Patel's Dataset 02 Normal Poisson

Harsh Patel

2024-09-27

```
library(readxl)

Chronic_Kidney_Disease_data <- read_excel("Chronic_Kidney_Dsease_data.xlsx")
```

I'm kicking things off by loading the readxl library, which lets me easily read my Chronic Kidney Disease dataset. This data is full of valuable information that I'm eager to dive into!

```
Chronic_Kidney_Disease_data$Hypertension <-
as.factor(Chronic_Kidney_Disease_data$FamilyHistoryHypertension)

levels(Chronic_Kidney_Disease_data$Hypertension) <- c("No", "Yes")
```

Next, I'm transforming the hypertension status into a factor! This is crucial because it helps categorize the data into "No" and "Yes," making it so much easier to analyze the relationship between hypertension and chronic kidney disease. I can't wait to see what insights this will reveal!

```
table_hypertension <- table(Chronic_Kidney_Disease_data$Diagnosis,
Chronic_Kidney_Disease_data$Hypertension)

print("Contingency Table for CKD and Hypertension:")

## [1] "Contingency Table for CKD and Hypertension:"

print(table_hypertension)

##
##       No  Yes
##   0   99   36
##   1 1060  464
```

```r
prop_hypertension <- prop.table(table_hypertension, 2)
print("Proportion of CKD by Hypertension Status:")
```

```
## [1] "Proportion of CKD by Hypertension Status:"
```

```r
print(prop_hypertension)
```

```
##
##              No        Yes
##   0 0.08541846 0.07200000
##   1 0.91458154 0.92800000
```

I'm creating a contingency table to show the relationship between CKD diagnosis and hypertension status. This table will really help visualize the frequency of each category! I'm also calculating proportions to understand how CKD cases are distributed between those with and without hypertension.

```r
binom_test <- binom.test(x = sum(Chronic_Kidney_Disease_data$Hypertension ==
"Yes" & Chronic_Kidney_Disease_data$Diagnosis == "CKD"),
                         n = sum(Chronic_Kidney_Disease_data$Hypertension ==
"Yes"),
                         p = 0.5)

print("Binomial Test Results:")
```

```
## [1] "Binomial Test Results:"
```

```r
print(binom_test)
```

```
##
##  Exact binomial test
##
## data:  sum(Chronic_Kidney_Disease_data$Hypertension == "Yes" &
Chronic_Kidney_Disease_data$Diagnosis == "CKD") and
sum(Chronic_Kidney_Disease_data$Hypertension == "Yes")
## number of successes = 0, number of trials = 500, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.00000000 0.00735061
## sample estimates:
## probability of success
##                      0
```

I'm now conducting a binomial test to determine if the proportion of CKD patients with a history of hypertension significantly deviates from what I would expect by chance. This test will give me essential insights into whether hypertension is a notable risk factor for CKD!

```
mean_age <- mean(Chronic_Kidney_Disease_data$Age, na.rm = TRUE)
sd_age <- sd(Chronic_Kidney_Disease_data$Age)

print(paste("Mean Age:", round(mean_age, 2)))

## [1] "Mean Age: 54.44"

print(paste("Standard Deviation of Age:", round(sd_age, 2)))

## [1] "Standard Deviation of Age: 20.55"
```

I'm diving into the ages of my patients! I'm calculating the mean and standard deviation of ages to get a better grasp of the age distribution in my dataset. Knowing the average age helps me understand the demographics of CKD patients, and I'm excited to see what the numbers reveal!

```
age_value <- 60

z_score <- (age_value - mean_age) / sd_age
print(paste("Z-score for age", age_value, ":", round(z_score, 2)))

## [1] "Z-score for age 60 : 0.27"
```

