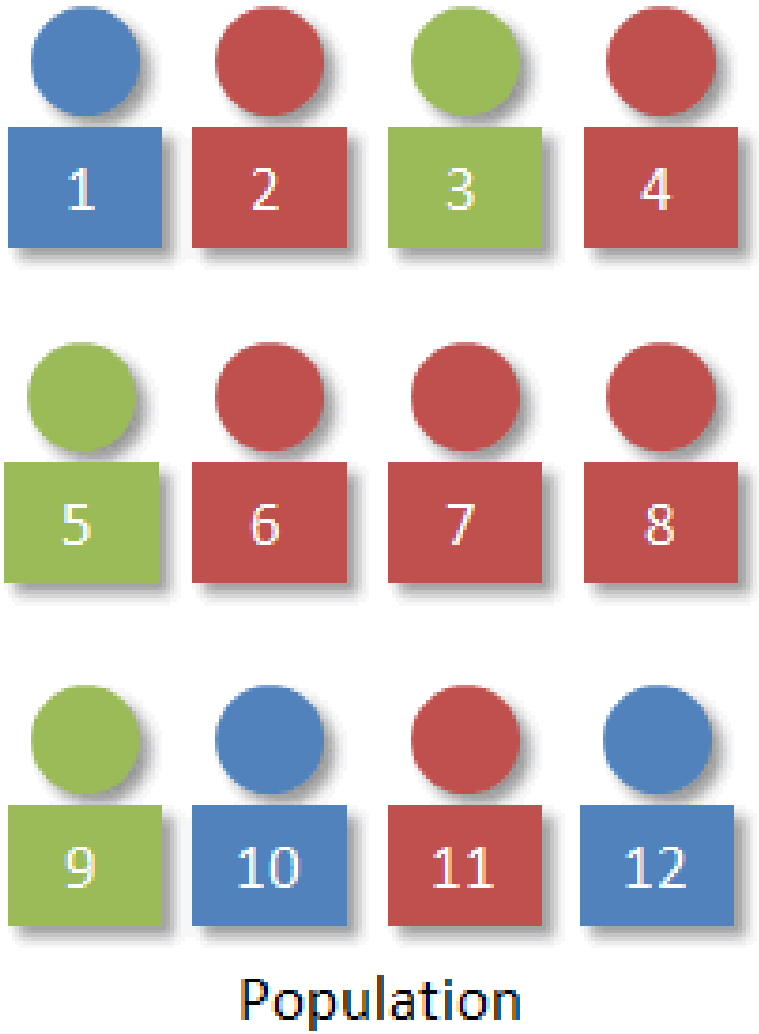
A short summary

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.

Simple Random Sampling

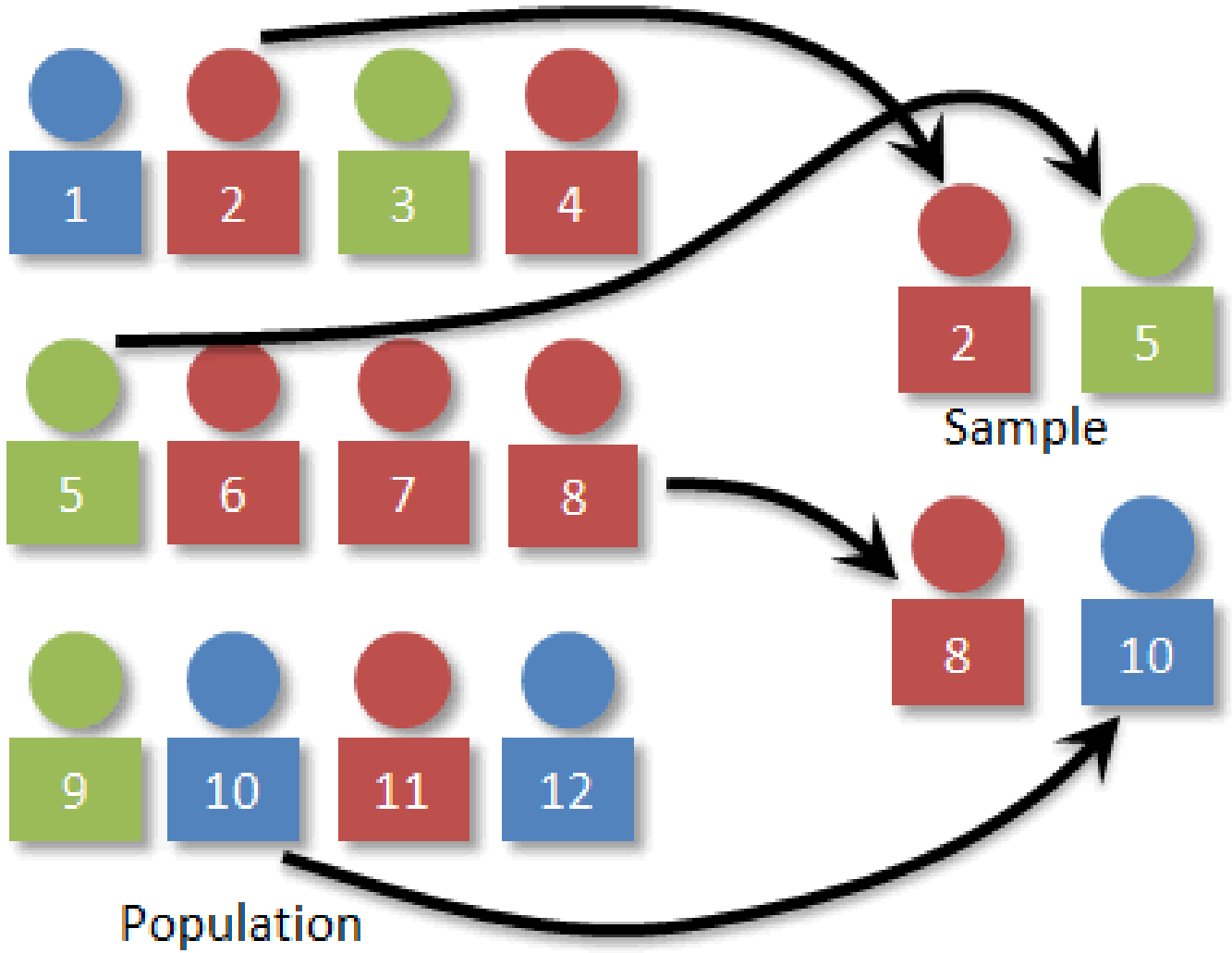
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Simple Random Sampling

