Statistical Analysis of Drug Prescription Patterns

Higher National Diploma in Software Engineering – 24.2F

Statistics Project Report

Submitted By :

P.A.A.I. Ponweera	KAHDSE242F- 002
N.M.O.M.B. Neinayake	KAHDSE242F-005
S.M.A.K.Semasinghe	KAHDSE242F-056

Date of submission: 02nd May 2025

School of Computing and Engineering

National Institute of Business Management Kandy

Contents

1. Introduction

This report presents an exploratory and inferential statistical analysis of a clinical dataset containing information about patients and the medications prescribed to them. Using RStudio we evaluate the influence of biological and demographic factors such as Age, Blood pressure, Cholesterol, Sodium and Potassium levels on the type of drug prescribed.

1.1 Objective of the study

The primary objective of this study is to examine the relationship between patient and characteristics and the type of drug prescribed. Specially, the study aims to:

- Analyze how demographic and biological factors – such as Age, Sex, Blood Pressure (BP), Cholesterol level, Sodium (Na), Potassium (K) influence the choice of medication.
- Identify which variables have the most significant impact on drug assignment.
- Use statistical techniques to detect meaningful patterns or associations within the dataset.
- Provide data-driven insights that could support clinical decision-making in drug prescription practices.

1.2 Description of the dataset

The data set consist of medical records for a group of patients, which each record containing both demographic and clinical variables. It includes 7 attributes for each patient.

1. Age – Numeric: The age of the patient years.
2. Sex – Categorical: The gender of the patient.
3. Blood pressure – Categorical: The patient's blood pressure level, recorded as low, high or normal.
4. Cholesterol – Categorical: Indicates whether the patient's cholesterol level is normal or high.
5. Sodium (Na) – Numeric: The sodium concentration level in the patient's blood.
6. Potassium (K) – Numeric: The potassium concentration level in the patient's blood.
7. Drug – Categorical: The type of drug prescribed to the patient.

The dataset is used to analyze whether these attributes influence the type of drug that is ultimately prescribed.

2. Methodology

A sample of 200 patient records was chosen a maintain a statistical validity while reducing computational load. This approach helps preserve the data's structure and supports reliable inferential analysis.

2.1 Data Collection Process

Data was collected through routine medical assessments, with demographic and clinical attributes recorded during check-ups and lab tests. The dataset was the organized into 7 structured variables – Numerical and Categorical.

2.2 Statistical Analysis Techniques

2.1.1 Descriptive Statistics

- Summarize key variables such as Age, Na and K.
- Frequency tables for categorical variables like Sex, BP, Cholesterol and Drug.

2.1.2 Correlation Analysis

- Measure the strength and direction of relationships between Age, Na and K using Pearson's correlation.

2.1.3 Multiple Regression Analysis

- Identifies how predictors like Age, BP, Cholesterol, Na, K influence the choice of drug using multinomial and logistic regression.

```
> mean(Drug$Na)
[1] 0.6970952
> mean(Drug$K)
[1] 0.0501739
> mean(Drug$Age)
[1] 44.315
```

Figure 2.1

The average sodium (Na) level in the drug dataset is approximately 0.70, while the average potassium (K) is around 0.50. The mean age of individuals in this dataset is approximately 44.3 years.

```
> median(Drug$Na)
[1] 0.721853
> median(Drug$K)
[1] 0.049663
> median(Drug$Age)
[1] 45
```

Figure 2.2

The median sodium (Na) level in the drug dataset is approximately 0.722, and the median potassium (K) level is around 0.05. The median age of individuals in this dataset is 45 years.

```
> var(Drug$Na)
[1] 0.01413895
> var(Drug$K)
[1] 0.0003101644
> var(Drug$Age)
[1] 273.7143
```

Figure2.3

The variance in sodium (Na) levels within the drug dataset is approximately 0.014, and the variance in potassium (K) level is considerably lower, at around 0.0003. The age of individuals in the dataset shows a larger variance of approximately 273.71.

```
> sd(Drug$Na)
[1] 0.1189073
> sd(Drug$K)
[1] 0.01761148
> sd(Drug$Age)
[1] 16.54431
```