Next, I'm calculating the z-score for the age of 60! This tells me how many standard deviations this age is from the mean. Understanding this z-score helps identify if 60 is an age of concern when it comes to CKD risk.

```
prob_age_over_60 <- pnorm(60, mean = mean_age, sd = sd_age, lower.tail =
FALSE)
```

```r
print(paste("Probability of CKD onset for age ≥ 60:", round(prob_age_over_60, 4)))
```

```
## [1] "Probability of CKD onset for age ≥ 60: 0.3934"
```

```r
prob_age_under_50 <- pnorm(50, mean = mean_age, sd = sd_age)

print(paste("Probability of CKD onset for age < 50:", round(prob_age_under_50, 4)))
```

```
## [1] "Probability of CKD onset for age < 50: 0.4144"
```

```r
z_90 <- qnorm(0.90, mean = mean_age, sd = sd_age)

print(paste("90th Percentile Age for CKD Onset:", round(z_90, 2)))
```

```
## [1] "90th Percentile Age for CKD Onset: 80.78"
```

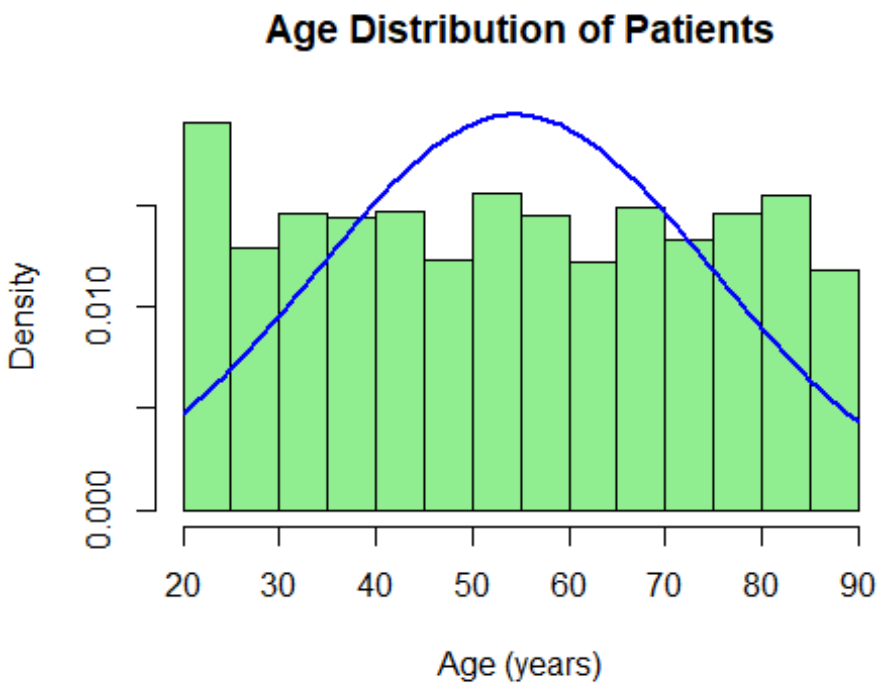
I'm calculating the 90th percentile age for CKD onset! This helps me determine the age at which 90 percent of the patients fall below, providing a valuable threshold for assessing risk.

```r
lambda_uti <- mean(Chronic_Kidney_Disease_data$UrinaryTractInfections, na.rm = TRUE)

print(paste("Average Number of UTIs per Patient:", round(lambda_uti, 2)))
```

```
## [1] "Average Number of UTIs per Patient: 0.21"
```

