Yahriel\_Salinas-Reyes\_7340\_HW1\_FA24.rmd

Yahriel Salinas-Reyes

2024-09-05

## clean R environment

rm(list = ls(all=TRUE))  
graphics.off()  
shell("cls")

## Include Packages

#library(datasets)   
library(ggplot2)  
#library(MASS)  
#library(dplyr)

# Question 1

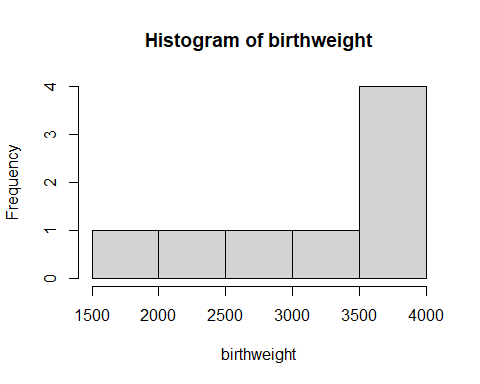
birthweight <- c(3600,1700,4000,3900,3100,3800,2200,3000)  
summary(birthweight)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1700 2800 3350 3162 3825 4000

sd(birthweight)

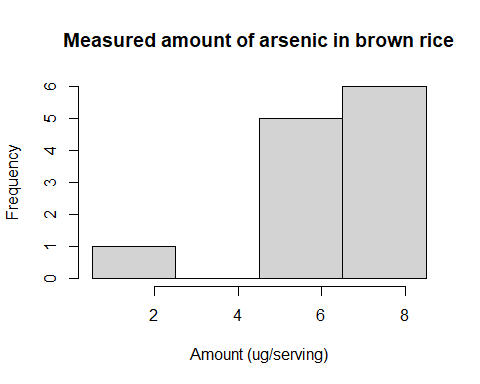
## [1] 839.9617

hist(birthweight)



# Question 2

Ar\_rice <- c(6.1, 5.4, 6.9, 4.9, 6.6, 6.3, 6.7, 8.2, 7.8, 1.5, 5.4, 7.3)  
hist(Ar\_rice,   
 main = "Measured amount of arsenic in brown rice",   
 xlab = "Amount (ug/serving)",  
 ylab = "Frequency",  
 breaks = seq(min(Ar\_rice)-1, max(Ar\_rice)+1, by=2))



# Question 3

father <- c(73.0, 75.5, 75.0, 75.0, 75.0, 74.0, 74.0, 73.0, 73.0, 78.5)  
son <- c(74.0, 73.5, 71.0, 70.5, 72.0, 76.5, 74.0, 71.0, 72.0, 73.2)  
summary(father)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 73.00 73.25 74.50 74.60 75.00 78.50

sd(father)

## [1] 1.66333

var(father)

## [1] 2.766667

str(father)

## num [1:10] 73 75.5 75 75 75 74 74 73 73 78.5

summary(son)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 70.50 71.25 72.60 72.77 73.88 76.50

sd(son)

## [1] 1.833667

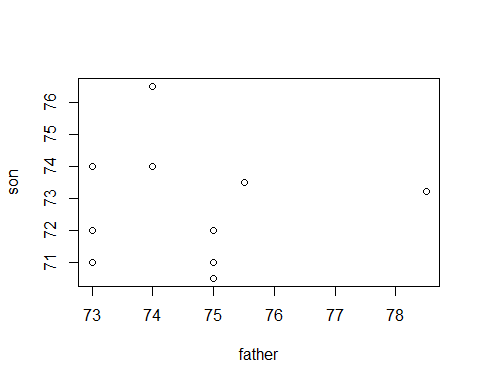
var(son)

## [1] 3.362333

str(son)

## num [1:10] 74 73.5 71 70.5 72 76.5 74 71 72 73.2

plot(father,son)

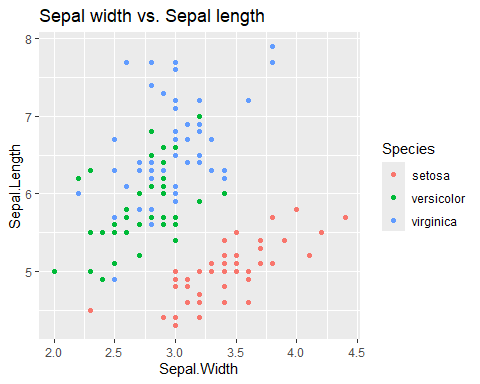


cor(father,son)

## [1] -0.01712208

# Question 4

mydata <- iris  
ggplot(mydata, aes(x=Sepal.Width, y=Sepal.Length, col=Species))+  
 geom\_point()+  
 ggtitle("Sepal width vs. Sepal length")