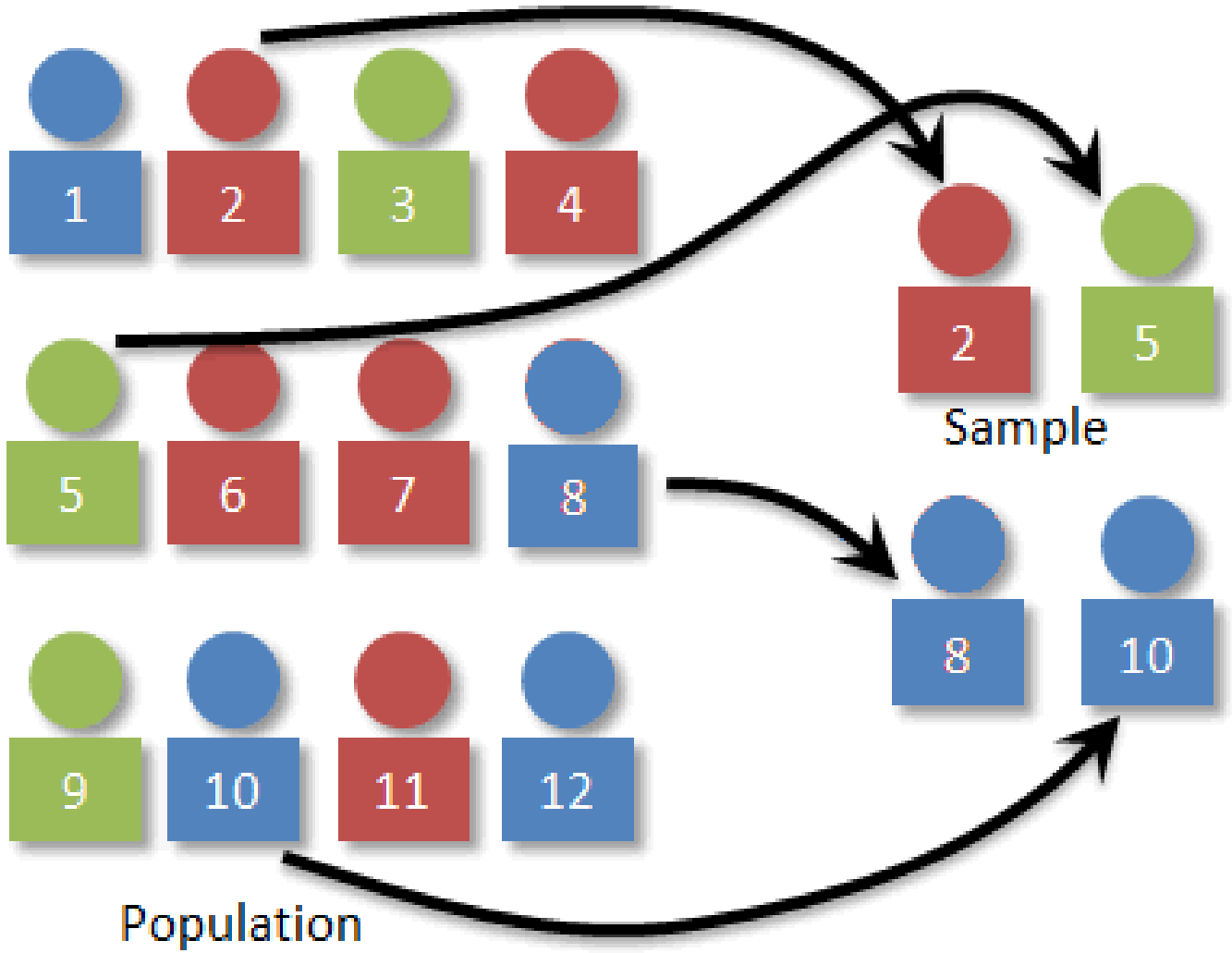
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Simple Random Sampling

