Sampling & Estimation

Nazgul Rakhimzhanova

IITU 2021

COURSE SCHEDULE

week	Mid Term (weeks 01-07)	End Term (weeks 08-14)	week
01	Intro: Data Science Area and open source tools for Data Science	Statistics: Distribution – Lognormal, Exponential	08
02	NumPy package for data science	Correlation and Covariance	09
03	Pandas package for data science	Sampling	10
04	Visualization with matplotlib	Hypothesis testing	11
05	Statistics: Distribution – Normal	Linear Regression	12
06	Exploratory Data Analysis (EDA)	Linear Regression	13
07	**Summary for 6 weeks QA session**	**Summary for 6 weeks QA session**	**14**
15	**Course summary**		

Outline of lecture

- Sampling
- Estimation

Sampling & Estimation

- The goal of this **lecture** is to introduce the estimation theory, but we'll talk about sampling theory first because estimation theory doesn't make sense until you understand sampling.

Sampling

- In almost every situation of interest, what we have available to us as DS is **a sample of data**.
- The data set available to us is **finite**, and **incomplete**.

Sampling

- **A sample** is a concrete thing. You can open up a data file, and there's the data from your sample.
- **A population**, is a more abstract idea.
- Sample mean, sample variance, sample std
- Population mean, population variance, population std

Sampling

Sampling methods

- **Simple** random sampling
- **Systematic** sampling
- **Stratified** sampling
- **Cluster** sampling

Simple random sampling

- **Population** – N
- **Sample** – n, n < N
- **Equal possibilities** to occur for every object

Simple random sampling

Systematic sampling

- **Population** – N
- **Sample** – n, n < N
- **Known probability**, every k-th element

Stratified sampling

- **Population** – N{A(60%),B(20%),C(20%)}
- **Sample** – n, n < N
- **The same** partition ration of different stratum

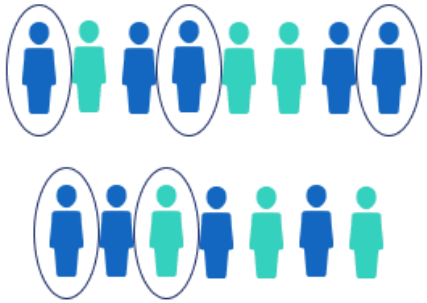
Stratified sampling

Cluster sampling

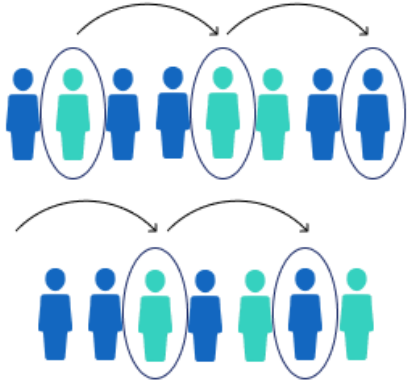
- **Population** – N{A(50%),B(50)}
- **Sample** – n, n < N
- **The same** probability for each cluster's objects

Cluster sampling

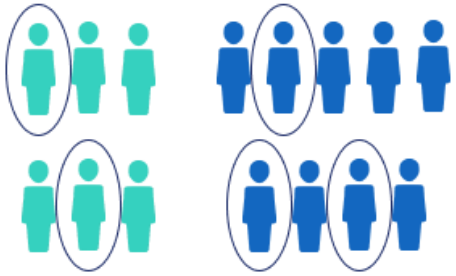
Simple random sample



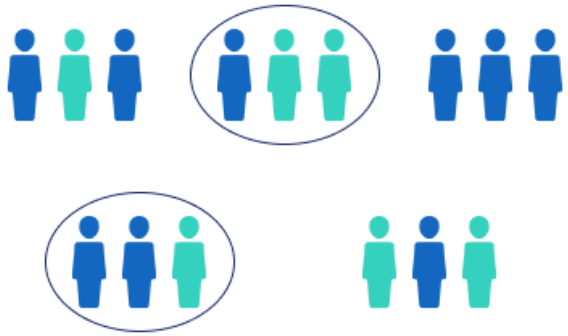
Systematic sample



Stratified sample



Cluster sample



Scribbr

Sampling Bias

- If you asking kids how many siblings do they have – your data will never show the families without children.

Sampling

- The best sampling technique provides more **precise conclusions** during descriptive analytics

- The best sampling techniques increases the **accuracy** of the **model**

Estimation

- Population mean-μ, variance – α
- Sample mean – $\bar{x}$, variance – S^2

- Estimation when we assume that

$$\mu = \bar{x} \text{ and } \alpha = S^2$$

- **Issue?**

Central Limit Theorem

- The Central Limit Theorem states that the **sampling distribution** of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution. This fact holds especially true for sample sizes over 30.

Central Limit Theorem

- Average of your sample means **will be the population mean.** In other words, add up the means from all of your samples, find the average and that average will be your actual population mean.
- Average of all of the standard deviations in your sample, will be the actual standard deviation for your population.

