WESTMINSTER
International University in Tashkent

Week 5
Sampling methods and sampling distribution

By

Olmos Isakov

Office hours: Tuesday, 09:00 – 11:00 (ATB 216)

AGENDA

1. Probability vs Non-Probability sampling
2. Probabilistic Sampling Methods
3. Sampling Distribution
4. Central Limit Theorem (CLT)

Reasons to sample

1. To contact the whole population would be time-consuming.

   E.g. Contacting each voter for election survey.

2. The cost of studying all the items in a population may be out of the budget constraint. *E.g. Product test using 37 million population in Uzbekistan.*

3. The physical impossibility of checking all items in the population.

   E.g. Studying the population of birds, fish, etc.

4. The destructive nature of some tests. *E.g. Car crash test.*

5. The sample results are adequate.

   E.g. To study the price of bread, we do not need to include all stores in the nation.

Sampling methods

Probability Sampling

- selected at random
- everyone has known chance
- useful for diverse populations
- represents the population accurately
- finding respondent is difficult and expensive

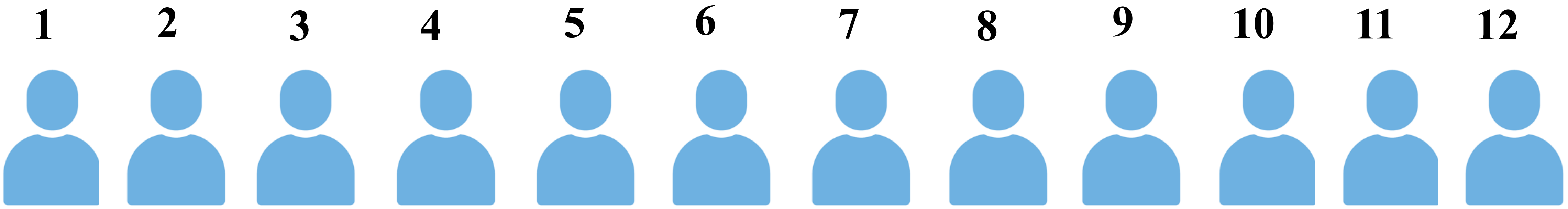
Non-probability Sampling

- selected based on judgement of the researcher
- everyone does not have known chance
- useful for populations with similar traits
- unrepresentative sample
- finding respondent is easy and inexpensive

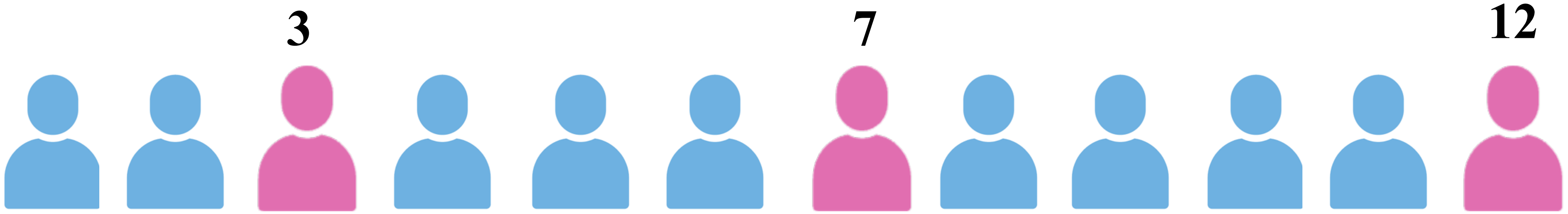
Sampling methods

- Non-probability
  - Convenience
  - Quota
  - Purposive
  - Snowball
  - Self-selection
- Probability
  - Simple Random Sampling
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling

SIMPLE RANDOM SAMPLE A sample selected so that each item or person in the population has the same chance of being included.