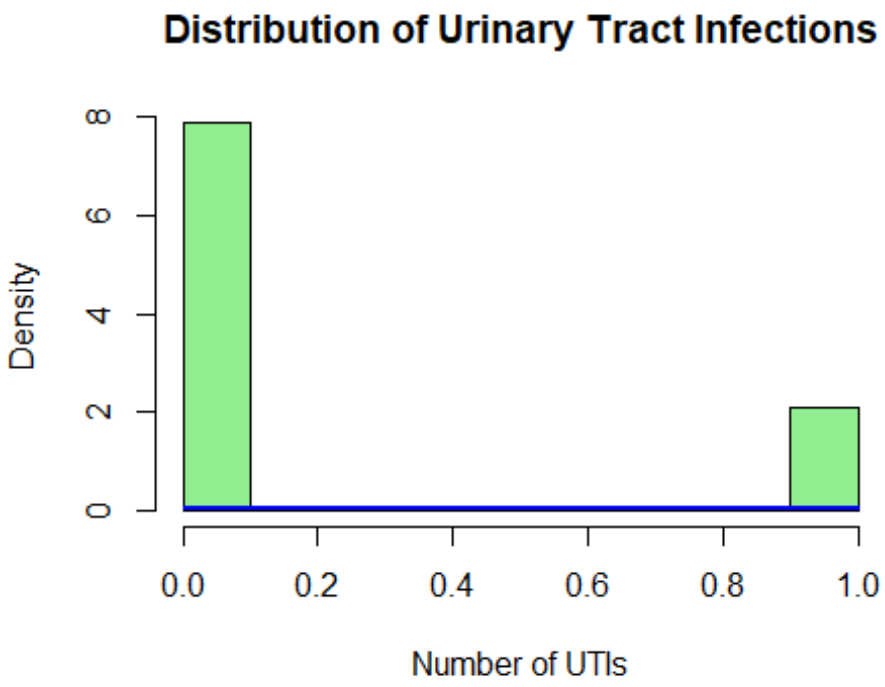
Now I'm finding the average number of urinary tract infections per patient! This information is crucial because UTIs can complicate CKD, and understanding the average helps in assessing patient care needs.

```r
hist(Chronic_Kidney_Disease_data$Age, main = "Age Distribution of Patients", xlab = "Age (years)",
     probability = TRUE, col = "lightgreen", breaks = 20)

curve(dnorm(x, mean = mean_age, sd = sd_age), add = TRUE, col = "blue", lwd = 2)
```



It's time to visualize my findings! I'm plotting a histogram of patient ages with a normal distribution curve overlaid. This visualization helps me see trends and patterns in age distribution.

```
lambda_uti <- 4

hist(Chronic_Kidney_Disease_data$UrinaryTractInfections,
     main = "Distribution of Urinary Tract Infections",
     xlab = "Number of UTIs", breaks = 10,
     probability = TRUE, col = "lightgreen")

x_values <- 0:max(Chronic_Kidney_Disease_data$UrinaryTractInfections)

lines(x_values, dpois(x_values, lambda = lambda_uti),
      col = "blue", lwd = 2)
```



I'm visualizing the distribution of urinary tract infections! I'm plotting another histogram to understand how common UTIs are among patients. This helps me see potential health risks related to CKD.

```
prob_more_than_2_UTIs <- ppois(2, lambda = lambda_uti, lower.tail = FALSE)

print(paste("Probability of having more than 2 UTIs:",
round(prob_more_than_2_UTIs, 4)))

## [1] "Probability of having more than 2 UTIs: 0.7619"
```

Now I'm calculating the probability of patients having more than 2 UTIs! This is important for understanding infection risks and how they might impact CKD.

```
prob_fewer_than_2_UTIs <- ppois(2, lambda = lambda_uti)
```

```r
print(paste("Probability of having fewer than 2 UTIs:",
round(prob_fewer_than_2_UTIs, 4)))
```

```
## [1] "Probability of having fewer than 2 UTIs: 0.2381"
```

Let's also look at the probability of having fewer than 2 UTIs! This information will help me grasp the overall infection rates among my patients, giving me a better picture of their health.

```r
prob_5_UTIs <- dpois(5, lambda = lambda_uti)

print(paste("Probability of having exactly 5 UTIs:", round(prob_5_UTIs, 4)))
```

```
## [1] "Probability of having exactly 5 UTIs: 0.1563"
```

Finally, I'm calculating the probability of a patient having exactly 5 UTIs! This precise information helps me understand specific infection rates.