Week5-Lab-Assignment

Chronic Kidney Disease Data Analysis

Kasra Vand

2024-09-27

Data Loading and Preparation

First, let's load the dataset and prepare it for analysis.

```
# Load required libraries
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Read the dataset
ckd_data <- read.csv("../Chronic_Kidney_Dsease_data.csv")

# Display the first few rows and summary
# head(ckd_data)
# summary(ckd_data)
```

Binomial Distribution

Let's look at the prevalence of certain risk factors in our dataset.

1. Family History of Kidney Disease

```
prob_family_history <- mean(ckd_data$FamilyHistoryKidneyDisease)

# Probability of exactly 10 out of 50 patients having family history
dbinom(10, 50, prob_family_history)
```

```
## [1] 0.07313085
```

```
# Probability of at most 10 out of 50 patients having family history
pbinom(10, 50, prob_family_history)
```

```
## [1] 0.9140657
```

```
# Probability of more than 10 out of 50 patients having family history
pbinom(10, 50, prob_family_history, lower.tail = FALSE)
```

```
## [1] 0.08593431
```

Questions: a) What is the probability that exactly 10 out of 50 patients have a family history of kidney disease? b) What is the probability that at most 10 out of 50 patients have a family history of kidney disease? c) What is the probability that more than 10 out of 50 patients have a family history of kidney disease?

Answers: a) The probability that exactly 10 out of 50 patients have a family history of kidney disease is approximately 0.0731 or 7.31%. b) The probability that at most 10 out of 50 patients have a family history of kidney disease is approximately 0.9141 or 91.41%. c) The probability that more than 10 out of 50 patients have a family history of kidney disease is approximately 0.0859 or 8.59%.

2. Simulation: Smoking Status

Let's simulate sampling 100 patients and count how many are smokers.

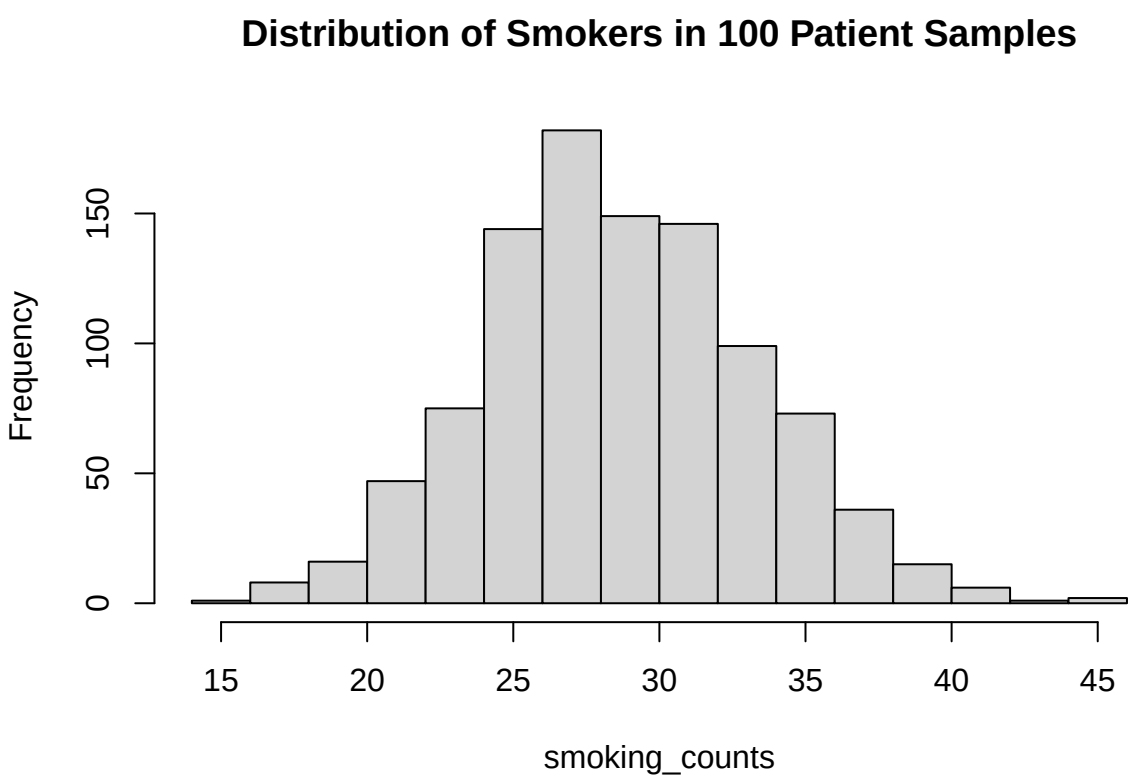
```
prob_smoking <- mean(ckd_data$Smoking)
num_patients <- 100
num_replicates <- 1000

set.seed(2023)

simulate_smoking <- function() {
  sum(sample(c(0,1), size = num_patients,
             prob = c(1 - prob_smoking, prob_smoking),
             replace = TRUE))
}

smoking_counts <- replicate(num_replicates, simulate_smoking())

hist(smoking_counts, breaks = 20, main = "Distribution of Smokers in 100 Patient Samples")
```