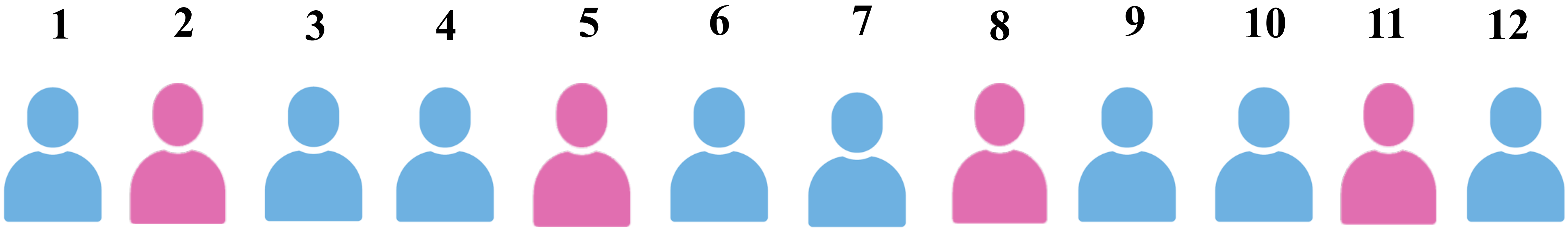


Random selection: 3, 7, 12

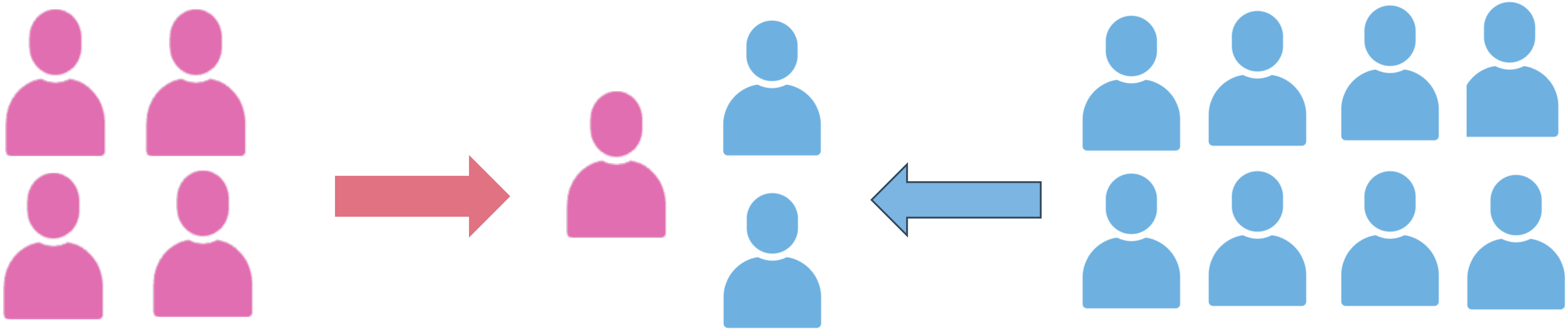


SYSTEMATIC RANDOM SAMPLE A random starting point is selected, and then every *k*th member of the population is selected.

Starting point is randomly selected (1-3): 2



STRATIFIED RANDOM SAMPLE A population is divided into subgroups, called strata, and a sample is randomly selected from each stratum.

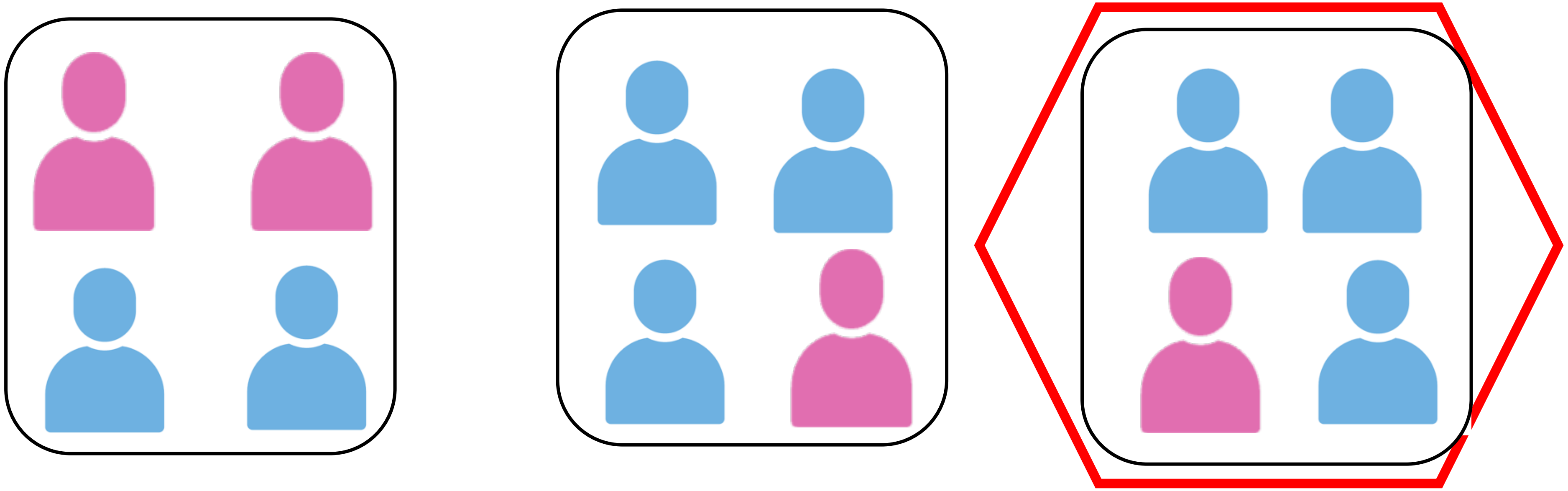


Example: Stratified sampling

A university wants to survey student satisfaction with online learning. The sample size needs to be **200** students. There are **5,000** students in total, divided by faculty as follows:

Faculty	Total students	Percentage in population	Sample size
Business	2,000	40%	80
Engineering	1,500	30%	60
Medicine	1,000	20%	40
Arts	500	10%	20
TOTAL	**5,000**	**100%**	**200**

CLUSTER SAMPLING A population is divided into clusters using naturally occurring geographic or other boundaries. Then, clusters are randomly selected and a sample is collected by randomly selecting from each cluster.