Figure 2.4

The standard deviation of sodium (Na) level in the drug dataset is approximately 0.11. The standard deviation of potassium (K) levels is notably smaller, at around 0.18. The ages of individuals in the dataset have a standard deviation of approximately 16.5 years.

```
> summary(Drug)
      Age            Sex                BP             Cholesterol
 Min.   :15.00   Length:200         Length:200         Length:200
 1st Qu.:31.00   Class :character   Class :character   Class :character
 Median :45.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :44.31
 3rd Qu.:58.00
 Max.   :74.00
       Na               K               Drug
 Min.   :0.5002   Min.   :0.02002   Length:200
 1st Qu.:0.5839   1st Qu.:0.03505   Class :character
 Median :0.7219   Median :0.04966   Mode  :character
 Mean   :0.6971   Mean   :0.05017
 3rd Qu.:0.8015   3rd Qu.:0.06600
 Max.   :0.8961   Max.   :0.07979
```

Figure 2.5

The dataset contains 200 observations for each of the variables: Age, Sex, BP, Cholesterol, Na, K and Drug. Age is numerical variable with a range from 15 to 74 years, a median of 45 years, and a mean of a 44.31 years. Sodium levels range from 0.5002 to 0.8961, with a median of 0.7219 and a mean of 0.6971. potassium levels range from 0.02002 to 0.07979, with a median of 0.04966 and a mean of 0.05017. Sex, BP, Cholesterol, and Drug are categorical variables, each represented as character data.

3. Results and Discussion

3.1 Potassium Level (K)

This histogram shows the distribution of Potassium levels (ranging from – 0.02 to 0.08) in the Drug dataset. The irregular, multimodal shape suggests potential subgroups with varying potassium concentrations, which could be explored in relation to other variables like drug type or health indicators. Further analysis of the distribution's characteristics and comparisons to healthy ranges might reveal clinically relevant insights.

