

№7 Дәріс. Деректерді сипаттайтын статистикалық әдістер

Дәрістің мақсаты:

Деректерді сипаттайтын негізгі статистикалық әдістермен таныстыру, деректердің орталық тенденциясы, таралуы және өзгергіштігін сипаттайтын көрсеткіштерді түсіндіру, олардың есептелу жолдарын және практикалық қолданылуын көрсету. Сонымен қатар студенттерді деректердің құрылымын талдауға, шеткі мәндерді анықтауға және алынған нәтижелерге негізделген қорытынды жасауға үйрету.

Кілттік сөздер:

Деректер, статистика, дескриптивті статистика, орташа арифметикалық (mean), медиана (median), мода (mode), дисперсия (variance), стандартты ауытқу (standard deviation), размах, квартильдер, квантильдер, IQR, шеткі мәндер (outliers), асимметрия (skewness), куртозис (kurtosis), таралу, вариация, Box-plot, exploratory data analysis (EDA), деректерді талдау.

Дәріс жоспары:

1. Кіріспе
 - Дескриптивті статистика ұғымы
 - Деректерді сипаттаудың маңыздылығы
2. Орталық тенденция көрсеткіштері
 - Орташа арифметикалық (Mean)
 - Медиана (Median)
 - Мода (Mode)
3. Таралу және вариация көрсеткіштері
 - Дисперсия (Variance)
 - Стандартты ауытқу (Standard deviation)
 - Размах (Range)
 - Қателік және вариация түсінігі
4. Квартильдер және квантильдер
 - Бірінші, екінші, үшінші квартильдер (Q1, Q2, Q3)
 - Интерквартильдік диапазон (IQR)
5. Шеткі мәндер (Outliers)
 - Tukey әдісі
 - Box-plot арқылы анықтау
6. Таралу формасын сипаттау
 - Скейнес (Asymmetry)
 - Куртозис (Kurtosis)
7. Нәтижелерді интерпретациялау және қолдану салалары
8. Қорытынды

1. Кіріспе

Қазіргі ақпараттық қоғамда деректерді жинау, өңдеу және талдау маңызды орын алады. Деректердің бастапқы сипаттамалары оларды әрі қарай

талдаудың, болжам жасаудың немесе модель құрудың негізі болып табылады. Деректерді сипаттайтын статистикалық әдістер зерттеушіге деректердің негізгі ерекшеліктерін түсінуге, олардың құрылымын көруге, ұқсастықтар мен айырмашылықтарды анықтауға мүмкіндік береді.

Бұл әдістер Exploratory Data Analysis (EDA) кезеңінің маңызды компоненті болып табылады. Есептелген статистикалық көрсеткіштер деректердің орташа шамасын, таралу деңгейін, өзгергіштігін, симметриясын және ерекшелігін анықтауға көмектеседі.

2. Дескриптивті статистика дегеніміз не?

Дескриптивті (сипаттамалық) статистика — деректер жиынының негізгі параметрлерін есептеп, оның қасиеттерін қысқаша сипаттайтын әдістер жиыны.

Негізгі міндеттері:

- Деректерді жүйелеу және түсіндіру
- Жалпы тенденцияларды анықтау
- Таралу сипаттарын зерттеу
- Әрі қарайғы талдау үшін база құру

3. Деректердің негізгі сипаттамалары

3.1. Орталық тенденция көрсеткіштері

Орташа арифметикалық (Mean)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Деректердің орташа мәнін көрсетеді.

Мысал:

Біздің мәндер: 5, 7, 8, 10

Орташа = $(5+7+8+10) / 4 = 7.5$

Медиана (Median)

Деректерді өсу ретімен жазғанда ортасында тұрған мән. Шектен шығатын мәндерге төзімді.

Мысал:

1, 2, 100 → медиана = 2

Мода (Mode)

Деректерде ең жиі кездесетін мән.

3.2. Дисперсия және өзгергіштік

Дисперсия (Variance)

$$\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Стандартты ауытқу (Standard Deviation)

$$\sigma = \sqrt{\sigma^2}$$

Өлшем бірлігі деректермен бірдей.

Размах (Range)

$$Range = Max - Min$$

3.3. Жиілік және таралу

- Гистограмма
- Жиілік кестесі
- Пайыздық үлестер
- Квантильдер, квартильдер

Квартильдер

Массивті 4 бөлікке бөледі.