# Question 5

## 5.1  
# Given  
mean <- 0  
sd <- 1  
# Z0.01 is the z-score separating the bottom 99% from the top 1%  
Z0.01 <- qnorm(p=.01, mean, sd, lower.tail=F)  
Z0.01

## [1] 2.326348

## 5.2  
# Find the Z score for the bottom 2%  
cutoff\_low <- qnorm(p=0.02, mean, sd, lower.tail=T)  
# Find the Z score for the top 2%  
cutoff\_high <- qnorm(p=0.98, mean, sd, lower.tail=T)  
# Output the results  
cutoff\_low

## [1] -2.053749

cutoff\_high

## [1] 2.053749

# Question 6

# 6.1 Probability of the bone density test score between -2.75 and 2.75  
prob\_1 <- pnorm(2.75, mean, sd, lower.tail=T) - pnorm(-2.75, mean, sd, lower.tail=T)  
prob\_1

## [1] 0.9940405

# 6.2 Probability of the bone density test score between -2.0 and 2.0  
prob\_2 <- pnorm(2.0, mean, sd, lower.tail=T) - pnorm(-2.0, mean, sd, lower.tail=T)  
prob\_2

## [1] 0.9544997

# Question 7

Given that the general characteristics of a normal distrubution are with mean=1 and std=1, then we expect anything to that is above the value of ‘2’ and below the value of’0’ to be an outlier. This is abiding by the “Range Rule of Thumb for Identifying Significant Values” as referenced in the lecture slides (wk 2 Tuesday)

# Question 8

prob\_1 <- 0.10  
prob\_2 <- 0.97  
prob\_3 <- 0.025  
  
# Finding the Z critical values  
z\_0.10 <- round(qnorm(prob\_1, mean = 0, sd = 1), 2)  
z\_0.97 <- round(qnorm(prob\_2, mean = 0, sd = 1), 2)  
z\_0.025 <- round(qnorm(prob\_3, mean = 0, sd = 1), 2)  
  
# Display the results  
z\_0.10

## [1] -1.28

z\_0.97

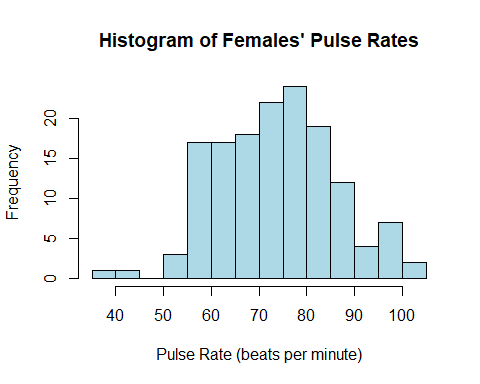
## [1] 1.88

z\_0.025

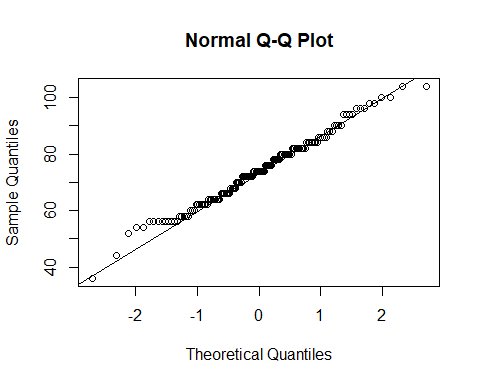
## [1] -1.96

# Question 9

library(readxl)  
Body\_Data <- read\_excel("Body Data.xlsx")  
# Filter the data to include only females (0 = Female, 1 = Male)  
females <- subset(Body\_Data, GENDER == 0)  
  
# Extract the pulse rates of females  
female\_pulse\_rates <- females$PULSE  
  
# Construct a histogram of females' pulse rates  
hist(female\_pulse\_rates,   
 main = "Histogram of Females' Pulse Rates",  
 xlab = "Pulse Rate (beats per minute)",   
 ylab = "Frequency",   
 col = "lightblue",   
 border = "black",  
 breaks = 10)



qqnorm(female\_pulse\_rates)  
qqline(female\_pulse\_rates)  
  
# Add a normal distribution curve for comparison  
x <- seq(min(female\_pulse\_rates), max(female\_pulse\_rates), length = 100)  
y <- dnorm(x, mean = mean(female\_pulse\_rates), sd = sd(female\_pulse\_rates))  
y <- y \* diff(hist(female\_pulse\_rates, plot = FALSE)$mids[1:2]) \* length(female\_pulse\_rates)  
lines(x, y, col = "darkgreen", lwd = 2)



In general, the female pulse rates do follow a normal distribution to some extent, as evidenced by the histogram with the overlayed bell curve. As we can see, the higher frequencies are concentrated toward the middle of the spread (~55-90) and flatten out toward the edges. We do see some outliers and see the highest frequency towards the middle of the spread (~70-80), which does fit the general behavior of a normal distribution.