Question: d) Based on the simulation results, describe the distribution of smokers in 100 patient samples.

Answer: d) The histogram shows the distribution of smokers in 100 patient samples over 1000 simulations. The distribution appears to be roughly symmetric and bell-shaped, centered around the mean number of smokers. This suggests that the number of smokers in a sample of 100 patients follows an approximately normal distribution.

Normal Distribution

3. Analysis of GFR (Glomerular Filtration Rate)

```
mean_gfr <- mean(ckd_data$GFR)
sd_gfr <- sd(ckd_data$GFR)

# Probability of a patient having GFR less than 60 (indicating kidney disease)
pnorm(60, mean_gfr, sd_gfr)
```

```
## [1] 0.4101127
```

```
# 95th percentile of GFR in the population
qnorm(0.95, mean_gfr, sd_gfr)
```

```
## [1] 116.2584
```

Questions: e) What is the probability of a patient having a GFR less than 60 (indicating kidney

disease)? f) What is the 95th percentile of GFR in the population?

Answers: e) The probability of a patient having a GFR less than 60 (indicating kidney disease) is approximately 0.4101 or 41.01%. f) The 95th percentile of GFR in the population is approximately 116.26 mL/min/1.73m².

4. Blood Pressure Analysis

```
# Assuming hypertension is defined as SystolicBP >= 140 or DiastolicBP >= 90
hypertension <- (ckd_data$SystolicBP >= 140) | (ckd_data$DiastolicBP >= 90)
prob_hypertension <- mean(hypertension)

# Probability of at least 40 out of 100 patients having hypertension
pbinom(39, 100, prob_hypertension, lower.tail = FALSE)
```

```
## [1] 1
```

Question: g) What is the probability of at least 40 out of 100 patients having hypertension?

Answer: g) The probability of at least 40 out of 100 patients having hypertension is 1, or 100%. This suggests that hypertension is very common in this patient population, with the prevalence being much higher than 40%.

Poisson Distribution

5. Urinary Tract Infections

Assuming UTIs are rare events, we can model them using a Poisson distribution.

```
mean_uti <- mean(ckd_data$UrinaryTractInfections)

# Probability of a patient having more than 2 UTIs in a year
ppois(2, mean_uti, lower.tail = FALSE)
```

```
## [1] 0.001326262
```

```
# Expected number of UTIs in a group of 100 patients
lambda_100 <- mean_uti * 100
lambda_100
```

```
## [1] 21.03677
```

```
# Probability of more than 50 UTIs in a group of 100 patients
ppois(50, lambda_100, lower.tail = FALSE)
```

```
## [1] 2.326752e-08
```

Questions: h) What is the probability of a patient having more than 2 UTIs in a year? i) What is the expected number of UTIs in a group of 100 patients? j) What is the probability of more than 50 UTIs occurring in a group of 100 patients?

Answers: h) The probability of a patient having more than 2 UTIs in a year is approximately 0.00133 or 0.133%. i) The expected number of UTIs in a group of 100 patients is approximately 21.04. j) The probability of more than 50 UTIs occurring in a group of 100 patients is extremely low, approximately 2.33e-08 or 0.00000233%. “‘