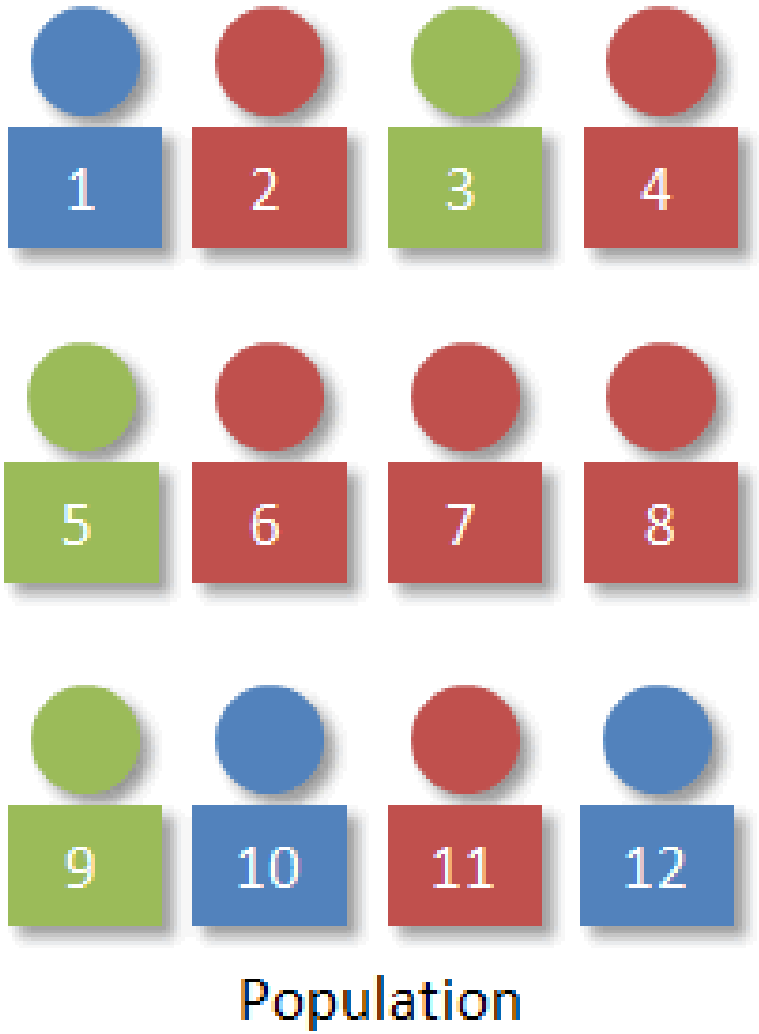
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Stratified Sampling

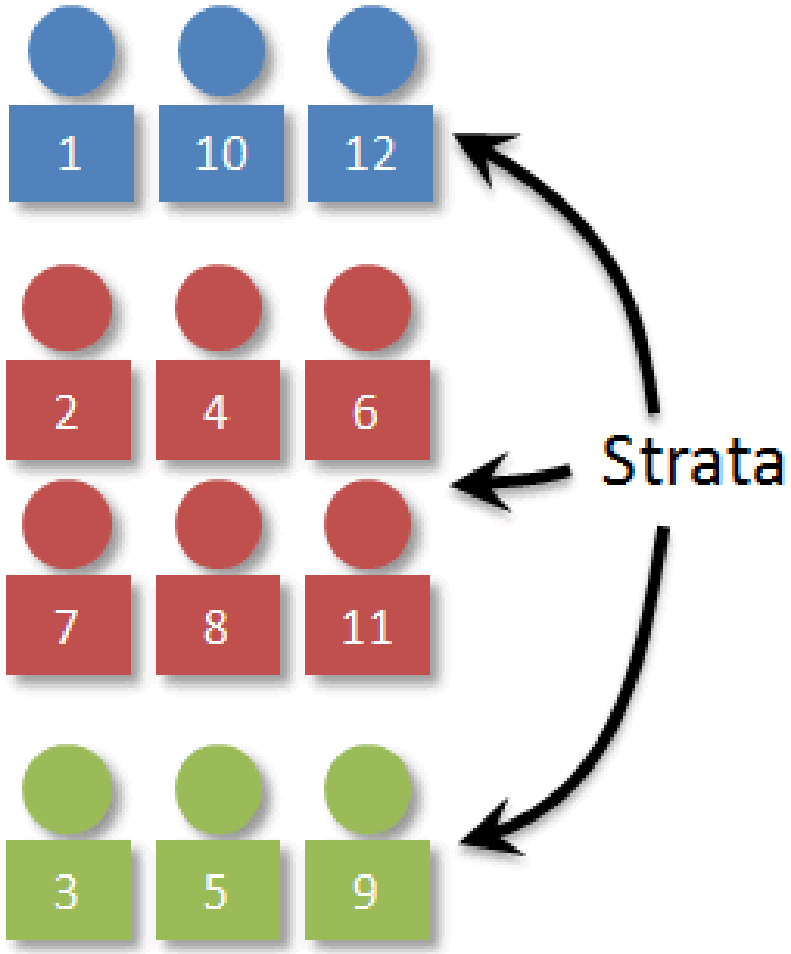
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Stratified Sampling