Central Limit Theorem

- **Central Limit Theorem (para phrased) –** regardless, the shape of the population's distribution, the distribution of **sample means** (N->∞) close to the normal distribution with mean **μ** and variance **σ^2/N**.
- Where, Standard error (SE) is a measure of how far we expect the estimate to be off, on average.

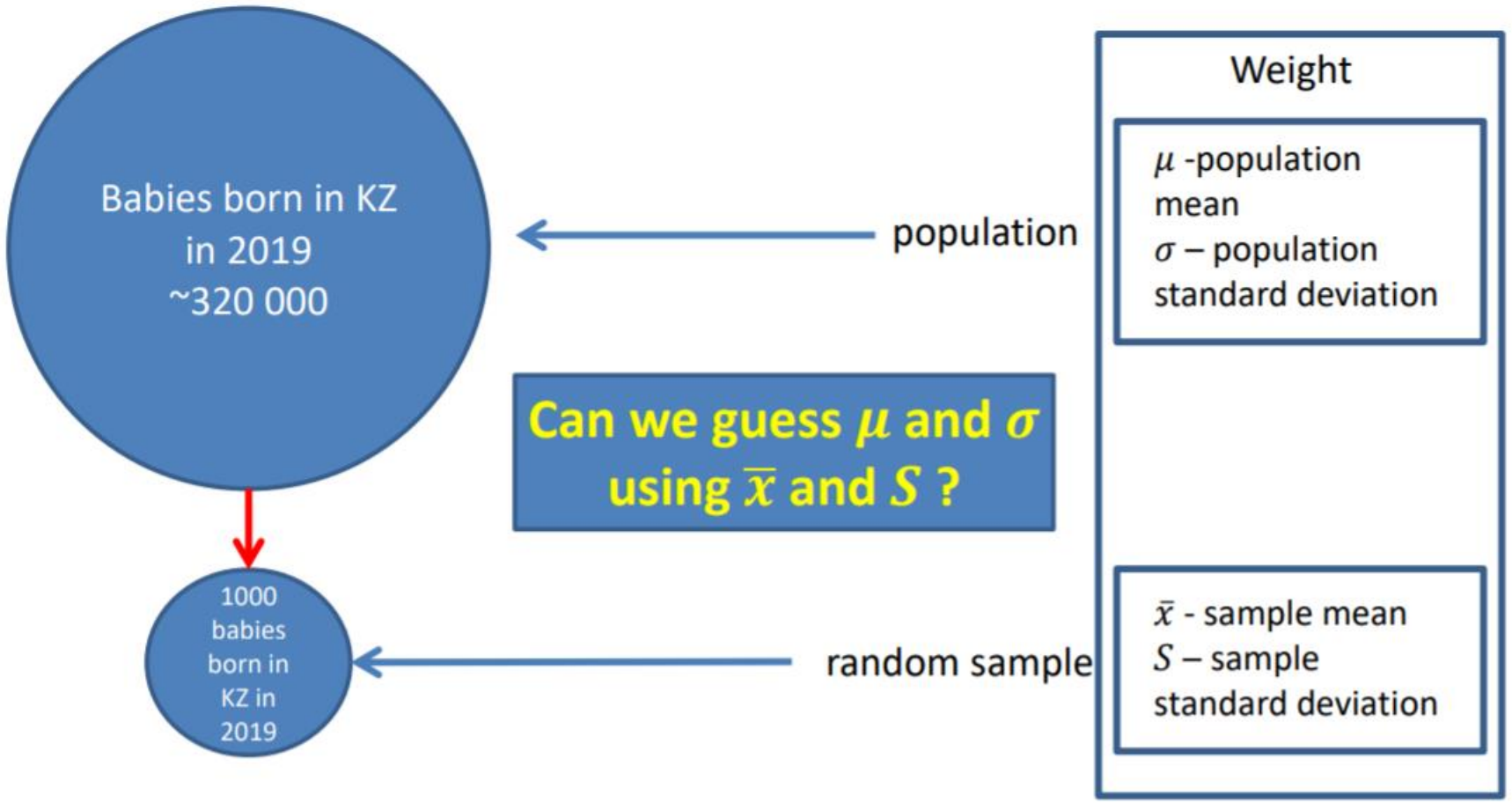
$$\mathbf{SEM = \sigma 2 / \sqrt{N}}$$

Central Limit Theorem

Summary

- If you are dealing with a sample but want to make inferences about a population – you can **experiment** with sample mean using **CLT.**

Example



https://countrymeters.info/ru/Kazakhstan#population_2019

Example

https://countrymeters.info/ru/Kazakhstan#population_2019

KEY TAKEAWAYS

- **The central limit** theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger.
- **Sample** sizes equal to or greater than 30 are considered sufficient for the CLT to hold.
- **A key** aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
- **A sufficiently** large sample size can predict the characteristics of a population accurately.

Reading

- https://www.khanacademy.org/math/statistics-probability/designing-studies/sampling-methods-stats/a/sampling-methods-review