HW3

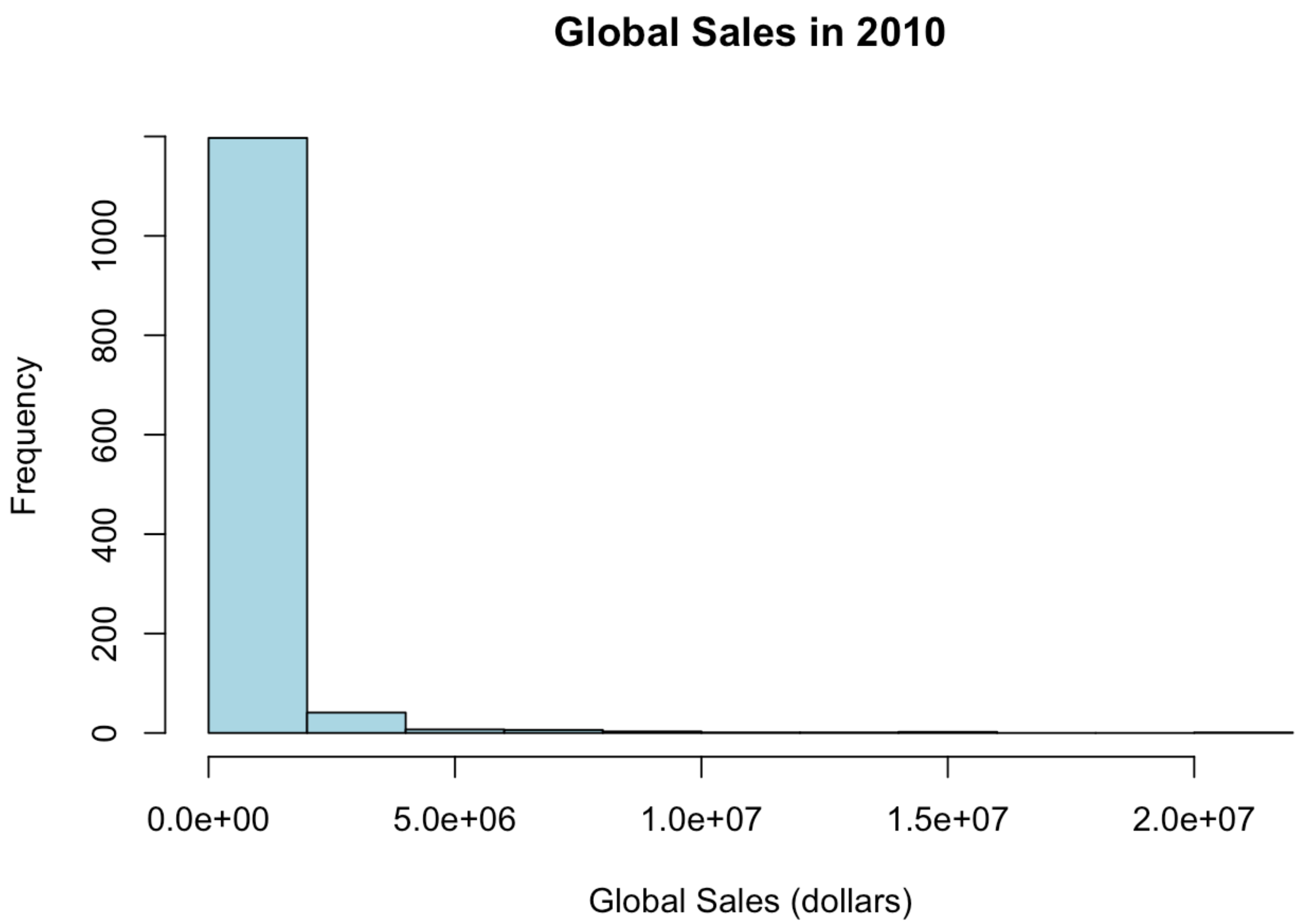
2024-04-27

Problem 1

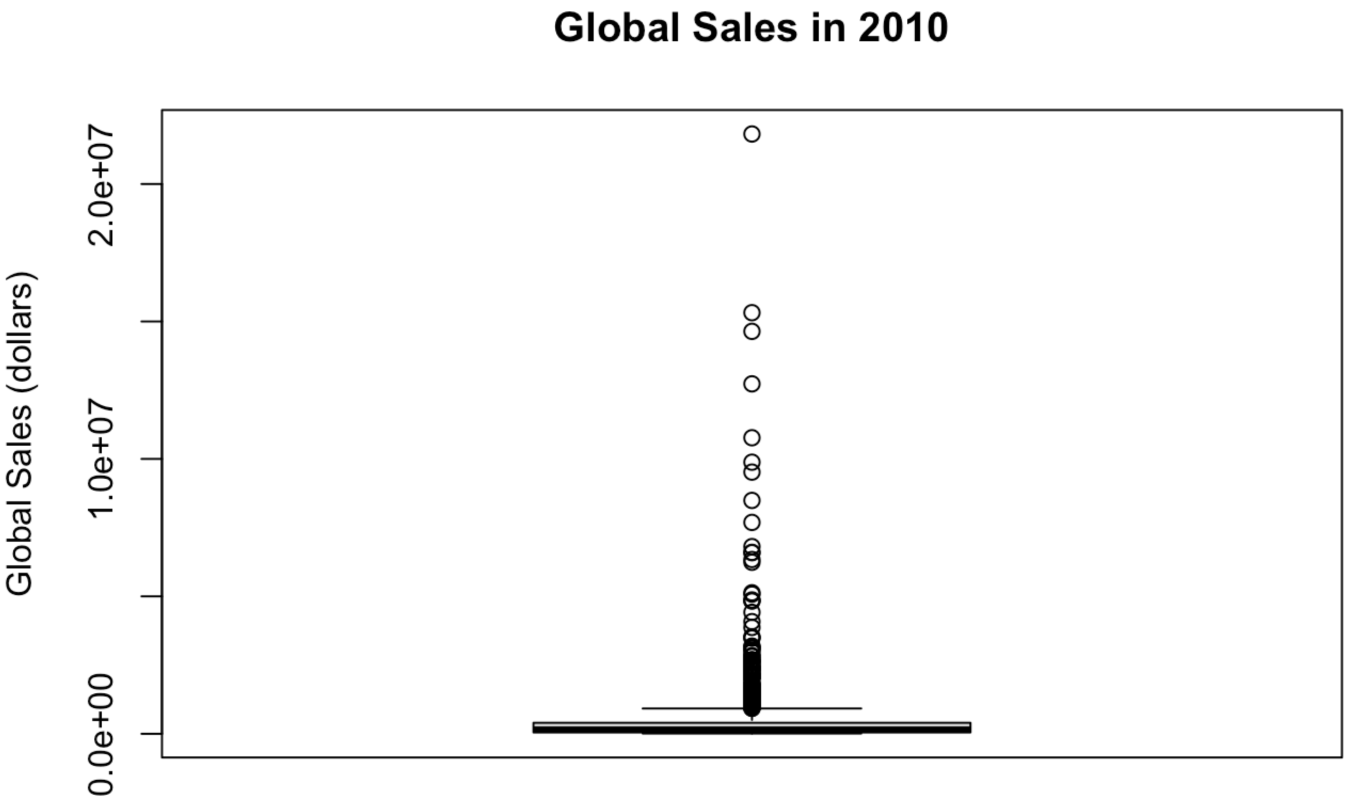
```
# A
data <- read.csv('vgsales.csv')

# covert to dollars
data$Global_Sales <- data$Global_Sales * 1000000

# focus on 2010 and global sales
global_sales <- subset(data, Year == 2010, select = Global_Sales)

# Show summary including median and mean
summary_stats <- summary(global_sales)
print(summary_stats)
```

```
##   Global_Sales
##  Min.   :   10000
##  1st Qu.:   50000
##  Median :  150000
##  Mean   :  476926
##  3rd Qu.:  400000
##  Max.   :21820000
```

```
# B
data_2010_Glo <- subset(data, Year == 2010, select = Global_Sales)

# Convert Global_Sales to numeric if it's not already
data_2010_Glo$Global_Sales <- as.numeric(data_2010_Glo$Global_Sales )

# Plot histogram
datafram <- data.frame(data_2010_Glo$Global_Sales)
hist(data_2010_Glo$Global_Sales,
     main = "Global Sales in 2010",
     xlab = "Global Sales (dollars)",
     ylab = "Frequency",
     col = "lightblue")
```

