---
title: "Assignment 1"
author: "Herthika Sivasuntharampillai"
date: "2024-11-14"
output: html_document

Question 1 - Comment all code blocks

```{r bivariate}
# Generate normally distributed data for personality and looks. Each vector has 100 observations
personality <- rnorm(100, 10, 4)   # Mean 10, SD 4
looks <- rnorm(100, 10, 3)         # Mean 10, SD 3

# Create a data frame to store the generated data
df.pl <- data.frame(looks, personality)

# Scatter plot of looks vs. personality
plot(df.pl$looks, df.pl$personality)

# Add a line of best fit with intercept 15 and slope -0.9
abline(15, -0.9)   # Creates a hypothetical trend line
```


```{r p-value}

set.seed(13)   # Set seed for reproducibility of random number functions

# Generate a matrix and label p.mat0 with 1000 rows and 6 columns of random uniform data. Runif is generating a vector of random numbers 0 to 1
nrow = 1000
ncol = 6
p.mat0 <- matrix(data=runif(nrow * ncol), nrow=nrow, ncol=ncol)

# For each row, performs a t-test comparing the first three columns [1:3] with the last three columns [1:6] and extracts the p-value.$ is the subset of anything and 1 means apply to every row
p.out0 <- apply(p.mat0, 1, function(x) as.numeric(t.test(x[1:3], x[4:6])$p.value))

# Plot a histogram of the p-values obtained from the t-tests. Should be flat if no true difference
hist(as.numeric(p.out0), breaks=99)

# Add a vertical red dashed line at 0.05, marking the threshold for significance.
abline(v=0.05, col='red', lty=2, lwd=3)

# Calculate the proportion of p-values below the 0.05 significance threshold.
sum(p.out0 < 0.05) / nrow

# Find the minimum p-value from the simulated tests
min(p.out0)

# This adjusts the p-values using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR)
min(p.adjust(p.out0))


# Repeat with a small true difference between columns 1-3 and 4-6
# Generate a matrix and label p.mat1 with 1000 rows and 6 columns of random uniform data. Runif is generating a vector of random numbers 0 to 1
nrow = 1000
ncol = 6
```

```
p.mat1 <- matrix(data=runif(nrow * ncol), nrow=nrow, ncol=ncol)

# Modify columns 4-6 to create a slight true difference. Adjusted values to introduce true difference
p.mat1[,4:6] <- runif(n=3000, min=0.2, max=1.2)

# For each row, performs a t-test comparing the first three columns [1:3] with the last three columns [1:6] and extracts the p-value.$ is the subset of anything and 1 means apply to every row
p.out1 <- apply(p.mat1, 1, function(x) as.numeric(t.test(x[1:3], x[4:6])$p.value))

# Plot a histogram of the p-values obtained from the t-tests. Should be flat if no true difference
hist(as.numeric(p.out1), breaks=99)

# Add a vertical red dashed line at 0.05, marking the threshold for significance.
abline(v=0.05, col='red', lty=2, lwd=3)

# Calculate the proportion of p-values below the 0.05 significance threshold.
sum(p.out1 < 0.05) / nrow

# Find the minimum p-value from the simulated tests
min(p.out1)

# This adjusts the p-values using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR)
min(p.adjust(p.out1))


```

Question 2 - make two new code blocks with comments where the number of samples in each test is 10, and is 20. What is the number of statistically significant values (relative to 3x3) in each with either raw p-values or with q-values? why? How does this relate to the concept of power?

```{r}
# Generating code block for sample size 10
set.seed(13)   # Set seed for reproducibility

#Generate a matrix and label p.matx with 1000 rows and 20 columns of random uniform data. Runif is generating a vector of random numbers 0 to 1
nrow = 1000
ncol = 20
p.matx <- matrix(data=runif(nrow * ncol), nrow=nrow, ncol=ncol)

# Modify columns 11-20 to create a slight true difference. Adjusted values to introduce true difference
p.matx[,11:20] <- runif(n=10000, min=0.2, max=1.2)

# For each row, performs a t-test comparing the first three columns [1:10] with the last three columns [11:20] and extracts the p-value.$ is the subset of anything and 1 means apply to every row
p.outx <- apply(p.matx, 1, function(x) as.numeric(t.test(x[1:10], x[11:20])$p.value))

# Plot a histogram of the p-values obtained from the t-tests.
hist(as.numeric(p.outx), breaks=99)

# Add a vertical red dashed line at 0.05, marking the threshold for significance.
abline(v=0.05, col='red', lty=2, lwd=3)

# Calculate the proportion of p-values below the 0.05 significance threshold.
sum(p.outx < 0.05) / nrow

# Find the minimum p-value from the simulated tests
```

```
min(p.outx)

# This adjusts the p-values using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR)
min(p.adjust(p.outx))

```
```{r}
# Generating code block for sample size 20
set.seed(13)   # Set seed for reproducibility

#Generate a matrix and label p.matx with 1000 rows and 40 columns of random uniform data. Runif is generating a vector of random numbers 0 to 1
nrow = 1000
ncol = 40
p.matx <- matrix(data=runif(nrow * ncol), nrow=nrow, ncol=ncol)

# Modify columns 21-40 to create a slight true difference. Adjusted values to introduce true difference
p.matx[,21:40] <- runif(n=20000, min=0.2, max=1.2)

# For each row, performs a t-test comparing the first three columns [1:20] with the last three columns [21:40] and extracts the p-value.$ is the subset of anything and 1 means apply to every row
p.outx <- apply(p.matx, 1, function(x) as.numeric(t.test(x[1:20], x[21:40])$p.value))

# Plot a histogram of the p-values obtained from the t-tests.
hist(as.numeric(p.outx), breaks=99)

# Add a vertical red dashed line at 0.05, marking the threshold for significance.
abline(v=0.05, col='red', lty=2, lwd=3)

# Calculate the proportion of p-values below the 0.05 significance threshold.
sum(p.outx < 0.05) / nrow

# Find the minimum p-value from the simulated tests
min(p.outx)

# This adjusts the p-values using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR)
min(p.adjust(p.outx))


```
The number of statistically significant values (p-values below 0.05) increases as the sample size grows from the 3x3 example to the sample sizes of 10 and 20. With larger sample sizes, the power of the test improves, meaning it becomes easier to detect true differences between groups. Raw p-values are lower with bigger samples because the tests have less variability and more precision. Adjusted p-values (q-values), which correct for multiple comparisons, also show more significant results as sample size increases, reflecting stronger evidence. This demonstrates how larger sample sizes reduce errors and make it easier to find real effects.

Question 3 - Using the information in the readings explain why the following statement is false? "If one observes a small P-value, there is a good chance that the next study will produce a P-value at least as small for the same hypothesis."

A small p-value suggests that the observed result is unlikely to have occurred due to chance under the null hypothesis. However, the likelihood of reproducing a similarly small p-value in another study depends on factors like the study's effect size, sample size, and statistical power.As the articles highlight, p-values alone are not enough to predict the reproducibility of results. They depend heavily on sample size and variability. A small p-

value in a study with a large sample size may indicate a very small, but statistically significant, effect that could be difficult to replicate with smaller samples. Effect size is crucial here: it measures the actual magnitude of the observed difference, providing a more reliable indicator of whether a similar result will be observed in subsequent studies. High statistical power, determined by sufficient sample size and a meaningful effect size, increases the chance of consistently observing small p-values in replication studies. Low power or marginal effect sizes, on the other hand, increase the risk of variability in p-values, even for the same hypothesis. Thus, focusing on effect size alongside p-values gives a more complete picture of the reliability and reproducibility of the findings.