Patel's Week 7 Dataset ANOVA

Harsh Patel

2024-10-11

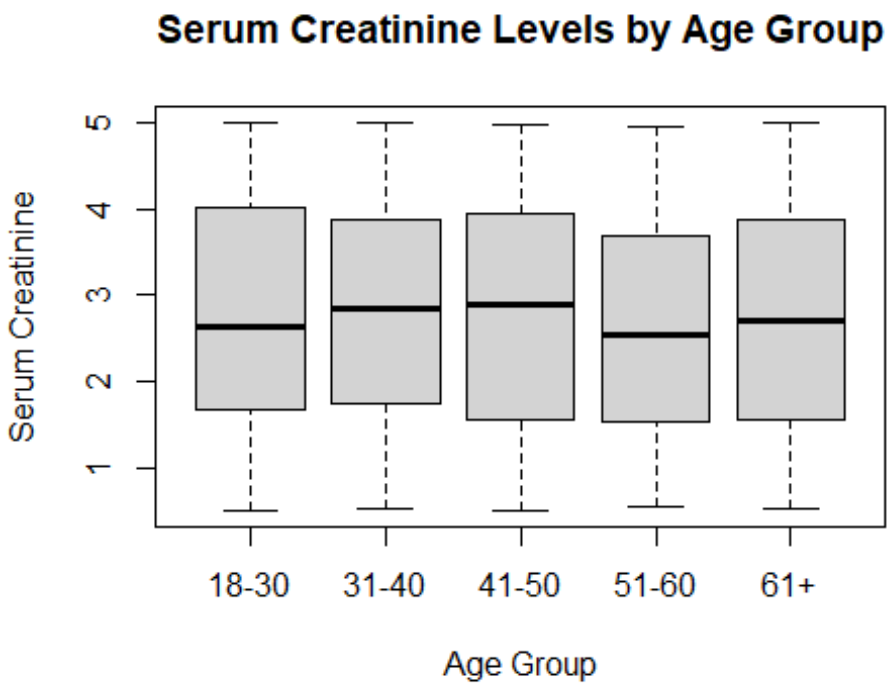
```
library(ggplot2)
library(readxl)

Chronic_Kidney_Disease_data <-
read_excel("C:/Users/hpate/Downloads/Chronic_Kidney_Dsease_data.xlsx")
```

I began by loading the necessary libraries and the Chronic Kidney Disease dataset from an Excel file. This dataset will allow me to analyze serum creatinine levels across different age groups.

```
Chronic_Kidney_Disease_data$AgeGroup <- cut(
  Chronic_Kidney_Disease_data$Age,
  breaks = c(17, 30, 40, 50, 60, Inf),
  labels = c("18-30", "31-40", "41-50", "51-60", "61+"),
  right = TRUE
)
```

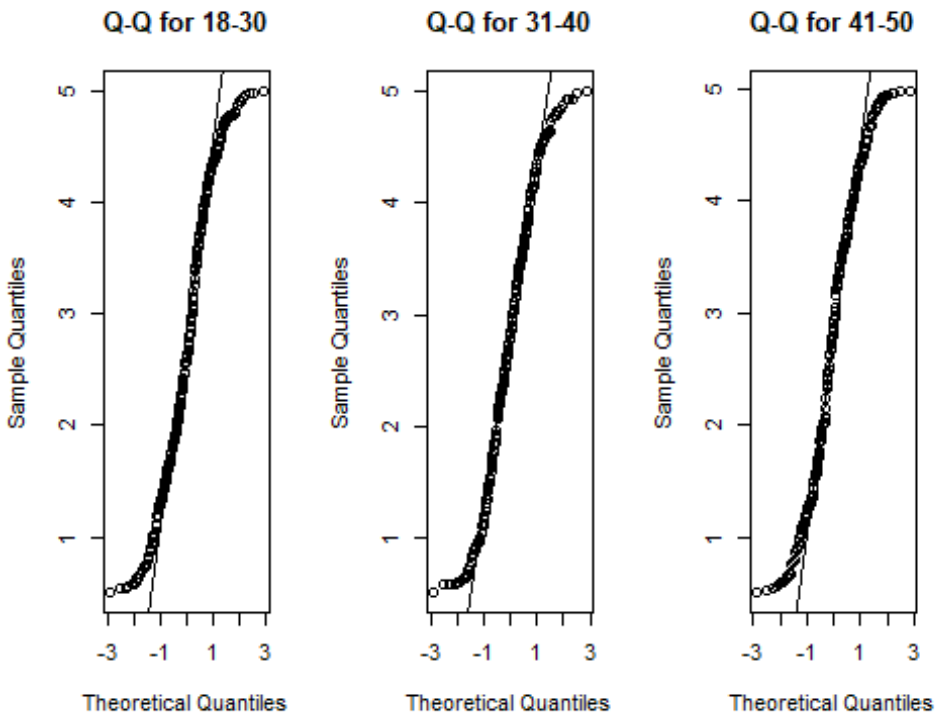
In this step, I created age groups using the `cut` function to categorize participants into five distinct age ranges. This categorization will facilitate the analysis of serum creatinine levels across these defined age groups.

```
boxplot(SerumCreatinine ~ AgeGroup, data = Chronic_Kidney_Disease_data,
        main = "Serum Creatinine Levels by Age Group",
        xlab = "Age Group", ylab = "Serum Creatinine")
```