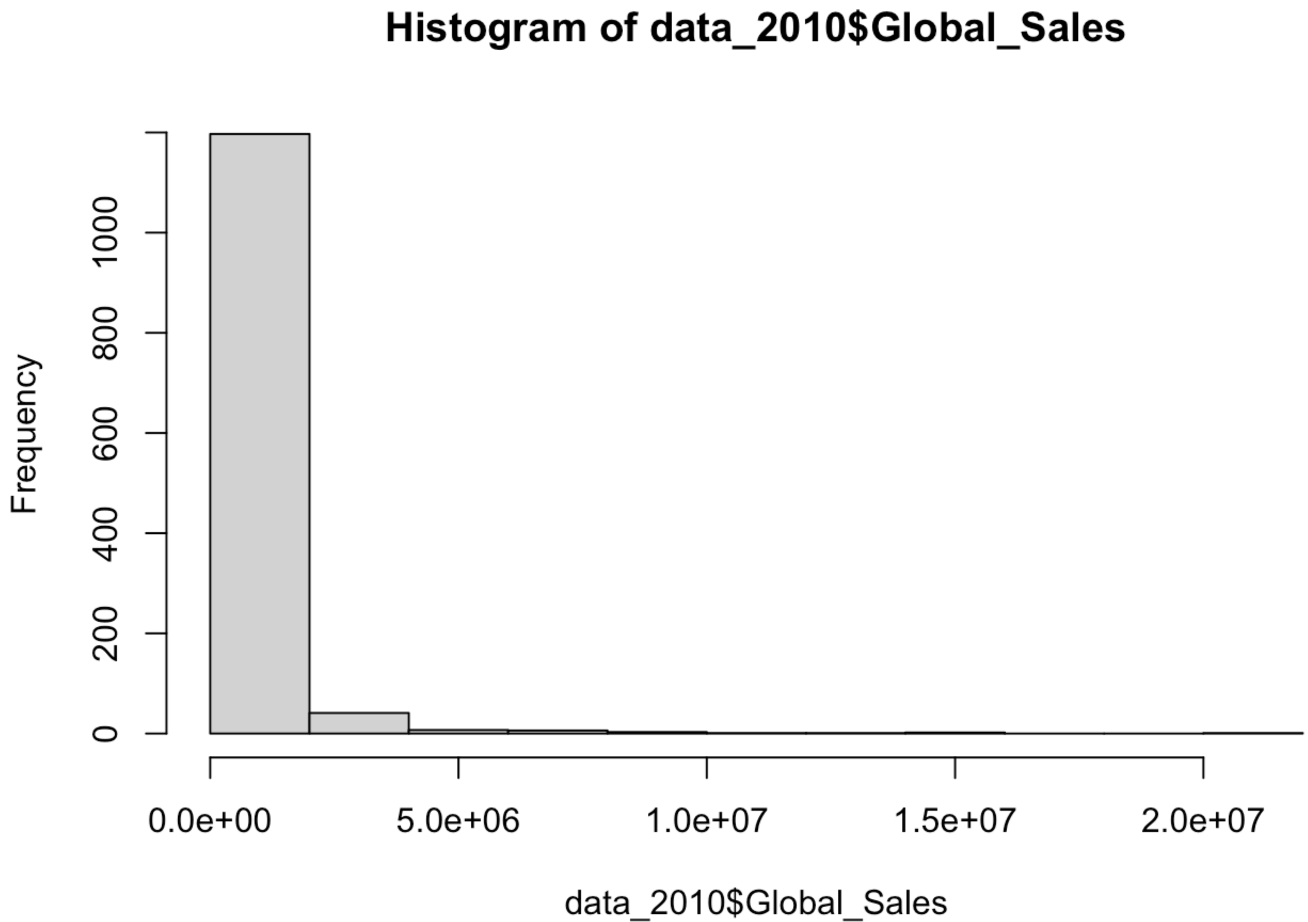



```
# plot boxplot
boxplot(data_2010_Glo$Global_Sales,
        main = "Global Sales in 2010",
        ylab = "Global Sales (dollars)")
```