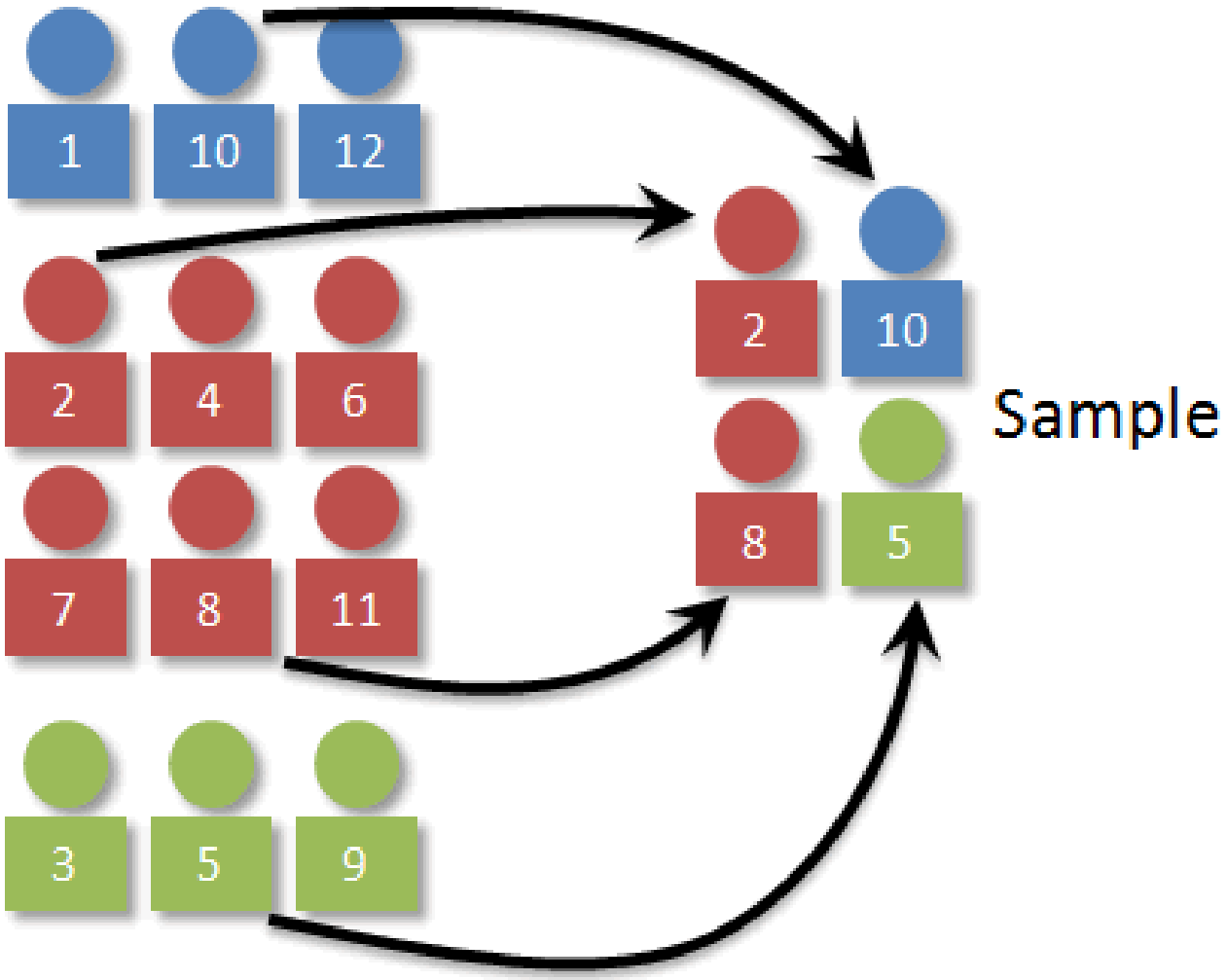
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Stratified Sampling

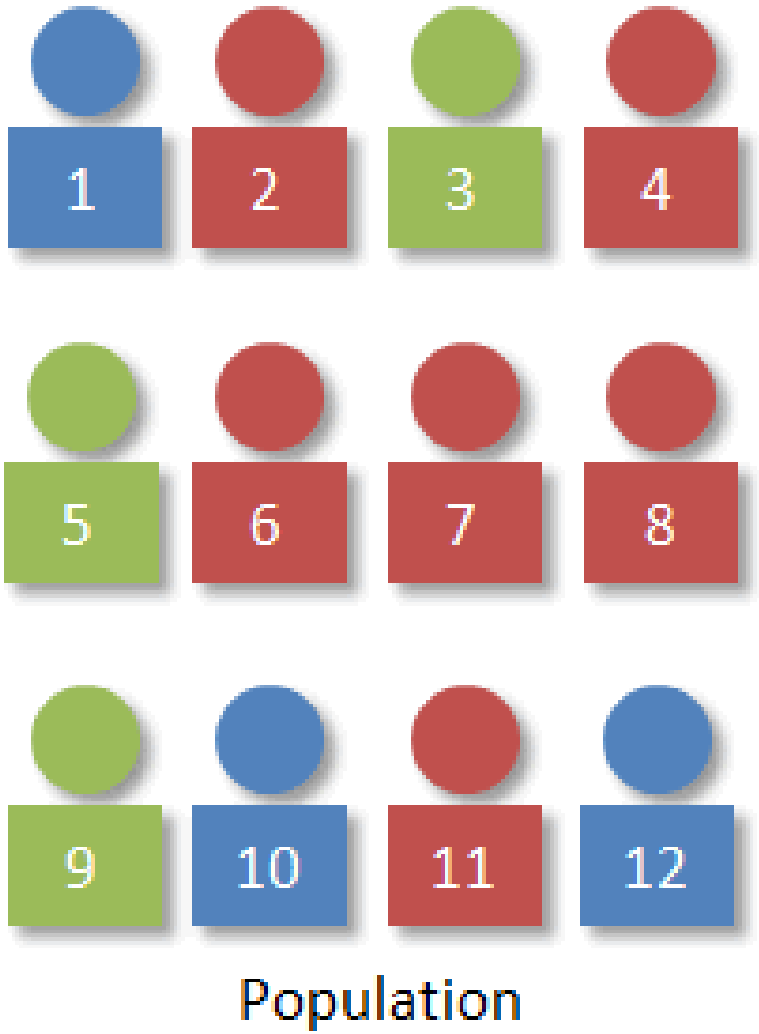
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Systematic Random Sampling

