Worksheet I

Group 1: Antonio Alfaro de Prado, Hyunchang Oh, Tanya Toluay

2024-04-21

1st Question: t-test

1.1 Sample and Calculate P-Values

```
# Set the parameters
N <- c(5, 10, 20, 50, 100)
num_samples <- 5000

# Function to conduct t-test and return p-value
conduct_t_test <- function(data) {
  if (abs(mean(data))==1){
    return(0)
  }
  t_test <- t.test(data)
  return(t_test$p.value)
}

# Function to generate data from different distributions
generate_data <- function(dist, size) {
  if (dist == "continuous_uniform") {
    return(runif(size, min = -1, max = 1))
  } else if (dist == "student_t") {
    return(rt(size, df = 1))
  } else if (dist == "discrete_uniform") {
    return(sample(c(-1, 1), size, replace = TRUE))
  }
}


# Function to plot p-value distributions
plot_pvalue_distribution <- function(p_values, dist_name) {
  hist(p_values,
       main = paste("P-value Distribution for", dist_name),
       xlab = "P-value", ylab = "Frequency", col = "lightblue")
}


# Loop through each distribution and sample size
below_alpha_count_store = c()
dist_store <-c()
for (dist in c("continuous_uniform", "student_t", "discrete_uniform")) {
```

```
    cat("\nDistribution: ",dist,'\n')
    below_alpha_counts = c()
  for (sample_size in N) {
    # Initialize vector to store p-values
    p_values <- numeric(num_samples)

    # Generate 1000 different sets of size-N samples and calculate p-values
    below_alpha_count <- 0
    for (i in 1:num_samples) {
      sampled_values <- generate_data(dist, sample_size)
      p_value<-conduct_t_test(sampled_values)
      p_values[i] <- p_value
      if(p_value<=0.05){
        below_alpha_count <- below_alpha_count+1
      }
    }
    below_alpha_counts <- c(below_alpha_counts,below_alpha_count)
    cat("Sample Size: ",sample_size, "\tMean p-value:", mean(p_values), "\n")
    if(sample_size==5){
        dist_store<-cbind(dist_store,p_values)
    }
  }
  below_alpha_count_store <- c(below_alpha_count_store,below_alpha_counts)
}
```

```
##
## Distribution:  continuous_uniform
## Sample Size:  5  Mean p-value: 0.5126548
## Sample Size:  10     Mean p-value: 0.5123674
## Sample Size:  20     Mean p-value: 0.4999248
## Sample Size:  50     Mean p-value: 0.495236
## Sample Size:  100    Mean p-value: 0.5065521
##
## Distribution:  student_t
## Sample Size:  5  Mean p-value: 0.4544577
## Sample Size:  10     Mean p-value: 0.4453488
## Sample Size:  20     Mean p-value: 0.4480026
## Sample Size:  50     Mean p-value: 0.4344147
## Sample Size:  100    Mean p-value: 0.4429207
##
## Distribution:  discrete_uniform
## Sample Size:  5  Mean p-value: 0.5077344
## Sample Size:  10     Mean p-value: 0.5198732
## Sample Size:  20     Mean p-value: 0.5109005
## Sample Size:  50     Mean p-value: 0.5090125
## Sample Size:  100    Mean p-value: 0.4971066
```