Global Sales in 2010
Global Sales (dollars)
0.0e+00
1.0e+07
2.0e+07

```
# poisson.test <- function(x) {
#   # Assuming "x" is your data vector with sales in millions (adjust units if needed
)
#   lambda <- mean(x)  # Estimate Poisson parameter (average sales)
#   D0 <- max(abs(ecdf(x) - ppois(x, lambda)))  # KS statistic
#
#   # Perform actual KS test with p-value calculation
#   ks_test_result <- ks.test(x, ppois, lambda = lambda, alternative = "two.sided")
#   p.value <- ks_test_result$p.value
#
#   return(p.value)
# }
#
# # Example usage
# sales_data <- c(2.5, 0.8, 1.2, ..., your data)  # Replace with your actual data
# p_value <- poisson.test(sales_data)
#
# cat("p-value from KS test for Poisson distribution: ", p_value, "\n")
#
# # Companion plot (optional)
# theoretical_poisson <- rpois(length(sales_data), lambda)
# plot(ecdf(sales_data), xlab = "Sales (in millions)", ylab = "Cumulative Probability
")
# lines(ppois(unique(sales_data), lambda), pch = 16, col = "red")
# legend("topright", legend = c("Data", "Theoretical Poisson"), col = c("black", "red
"), pch = c(1, 16))
# title(main = "Empirical CDF vs. Theoretical Poisson CDF")
```

```
# Filter data for 2010, global sales
data_2010 <- subset(data, Year == 2010)

# Test if the data follows a Poisson distribution using chi-square goodness-of-fit te
st
observed <- hist(data_2010$Global_Sales)$counts
```




```
expected <- rpois(length(observed), mean(data_2010$Global_Sales))

# Perform the chi-square test
chisq_test <- chisq.test(table(observed, expected))
```

```
## Warning in chisq.test(table(observed, expected)): Chi-squared approximation may
## be incorrect
```

```
# Print the chi-square test results
cat("Chi-Square Test for Poisson Distribution:\n")
```

```
## Chi-Square Test for Poisson Distribution:
```