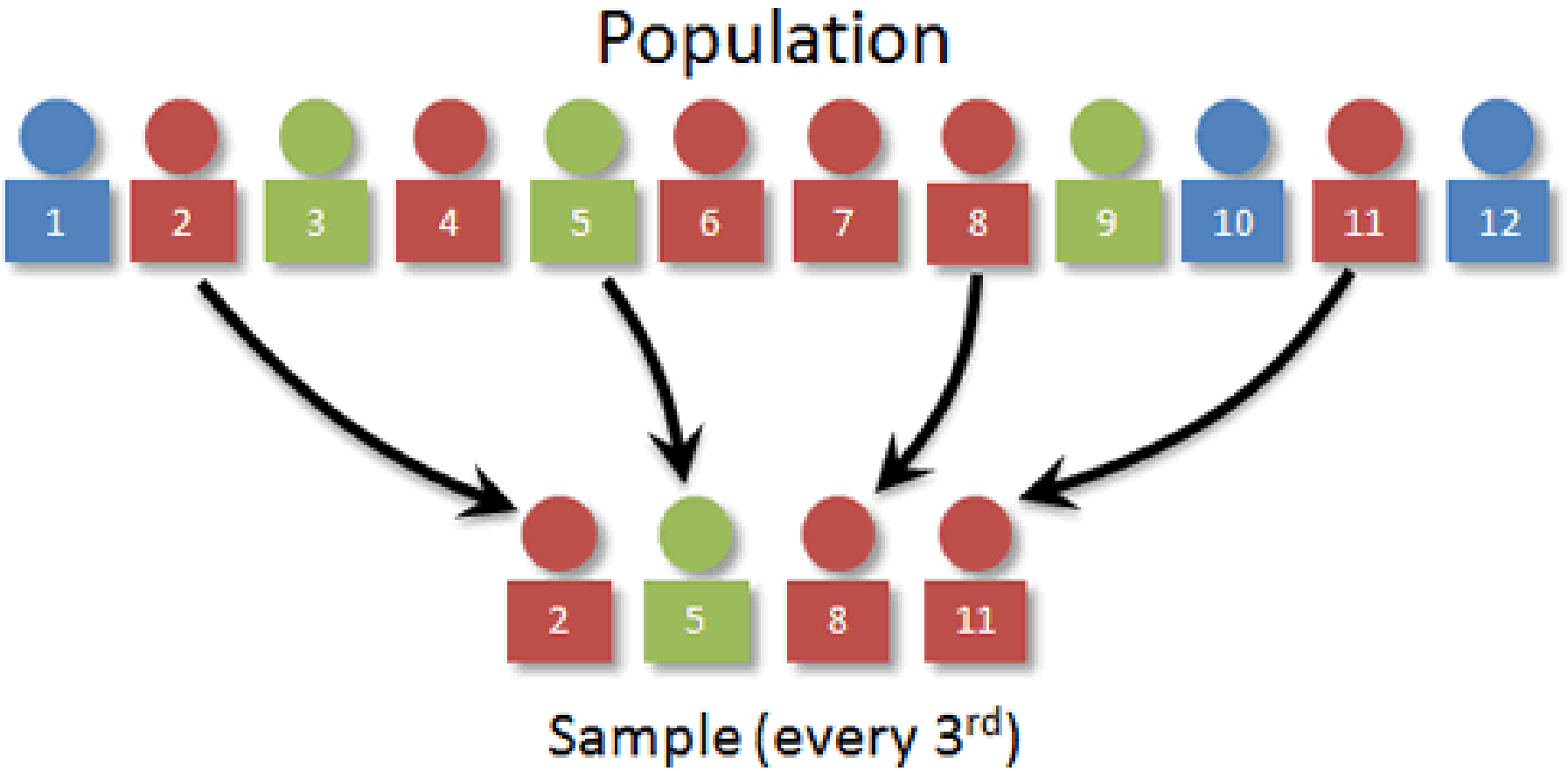
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Systematic Random Sampling

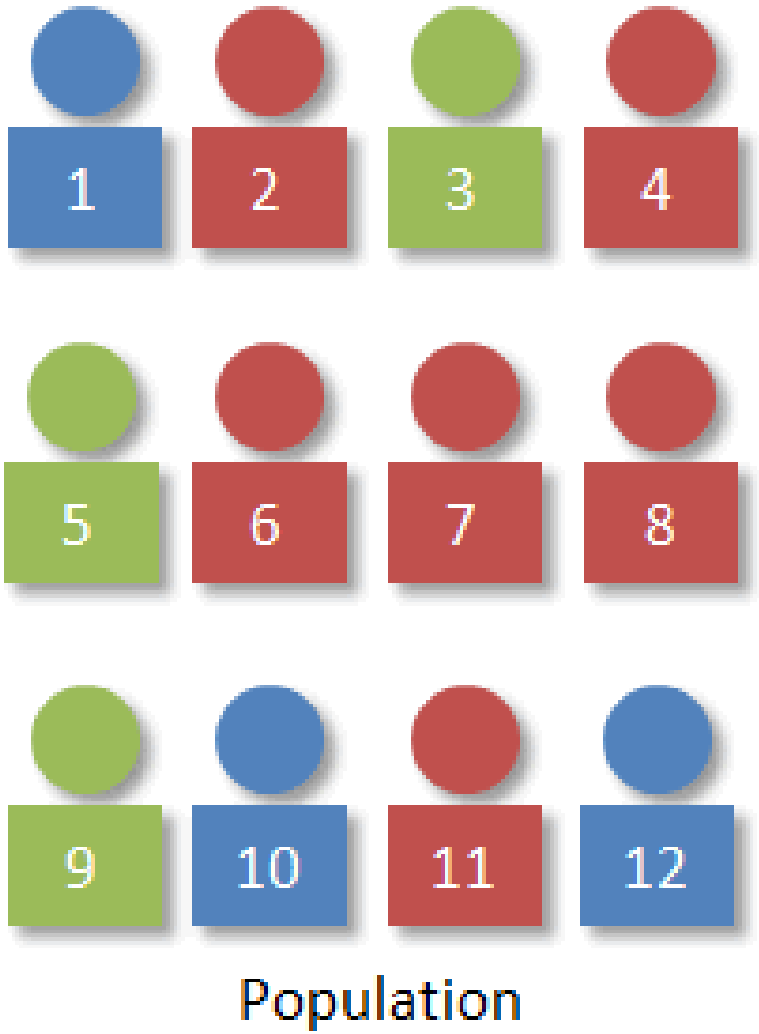
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Multi-Stage Sampling

