Lab

AUTHOR
JP Garcia

Part 1a)

```
library(readr)
salary_data <- read_csv("jp-garcia-131a/SFSalaries2014.csv")
head(salary_data)
```

```
# A tibble: 6 × 10
    ...1     Id JobTitle          BasePay OvertimePay OtherPay Benefits TotalPay
   <dbl>  <dbl> <chr>               <dbl>       <dbl>    <dbl>    <dbl>    <dbl>
1 110532 110532 Deputy Chief 3    129150.          0   342803.   38780.  471953.
2 110533 110533 Asst Med Examiner 318835.      10713.   60564.   89540.  390112.
3 110534 110534 Chief Investment… 257340           0    82314.   96571.  339654.
4 110535 110535 Chief of Police   307450.          0    19267.   91302.  326717.
5 110536 110536 Chief, Fire Depa… 302068           0    24165.   91202.  326233.
6 110537 110537 Asst Med Examiner 270222.       6009.   67956.   71580.  344187.
# ℹ 2 more variables: TotalPayBenefits <dbl>, Status <chr>
```

```
print("Null Hypothesis: The median of the log total pay is the same for both part-time an
```

```
[1] "Null Hypothesis: The median of the log total pay is the same for both part-time and
full-time workers. Alternative Hypothesis: The median of the log total pay is different
for part-time and full-time workers."
```

Part 1b)

```
clean_data <- salary_data[!is.na(salary_data$TotalPay) & !is.na(salary_data$Status), ]

obs_stat <- median(log(clean_data$TotalPay[clean_data$Status == 'PT'] + 1)) - median(log(

print(obs_stat)
```

```
[1] -1.43664
```

Part 1c)

```
n_perm <- 10000
perm_stats <- replicate(n_perm, {
  shuffled <- sample(clean_data$Status)
  median(log(clean_data$TotalPay[shuffled == 'PT'] + 1)) - median(log(clean_data$TotalPay
})
```

Part 1d)

```
p_value <- mean(abs(perm_stats) >= abs(obs_stat))


print(p_value)
```

```
[1] 0
```

Part 1e)

```
if (p_value <= 0.05) {
  print("We reject the null hypothesis: There is a significant difference in the median l
} else {
  print("We do not reject the null hypothesis: There isn't enough evidence to suggest a s
}
```

```
[1] "We reject the null hypothesis: There is a significant difference in the median log
TotalPay between PT and FT workers."
```

Part 2

```
library(readr)
heartDisease <- read_csv("jp-garcia-131a/heartDisease.csv")
head(heartDisease)
```

```
# A tibble: 6 × 14
    age   sex    cp trestbps  chol   fbs restecg thalach exang oldpeak slope
  <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>
1    63     1     1      145   233     1       2     150     0     2.3     3
2    67     1     4      160   286     0       2     108     1     1.5     2
3    67     1     4      120   229     0       2     129     1     2.6     2
4    37     1     3      130   250     0       0     187     0     3.5     3
5    41     0     2      130   204     0       2     172     0     1.4     1
6    56     1     2      120   236     0       0     178     0     0.8     1
# ℹ 3 more variables: ca <dbl>, thal <dbl>, num <dbl>
```

Part 2a)

```
angina_group <- heartDisease$chol[heartDisease$cp %in% c(1, 2)]
non_angina_group <- heartDisease$chol[heartDisease$cp %in% c(3, 4)]
t_test_result <- t.test(angina_group, non_angina_group)

print(t_test_result$p.value)
```

```
[1] 0.3087863
```

```
if (t_test_result$p.value < 0.05) {
  print("There is a significant difference in serum cholesterol levels between patients w
```

```
} else {
  print("There is no significant difference in serum cholesterol levels between pateins w
}
```

```
[1] "There is no significant difference in serum cholesterol levels between pateins with
angina and those with non anginal pain."
```

Part 2b)

```
angina_group <- heartDisease$chol[heartDisease$cp %in% c(1, 2)]
non_angina_group <- heartDisease$chol[heartDisease$cp %in% c(3, 4)]


mean_angina <- mean(angina_group)
mean_non_angina <- mean(non_angina_group)

var_angina <- var(angina_group)
var_non_angina <- var(non_angina_group)


n_angina <- length(angina_group)
n_non_angina <- length(non_angina_group)


t_statistic <- (mean_angina - mean_non_angina) / sqrt((var_angina / n_angina) + (var_non_

print((paste("t-stat:", t_statistic)))
```

```
[1] "t-stat: -1.02098914939664"
```

```
df <- min(n_angina - 1, n_non_angina - 1)

p_value <- 2 * pt(-abs(t_statistic), df)

print(paste("p-value:",(p_value)))
```

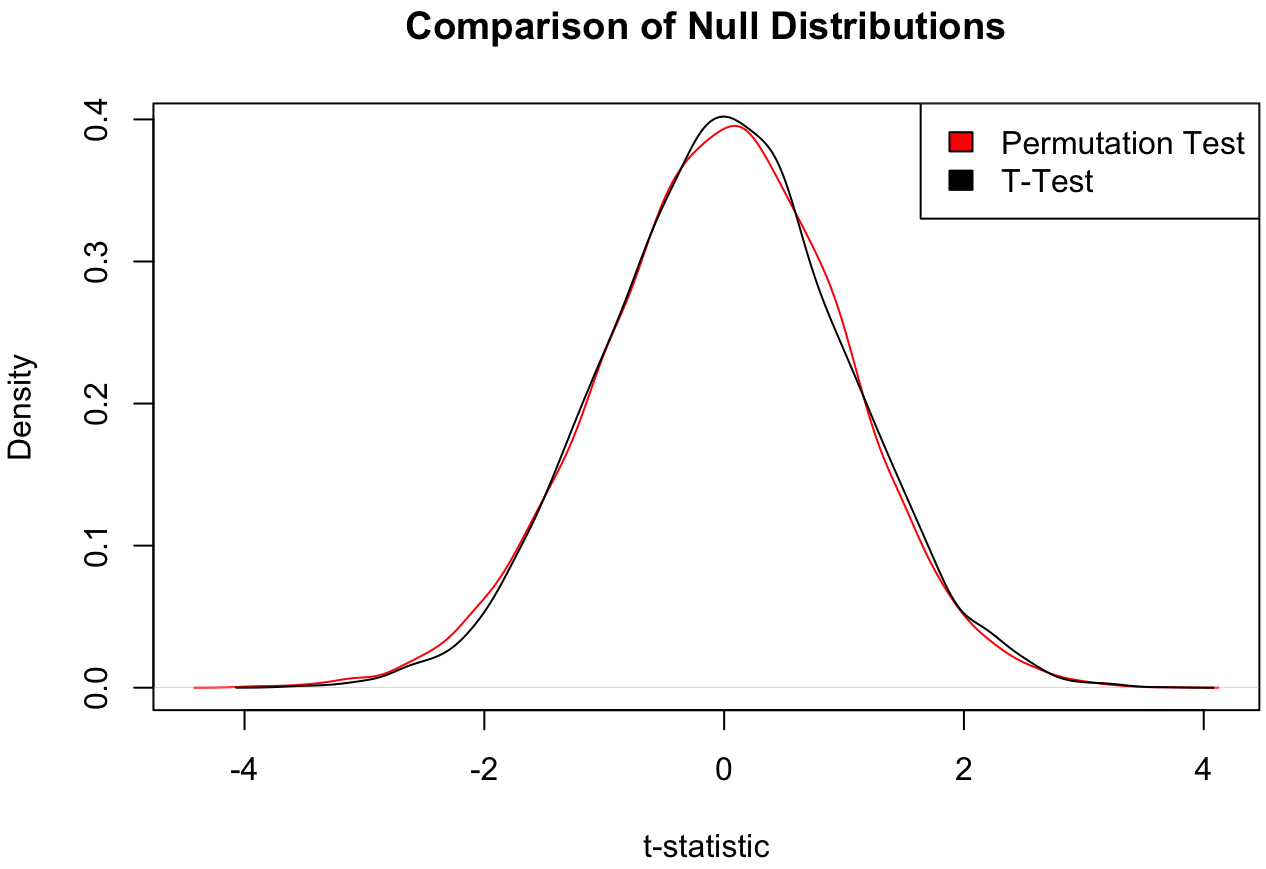
```
[1] "p-value: 0.310725936700205"
```

```
if (p_value < 0.05) {
  cat("There is a significant difference in serum cholesterol levels between patients wit
} else {
  cat("There is no significant difference in serum cholesterol levels between patients wi
}
```

```
There is no significant difference in serum cholesterol levels between patients with any
type of angina and those with non-anginal pain.
```

Part 2c

```r
library(ggplot2)


n_perm <- 10000


perm_t_stats <- replicate(n_perm, {
  shuffled_labels <- sample(heartDisease$cp)
  perm_angina_group <- heartDisease$chol[shuffled_labels %in% c(1, 2)]
  perm_non_angina_group <- heartDisease$chol[shuffled_labels %in% c(3, 4)]

  t.test(perm_angina_group, perm_non_angina_group)$statistic
})



common_sd <- sd(heartDisease$chol)
n_angina <- sum(heartDisease$cp %in% c(1, 2))
n_non_angina <- sum(heartDisease$cp %in% c(3, 4))

t_test_stats <- replicate(n_perm, {
  sim_angina_group <- rnorm(n_angina, mean = mean(heartDisease$chol), sd = common_sd)
  sim_non_angina_group <- rnorm(n_non_angina, mean = mean(heartDisease$chol), sd = common

  t.test(sim_angina_group, sim_non_angina_group)$statistic
})



plot(density(perm_t_stats), col = "red", main = "Comparison of Null Distributions", xlab
lines(density(t_test_stats), col = "black")
legend("topright", legend = c("Permutation Test", "T-Test"), fill = c("red", "black"))
```