# Question 10

# Data: Measured radiation emissions (W/kg) from the selected cell phones  
radiation\_data <- c(0.38, 0.55, 1.54, 1.55, 0.50, 0.60, 0.92, 0.96, 1.00, 0.86, 1.46)  
  
# Calculate the mean and standard deviation of the sample  
mean\_radiation <- mean(radiation\_data)  
sd\_radiation <- sd(radiation\_data)  
n <- length(radiation\_data)  
  
# Calculate the 90% confidence interval  
confidence\_level <- 0.90  
alpha <- 1 - confidence\_level  
t\_value <- qt(1 - alpha/2, df = n - 1)  
  
# Margin of error  
margin\_of\_error <- t\_value \* (sd\_radiation / sqrt(n))  
  
# Confidence interval  
lower\_bound <- mean\_radiation - margin\_of\_error  
upper\_bound <- mean\_radiation + margin\_of\_error  
confidence\_interval <- c(lower\_bound, upper\_bound)  
  
# Output the confidence interval  
cat("90% Confidence Interval for the population mean:", confidence\_interval, "\n")

## 90% Confidence Interval for the population mean: 0.7070946 1.169269

# Compare with FCC standard of 1.6 W/kg  
if (upper\_bound < 1.6) {  
 cat("\nThe result suggests that the population mean radiation is less than the FCC standard of 1.6 W/kg.\n")  
} else {  
 cat("The result does not conclusively suggest that the population mean radiation is less than the FCC standard of 1.6 W/kg.\n")  
}

##   
## The result suggests that the population mean radiation is less than the FCC standard of 1.6 W/kg.

# Question 11

# Data given  
n\_placebo <- 43  
mean\_placebo <- 21.57  
sd\_placebo <- 3.87  
  
n\_treatment <- 33  
mean\_treatment <- 20.38  
sd\_treatment <- 3.91  
  
# Perform a two-sample t-test  
t\_test\_result <- t.test(  
 x = c(rnorm(n\_placebo, mean = mean\_placebo, sd = sd\_placebo)),  
 y = c(rnorm(n\_treatment, mean = mean\_treatment, sd = sd\_treatment)),  
 var.equal = TRUE, # Assume equal variances  
 alternative = "two.sided", # Two-tailed test  
 conf.level = 0.05 # Significance level  
)  
  
# Output the result  
t\_test\_result

##   
## Two Sample t-test  
##   
## data: c(rnorm(n\_placebo, mean = mean\_placebo, sd = sd\_placebo)) and c(rnorm(n\_treatment, mean = mean\_treatment, sd = sd\_treatment))  
## t = 0.53241, df = 74, p-value = 0.596  
## alternative hypothesis: true difference in means is not equal to 0  
## 5 percent confidence interval:  
## 0.3857201 0.4891054  
## sample estimates:  
## mean of x mean of y   
## 22.03981 21.60239

if (t\_test\_result$p.value < 0.05) {  
 cat("There is significant evidence to suggest that paroxetine has a different effect on bipolar depression compared to the placebo.\n")  
} else {  
 cat("There is not enough evidence to suggest that paroxetine is more effective than the placebo for treating bipolar depression.\n")  
}

## There is not enough evidence to suggest that paroxetine is more effective than the placebo for treating bipolar depression.

# Question 12

library(readxl)  
Body\_Data <- read\_excel("Body Data.xlsx")  
female\_BP <- Body\_Data$DIASTOLIC[Body\_Data$GENDER == 0]  
  
# Perform a one-sample t-test to test the claim that the mean diastolic blood pressure is less than 90 mmHg  
t\_test\_result <- t.test(female\_BP, mu = 90, alternative = "less")  
  
# Print the test results  
print(t\_test\_result)

##   
## One Sample t-test  
##   
## data: female\_BP  
## t = -21.435, df = 146, p-value < 2.2e-16  
## alternative hypothesis: true mean is less than 90  
## 95 percent confidence interval:  
## -Inf 71.69517  
## sample estimates:  
## mean of x   
## 70.16327

# Conclusion based on the p-value  
if (t\_test\_result$p.value < 0.05) {  
 cat("At the 0.05 significance level, we reject the null hypothesis.\n")  
 cat("This suggests that the mean diastolic blood pressure for adult females is significantly less than 90 mmHg.\n")  
} else {  
 cat("At the 0.05 significance level, we fail to reject the null hypothesis.\n")  
 cat("This suggests that we do not have enough evidence to claim that the mean diastolic blood pressure for adult females is less than 90 mmHg.\n")  
}

## At the 0.05 significance level, we reject the null hypothesis.  
## This suggests that the mean diastolic blood pressure for adult females is significantly less than 90 mmHg.