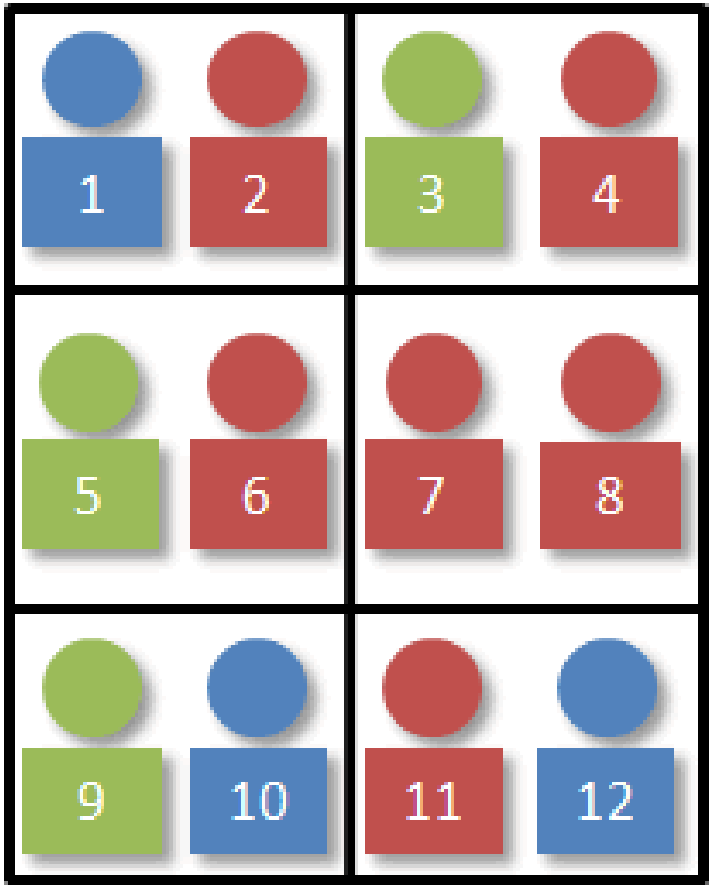
Drawing a sample



Source: Dan Kerlner, Elgin Community College

Multi-Stage Sampling

Drawing a sample

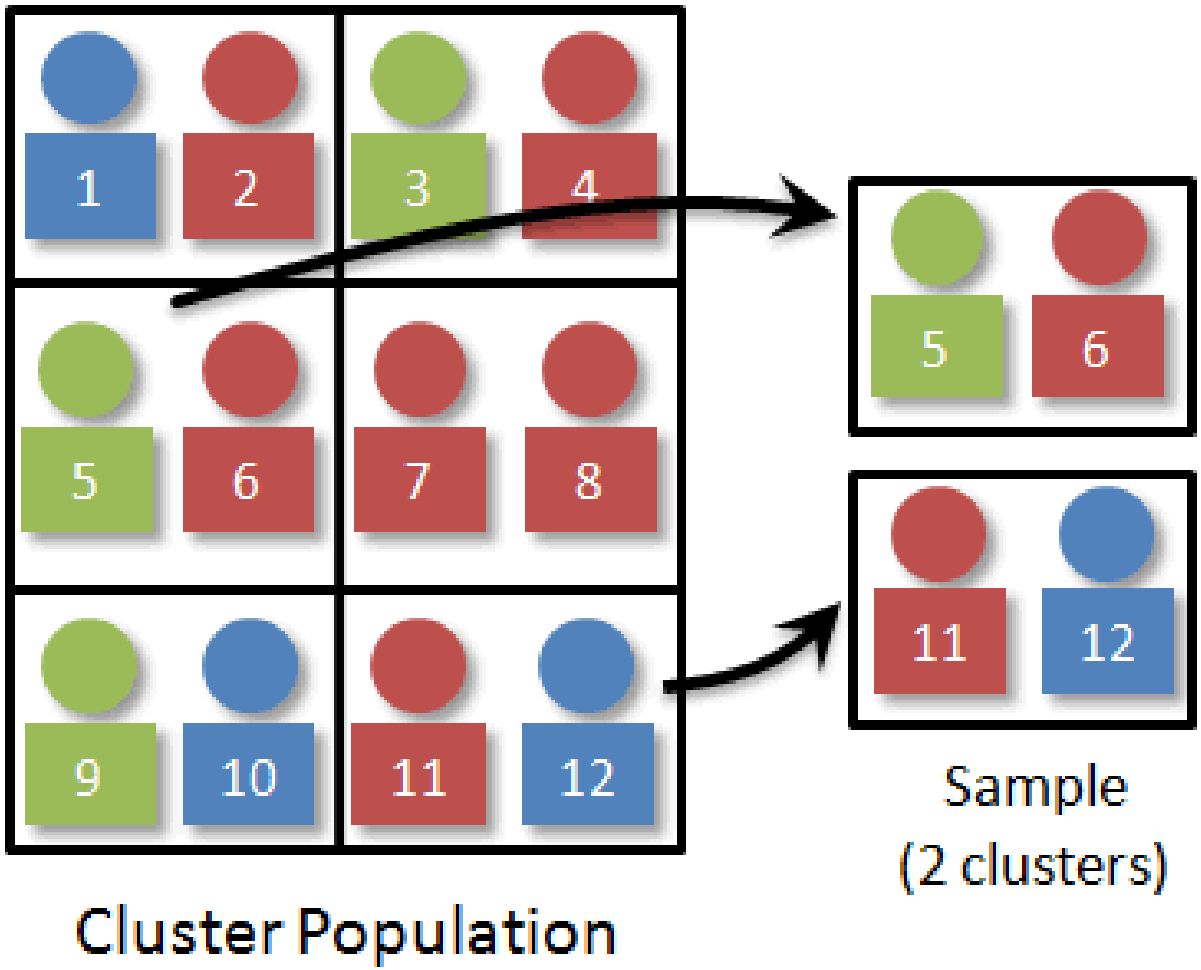


Cluster Population

Source: Dan Kerlner, Elgin Community College

Multi-Stage Sampling

Drawing a sample



Source: Dan Kerlner, Elgin Community College

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%)