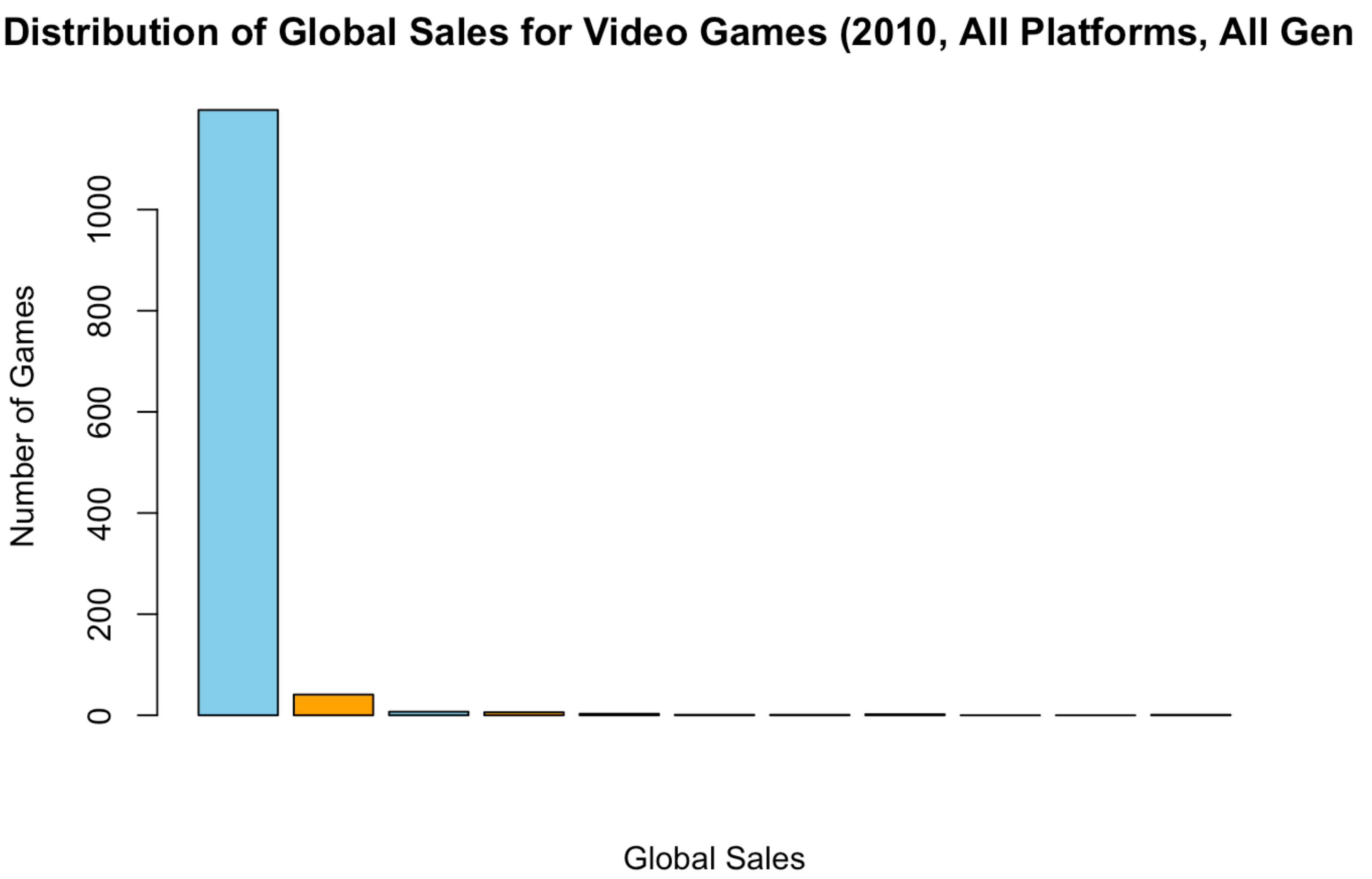
```
cat("Chi-Square: ", chisq_test$statistic, "\n")
```

```
## Chi-Square:  77
```

```
cat("p-value: ", chisq_test$p.value, "\n")
```

```
## p-value:  0.2647101
```

```
barplot(observed, col = c("skyblue", "orange"),
        main = "Distribution of Global Sales for Video Games (2010, All Platforms, Al
l Genres)",
        xlab = "Global Sales", ylab = "Number of Games")
```



Problem 2

```
# A
# Function to perform flip sign test
flipSignTest <- function(x, B = 10000) {
  n <- length(x)
  observed_mean <- mean(x)
  observed_abs_mean <- abs(observed_mean)

  # Track the number of |Ye| >= |Y*|
  count <- 0

  for (i in 1:B) {
    # Generate random sign
    sign <- sample(c(-1, 1), size = n, replace = TRUE)

    # Y_epsilon
    Y_epsilon <- mean(sign * x)

    # Check if |Ye| >= |Y*|
    if (abs(Y_epsilon) >= observed_abs_mean) {
      count <- count + 1
    }
  }

  # Calculate p-value
  p_value <- count / (2 ^ n)
  return(p_value)
}
```

```
# B
data_2010 <- subset(data, Year == 2010)
p_value <- flipSignTest(data_2010$EU_Sales - data_2010$JP_Sales)

# Print the p-value
print(p_value)
```

```
## [1] 0
```

Problem 3

A. Hypothesis test problem: Do action game have a significantly higher or lower median Global Sales compared to Fighting video games?

Null hypothesis (H0): he median Global Sales are the same for action and fighting video games.

Alternative hypothesis (H1): Action video game has not have an equal median global sales compared to fighting video games.

