# Exploring the Impact of Extracurricular Activities on Secondary School Student Performance in Portugal

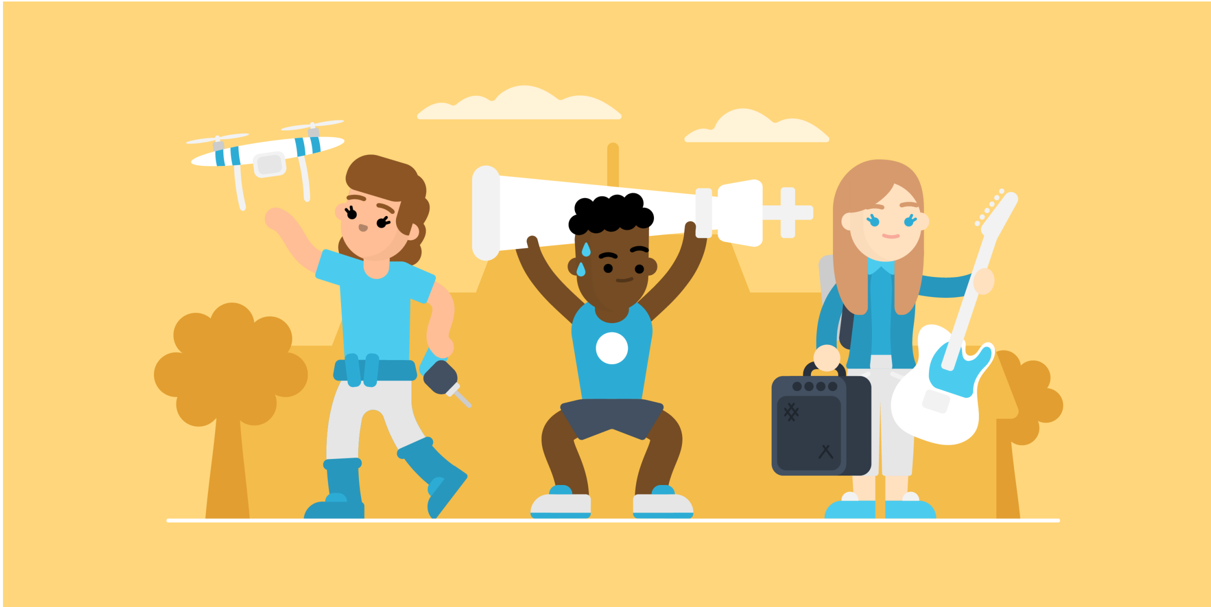

- **Authors**: Nariman Avakoli, Marcela Flaherty, Yi Tang, Brianna Zhou
- **Group**: 13


## 1. Introduction

Portugal's education system is characterized by low levels of education, high failure rates, and a significant early school dropout rate (Caetano, 2012). A strategic allocation of resources towards extracurricular activities could help improve academic performance among secondary students. Given the impact of education levels on GDP, growth, and child nutrition (Mukesh, Acharya & Pillai, 2023), it is crucial for Portugal to prioritize improving its educational system's performance.

This project aims to investigate the relationship between participation in extracurricular activities and academic performance among secondary school students in Portugal. The focus will be on students enrolled in core Mathematics or Portuguese language courses. By conducting two separate hypothesis tests, we will examine whether there is a significant difference in the final grades between students who participate in extracurricular activities and those who do not.

- Our population of interest: all secondary school students in Portugal who are enrolled in the Mathematics or Portuguese language core courses.
- The point estimate: the mean difference in final grades between students who participate in extracurricular activities and those who do not.
- Location parameter: mean final grades of students.
- Scale parameter: no scale parameter is needed because we are not contrasting the two groups directly comparing students in mathematics to those in Portuguese.



The data for this study, [Student Performance Data Set (https://archive.ics.uci.edu/dataset/320/student+performance)](https://archive.ics.uci.edu/dataset/320/student+performance), was sourced from the University of California, Irvine machine learning repository. The dataset comprises 33 attributes and 395 observations for students in the Mathematics course and 649 observations for students in the Portuguese course, including data on participation in extracurricular activities. There are no missing values in both datasets.

Of the variables, below are the two of interest:

- `activities` : extra-curricular activities (binary: yes or no)
- `G3` : final grade (numeric: from 0 to 20, output target)

**Is there a significant difference in the final grades between Portuguese secondary students enrolled in mathematics or Portuguese who participate in extracurricular activities and those who do not?**

- $H_0$: Mean final grades of students who participate in extracurricular activities and those who do not, are **equal**.
- $H_A$: Mean final grades of students who participate in extracurricular activities are **greater** than those who do not.

In this study, we expect to find a significant difference in final grades between students participating in extracurricular activities to those who do not in both the mathematics and Portuguese language courses. The analysis may show a correlation between extracurricular activities and final grades, but establishing causality may be a challenge due to potential confounding variables and interdependencies. Our analysis involves random processes, using fixed seeds and specifies the sampling procedure explicitly to ensure reproducibility of the analysis.