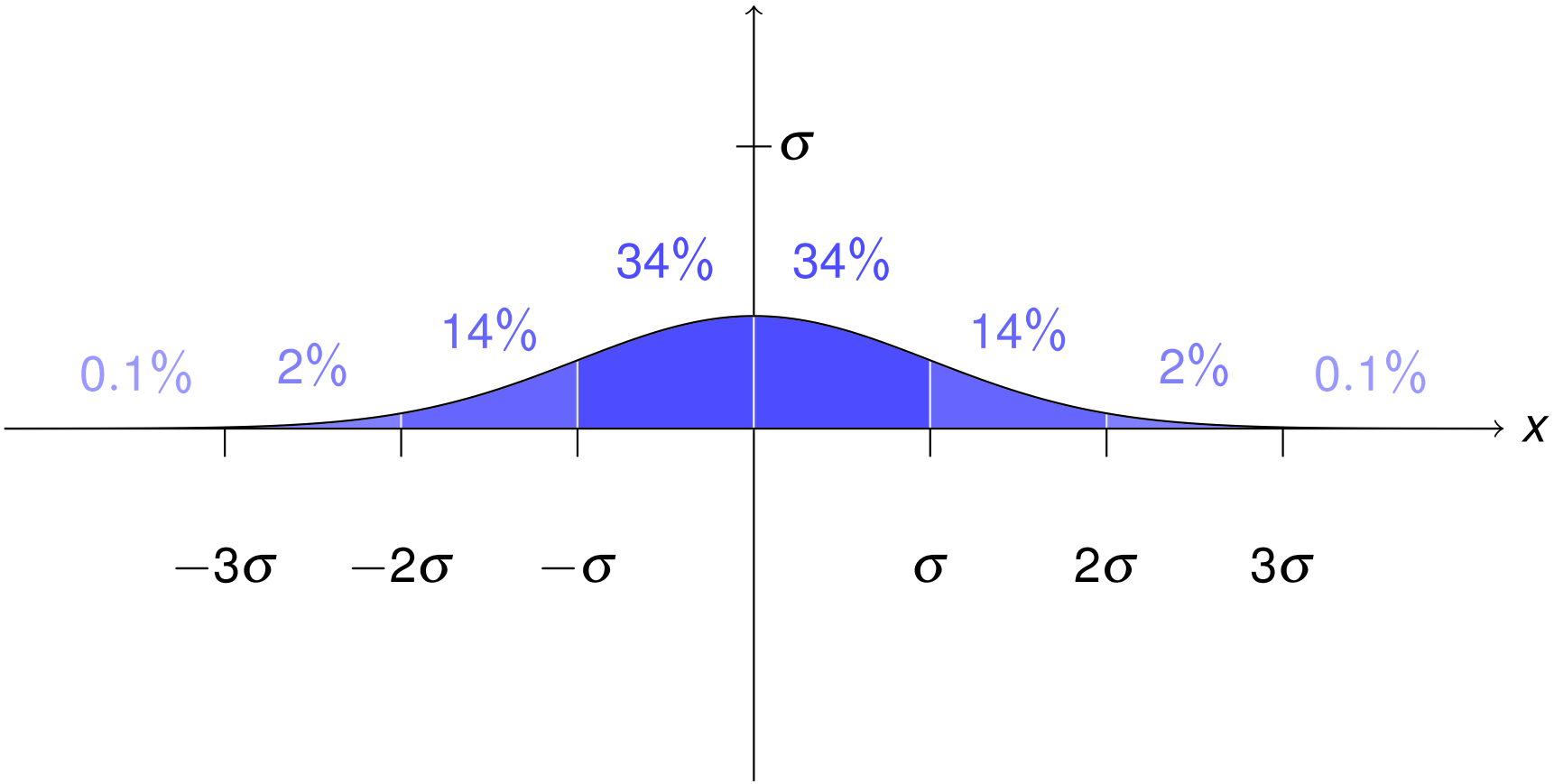
$$FPC = \sqrt{1 - \frac{n}{N}}$$

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

Computation

A short summary on Standard error

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



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

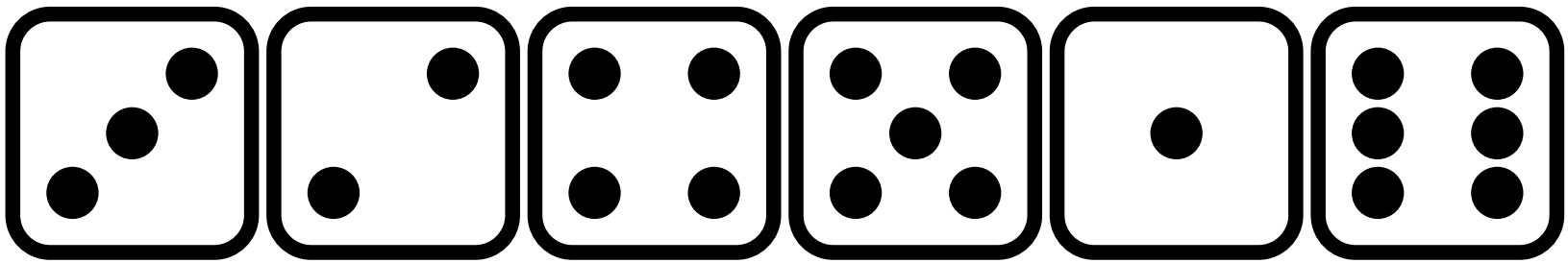
Game rules

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.