1.2 Draw Histogram of P-Values

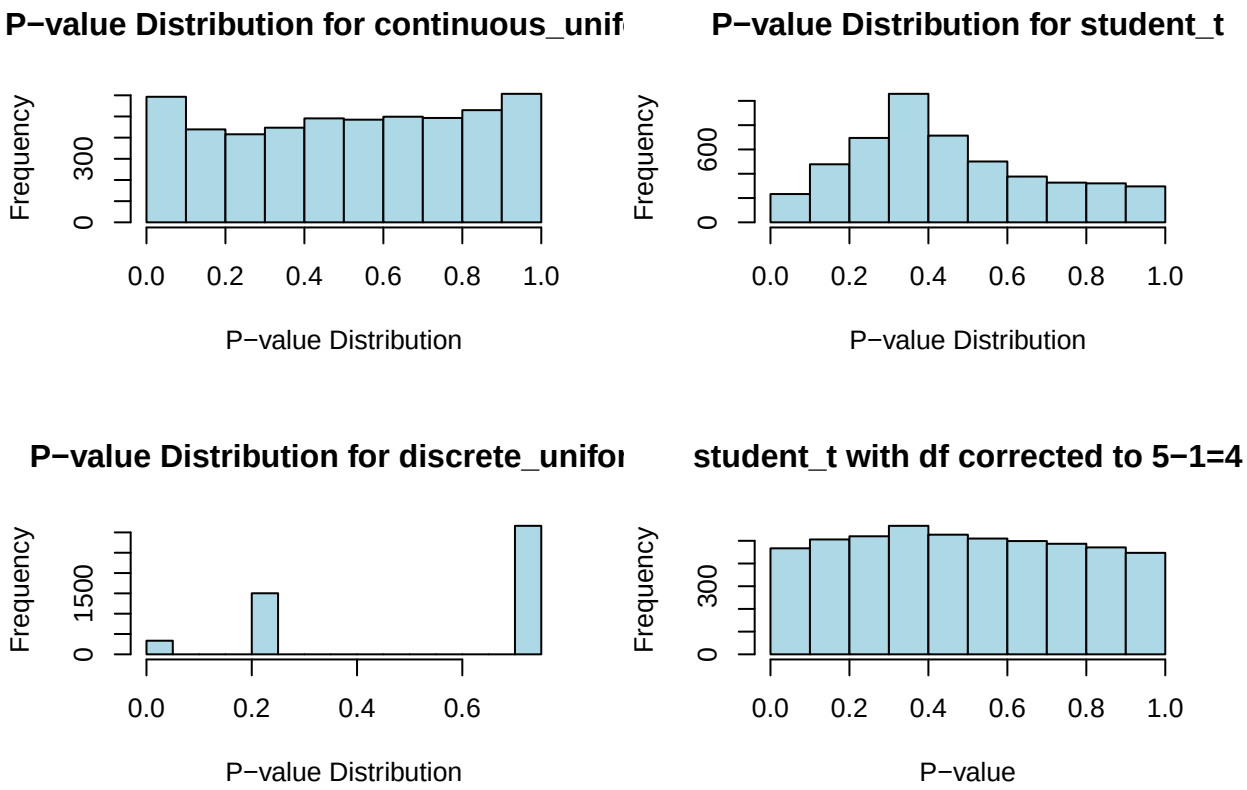
```
# Function to plot p-value distribution
plot_pvalue_distribution <- function(p_values, title) {
```

```
  hist(p_values, main = paste(title), xlab = "P-value",
       ylab = "Frequency", col = "lightblue")
}

# Set up the layout for the plots
par(mfrow=c(2, 2))

# Draw Plots from saved p_values
dist_names <- c("continuous_uniform", "student_t", "discrete_uniform")
for(i in 1:3){
  dist_name <- dist_names[i]
  hist(dist_store[,i],
       main = paste("P-value Distribution for", dist_name),
       xlab = "P-value Distribution", ylab = "Frequency", col = "lightblue")
}

# Draw an extra plot for df = 4 corrected version
p_values <- numeric(5000)
for (i in 1:5000) {
  sampled_values <- rt(5, df = 4)
  p_value <- conduct_t_test(sampled_values)
  p_values[i] <- p_value
}
plot_pvalue_distribution(p_values, "student_t with df corrected to 5-1=4")
```



1.3 Fraction of P-Values Below Alpha

```
N <- c(5, 10, 20, 50, 100)

cat("\n**Continuous Uniform**\n")
```

```
##
## **Continuous Uniform**
```

```
for(i in 1:5){
    cat("Sample Size: ",N[i],"\tFraction of p-values under alpha: ",below_alpha_count_store[i]/5000,"\n
}
```

```
## Sample Size:  5  Fraction of p-values under alpha:  0.0696
## Sample Size:  10     Fraction of p-values under alpha:  0.0544
## Sample Size:  20     Fraction of p-values under alpha:  0.0498
## Sample Size:  50     Fraction of p-values under alpha:  0.0438
## Sample Size:  100    Fraction of p-values under alpha:  0.0488
```

```
cat("\n**Student's t**\n")
```

```
##
## **Student's t**
```

```
for(j in 1:5){
    cat("Sample Size: ",N[j],"\tFraction of p-values under alpha: ",below_alpha_count_store[j+5]/5000,"
}
```

```
## Sample Size:  5  Fraction of p-values under alpha:  0.0156
## Sample Size:  10     Fraction of p-values under alpha:  0.0164
## Sample Size:  20     Fraction of p-values under alpha:  0.0228
## Sample Size:  50     Fraction of p-values under alpha:  0.0196
## Sample Size:  100    Fraction of p-values under alpha:  0.0204
```

```
cat("\n**Discrete Uniform**\n")
```

```
##
## **Discrete Uniform**
```

```
for(k in 1:5){
    cat("Sample Size: ",N[k],"\tFraction of p-values under alpha: ",below_alpha_count_store[k+10]/5000,
}
```

```
## Sample Size:  5  Fraction of p-values under alpha:  0.067
## Sample Size:  10     Fraction of p-values under alpha:  0.0214
## Sample Size:  20     Fraction of p-values under alpha:  0.0412
## Sample Size:  50     Fraction of p-values under alpha:  0.0638
## Sample Size:  100    Fraction of p-values under alpha:  0.0592
```

1.4 Discussion

1.4-1. Mean p-value:

In all cases, the mean p-value was close to 0.5 as it should be. The data was drawn from distributions with mean value of 0, so making a decision to accept or reject the hypothesis is as random as deciding it by throwing a coin.