# 2. Methods and Results

## 2.1 Preliminary Results

### 2.1.1 Loading relevant libraries

In [1]:
```
# Run this cell before continuing.
library(cowplot)
library(digest)
library(gridExtra)
library(infer)
library(repr)
library(tidyverse)
library(datateachr)
library(broom)

# Set.seed for reproducibility
set.seed(50)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
✔ ggplot2 3.3.6     ✔ purrr   0.3.4
✔ tibble  3.1.8     ✔ dplyr   1.0.10
✔ tidyr   1.2.1     ✔ stringr 1.4.1
✔ readr   2.1.2     ✔ forcats 0.5.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::combine() masks gridExtra::combine()
✖ dplyr::filter()  masks stats::filter()
✖ dplyr::lag()     masks stats::lag()
```

### 2.1.2 Reading & Wrangling the datasets from the web into R

- Since we were dealing with 2 data sets, it is decided to read both the data sets separately: `mat_data` and `por_data`.

In [2]:
```
# Reading Math course data and convert it to a dataframe.
mat_data <- read_delim("https://raw.githubusercontent.com/Nariman-av/STAT-201-Group-13/m
    as.data.frame()
head(mat_data, 4)
colnames(mat_data)
mat_data %>% nrow()
```

```
Rows: 395 Columns: 33
── Column specification ────────────────────────────────────────────────────────
Delimiter: ";"
chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
dbl (16): age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

A data.frame: 4 × 33

| | school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | ⋯ | famrel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | <chr> | <chr> | <dbl> | <chr> | <chr> | <chr> | <dbl> | <dbl> | <chr> | <chr> | ⋯ | <dbl> |
| **1** | GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | ⋯ | 4 |
| **2** | GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | ⋯ | 5 |
| **3** | GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | ⋯ | 4 |
| **4** | GP | F | 15 | U | GT3 | T | 4 | 2 | health | services | ⋯ | 3 |

'school' 'sex' 'age' 'address' 'famsize' 'Pstatus' 'Medu' 'Fedu' 'Mjob' 'Fjob' 'reason' 'guardian' 'traveltime' 'studytime' 'failures' 'schoolsup' 'famsup' 'paid' 'activities' 'nursery' 'higher' 'internet' 'romantic' 'famrel' 'freetime' 'goout' 'Dalc' 'Walc' 'health' 'absences' 'G1' 'G2' 'G3'

395

Data Frame 1

In [3]:
```
# Reading Portuguese course data and convert it to a dataframe.
por_data <- read_delim("https://raw.githubusercontent.com/Nariman-av/STAT-201-Group-13/m
    as.data.frame()
head(por_data, 4)
colnames(por_data)
por_data %>% nrow()
```

```
Rows: 649 Columns: 33
── Column specification ────────────────────────────────────────────────────────
Delimiter: ";"
chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
dbl (16): age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

A data.frame: 4 × 33

| | school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | ⋯ | famrel |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | <chr> | <chr> | <dbl> | <chr> | <chr> | <chr> | <dbl> | <dbl> | <chr> | <chr> | ⋯ | <dbl> |
| **1** | GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | ⋯ | 4 |
| **2** | GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | ⋯ | 5 |
| **3** | GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | ⋯ | 4 |
| **4** | GP | F | 15 | U | GT3 | T | 4 | 2 | health | services | ⋯ | 3 |

'school' 'sex' 'age' 'address' 'famsize' 'Pstatus' 'Medu' 'Fedu' 'Mjob' 'Fjob' 'reason' 'guardian' 'traveltime' 'studytime' 'failures' 'schoolsup' 'famsup' 'paid' 'activities' 'nursery' 'higher' 'internet' 'romantic' 'famrel' 'freetime' 'goout' 'Dalc' 'Walc' 'health' 'absences' 'G1' 'G2' 'G3'

649

Data Frame 2

In [4]:
```
# Select the two attribes, final grade and extracurricular activity participation.
mat_data <- mat_data %>%
    select(activities, G3)


head(mat_data)
```

A data.frame: 6 × 2

| | activities | G3 |
|---|---|---|
| | <chr> | <dbl> |
| **1** | no | 6 |
| **2** | no | 6 |
| **3** | no | 10 |
| **4** | yes | 15 |
| **5** | no | 10 |
| **6** | yes | 15 |

Data Frame 3

In [5]:
```
# Select the two attribes, final grade and extracurricular activity participation.
por_data <- por_data %>%
    select(activities, G3)


head(por_data)
```

A data.frame: 6 × 2

| | activities | G3 |
|---|---|---|
| | <chr> | <dbl> |
| **1** | no | 11 |
| **2** | no | 11 |
| **3** | no | 12 |
| **4** | yes | 14 |
| **5** | no | 13 |
| **6** | yes | 13 |

Data Frame 4

### 2.1.3 Exploratory Data Analysis

- We begin with the summary tables for `mat_data` and `por_data` .

In [6]:
```
# Summary table for mathematics course.
summary(mat_data)

# Summary table for Portuguese course.
summary(por_data)
```

```
  activities              G3       
 Length:395         Min.   : 0.00  
 Class :character   1st Qu.: 8.00  
 Mode  :character   Median :11.00  
                    Mean   :10.42  
                    3rd Qu.:14.00  
                    Max.   :20.00  

  activities              G3       
 Length:649         Min.   : 0.00  
 Class :character   1st Qu.:10.00  
 Mode  :character   Median :12.00  
                    Mean   :11.91  
                    3rd Qu.:14.00  
                    Max.   :19.00  
```

In [7]:
```
# Point estimates of mean, median, and standard deviation for math course
mat_data_summary <- mat_data %>%
    group_by(activities) %>%
    summarize(Mean_mat_grade = mean(G3),
              Median_mat_grade = median(G3),
              SD_mat_grade = sd(G3),
              Num_mat_students = n())

mat_data_summary
```

A tibble: 2 × 5

| activities | Mean_mat_grade | Median_mat_grade | SD_mat_grade | Num_mat_students |
|---:|---:|---:|---:|---:|
| <chr> | <dbl> | <dbl> | <dbl> | <int> |
| no | 10.34021 | 11 | 4.488065 | 194 |
| yes | 10.48756 | 11 | 4.679861 | 201 |

Table 1

In [8]:
```
# Point estimates of mean, median, and standard deviation for portuguese course
por_data_summary <- por_data %>%
    group_by(activities) %>%
    summarize(Mean_por_grade = mean(G3),
              Median_por_grade = median(G3),
              SD_por_grade = sd(G3),
              Num_por_students = n())

por_data_summary
```

A tibble: 2 × 5

| activities | Mean_por_grade | Median_por_grade | SD_por_grade | Num_por_students |
|---:|---:|---:|---:|---:|
| <chr> | <dbl> | <dbl> | <dbl> | <int> |
| no | 11.71856 | 11 | 3.235290 | 334 |
| yes | 12.10476 | 12 | 3.218944 | 315 |

Table 2

In [9]:
```
# Create a distribution of mathematics grades.
mat_data %>%
    ggplot(aes(x = G3)) +
    geom_histogram(bins = 20, color = "white") +
    facet_grid(activities ~ .) +
    labs(title = "Math Grade Distribution Based on Participation in Extracurricular Acti
    theme(text = element_text(size = 10),
          title = element_text(size = 11), plot.caption = element_text(size = 10, hjust
```

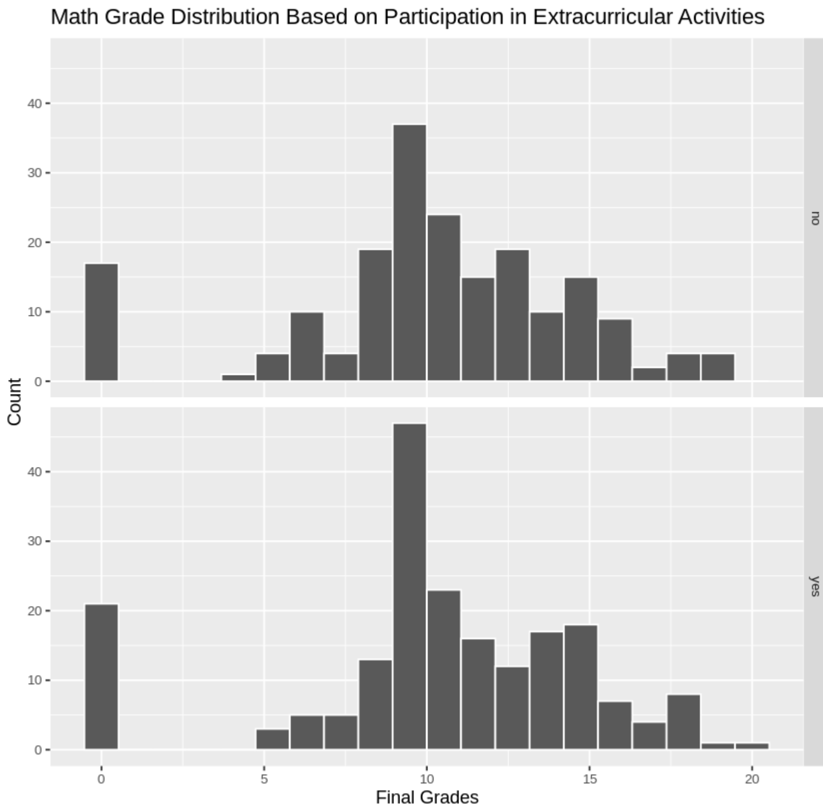


Figure 1

In [10]:
```
# Create a distribution of Portuguese grades.
por_data %>%
    ggplot(aes(x = G3)) +
    geom_histogram(bins = 20, color = "white") +
    facet_grid(activities ~ .) +
    labs(title = "Portuguese Grade Distribution Based on Participation in Extracurricula
    theme(text = element_text(size = 10),
          title = element_text(size = 11),plot.caption = element_text(size = 10, hjust =
```

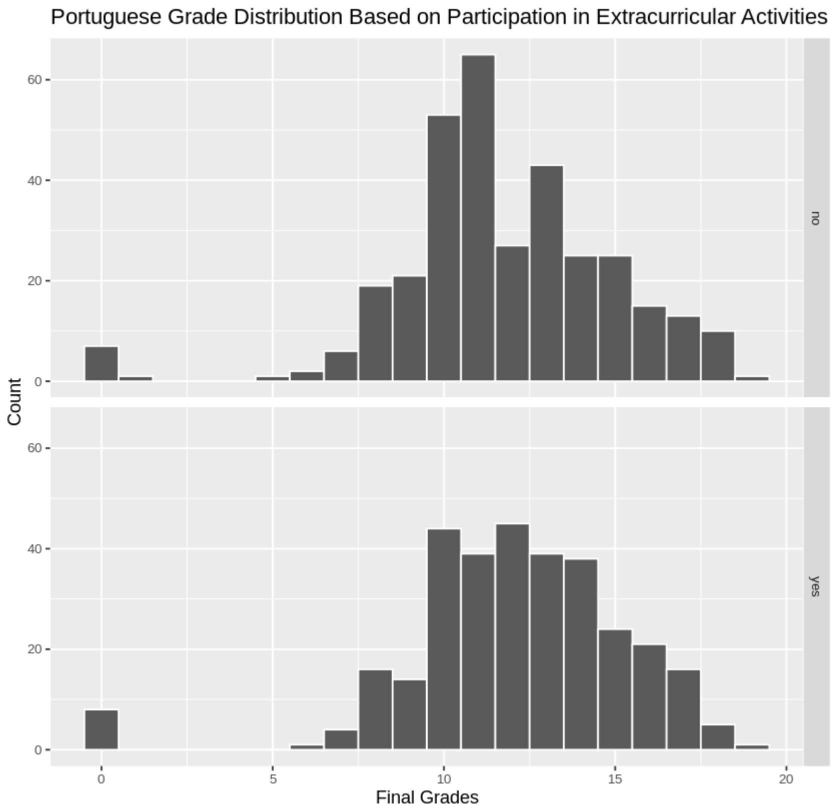


Figure 2

- Final Grades (G3) for both courses, ranging from 0 to 20, mainly fall between 7.5 and 16, with notable outliers at 0. The distributions are roughly bell-shaped and symmetric, suggesting normality, with the unimodal peak indicating the data's central tendency.

In [11]:
```
# Create a boxplot of Math grades based on participation in extracurricular activities
mat_box_plot<-
    mat_data %>%
        ggplot(aes(x = activities, y = G3, fill = activities)) +
        geom_boxplot() +
        labs(x = "Participation in Extracurricular Activities", y = "Math Final Grade (G
        title = "Impact of Extracurricular Activities on Student Performance on Math",ca
        theme(text = element_text(size = 11),plot.caption = element_text(size = 10, hjus
mat_box_plot
```

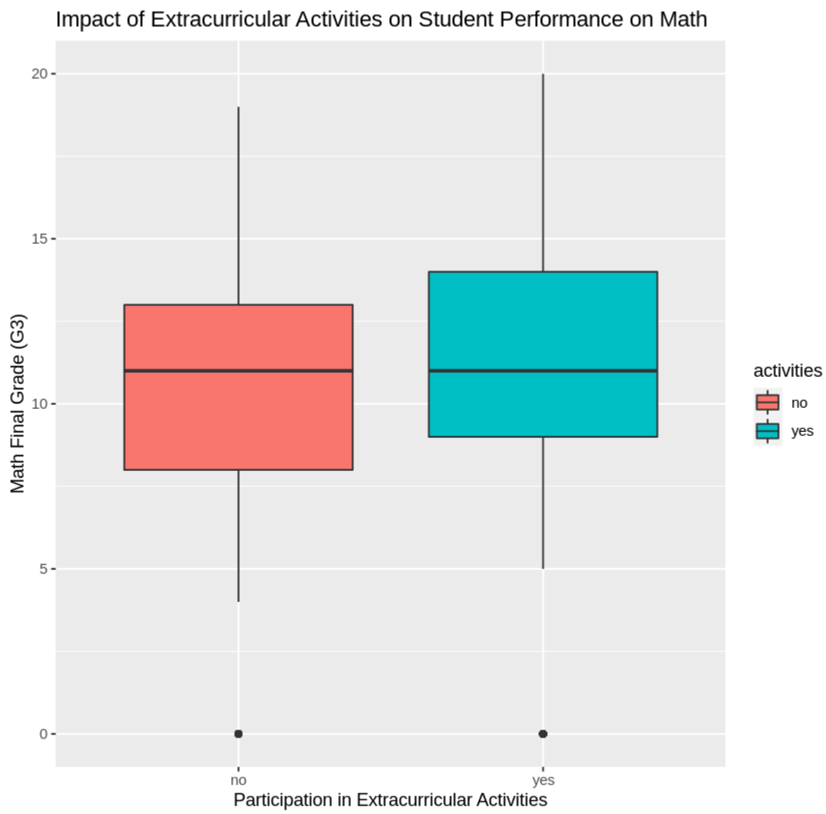


Figure 3

In [12]:
```
mat_diff <-
    mat_data_summary %>%
    ungroup() %>%
    summarize(Math_mean_yes = Mean_mat_grade[activities == "yes"],
              Math_mean_no = Mean_mat_grade[activities == "no"],
              Math_mean_difference = Mean_mat_grade[activities == "yes"] - Mean_mat_grac

mat_diff
```

A tibble: 1 × 3

| Math_mean_yes | Math_mean_no | Math_mean_difference |
|---:|---:|---:|
| <dbl> | <dbl> | <dbl> |
| 10.48756 | 10.34021 | 0.147356 |

Table 3

In [13]:
```
# Create a boxplot of Portuguese grades based on participation in extracurricular activi
por_box_plot <-
    por_data %>%
        ggplot(aes(x = activities, y = G3, fill = activities)) +
        geom_boxplot() +
        labs(x = "Participation in Extracurricular Activities", y = "Portuguese Final Gr
        title = "Impact of Extracurricular Activities on Student Performance on Portugue
        theme(text = element_text(size = 11),plot.caption = element_text(size = 10, hjus
por_box_plot
```

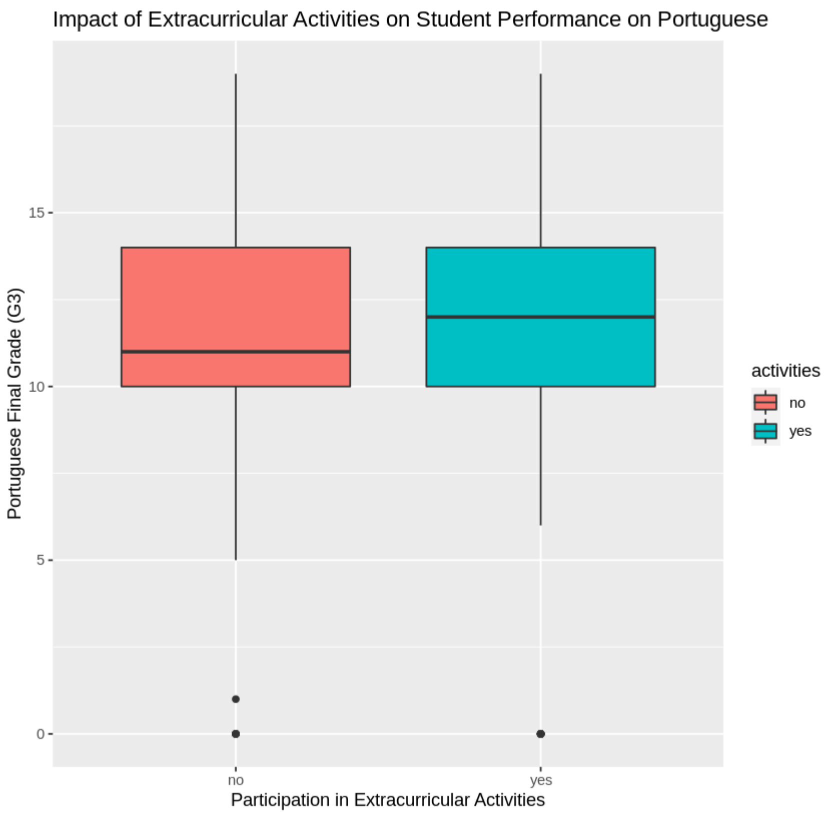


Figure 4

In [14]:
```
por_diff <-
    por_data_summary %>%
    ungroup() %>%
    summarize(Por_mean_yes = Mean_por_grade[activities == "yes"],
              Por_mean_no = Mean_por_grade[activities == "no"],
              Por_mean_difference = Mean_por_grade[activities == "yes"] - Mean_por_grade

por_diff
```

A tibble: 1 × 3

| Por_mean_yes | Por_mean_no | Por_mean_difference |
|---|---|---|
| <dbl> | <dbl> | <dbl> |
| 12.10476 | 11.71856 | 0.386199 |

Table 4

- In both Mathematics and Portuguese courses, students participating in extracurricular activities have slightly higher mean final grades than those not participating, with mean differences of 0.15 and 0.39 respectively. The boxplots (Figures 3 and 4) show that the grade ranges for students in activities are higher in Mathematics and slightly narrower in Portuguese compared to those not participating, with outliers at 0 in both courses.

## 2.2 Methods

- Our plots and estimates produced offer valuable analysis, but they may not be comprehensive enough for stakeholders. Therefore, we will implement a bootstrap distribution, an asymptoptic distribution using the Central Limit Theorem, and significance testing at a 5% level alongside confidence intervals for a more precise evaluation of the mean difference. We chose a 5% significance level because it is a convention that has been widely adopted in many scientific fields, and it is the probability of a type I error that we are willing to accept.
- $\mu_{\text{activity\_yes}}$ is the mean final grade in a given course for student who participate in extracurricular activities $\texttt{activities == "yes"}$
- $\mu_{\text{activity\_no}}$ is the mean final grade in a given course for student who do not participate in extracurricular activities $\texttt{activities == "no"}$ .

$$
\begin{aligned}
H_0 &: \mu_{\text{activity\_yes}} - \mu_{\text{activity\_no}} = 0 \\
H_1 &: \mu_{\text{activity\_yes}} - \mu_{\text{activity\_no}} > 0
\end{aligned}
$$

- $\alpha$ is the significance level.

$$\alpha = 0.05$$

### 2.2.1 Bootstrapped Hypothesis Testing for Mathematics Final Grades Using Mean Differences

- In this section, we will work with two samples: secondary school students in Portugal who are enrolled in the mathematics that participate in extracurricular activities; and secondary school students in Portugal who are enrolled in mathematics that do not participate in extracurricular activities.
- $\mu_{\text{activity\_yes}}$ is the mean mathematics final grade for student who participate in extracurricular activities $\texttt{activities == "yes"}$
- $\mu_{\text{activity\_no}}$ is the mean mathematics final grade for student who do not participate in extracurricular activities $\texttt{activities == "no"}$ .

- Assumptions & conditions:
  - The sample is randomly drawn from the population
  - The sample values are independent
  - The sample size is large enough (n > 30)

- The sample is representative of the population.
- Resampling is done with replacement.
- The bootstrap sample size is the same as the original sample size.