B. Summary statistic and plots:

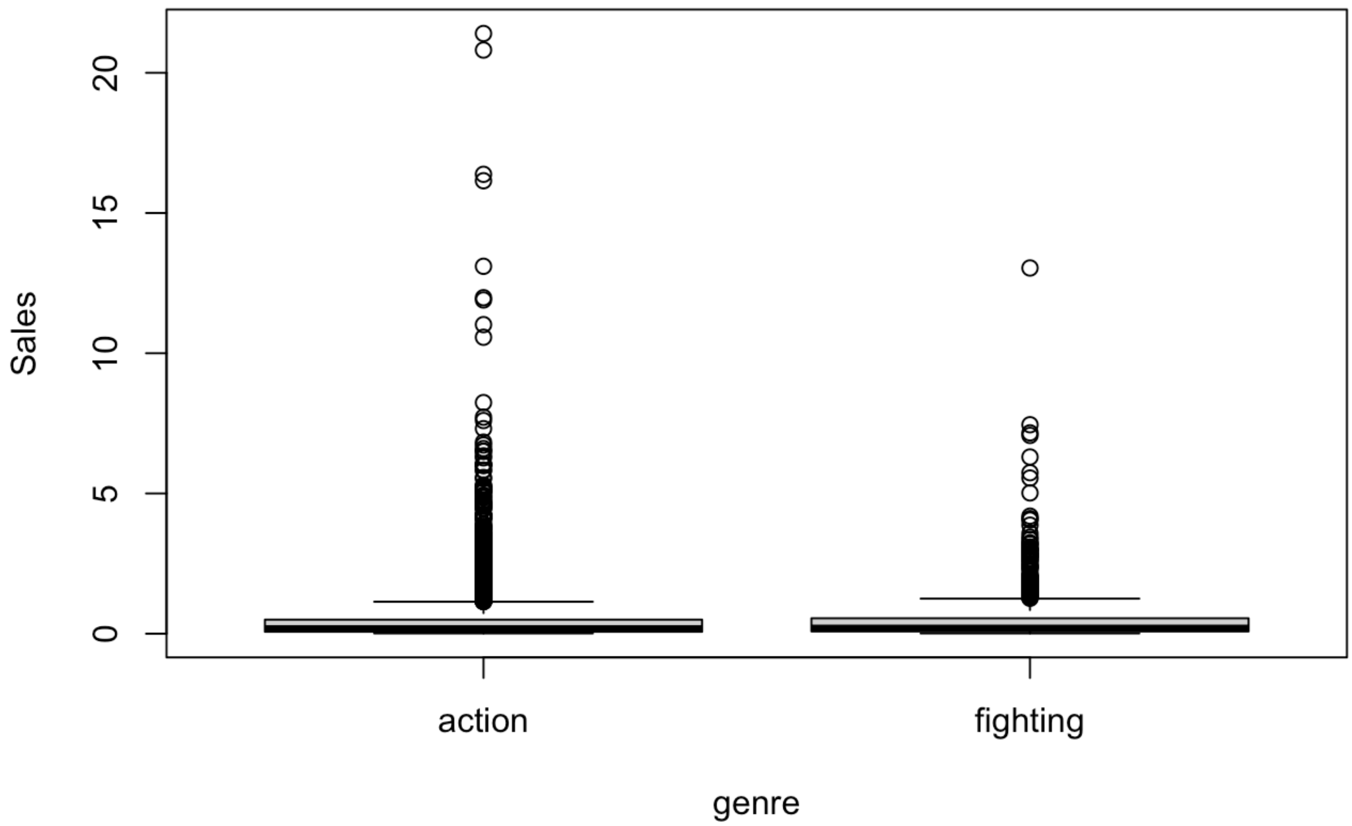
```
data3 <- read.csv('vgsales.csv')

fighting <- data3[data3$Genre == "Fighting",]$Global_Sales
action <- data3[data3$Genre == "Action",]$Global_Sales

summary(action)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.0700  0.1900  0.5281  0.5000 21.4000
```

```
summary(fighting)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.0800  0.2100  0.5294  0.5500 13.0400
```

```
boxplot(action, fighting, names = c("action", "fighting"),
        xlab = "genre", ylab = "Sales")
```



From the summary and boxplot above, we get to see that there is a difference in median of the global sale between action and fighting video game.

C. Apply a test:

```
wilcox.test(action, fighting, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  action and fighting
## W = 1353228, p-value = 0.0912
## alternative hypothesis: true location shift is not equal to 0
```

From here, since p value > 0.05, we will fail to reject the null hypothesis. Therefore, there is no significant different in median sale between Japan and Europe.