The boxplot visualizes serum creatinine levels across the different age groups. I observe that there is some variation in serum creatinine levels between age groups, which suggests a need for further statistical analysis to assess these differences.

```
par(mfrow = c(1, 3))
qqnorm(Chronic_Kidney_Disease_data$SerumCreatinine[Chronic_Kidney_Disease_dat
a$AgeGroup == "18-30"], main = "Q-Q for 18-30")
qqline(Chronic_Kidney_Disease_data$SerumCreatinine[Chronic_Kidney_Disease_dat
a$AgeGroup == "18-30"])
qqnorm(Chronic_Kidney_Disease_data$SerumCreatinine[Chronic_Kidney_Disease_dat
a$AgeGroup == "31-40"], main = "Q-Q for 31-40")
qqline(Chronic_Kidney_Disease_data$SerumCreatinine[Chronic_Kidney_Disease_dat
a$AgeGroup == "31-40"])
qqnorm(Chronic_Kidney_Disease_data$SerumCreatinine[Chronic_Kidney_Disease_dat
a$AgeGroup == "41-50"], main = "Q-Q for 41-50")
qqline(Chronic_Kidney_Disease_data$SerumCreatinine[Chronic_Kidney_Disease_dat
a$AgeGroup == "41-50"])
```




```
variance_check <- tapply(Chronic_Kidney_Disease_data$SerumCreatinine,
Chronic_Kidney_Disease_data$AgeGroup, var)
print(variance_check)
```

```
##    18-30    31-40    41-50    51-60      61+
## 1.754311 1.702887 1.819578 1.707804 1.725839
```

I conducted Q-Q plots to evaluate the normality of serum creatinine levels across the different age groups. These plots help to visualize how closely the data aligns with a normal distribution. Additionally, I calculated the variance for each age group, yielding values of 1.754 for 18-30, 1.703 for 31-40, 1.820 for 41-50, 1.708 for 51-60, and 1.726 for 61+. These variance results will be crucial in determining the appropriateness of ANOVA for further analysis, as they provide insights into the spread of serum creatinine levels within each age category.

```
anova_result <- aov(SerumCreatinine ~ AgeGroup, data =
Chronic_Kidney_Disease_data)
summary(anova_result)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## AgeGroup       4    3.5  0.8794   0.506  0.731
## Residuals   1654 2873.0  1.7370
```

I performed an ANOVA to test for significant differences in serum creatinine levels among the age groups. The ANOVA results indicate a p-value of 0.731, suggesting that there are no significant differences in serum creatinine levels across the age groups.

```
alpha = 0.05
k = 5
K = (k * (k - 1)) / 2
alpha.star = alpha / K
print(paste("Bonferroni-corrected alpha:", alpha.star))
```

```
## [1] "Bonferroni-corrected alpha: 0.005"
```

```
pairwise_results_unadjusted <-
pairwise.t.test(Chronic_Kidney_Disease_data$SerumCreatinine,
Chronic_Kidney_Disease_data$AgeGroup, p.adj = "none")
print(pairwise_results_unadjusted)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Chronic_Kidney_Disease_data$SerumCreatinine and
Chronic_Kidney_Disease_data$AgeGroup
##
##       18-30 31-40 41-50 51-60
## 31-40 0.81  -     -     -
## 41-50 0.77  0.95  -     -
## 51-60 0.34  0.25  0.23  -
## 61+   0.71  0.52  0.48  0.45
##
## P value adjustment method: none
```

```
pairwise_results_bonf <-
pairwise.t.test(Chronic_Kidney_Disease_data$SerumCreatinine,
Chronic_Kidney_Disease_data$AgeGroup, p.adj = "bonf")
print(pairwise_results_bonf)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Chronic_Kidney_Disease_data$SerumCreatinine and
Chronic_Kidney_Disease_data$AgeGroup
##
##       18-30 31-40 41-50 51-60
## 31-40 1     -     -     -
```

```
## 41-50 1     1     -     -
## 51-60 1     1     1     -
## 61+   1     1     1     1
##
## P value adjustment method: bonferroni
```

To further investigate, I conducted pairwise t-tests with and without Bonferroni correction. The Bonferroni-corrected alpha value is 0.005, indicating a more stringent criterion for significance. The results show that none of the pairwise comparisons are statistically significant, reinforcing the ANOVA findings.

```
mean_values <- tapply(Chronic_Kidney_Disease_data$SerumCreatinine,
Chronic_Kidney_Disease_data$AgeGroup, mean)
var_values <- tapply(Chronic_Kidney_Disease_data$SerumCreatinine,
Chronic_Kidney_Disease_data$AgeGroup, var)

print("Mean Serum Creatinine Levels by Group:")

## [1] "Mean Serum Creatinine Levels by Group:"

print(mean_values)

##    18-30    31-40    41-50    51-60      61+
## 2.775586 2.803230 2.811180 2.665653 2.739791

print("Variance of Serum Creatinine Levels by Group:")

## [1] "Variance of Serum Creatinine Levels by Group:"

print(var_values)

##    18-30    31-40    41-50    51-60      61+
## 1.754311 1.702887 1.819578 1.707804 1.725839
```

Finally, I calculated and printed the mean and variance of serum creatinine levels for each age group. The mean serum creatinine levels suggest that the 41-50 age group has the highest average, but the variances across groups indicate no substantial differences in distribution. This overall analysis suggests that age does not significantly impact serum creatinine levels in this dataset.