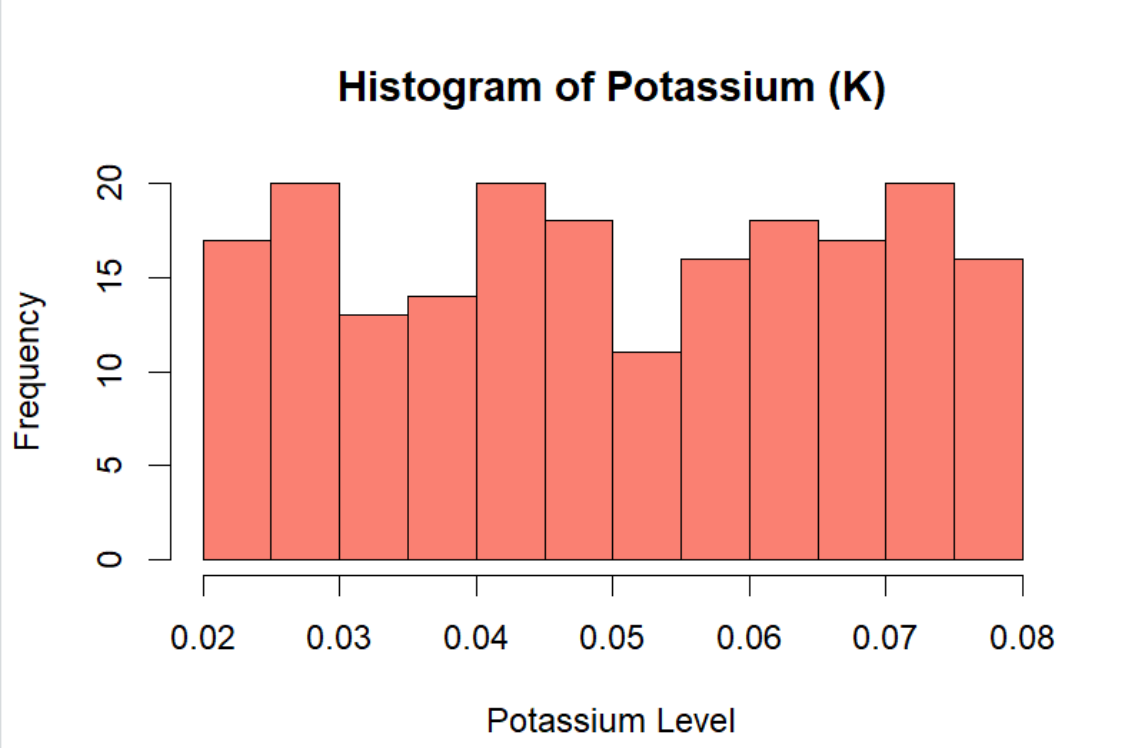

Figure 3.1

3.2 Sodium level (Na)

Scatter plots displays distribution of sodium levels (Na) across the observations in the Drug dataset. Each circle represents a single observation, with its vertical position indicating the sodium level and its horizontal position representing the index or order of that observation in the dataset. The plot shows the range of sodium values and how they are spread across the dataset, without revealing any obvious trends or patterns related to the observation order.

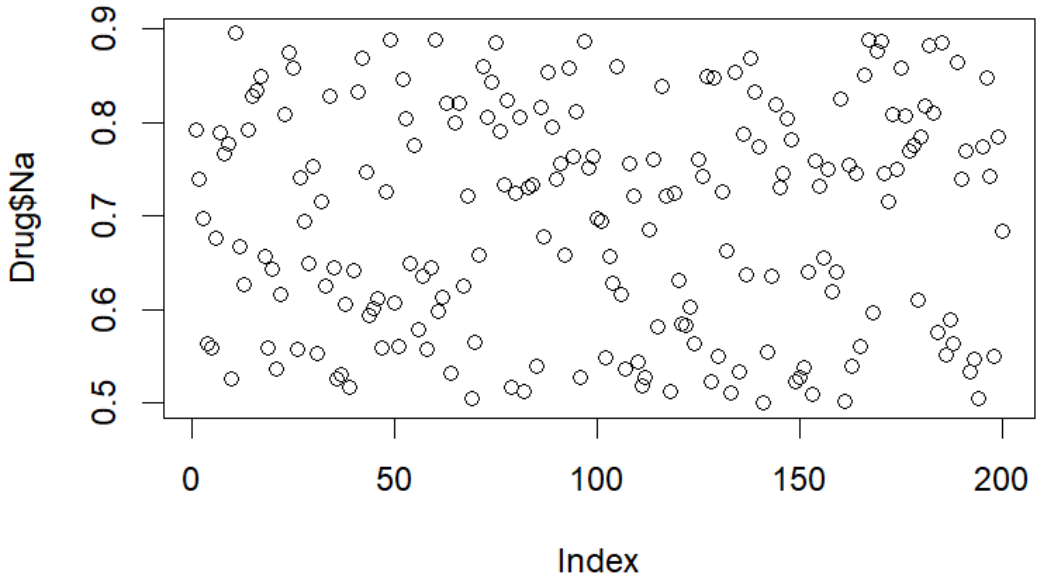

Figure 3.2

The histogram illustrates the distribution of sodium (Na) levels within the Drug dataset. The X-axis represents the range of sodium levels, spanning from approximately 0.5 to 0.9. The Y-axis indicates the frequency, or the number of observations falling into each sodium level bin. The distribution appears somewhat uneven, with the highest frequency of sodium levels observed in the range of 0.5 to 0.55. This visualization provides an overview of how sodium levels are distributed among the individuals in the dataset.