1.4-2. Histogram of P-values:

Continuous Uniform Distribution: The distribution of p-values resemble the original uniform distribution.

Student's T-Distribution: The distribution of p-values does not resemble the original uniform distribution, because the degree of freedom was set to 1. If it was corrected to be 4 (sample size - 1), it followed the uniform distribution again.

Discrete Uniform Distribution: The distribution of p-values is discrete, just like the original t-distribution. There are three columns in the histogram, representing the 3 possible combinations of the 5 samples: (0,5), (1,4), (2,3).

Overall, if the distribution is continuous, the p-values under the null hypothesis follow uniform distribution.

1.4-3. Fraction of P-values below Alpha: For continuous distributions, as discussed in 1-2, the fraction of p-values smaller than alpha is equal to alpha, as they follow uniform distributions. Student's t-distribution did not show this however, because the degree of freedom was not set accordingly and just fixed to 1.

2nd Question: t-test

2.1 Importing, Visualization & Analysis

```
# Read in the trial data
lung_data <- read.csv("C:/Users/tanya/Downloads/lung_data.csv")

# View the first few rows of the data
kable(head(lung_data))
```

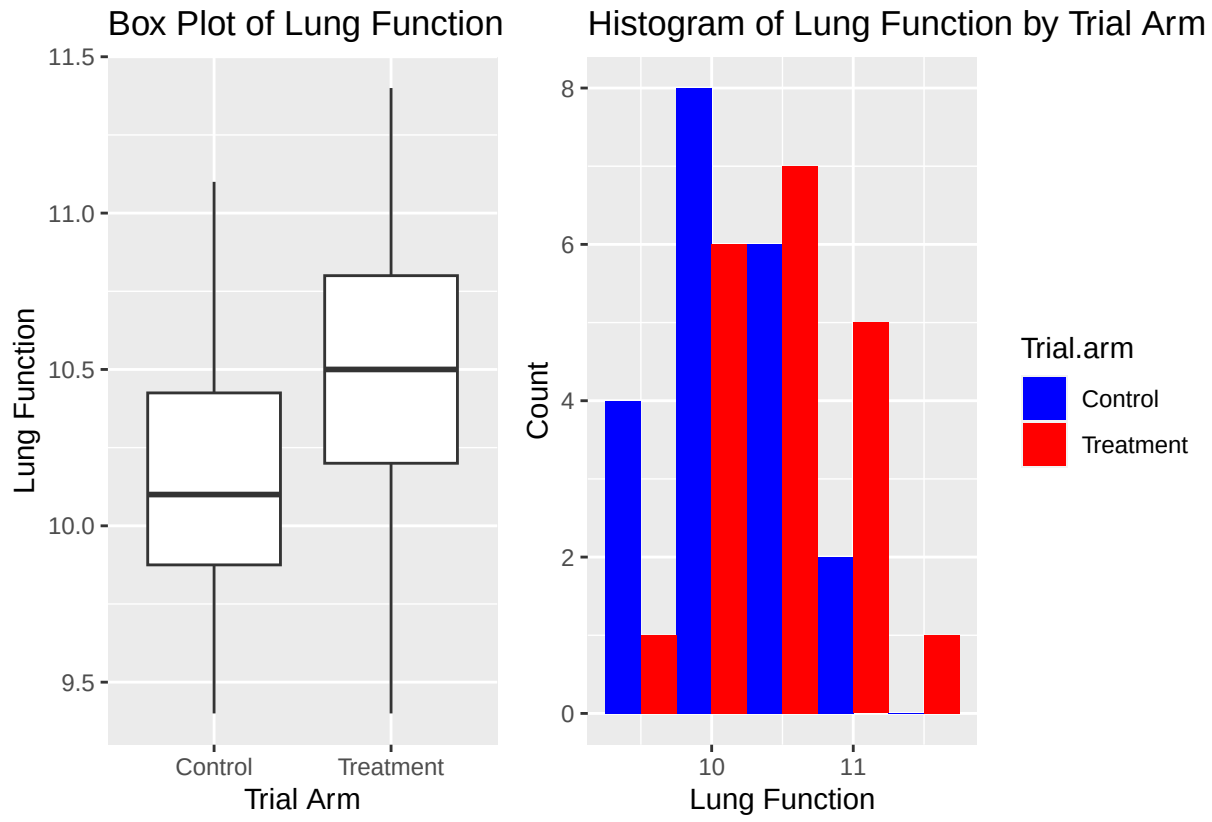
Subject.id	Lung.function	Trial.arm
1	11.0	Control
2	11.1	Control
3	9.5	Control
4	10.1	Control
5	9.7	Control
6	9.4	Control

```
# Box plot
box_plot <- ggplot(lung_data, aes(x = Trial.arm, y = Lung.function)) +
  geom_boxplot() +
  labs(title = "Box Plot of Lung Function by Trial Arm",
       x = "Trial Arm",
       y = "Lung Function")

# Histogram
histogram <- ggplot(lung_data, aes(x = Lung.function, fill = Trial.arm)) +
```

```
  geom_histogram(binwidth = 0.5, position = "dodge") +
  labs(title = "Histogram of Lung Function by Trial Arm",
       x = "Lung Function",
       y = "Count") +
  scale_fill_manual(values = c("Control" = "blue", "Treatment" = "red"))

# Arrange plots side by side
box_plot + histogram
```




```
# Separate lung function data for control and treatment groups
lung_function_control <- lung_data$Lung.function[lung_data$Trial.arm == "Control"]
lung_function_treatment <- lung_data$Lung.function[lung_data$Trial.arm == "Treatment"]

# Perform independent samples t-test
t_test_result <- t.test(lung_function_control, lung_function_treatment)

# Print the result
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  lung_function_control and lung_function_treatment
## t = -2.1569, df = 37.988, p-value = 0.0374
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.63003538 -0.01996462
## sample estimates:
## mean of x mean of y
##    10.150    10.475
```