- We will calculate the observed test statistic as the difference in the mean mathematics final grades between `activities == "yes"` & `activities == "no"`:

In [15]:
```
# Calculate the observed test statistic, mean difference
obs_mean_mat_diff <-
    mat_diff %>%
    pull(Math_mean_difference)

obs_mean_mat_diff
```

0.147356003487715

- Using the infer workflow to obtain the bootstrap sampling distribution of mean differences of math final grades.
  - The bootstrap sample size is used for sampling with replacement, and it is set to be the same as the actual sample size.
  - There are 5000 replications to obtain 5000 separate bootstrap samples.
    - For each replicated bootstrap sample, we calculated the point estimate to obtain a bootstrap sampling distribution of the difference in mean math final grades.

In [16]:
```
# Create a bootstrap distribution of the difference in means with replacement and 5000 r
bootstrap_distribution_mat <-
   mat_data %>%
   specify(formula = G3 ~ activities) %>%
   generate(reps = 5000, type = "bootstrap") %>%
   calculate(stat="diff in means", order = c("yes", "no"))

head(bootstrap_distribution_mat)
```

A infer: 6 × 2

| replicate | stat |
|---|---|
| <int> | <dbl> |
| 1 | 0.77627302 |
| 2 | 0.67416076 |
| 3 | 0.26440261 |
| 4 | 0.49819429 |
| 5 | 0.02700689 |
| 6 | 0.33017268 |

Table 5

- Visualize the results using a histogram and overlay the observed statistic `obs_mean_mat_diff` to see where it lies in the distribution.
  - `visualize()` was used to visualize the resulting bootstrap distribution.
  - `shade_p_value()` with `obs_stat` set to the observed test statistic value `obs_mean_mat_diff` and the `direction = “right”` reflecting our one-sided alternative hypothesis.

In [17]:
```
# Create a plot of the bootstrap distribution highlighting the p value and tail
mat_result_plot <-
   bootstrap_distribution_mat %>%
   visualize() +
   shade_p_value(obs_stat = obs_mean_mat_diff, direction = "right") +
   labs(x = "Difference in Means for Final Grades", y = "Count",
        title = "Mathematics 5000 Sample Bootstrap Distribution and Observed Statistic",
    theme(text = element_text(size = 11),plot.caption = element_text(size = 10, hjust =

mat_result_plot
```

Mathematics 5000 Sample Bootstrap Distribution and Observed Statistic

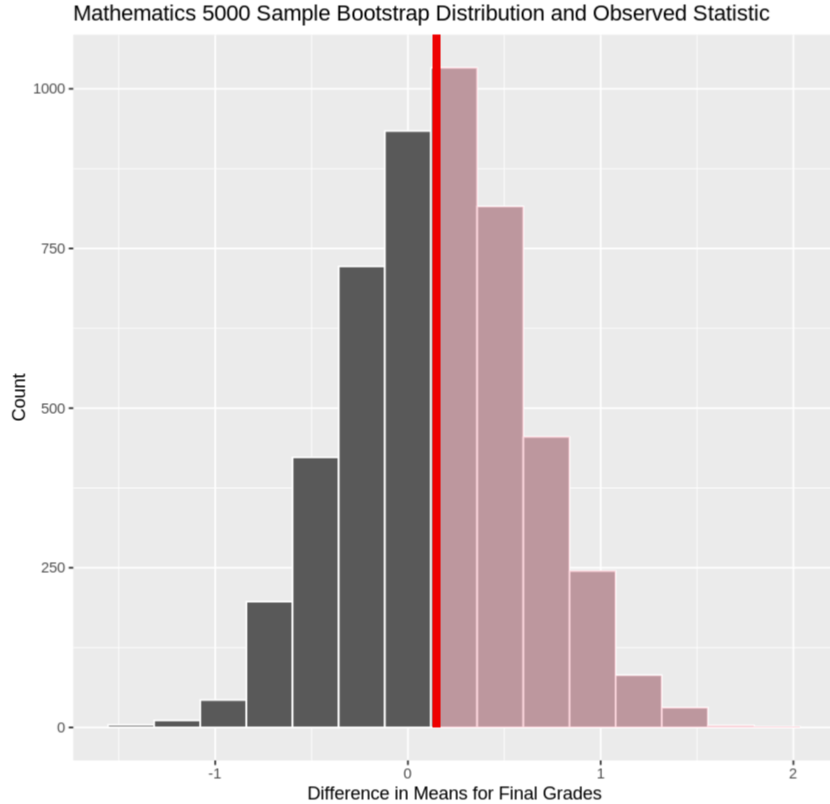


Figure 5

- Obtain the p-value from `bootstrap_distribution_mat` using `get_p_value()` following our set arguments for `obs_stat` and `direction`.
- We will assume the null hypothesis is true.

In [18]:
```
# Calculate the p value with a one tailed test
bootstrap_distribution_mat %>%
  get_p_value(obs_stat = obs_mean_mat_diff, direction = "right")
```

A tibble: 1 × 1

| p_value |
| --- |
| <dbl> |
| 0.51 |

Table 6

- Assuming the null hypothesis is true, there is a probability of approximately 0.51 of observing a t-value as extreme as, or more extreme than the one computed from the sample.
- Since the p-value is larger than our pre-specified significance level $\alpha = 0.05$, **we fail to reject the null hypothesis** $H_0$.

## 2.2.2 Estimating Confidence Intervals for Mean Differences Using Bootstrap Distribution

- We obtain a 95% confidence interval for the mean difference in mathematics final grades across the two groups.