# Interpretation regarding hypertension  
if (t\_test\_result$conf.int[2] < 90) {  
 cat("Based on the result, we can conclude that it is unlikely that any of the adult females in the sample have hypertension (diastolic BP > 90 mmHg).\n")  
} else {  
 cat("Based on the result, we cannot conclusively state that none of the adult females in the sample have hypertension.\n")  
}

## Based on the result, we can conclude that it is unlikely that any of the adult females in the sample have hypertension (diastolic BP > 90 mmHg).

# Question 13

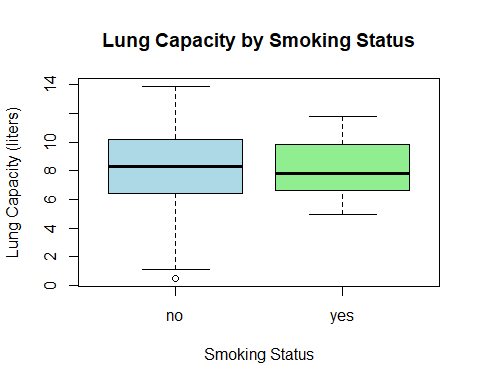
#Import LC Data  
library(readxl)  
LCData <- read\_xlsx("LCData.xlsx")  
# View the data structure  
str(LCData)

## tibble [200 × 6] (S3: tbl\_df/tbl/data.frame)  
## $ LC : num [1:200] 6.47 10.12 9.55 4.8 6.22 ...  
## $ Age : num [1:200] 6 18 16 5 11 11 15 11 19 17 ...  
## $ Height : num [1:200] 62.1 74.7 69.7 56.9 58.7 70.4 70.5 59.2 76.4 71.7 ...  
## $ Smoke : chr [1:200] "no" "yes" "no" "no" ...  
## $ Gender : chr [1:200] "male" "female" "female" "male" ...  
## $ Caesarean: chr [1:200] "no" "no" "yes" "no" ...

summary(LCData)

## LC Age Height Smoke   
## Min. : 0.507 Min. : 3.00 Min. :48.70 Length:200   
## 1st Qu.: 6.450 1st Qu.:10.00 1st Qu.:61.10 Class :character   
## Median : 8.225 Median :13.00 Median :65.65 Mode :character   
## Mean : 8.131 Mean :12.68 Mean :65.67   
## 3rd Qu.:10.137 3rd Qu.:16.00 3rd Qu.:71.12   
## Max. :13.875 Max. :19.00 Max. :81.80   
## Gender Caesarean   
## Length:200 Length:200   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

# a. Boxplot to examine the relationship between lung capacity and smoking status  
boxplot(LC ~ Smoke, data = LCData,   
 main = "Lung Capacity by Smoking Status",  
 xlab = "Smoking Status",  
 ylab = "Lung Capacity (liters)",  
 col = c("lightblue", "lightgreen"))



# b. State the hypotheses  
cat("\n Null Hypothesis (Ho): There is no difference in mean lung capacity between smokers and nonsmokers (μ\_smokers = μ\_nonsmokers).\n")

##   
## Null Hypothesis (Ho): There is no difference in mean lung capacity between smokers and nonsmokers (μ\_smokers = μ\_nonsmokers).

cat("Alternative Hypothesis (Ha): There is a difference in mean lung capacity between smokers and nonsmokers (μ\_smokers ≠ μ\_nonsmokers).\n")

## Alternative Hypothesis (Ha): There is a difference in mean lung capacity between smokers and nonsmokers (μ\_smokers ≠ μ\_nonsmokers).