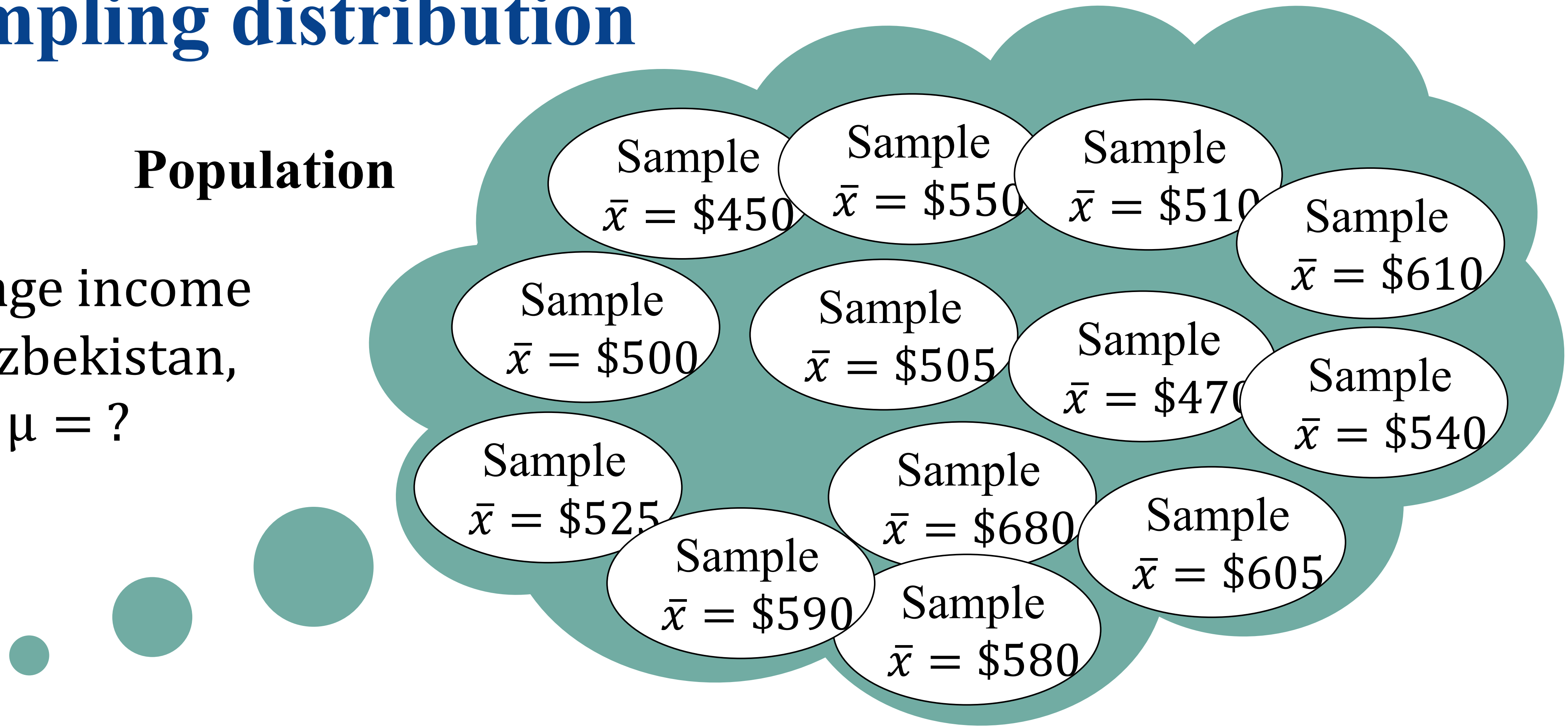
Example: Cluster sampling

Your company has offices in all 14 regions across Uzbekistan and you want to study employee productivity in your company. You randomly select 3 offices using SRS. Then study all employees in these 3 offices.