In [19]:
```
# Calculate the confidence interval with level 0.95
percentile_ci_mat <- bootstrap_distribution_mat %>%
  get_confidence_interval(level = 0.95, type = "percentile")

percentile_ci_mat
```

A tibble: 1 × 2

| lower_ci | upper_ci |
| --- | --- |
| <dbl> | <dbl> |
| -0.7265731 | 1.069669 |

Table 7

- We are 95% confident that the true mean difference in mathematics final grades between the two groups of students participating in extracurricular activities and those not participating is captured in the interval [-0.7265731, 1.069669].
- If we repeated our sampling many times, we'd expect the true grade difference to fall in this range about 95% of the time. However, since 0 is within this interval, it's possible that extracurriculars might not have a clear impact on grades.

- We will visualize the percentile-based 95% confidence interval with the bootstrapping distribution of mean differences.

In [20]:
```
# Create a plot of the bootstrap distribution highlighting the confidence interval
visualize(bootstrap_distribution_mat) +
  shade_confidence_interval(endpoints = percentile_ci_mat)+
  labs(x = "Difference in Means for Final Grades", y = "Count",
        title = "Mathematics 5000 Sample Bootstrap Distribution and Confidence Interval"
    theme(text = element_text(size = 11),plot.caption = element_text(size = 10, hjust =
```

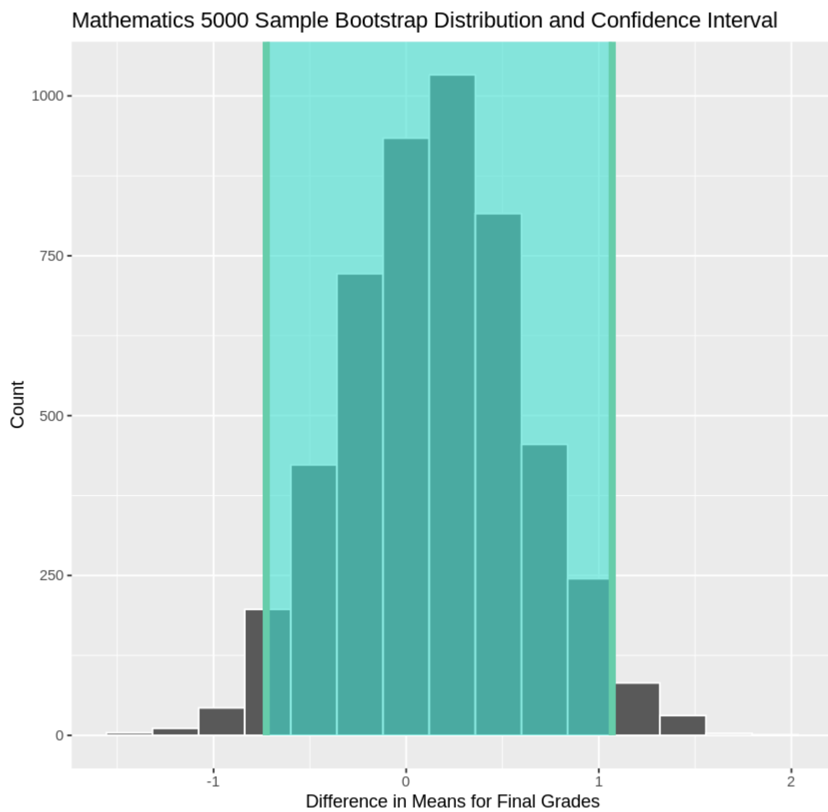


Figure 6

- Given our one-tailed test, the confidence interval indicates that while students participating in extracurriculars might have higher mean mathematics grades, a difference of 0 suggests there could also be no difference.

### 2.2.3 Asymptoptic Hypothesis Testing for Portuguese Final Grades Using Mean Differences

- In this section, we will work with two samples: secondary school students in Portugal who are enrolled in the Portuguese that participate in extracurricular activities; and secondary school students in Portugal who are enrolled in Portuguese that do not participate in extracurricular activities.
- $\mu_{\text{activity\_yes}}$ is the mean Portuguese final grade for student who participate in extracurricular activities `activities == "yes"`
- $\mu_{\text{activity\_no}}$ is the mean Portuguese final grade for student who do not participate in extracurricular activities `activities == "no"` .

- Assumptions & conditions:
  - The sample is randomly drawn from the population.
  - The sample values are independent.
  - The sample size is large enough (n > 30).

- We will use a one tailed two-sample t-test, and first calculate mean, standard deviation, and the number of rows for each activity group.

In [21]:
```
# Extract the values from por_data_summary
mean_yes <- por_data_summary$Mean_por_grade[2]
mean_no <- por_data_summary$Mean_por_grade[1]
sd_yes <- por_data_summary$SD_por_grade[2]
sd_no <- por_data_summary$SD_por_grade[1]
n_yes <- por_data_summary$Num_por_students[2]
n_no <- por_data_summary$Num_por_students[1]
```

- Calculate this test statistic:

$$T = \frac{\bar{x}_{\text{activity\_yes}} - \bar{x}_{\text{activity\_no}}}{\sqrt{\frac{s^2_{\text{activity\_yes}}}{n_1} + \frac{s^2_{\text{activity\_no}}}{n_2}}}$$

In [22]:
```
# Calculate the test statistic
test_stat <- (mean_yes - mean_no) / sqrt((sd_yes^2 / n_yes) + (sd_no^2 / n_no))

test_stat
```

1.52381914123355

- We will assume the null hypothesis is true.
- The $T$ statistic follows a $t$ distribution with approximately

$$v = \frac{\left(\frac{s^2_{activity yes}}{n_1} + \frac{s^2_{activity no}}{n_2}\right)^2}{\frac{s^4_{activity yes}}{\cdots} \quad \frac{s^2_{activity\ldots}}{\cdots}}$$

In [23]:
```
# Calculate degrees of freedom
df <- ((sd_yes^2/n_yes + sd_no^2/n_no)^2) / ((sd_yes^4 / (n_yes^2*(n_yes-1))) + (sd_no

df
```

645.14626605344

- Obtain the p-value for this hypothesis test using $v \approx 645$.