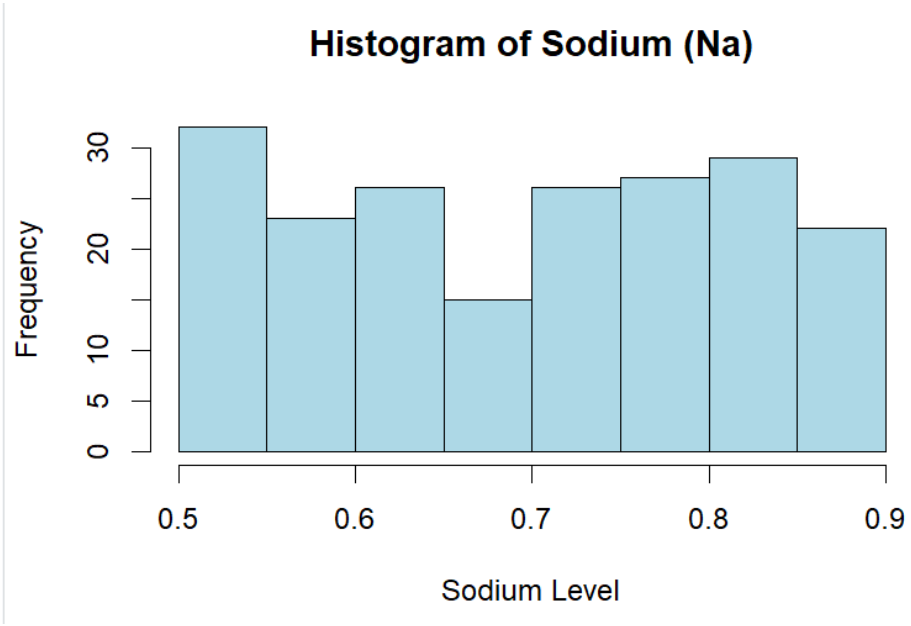

Figure 3.3

3.3 Age Distribution

Most patients, with a median age of 45.

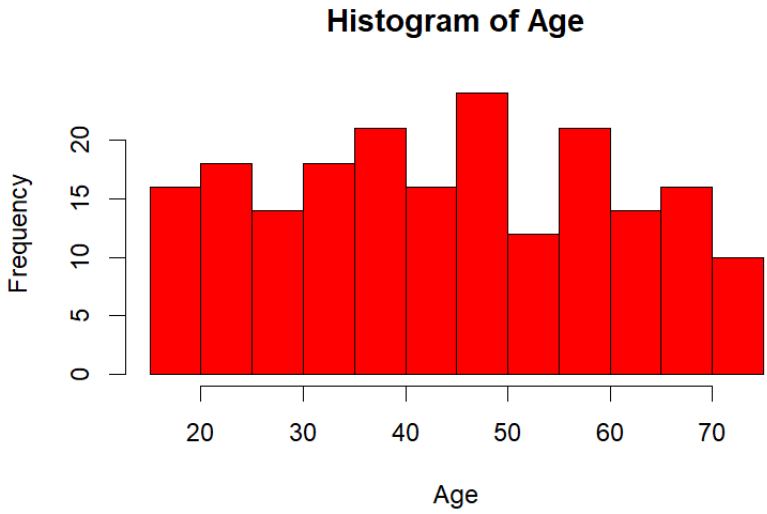

Figure 3.4

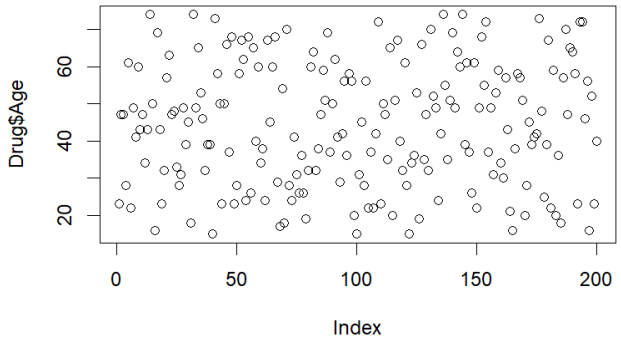

Figure 3.5

This histogram illustrates the age distribution within the Drug dataset. The X-axis represents the age of the individuals ranging from approximately 15 to 75 years. The Y-axis indicates the frequency or number of individuals, within each age group. The distribution shows that individuals are spread across various age ranges, with notable concentrations around the 40-50 and 50-55 age groups, suggesting a relatively diverse age profile within the study population.

4. Conclusion and Recommendations

4.1 Summary of the results

- Monitor Sodium and Potassium levels during patient assessment, as they appear to influence prescription patterns.
- BP and Cholesterol should be considered when determining medication plans.
- Future analysis can include patient outcomes and longer-term health indicators.