Sampling distribution

Population

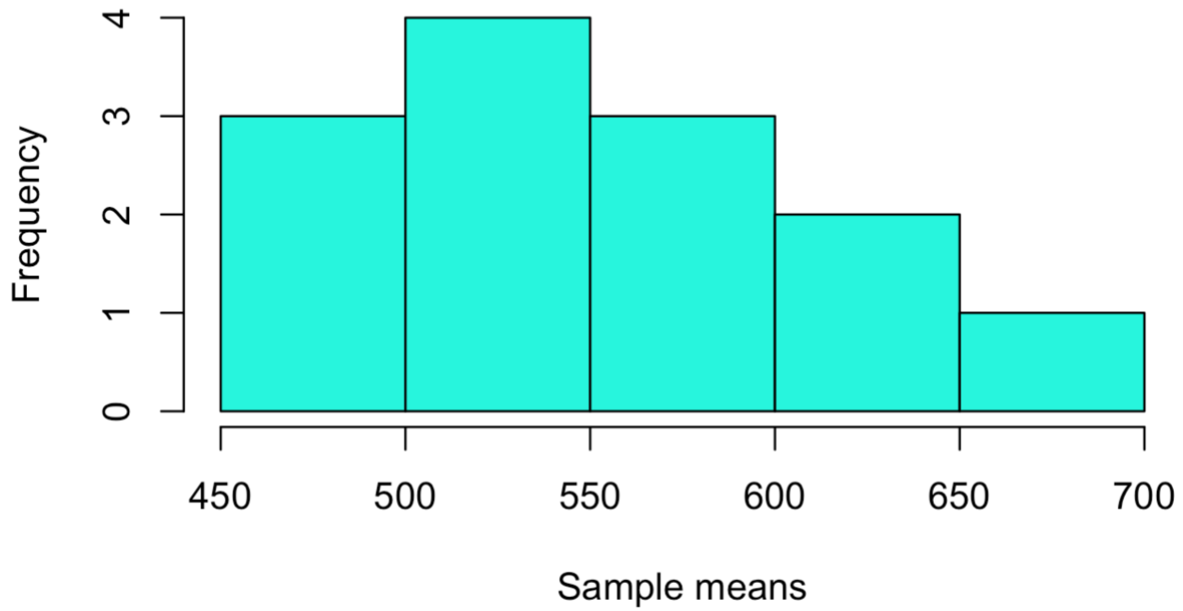
Average income
in Uzbekistan,
$\mu = ?$



A **sampling distribution** is the probability distribution of a sample statistic that is formed when samples of size *n* are repeatedly taken from a population.

Sampling distribution

If the sample statistic is the sample mean, then the distribution is the **sampling distribution of sample means**.



SAMPLING ERROR The difference between a sample statistic and its corresponding population parameter.

Example: Sampling distribution

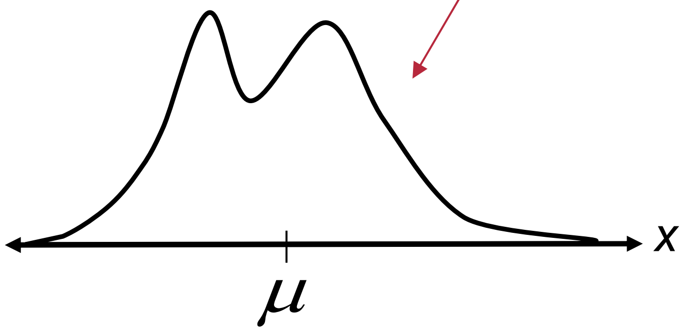
Example: Suppose we have a small population of the following values: **3, 5, 10, 12** where, $\mu = \frac{3+5+10+12}{4}$ = **7.5** (population parameter). Then you have the following sampling distribution for n = 2 (without replacement) and their means:

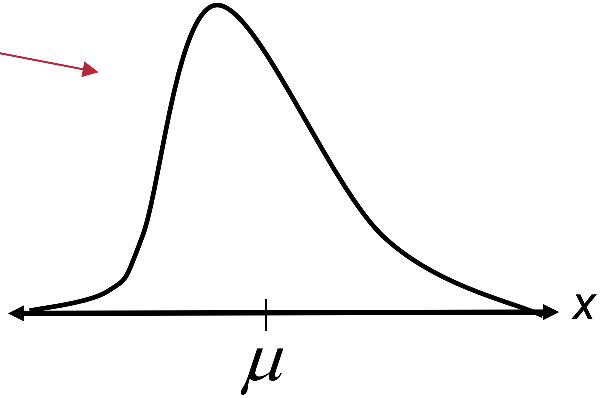
Sampling Distribution	Sampling Distribution of sample means	
3, 5	4	$\bar{x}_1$
3, 10	6.5	$\bar{x}_2$
3, 12	7.5	$\bar{x}_3$
5, 10	7.5	$\bar{x}_4$
5, 12	8.5	$\bar{x}_5$
10, 12	11	$\bar{x}_6$
$\boldsymbol{\mu_{\bar{x}}}$	**7.5**	

Sample statistics

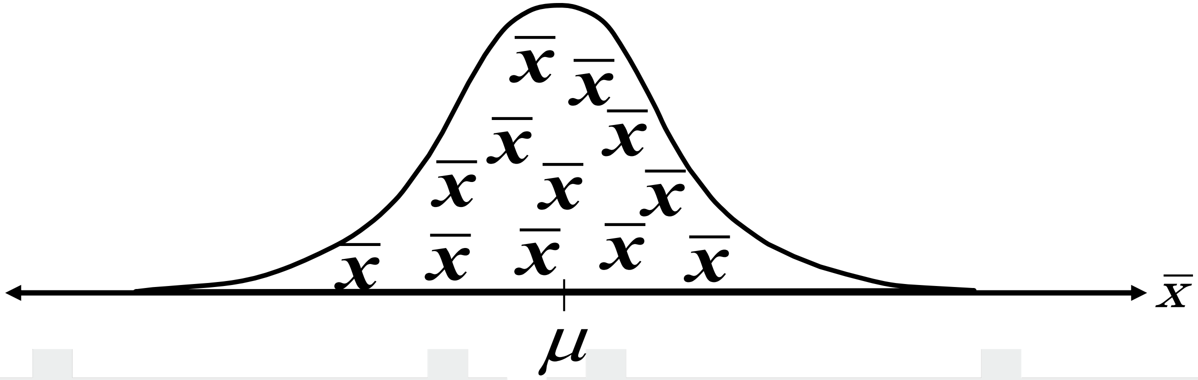
Central Limit Theorem (CLT)