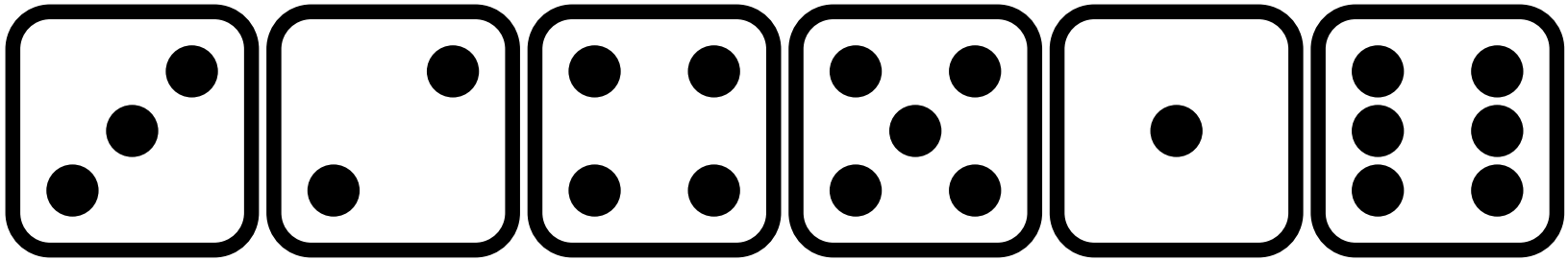
After rolling the below values, what would you think about the expected value of the game?



Would you continue playing?

Computation

Results



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

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

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

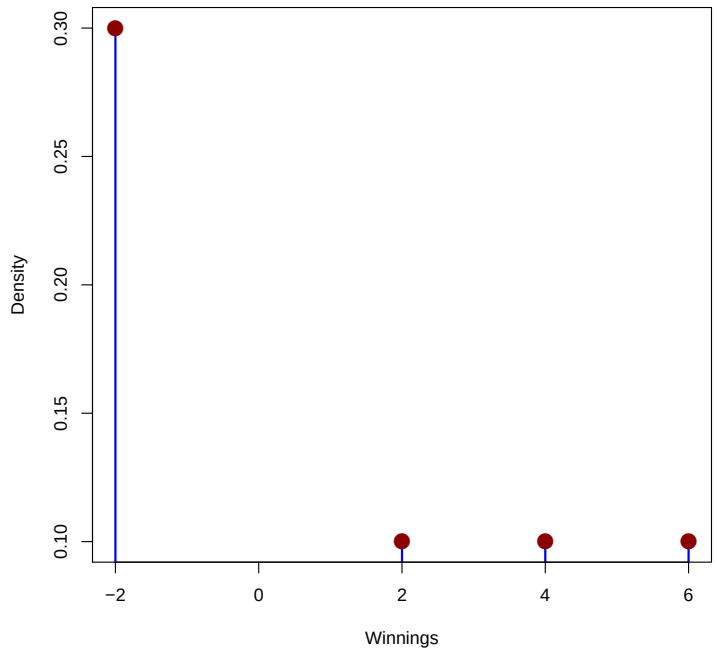
$$= \sqrt{\frac{9+1+9+9+9+25}{5}} = \sqrt{\frac{62}{5}} = \sqrt{12.4} = 3.521363$$

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

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

Good luck!

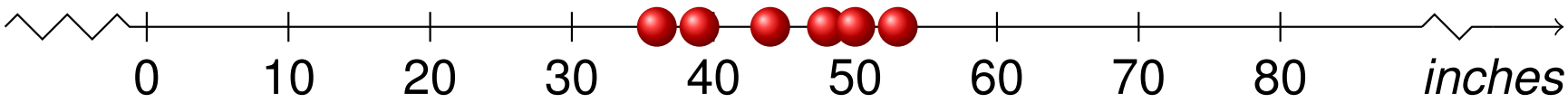
Forget about the experiment and try to determine the **real** expected value of the game!



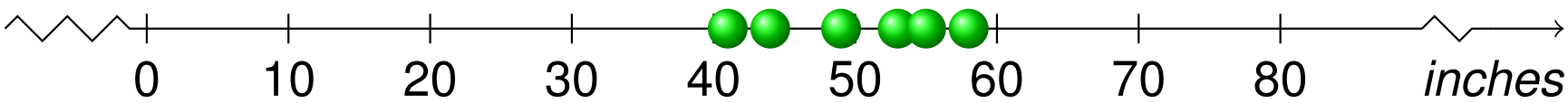
What is wrong with the above plot?

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



2. **2011 April 1**: 41 53 55 49 58 44

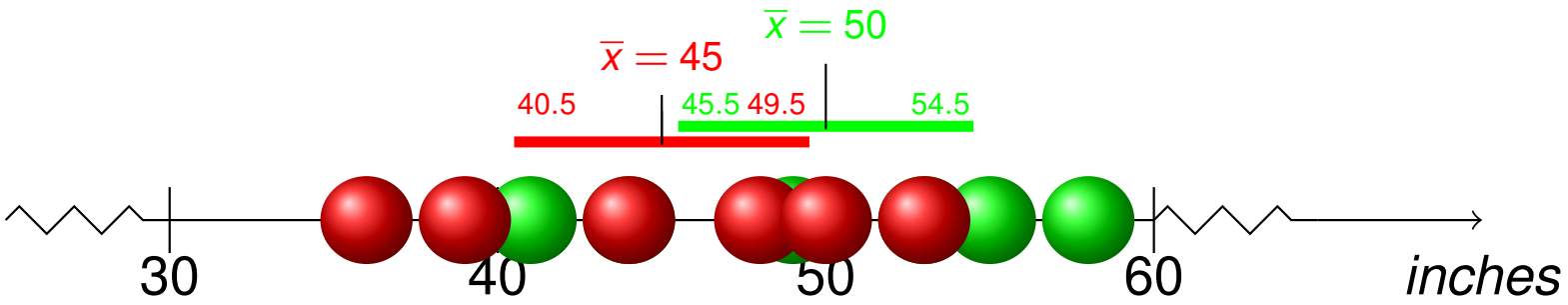


Computation

Results

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. **2010 November 22**: 36 48 50 44 53 39
2. **2011 April 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?

It was a pleasure!

Daróczi Gergely
daroczi.gergely@btk.ppke.hu