# Assumptions check  
cat("1. Normality: Check if lung capacities are approximately normally distributed for each group. (can use histograms or Shapiro-Wilk test)\n")

## 1. Normality: Check if lung capacities are approximately normally distributed for each group. (can use histograms or Shapiro-Wilk test)

cat("2. Equal variances: Check if the variances of lung capacities are equal for smokers and nonsmokers. (can use F-test)\n")

## 2. Equal variances: Check if the variances of lung capacities are equal for smokers and nonsmokers. (can use F-test)

# Normality check using Shapiro-Wilk test  
shapiro.test(LCData$LC[LCData$Smoke == "yes"])

##   
## Shapiro-Wilk normality test  
##   
## data: LCData$LC[LCData$Smoke == "yes"]  
## W = 0.95242, p-value = 0.4642

shapiro.test(LCData$LC[LCData$Smoke == "no"])

##   
## Shapiro-Wilk normality test  
##   
## data: LCData$LC[LCData$Smoke == "no"]  
## W = 0.98857, p-value = 0.1507

# Variance check using Bartlett's test  
bartlett.test(LC ~ Smoke, data = LCData)

##   
## Bartlett test of homogeneity of variances  
##   
## data: LC by Smoke  
## Bartlett's K-squared = 2.0558, df = 1, p-value = 0.1516

# c. Decide on a test type and significance level  
# Given the research question, a two-sided test is appropriate.  
# Set significance level (alpha)  
alpha <- 0.05  
  
# d. Perform t-test assuming equal variances  
t\_test\_result <- t.test(LC ~ Smoke, data = LCData, var.equal = TRUE)  
t\_test\_result

##   
## Two Sample t-test  
##   
## data: LC by Smoke  
## t = -0.32277, df = 198, p-value = 0.7472  
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0  
## 95 percent confidence interval:  
## -1.515432 1.089135  
## sample estimates:  
## mean in group no mean in group yes   
## 8.111852 8.325000

# e. Interpret the R output and identify key statistics  
test\_statistic <- t\_test\_result$statistic  
df <- t\_test\_result$parameter  
p\_value <- t\_test\_result$p.value  
CI <- t\_test\_result$conf.int  
means <- t\_test\_result$estimate  
  
# Print out the key statistics  
cat("Test Statistic (t):", test\_statistic, "\n")

## Test Statistic (t): -0.3227656

cat("Degrees of Freedom (df):", df, "\n")

## Degrees of Freedom (df): 198

cat("P-value:", p\_value, "\n")

## P-value: 0.7472134

cat("95% Confidence Interval:", CI, "\n")

## 95% Confidence Interval: -1.515432 1.089135

cat("Means for Smokers and Nonsmokers:", means, "\n")

## Means for Smokers and Nonsmokers: 8.111852 8.325

# f. Interpret the CI and use it to conclude  
# If the 95% CI does not include 0, we reject the null hypothesis (Ho).  
if(CI[1] > 0 | CI[2] < 0) {  
 cat("Reject Ho: There is a significant difference in lung capacity between smokers and nonsmokers.\n")  
} else {  
 cat("Fail to Reject Ho: There is no significant difference in lung capacity between smokers and nonsmokers.\n")  
}

## Fail to Reject Ho: There is no significant difference in lung capacity between smokers and nonsmokers.

# g. Manually calculate the mean difference’s 95% CI  
n1 <- length(LCData$LC[LCData$Smoke == "yes"])  
n2 <- length(LCData$LC[LCData$Smoke == "no"])  
mean\_diff <- means[1] - means[2]  
pooled\_sd <- sqrt(((n1-1)\*var(LCData$LC[LCData$Smoke == "yes"]) +   
 (n2-1)\*var(LCData$LC[LCData$Smoke == "no"])) / (n1 + n2 - 2))  
se\_diff <- pooled\_sd \* sqrt(1/n1 + 1/n2)  
manual\_CI <- mean\_diff + c(-1, 1) \* qt(1 - alpha/2, df) \* se\_diff  
  
cat("Manually Calculated 95% Confidence Interval for Mean Difference:", manual\_CI, "\n")

## Manually Calculated 95% Confidence Interval for Mean Difference: -1.515432 1.089135

# h. Manually calculate the t-statistic and p-value  
manual\_t\_statistic <- mean\_diff / se\_diff  
manual\_p\_value <- 2 \* pt(-abs(manual\_t\_statistic), df)  
  
cat("Manually Calculated t-Statistic:", manual\_t\_statistic, "\n")

## Manually Calculated t-Statistic: -0.3227656

cat("Manually Calculated P-value:", manual\_p\_value, "\n")

## Manually Calculated P-value: 0.7472134

# Conclusion based on manual calculations  
if(manual\_p\_value < alpha) {  
 cat("Reject Ho based on manual calculation: There is a significant difference in lung capacity between smokers and nonsmokers.\n")  
} else {  
 cat("Fail to Reject Ho based on manual calculation: There is no significant difference in lung capacity between smokers and nonsmokers.\n")  
}

## Fail to Reject Ho based on manual calculation: There is no significant difference in lung capacity between smokers and nonsmokers.