If a sample of size $n \geq 30$ is taken from a population with ***any type of distribution*** that has a mean = μ and standard deviation = σ,



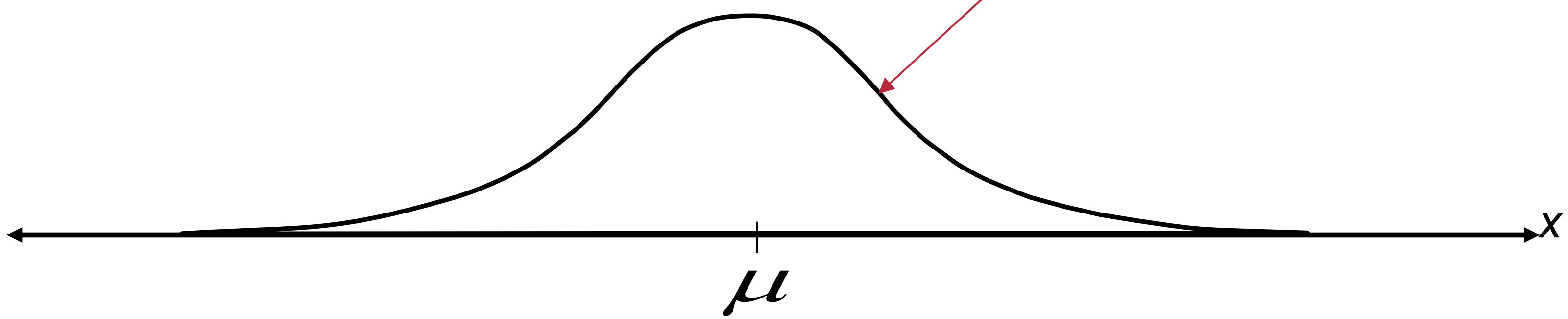


the ***sample means*** will have a **normal distribution**.

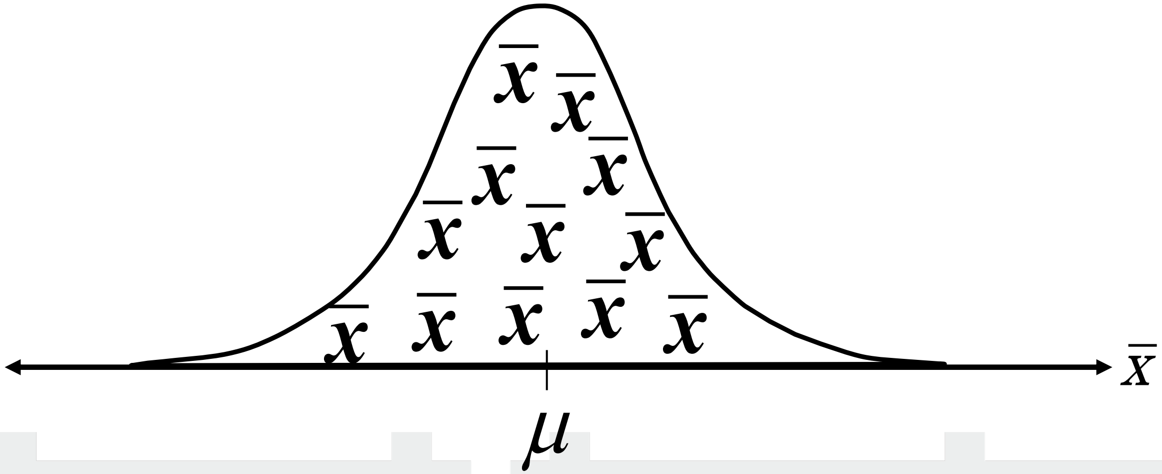


Central Limit Theorem (CLT)

If the population itself is ***normally distributed***, with mean = μ and standard deviation = σ,



the ***sample means*** will have a **normal distribution** for *any* sample size n.



Central Limit Theorem (CLT)

In either case, the sampling distribution of sample means has a mean equal to the population mean.

$$\mu_{\bar{x}} = \mu$$

Mean of the sample means

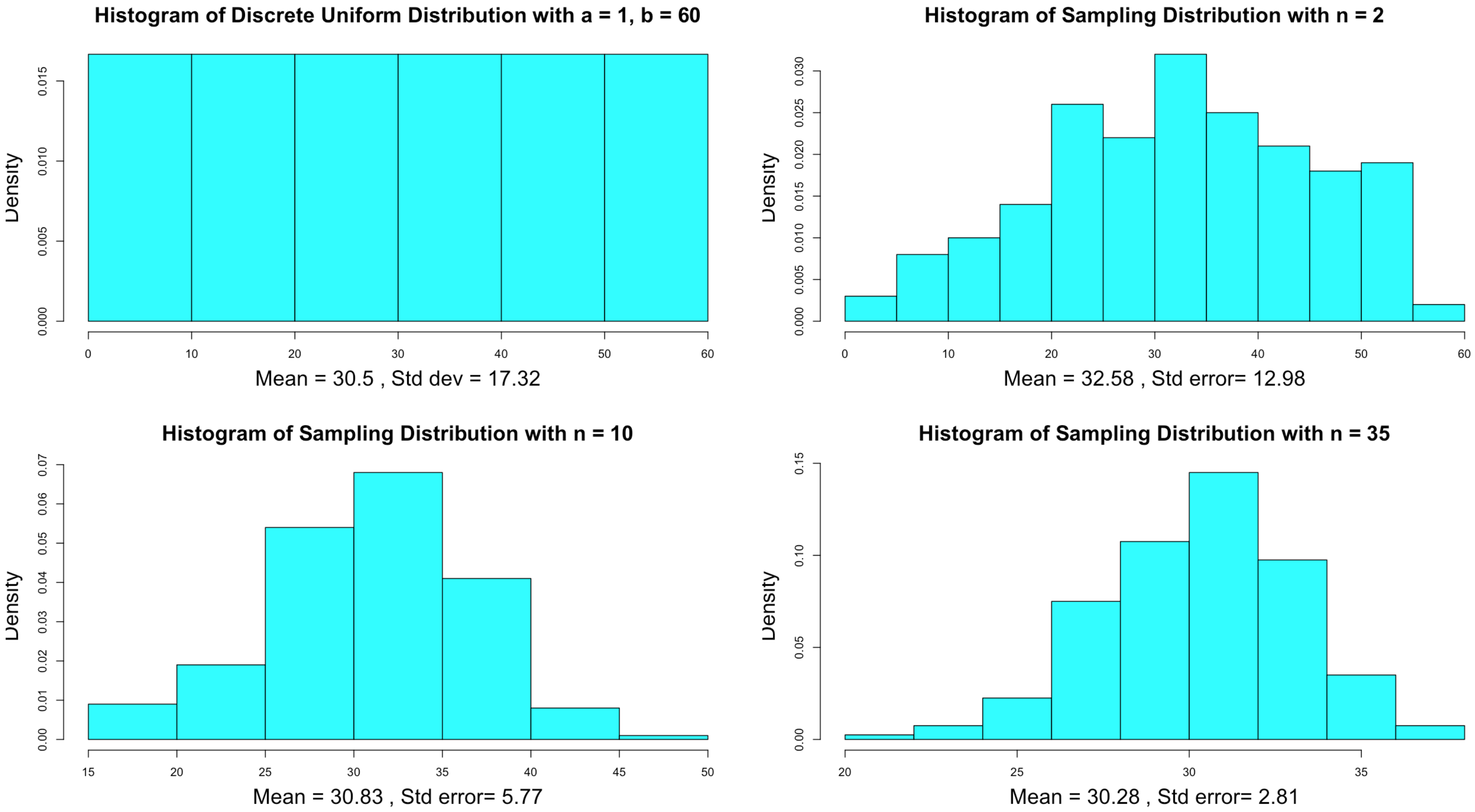
The sampling distribution of sample means has a standard deviation equal to the population standard deviation divided by the square root of *n*.

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

Standard deviation of the sample means (standard error)

This is also called the **standard error of the mean.**

Simulation in R



FINDING THE z VALUE OF $\bar{x}$ WHEN THE POPULATION STANDARD DEVIATION IS KNOWN

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

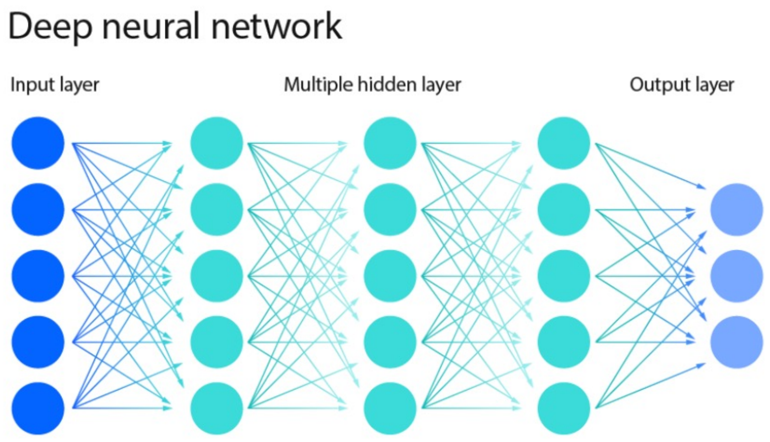
Example.

Training time for a neural network model varies due to hardware usage and random initialization and it follows a normal distribution. Assume population mean is 200 seconds and σ is 40 seconds.

a. Find the probability that the *average training time* from 36 runs exceeds 210 seconds.

b. Compare this with the probability that *a single training time* exceeds 210 seconds.

c. Find the 90th percentile of the sampling distribution of the *average training time*.



Solution: CLT

Training time for a neural network model varies due to hardware usage and random initialization and it follows a normal distribution. Assume population mean is 200 seconds and σ is 40 seconds.

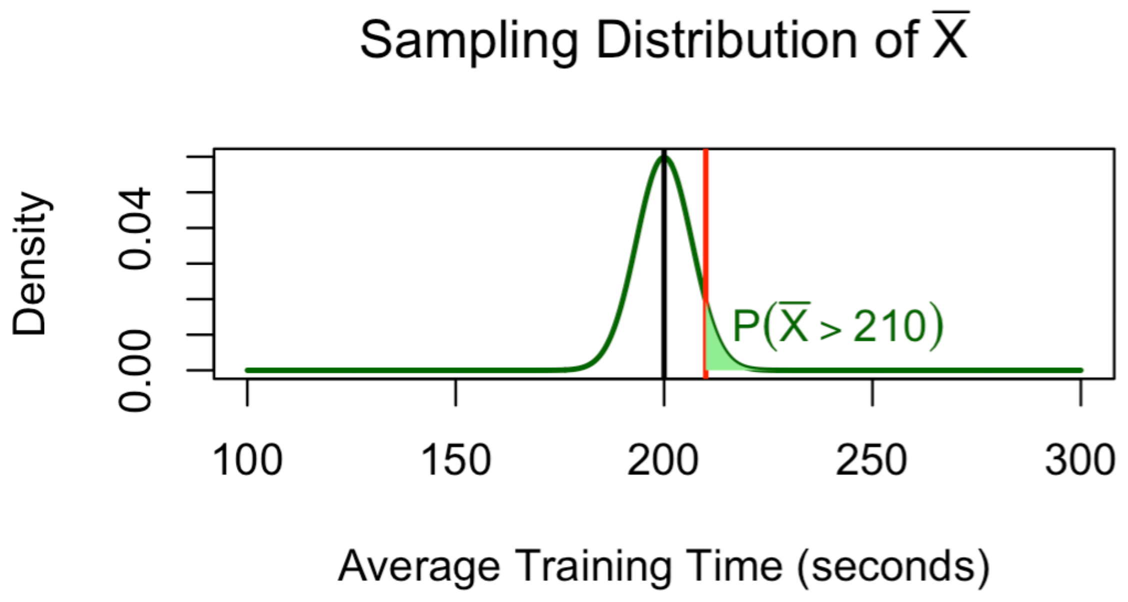
$\sigma_{\bar{x}} = \frac{40}{\sqrt{36}} \approx 6.67$

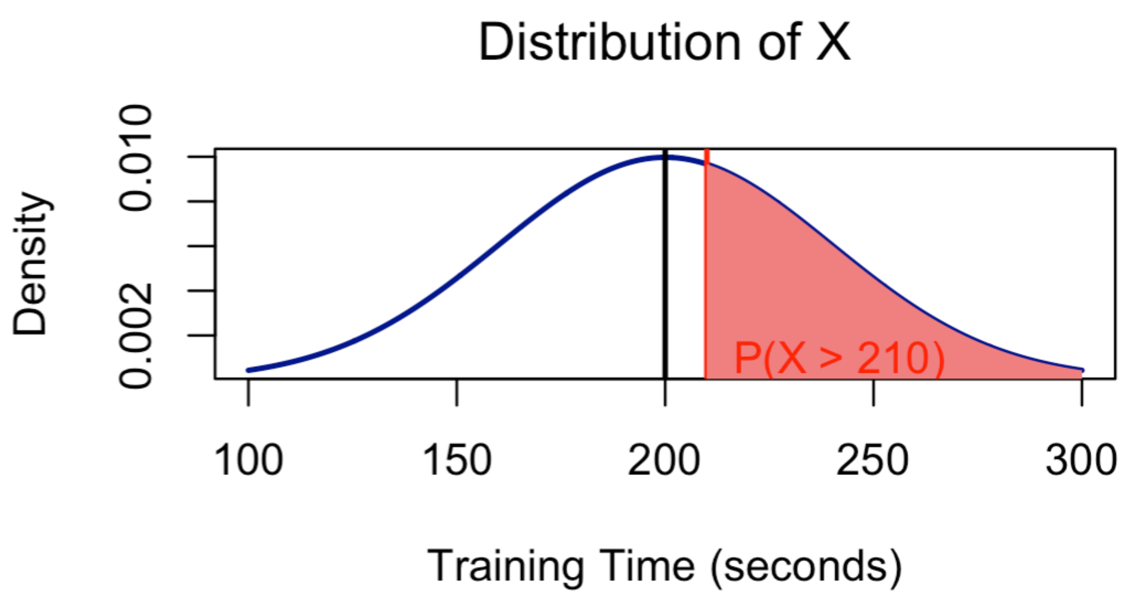
a. Find the probability that the *average training time* from 36 runs exceeds 210 seconds.