2.2 Discussion - Report:

Objective: The aim of this report is to assess whether there is a significant difference in lung function between a control group and a treatment group using the Welch Two Sample t-test.

Data Description: The analysis compares lung function data between two groups: a control group (lung_function_control) and a treatment group (lung_function_treatment).

Results: The Two Sample t-test yielded a t-value of -2.1569 with a degrees of freedom (df) of approximately 37.988. The associated p-value is 0.0374, indicating statistical significance at the 0.05 level.

Hypothesis Testing: The null hypothesis, which posits that there is no difference in means between the control and treatment groups, is rejected. The alternative hypothesis suggests that the true difference in means is not equal to zero, implying that there is a significant difference in lung function between the two groups.

Confidence Interval: The 95 percent confidence interval for the difference in means ranges from -0.6300 to -0.0199. This interval suggests that we can be 95 percent confident that the true difference in means lies within this range.

Sample Estimates:

The mean lung function for the control group (mean of x) is estimated to be 10.150. The mean lung function for the treatment group (mean of y) is estimated to be 10.475. Conclusion: Based on the results of the Welch Two Sample t-test, there is evidence to suggest a statistically significant difference in lung function between the control and treatment groups. The treatment group appears to have a higher mean lung function compared to the control group. Further investigation into the efficacy of the treatment on lung function is warranted.

3rd Question: True/False Statements

Mark True or False for each of the following statements: ## p-values are used to calculate the probability of the null hypothesis given the data. **False**, p-values say how likely it is to observe the data (or even more extreme) given the null hypothesis is true, not the other way around.

The significance level alpha used in statistical hypothesis testing is the probability of rejecting the null when it is true.

True, it represents the probability of making a Type I error, which is the error of concluding that there is a significant effect or difference when there is none in reality . Therefore when it is very low (usually below 0.05), it means that the probability of there not being a significant difference is very low.

The Central Limit Theorem only holds if the population from which we are sampling is normally distributed.

False, even if the original variables are not normally distributed, under certain conditions, the distribution of a normalized version of the sample mean converges to a standard normal distribution . However, there are certain conditions and assumptions that need to be met for the CLT to hold, such as random sampling, finite variance, and sufficiently large sample size.

As the sample size gets larger, the standard error of the sampling distribution of the sample mean gets larger as well.

False, as the sample size gets larger, the standard error of the sampling distribution of the sample mean gets smaller, because the standard error is inversely proportional to the square root of the sample size. As the sample size increases, the variability of the sample mean decreases, resulting in a smaller standard error.

The statistical power of a hypothesis test is the probability of not rejecting the null when H1 is true.

False

The statistical power of a hypothesis test is the probability of rejecting H1 when H1 is true.

False

To answer both statements, they are both false. The power of a hypothesis test is the probability of making the correct decision if the alternative hypothesis is true. That is, the power of a hypothesis test is the probability of rejecting the null hypothesis when the alternative hypothesis is the hypothesis that is true. As we can see on the table, it is the probability of the bottom right case, rejecting the null hypothesis, or accepting the alternative hypothesis H1, when the null hypothesis is false, or H1 is true.

Table	H_0 is True	H_0 is False
Do not reject H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision