Code ▾

In Class Activity - Srujana Vanka, 2020102005

19-02-2024

Hide

```
library(readxl)
```

2

a. Calculate the difference between group medians and perform a permutation test (10000 iterations) to calculate the significance of the observed statistic. Plot a histogram displaying the bootstrap distribution.

Hide

```
# Load the dataset
mosquito <- read_excel("BRSM_Results Visualization.xlsx", sheet = "Mosquito")

beer_median <- median(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"])
water_median <- median(mosquito$`No. of Mosquitoes`[mosquito$Group == "Water"])
observed_stat <- beer_median - water_median
cat("Observed difference in medians:", observed_stat)
```

```
Observed difference in medians: 4
```

Hide

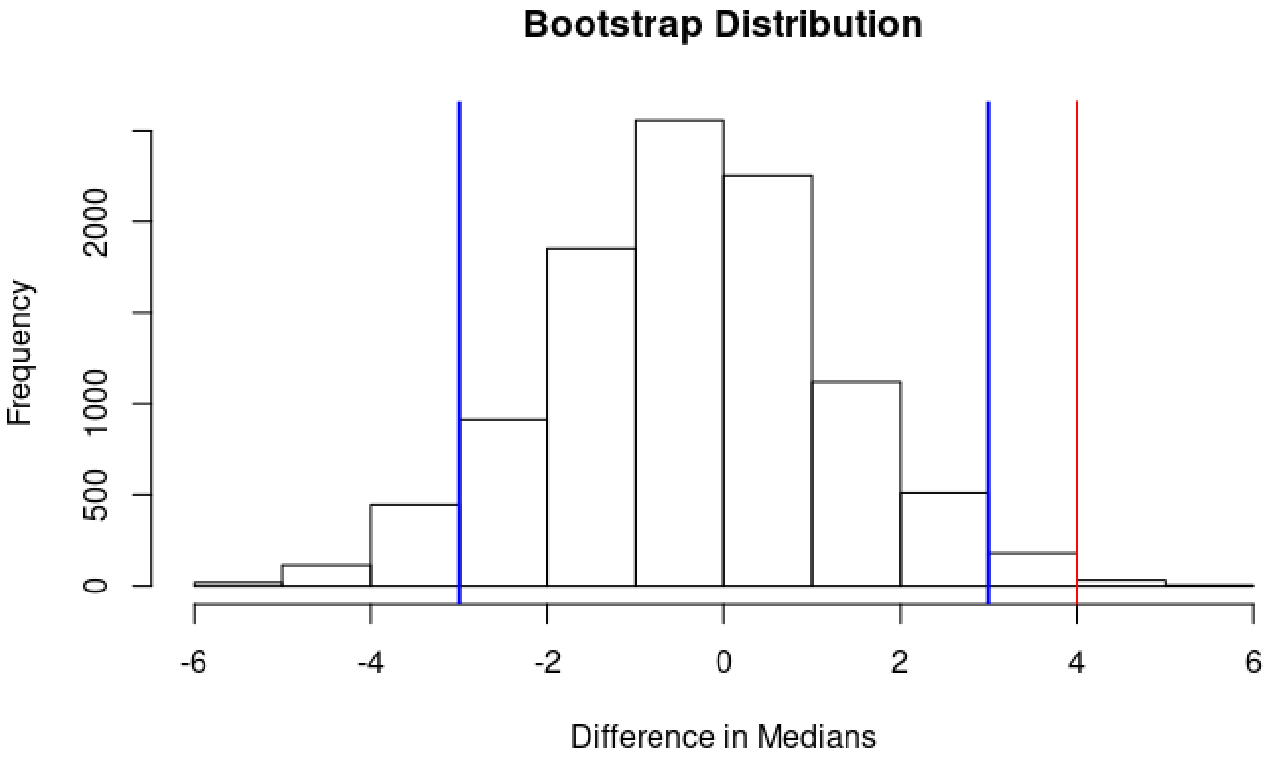
```
replacement <- 1;
iterations <- 10000;

data <- c(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"], mosquito$`No. of Mo
squitoes`[mosquito$Group == "Water"])
t_bootstrap <- c(1:iterations)
group1 <- c(1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"]))
group2 <- c(1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Water"]))

for (i in 1:iterations) {
  if (replacement == 1) {
for (m in 1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"])) {
randomized = sample(length(data), 1)
group1[m] <- data[randomized]
}
for (k in 1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Water"])) {
randomized = sample(length(data), 1)
group2[k] <- data[randomized]
}
  } else {
    randomized = sample(length(data), length(data));
    group1 <- randomized[1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Bee
r"])]
    group2 <- randomized[length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Bee
r"])+1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Water"])]
  }

  t_bootstrap[i] <- median(group1) - median(group2)
}

sorted_bootstrap = sort(t_bootstrap)
```

Hide

```
# Plotting histogram
hist(t_bootstrap, main="Bootstrap Distribution", xlab="Difference in Medians")
sorted_bootstrap = sort(t_bootstrap)
limit1 = sorted_bootstrap[round(0.025*iterations)]
limit2 = sorted_bootstrap[iterations - round(0.025*iterations)]
abline(v=limit1, col="blue", lwd=2)
```

Hide

```
abline(v=limit2, col="blue", lwd=2)
abline(v=observed_stat, col="red")
```



Hide

```
# Calculate two-tailed p-value
p_value <- sum(abs(sorted_bootstrap) >= abs(observed_stat)) / iterations
cat("Two-tailed P-value:", p_value, "\n")
```

```
Two-tailed P-value: 0.0261
```

b. Repeat step ‘a’ on the t-statistic instead of difference in medians

Hide

```
mosquito <- read_excel("BRSM_Results Visualization.xlsx", sheet = "Mosquito")

# Using t.test to calculate the observed statistic
observed_stat <- t.test(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"], mosqu
ito$`No. of Mosquitoes`[mosquito$Group == "Water"])
print(observed_stat)
```

```
	Welch Two Sample t-test

data:  mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"] and mosquito$`No. of Mo
squitoes`[mosquito$Group == "Water"]
t = 3.6582, df = 39.113, p-value = 0.0007474
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.957472 6.798084
sample estimates:
mean of x mean of y 
 23.60000  19.22222 
```

Hide

```
observed_stat = observed_stat$statistic
print(observed_stat)
```

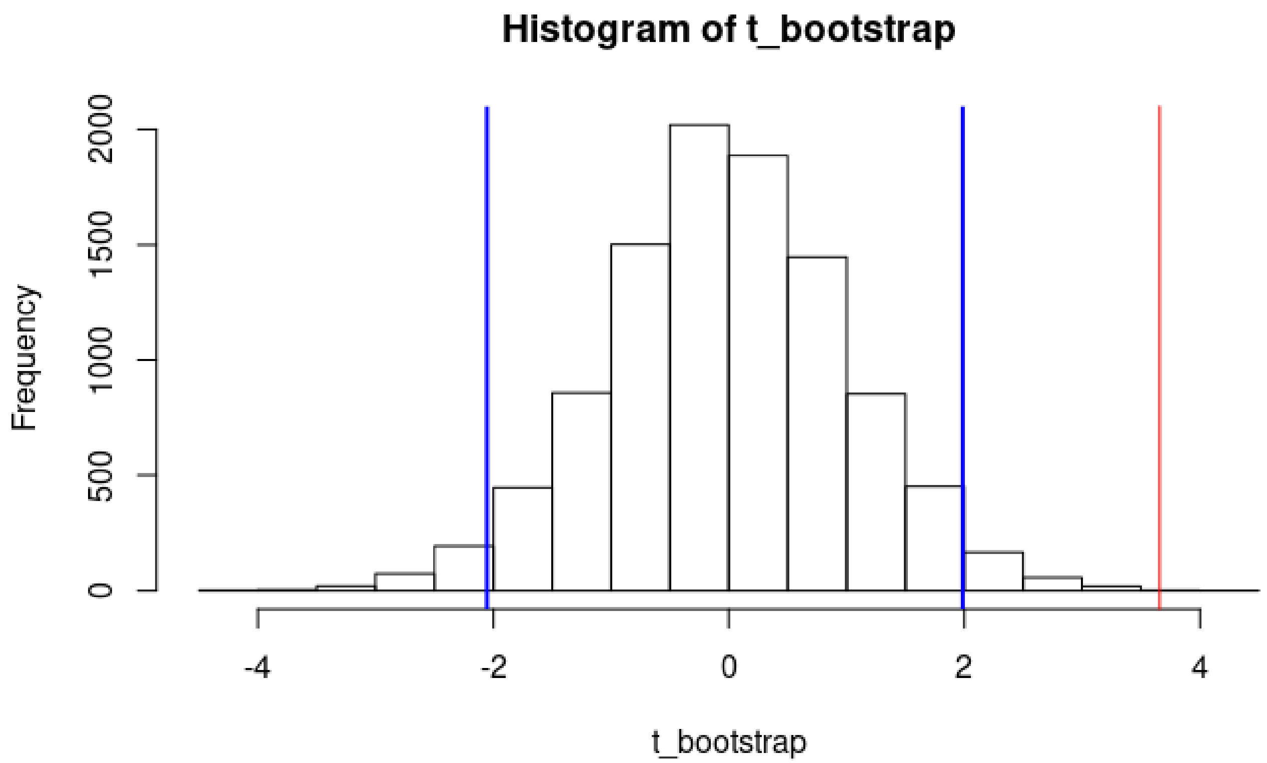
```
       t 
3.658245 
```

Hide

```
replacement <- 1;
iterations <- 10000;
group1 <- c(1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"]))
group2 <- c(1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Water"]))

data <- c(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"], mosquito$`No. of Mo
squitoes`[mosquito$Group == "Water"])
t_bootstrap <- c(1:iterations)

for (i in 1:iterations) {
  if (replacement == 1) {
    for (m in 1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Beer"])) {
      randomized <- sample(length(data), 1)
      group1[m] <- data[randomized]
    }
    for (k in 1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Water"])) {
      randomized <- sample(length(data), 1)
      group2[k] <- data[randomized]
    }
  } else {
    randomized <- sample(length(data), length(data));
    group1 <- data[randomized[1:length(mosquito$`No. of Mosquitoes`[mosquito$Group ==
"Beer"])]]
    group2 <- data[randomized[length(mosquito$`No. of Mosquitoes`[mosquito$Group ==
"Beer"])+1:length(mosquito$`No. of Mosquitoes`[mosquito$Group == "Water"])]]
  }

  t_bootstrap[i] <- t.test(group1, group2)$statistic;
}

# Plot histogram
hist(t_bootstrap)
sorted_bootstrap = sort(t_bootstrap)
limit1 = sorted_bootstrap[round(0.025*iterations)]
limit2 = sorted_bootstrap[iterations - round(0.025*iterations)]
abline(v=limit1, col="blue", lwd=2)
```

Hide

```
abline(v=limit2, col="blue", lwd=2)
abline(v=observed_stat, col="red")
```



Show

c. Assuming a non-directional HA (suggesting that there will be a difference in groups), calculate the new significance values of the above observed statistics.

We compute the p-value for both tails of the permutation distribution by summing the occurrences of values greater than or equal to the observed t-statistic and those less than or equal to the negative of the observed t-statistic. This sum is then divided by the total number of permutations, providing a measure of the extremeness of the observed t-statistic within the distribution of permuted values.

Hide

```
p_value_non_directional = sum((abs(sorted_bootstrap) >= observed_stat))/ iterations
print("NON DIRECTIONAL")
```

```
[1] "NON DIRECTIONAL"
```

Hide

```
print(p_value_non_directional)
```

```
[1] 7e-04
```

Hide

```
# One-tailed p-value for difference in medians (directional)
p_value_directional = sum((sorted_bootstrap >= observed_stat)) / iterations
print("DIRECTIONAL")
```

```
[1] "DIRECTIONAL"
```

Hide

```
print(p_value_directional)
```

```
[1] 2e-04
```

3 IQ dataset problem

Hide

```
data <- read_excel("data.xlsx")
```

```
New names:
* `` -> ...1
* GPA -> GPA...2
* GPA -> GPA...6
* `` -> ...7
* `` -> ...8
* … and 9 more problems
```

Hide

```
# Calculate the observed correlation
observed_corr <- cor(data$IQ, data$Placement_TESTSCORE)
cat("Observed correlation between IQ and Test Score:", observed_corr, "\n")
```

```
Observed correlation between IQ and Test Score: 0.4931479
```