Part 2d)

```
lower_limit <- 125
upper_limit <- 200


angina_group <- heartDisease$chol[heartDisease$cp %in% c(1, 2)]
non_angina_group <- heartDisease$chol[heartDisease$cp %in% c(3, 4)]

prop_angina_abnormal <- sum(angina_group < lower_limit | angina_group > upper_limit) / le
prop_non_angina_abnormal <- sum(non_angina_group < lower_limit | non_angina_group > upper


print(paste("Proportion of abnormal cholesterol levels in Angina group: ", prop_angina_ab
```

```
[1] "Proportion of abnormal cholesterol levels in Angina group:  0.861111111111111"
```

```
print(paste("Proportion of abnormal cholesterol levels in Non-Angina group: ", prop_non_a
```

```
[1] "Proportion of abnormal cholesterol levels in Non-Angina group:  0.826666666666667"
```

Part 3a and Part 3b

```r
n <- 20
group_1 <- rnorm(n, 10, 2.5)
group_2 <- rnorm(n, 10, 5)

p_value <- t.test(group_1, group_2)$p.value
print(paste("p-value:", p_value))
```

```
[1] "p-value: 0.24675661141831"
```

Part 3c

```r
n_simulations <- 10000

p_value <- replicate(n_simulations, {
  group_1 <- rnorm(n, 10, 2.5)
  group_2 <- rnorm(n, 10, 5)
  t.test(group_1, group_2)$p.value
})

type_I_error_rate <- mean(p_value < 0.05)
print(paste("type 1 error rate:", type_I_error_rate))
```

```
[1] "type 1 error rate: 0.0489"
```

Part 3d

```r
p_values_gamma <- replicate(n_simulations, {
  group_1 <- rgamma(n, 1, 3)
  group_2 <- rgamma(n, 1, 3)
  t.test(group_1, group_2)$p.value
})

type_I_error_rate_gamma <- mean(p_values_gamma < 0.05)
print(paste("type 1 error rate gamma:", type_I_error_rate_gamma))
```

```
[1] "type 1 error rate gamma: 0.0451"
```

Part 4

```r
MAKE_BOOTSTRAP_STATS <- function(values, FUN, B = 10000) {
    n <- length(values)  # Number of observations in the sample
    boot_values <- replicate(B, FUN(sample(values, n, replace = TRUE)))
  stat <- FUN(boot_values)
  return(stat)
}
```

Part 5a)

```
angina_chol <- heartDisease$chol[heartDisease$cp %in% c(1, 2)]
non_angina_chol <- heartDisease$chol[heartDisease$cp %in% c(3, 4)]


ci <- t.test(angina_chol, non_angina_chol, var.equal = TRUE)$conf.int


print(ci)
```

```
[1] -20.025356   7.697578
attr(,"conf.level")
[1] 0.95
```

```
print("Since the confidence interval contains zero (it spans from a negative value to a p
```

```
[1] "Since the confidence interval contains zero (it spans from a negative value to a
positive value), it suggests that, at the 95% confidence level, we do not have sufficient
evidence to conclude that there is a significant difference in mean serum cholesterol
levels between patients with any type of angina and those with non-anginal pain. This
means the difference in means is not statistically significant at the 0.05 significance
level."
```

Part 5b)

```
n_boot <- 10000

boot_diffs <- numeric(n_boot)

set.seed(123)
for (i in 1:n_boot) {
  boot_angina <- sample(angina_chol, replace = TRUE)
  boot_non_angina <- sample(non_angina_chol, replace = TRUE)
  boot_diffs[i] <- mean(boot_angina) - mean(boot_non_angina)
}

alpha <- 0.05
ci_low <- quantile(boot_diffs, alpha / 2)
ci_high <- quantile(boot_diffs, 1 - alpha / 2)

print(c(ci_low, ci_high))
```

```
     2.5%      97.5%
-18.053361   5.650708
```

```
print("There is no sufficient evidence to claim a significant difference in serum cholest
```

```
[1] "There is no sufficient evidence to claim a significant difference in serum
cholesterol between patients with any type of angina and those with non-anginal pain. The
```

```
true difference in means is uncertain; it could be negative, zero, or positive, based on
this confidence interval."
```

Part 5c)

```
MAKE_DIFFERENCE_BOOTSTRAP_STATS <- function(values_1, values_2, FUN)
{
  boot_values_1 <- FUN(sample(values_1, size = length(values_1), replace = TRUE))
  boot_values_2 <- FUN(sample(values_2, size = length(values_2), replace = TRUE))
  difference <- boot_values_1 - boot_values_2
  return(difference)
}
```