- Q2 — медиана
- $IQR = Q3 - Q1$ (интерквартильдік диапазон)

3.4. Шеткі мәндер (Outliers)

Аномалды мәндер статистикалық талдауды бұрмалайды.

Tukey әдісі:

$$\text{Outlier if } x < Q1 - 1.5 IQR \text{ or } x > Q3 + 1.5 IQR$$

3.5. Таралу формасы

- Скъюнес (Skewness) — асимметрия
- Куртозис (Kurtosis) — шыңдық деңгейі

4. Деректерді сипаттау құралдары

Әдіс	Мақсаты
Орташа, медиана, мода	Орталық тенденцияны анықтау
Дисперсия, SD, IQR	Таралу мен өзгергіштікті анықтау
Гистограмма, BoxPlot	Визуалды талдау
Skewness, Kurtosis	Таралу формасын бағалау

5. Практикалық мысал (Python код идеясы)

```
python

import pandas as pd

df = pd.read_csv('data.csv')
print(df.describe())
```

6. Қолдану салалары

- Экономика және қаржы
- Машиналық оқытуда деректерді алдын ала талдау
- Сапаны бақылау
- Медициналық зерттеулер
- Әлеуметтік ғылымдар

7. Қорытынды

Деректерді сипаттайтын статистикалық әдістер — аналитикалық ойлаудың және ғылыми-зерттеу жұмысының негізі. Бұл әдістер деректердің құрылымын түсінуге, негізгі тенденцияларды көруге және шынайы шешім қабылдауға мүмкіндік береді.

Негізгі ойлар:

- Орташа, медиана, мода — орталық тенденция көрсеткіштері
- Дисперсия, стандартты ауытқу — өзгергіштік көрсеткіштері
- Квантильдер мен IQR — шеткі мәндерді табуға көмектеседі
- Визуализация — нәтижені түсіндіруді жеңілдетеді

Деректерді дұрыс сипаттау — сапалы модельдің және дұрыс шешімнің негізі.

Бақылау сұрақтары

1. Дескриптивті статистика дегеніміз не және оның негізгі міндеттері қандай?
2. Орталық тенденция көрсеткіштерінің түрлерін атаңыз және олардың айырмашылығын түсіндіріңіз.
3. Орташа арифметикалық, медиана және моданы қандай жағдайда қолданған дұрыс?
4. Дисперсия және стандартты ауытқу қандай мақсатта қолданылады?
5. Размах (Range) дегеніміз не және ол қандай ақпарат береді?
6. Квартильдер дегеніміз не? Q1, Q2, Q3 анықтамаларын келтіріңіз.
7. Интерквартильдік диапазон (IQR) қалай есептеледі және оның қолданылуы қандай?
8. Шеткі мәндер (Outliers) дегеніміз не және оларды анықтау әдістерін атаңыз.
9. Tukey әдісін қолдану ерекшеліктерін түсіндіріңіз.
10. Box-plot диаграммасы қандай ақпарат береді?
11. Скьюнес (асимметрия) дегеніміз не және оның статистикалық мәнін түсіндіріңіз.
12. Куртозис (шыңдылық деңгейі) қандай жағдайды сипаттайды?
13. Деректердің таралуын сипаттау кезінде қандай көрсеткіштер жиі қолданылады?
14. Бұл әдістер деректер талдау мен машиналық оқытуда қандай рөл атқарады?

Пайдаланылған әдебиеттер

1. Montgomery, D.C., & Runger, G.C. *Applied Statistics and Probability for Engineers*. — Wiley, 2021.
2. James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning*. — Springer, 2023.
3. Mann, P.S. *Introductory Statistics*. — Wiley, 2018.
4. Alan Agresti, Christine Franklin. *Statistics: The Art and Science of Learning from Data*. — Pearson, 2020.
5. ISO/IEC Guide 98–3:2008. *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)*.

6. Т. Хайруллин, А. Қадырбеков. *Статистикалық әдістер және деректер талдау*. — Алматы, 2020.
7. Қазақстан Республикасының ЖОО-ға арналған статистика оқулықтары мен оқу құралдары.
8. Kaggle, StatQuest, and DataCamp — деректер талдау бойынша онлайн ресурстар.