Hide

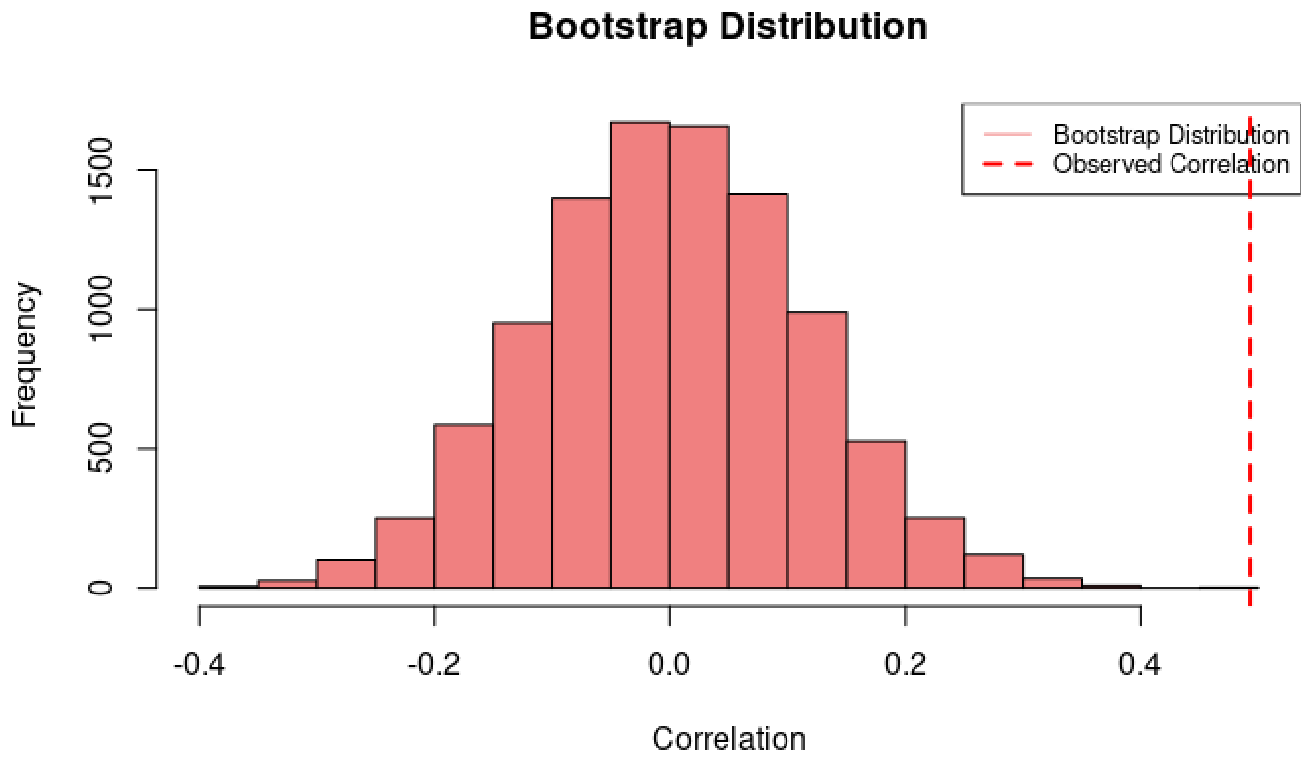
```
# Set the number of iterations for the permutation test
iterations <- 10000

# Create a vector to store bootstrap distribution of correlation
corr_bootstrap <- numeric(iterations)

# Permutation test
for (i in 1:iterations) {
  # Shuffle the "Placement_TESTSCORE" values while keeping IQ unchanged
  shuffled_testscore <- sample(data$Placement_TESTSCORE)

  # Calculate the correlation for the shuffled data
  corr_bootstrap[i] <- cor(data$IQ, shuffled_testscore)
}
```

Hide

```
# Plot histogram of the bootstrap distribution
hist(corr_bootstrap, main="Bootstrap Distribution", xlab="Correlation", col="lightcor
al")
# Add a legend
legend("topright", legend = c("Bootstrap Distribution", "Observed Correlation"),
       col = c("lightcoral", "red"), lty = c(1, 2), lwd = c(1, 2), cex = 0.8)
```

Hide

```
abline(v=observed_corr, col="red", lty=2, lwd=2)  # Add a dashed line at the observed
correlation
```



The histogram represents the distribution of permuted correlation coefficients, with the observed correlation coefficient indicated by a purple dashed line.

This visual representation allows us to assess where the observed correlation stands within the range of permuted correlations.

Hide

```
# Calculate p-value

p_value = sum(abs(corr_bootstrap) >= abs(observed_corr)) / iterations
cat("Two-tailed P-value:", p_value, "\n")
```

```
Two-tailed P-value: 0
```

Hide

```
# Determine whether to reject or accept the null hypothesis
alpha = 0.05
if (p_value < alpha) {
  cat("Reject the null hypothesis: There is a significant correlation between IQ and
Placement TESTSCORE.\n")
} else {
  cat("Accept the null hypothesis: There is no significant correlation between IQ and
Placement TESTSCORE.\n")
}
```

```
Reject the null hypothesis: There is a significant correlation between IQ and Placeme
nt TESTSCORE.
```

The p-value represents the probability of observing a correlation as extreme as the observed correlation under the null hypothesis which is observed to be 0.