4.2 Recommendation

- Improve clinical decision – making based on Sodium and Potassium levels
  Sodium and Potassium levels vary significantly across drug types. Clinicians should prioritize blood tests and use defined thresholds to guide more accurate and personalized drug prescriptions.
- Consider peer group treatment patterns
  If patients with similar profiles consistently receive the same drug, it may be useful to study these "peer clusters" to develop standardized yet flexible prescribing frameworks.
- Train healthcare staff on pattern-based prescribing
  Encourage training session for doctors and pharmacists on how different variables like Sodium, Potassium, and BP levels influence medication decisions-using evidence-based findings from this analysis.

5. Reference

Google Form

https://docs.google.com/forms/d/e/1FAIpQLSeVqYQaSHlPHlhVwMe9rJ079Gznb7xGaTZtGrVlirV0nDN_TA/viewform?usp=header

6. Appendices

6.1 Data Structures

Columns including participant age, sex, BP, cholesterol, Na, and K are included in the dataset used in this study. A sample of the raw data is displayed in the table below

	A	B	C	D	E	F	G
1	Age	Sex	BP	Cholesterc	Na	K	Drug
2	23	F	HIGH	HIGH	0.792535	0.031258	drugY
3	47	M	LOW	HIGH	0.739309	0.056468	drugC
4	47	M	LOW	HIGH	0.697269	0.068944	drugC
5	28	F	NORMAL	HIGH	0.563682	0.072289	drugX
6	61	F	LOW	HIGH	0.559294	0.030998	drugY
7	22	F	NORMAL	HIGH	0.676901	0.078647	drugX
8	49	F	NORMAL	HIGH	0.789637	0.048518	drugY
9	41	M	LOW	HIGH	0.766635	0.069461	drugC
10	60	M	NORMAL	HIGH	0.777205	0.05123	drugY
11	43	M	LOW	NORMAL	0.526102	0.027164	drugY
12	47	F	LOW	HIGH	0.896056	0.076147	drugC
13	34	F	HIGH	NORMAL	0.667775	0.034782	drugY
14	43	M	LOW	HIGH	0.626527	0.040746	drugY
15	74	F	LOW	HIGH	0.792674	0.037851	drugY
16	50	F	NORMAL	HIGH	0.82778	0.065166	drugX
17	16	F	HIGH	NORMAL	0.833837	0.053742	drugY
18	69	M	LOW	NORMAL	0.848948	0.074111	drugX
19	43	M	HIGH	HIGH	0.656371	0.046979	drugA
20	23	M	LOW	HIGH	0.55906	0.076609	drugC
21	32	F	HIGH	NORMAL	0.643455	0.024773	drugY
22	57	M	LOW	NORMAL	0.536746	0.028061	drugY
23	63	M	NORMAL	HIGH	0.616117	0.023773	drugY
24	47	M	LOW	NORMAL	0.809199	0.026472	drugY
25	48	F	LOW	HIGH	0.87444	0.058155	drugY
26	33	F	LOW	HIGH	0.858387	0.025634	drugY

Figure 6.1

6.2 R Studio code for data analysis

Descriptive statistics, histograms, and other statistical analysis of the dataset were produced using the following R Studio code.

```
setwd("search-ms:displayname=検索場所%3A%20Home&crumb=location:%3A%3A{F874310E
setwd
data=read.csv("Drug.csv")
data
mean(Drug$Na)
mean(Drug$K)
mean(Drug$Age)
median(Drug$Na)
median(Drug$K)
median(Drug$Age)
var(Drug$Na)
var(Drug$K)
var(Drug$Age)
sd(Drug$Na)
sd(Drug$K)
sd(Drug$Age)
summary(Drug)
hist(Drug$Na,
    main = "Histogram of Sodium (Na)",
    xlab = "Sodium Level",
    col = "lightblue",
    border = "black")
hist(Drug$K,
     main = "Histogram of Potassium (K)",
     xlab = "Potassium Level",
     col = "salmon",
     border = "black")
plot(Drug$K)
plot(Drug$Na)
```

Figure 6.2