$P(\bar{x} > 210) = P(z > 1.50) = 1 – P(z < 1.50) = 0.0668$

b. Compare this with the probability that *a single training time* exceeds 210 seconds.

$P(X > 210) = P(z > 0.25) = 0.4013$



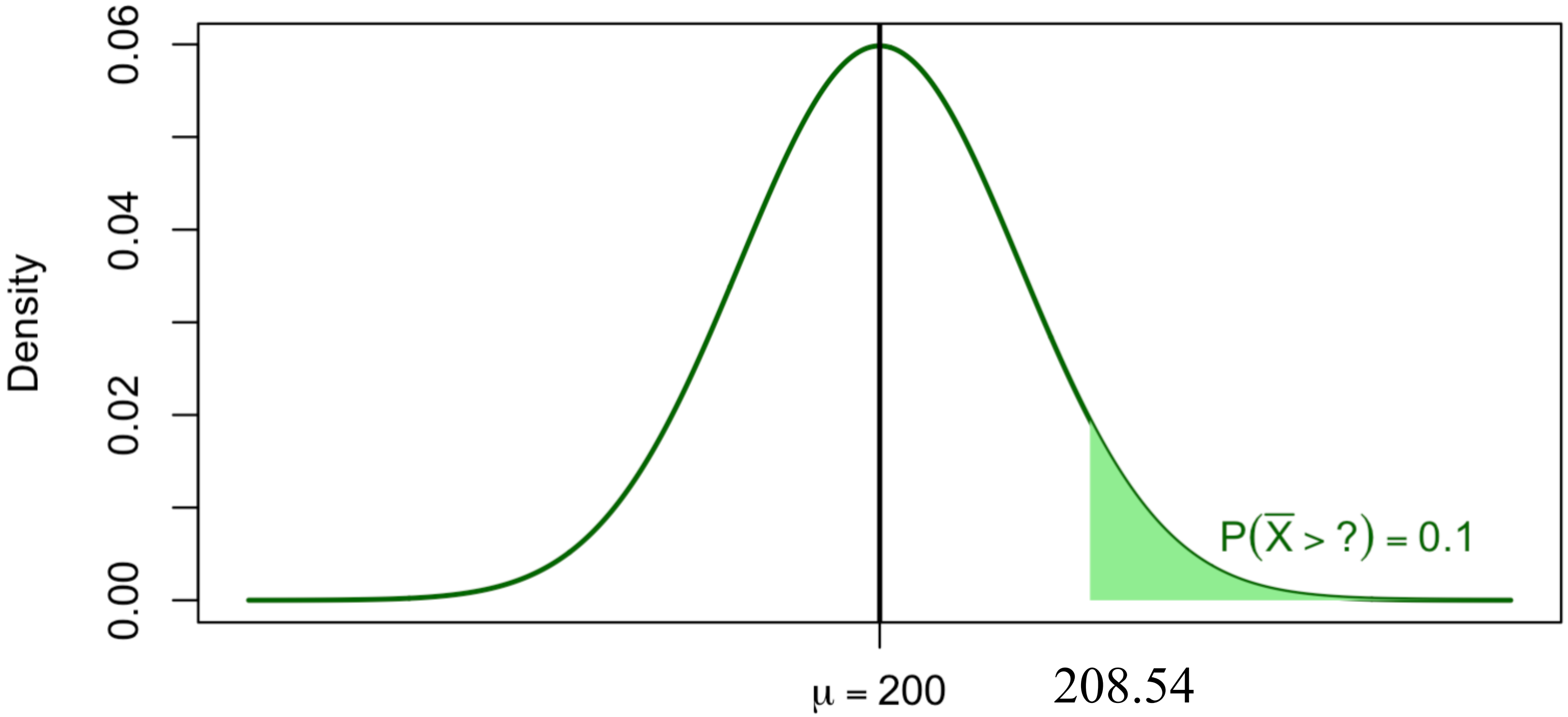


Solution: CLT

c. Find the 90th percentile of the sampling distribution of the *average training time*.

$$z = \frac{\bar{x} - 200}{6.67} = 1.28$$

$$\bar{x} = 200 + 1.28 * 6.67 =$$
$$= 208.54$$



REFERENCES

1. Lind et al. (ISBN 978-1-260-18750-2), Chapter 8.
2. McClave & Sincich (ISBN 978-0-321-75593-3), Chapter 6.
3. Ott & Longnecker (ISBN 978-0-495-01758-5), Chapter 2.

Thank You!