In [24]:
```
# P-value calculation for a one-sided test
p_value <- pt(test_stat, df = df, lower.tail = FALSE)

p_value
```

0.0640219571627865

- Assuming the null hypothesis is true, there is a probability of approximately 0.06 of observing a t-value as extreme as, or more extreme than approximately 1.52.
- At a 5% significance level, the `p-value` is greater than 0.05. Therefore, **we fail to reject the null hypothesis** $H_0$. This suggests that there is insufficient evidence to claim that the mean Portuguese final grade for students who participate in extracurricular activities is higher than that of students who do not participate in such activities.

### 2.2.4 Estimating Confidence Intervals of Portuguese Final Grades Using Central Limit Theorem

- Obtain a 95% confidence interval for the mean differences in Portuguese final grades between the two groups.

In [25]:
```
activity_yes_students <- por_data_summary$Mean_por_grade[2]
activity_no_students <- por_data_summary$Mean_por_grade[1]

# Sample variances and sizes
var_yes <- por_data_summary$SD_por_grade[2]^2
var_no <- por_data_summary$SD_por_grade[1]^2

# Difference in means
diff_means <- activity_yes_students - activity_no_students

# SD for the difference in means
std_error_diff <- sqrt(var_yes/n_yes + var_no/n_no)

# 95% confidence interval for the difference in means using CLT
ci_lower <- diff_means - qnorm(0.975) * std_error_diff
ci_upper <- diff_means + qnorm(0.975) * std_error_diff

# CI
por_diff_means_ci <- tibble(
  lower_ci = ci_lower,
  upper_ci = ci_upper
)
por_diff_means_ci
```

A tibble: 1 × 2

| lower_ci | upper_ci |
|---|---|
| <dbl> | <dbl> |
| -0.1105372 | 0.8829353 |

Table 8

- We are 95% confident that the true mean difference in Portuguese final grades between the two groups of students participating in extracurricular activities and those not participating is captured in the interval [-0.1105372, 0.8829353].
- If we repeated our sampling many times, we'd expect the true grade difference to fall in this range about 95% of the time. However, since 0 is within this interval, it's possible that extracurriculars might not have a clear impact on grades.

## 2.2.5 Bootstrap vs. Asymptotics

Bootstrapping:

- Results: The confidence interval for the difference in mathematics grades is [-0.726, 1.069] with a p-value of approxiamtely 0.51.

Central Limit Theorem (CLT):

- Results: The confidence interval for the difference in Portuguese grades is [-0.1105, 0.8829] with a p-value of approximately 0.06.

Analysis:

- The CLT states that if the sample size is large enough, sample values are independent, and the sample is randomly drawn, the resulting sampling distribution will converge to an approximately normal distribution. This is true even if the underlying population distribution isn't normal. With our substantial sample size (n > 50) and lack of skew, the t-distribution closely mirrors the normal distribution, potentially providing a better approximation of the sampling distribution.
- The bootstrap is beneficial regarding uncertainty about the distribution, offering flexibility and lack of assumption. Our sample size is large, and our data distribution follows normality, so a bootstrapping distribution is not necessarily more beneficial than the CLT.
- Both methods yielded confidence intervals that include zero, suggesting no significant difference in grades based on participation in extracurricular activities.
- The p-values differ, with bootstrapping suggesting a stronger lack of significance than the CLT. However, both p-values are above the 0.05 threshold, leading to the same conclusion of not rejecting the null hypothesis.

Which is more appropriate?:

- Based on the comparison of confidence intervals and p-values, **we consider the CLT method more appropriate than bootstrapping**.
- We also consider the CLT method, due to its alignment with the t-distribution for large samples, to be more trustworthy in this context. However, since bootstrapping, a method with fewer assumptions, produced the same conclusion adds reliability to our analysis.

# 3. Discussion

In summing up our findings, the p-values for the mean differences in final grades between students engaged in extracurricular activities and those who weren't are not statistically significant (p-value > 0.05). This failure to reject our null hypothesis suggests that within our data's specific context, extracurricular involvement does not significantly affect academic performance in terms of final grades. Educators and policymakers could interpret these findings as an opportunity to reevaluate the role of extracurricular activities, understanding that they may not enhance grades directly, but might still be beneficial in other aspects of student growth and development.

Our study was primarily designed to determine whether extracurricular activities influenced students' final grades in both mathematics and Portuguese courses. We expected to find a significant difference in final grades between students participating in extracurricular activities to those who do not in both the mathematics and Portuguese language courses. The results did not align with this expectation. While our analysis did imply a correlation, it did not establish a causative relationship between participation in these activities and improved academic performance.

The lack of statistical significance leads to further questions:

- Is there a positive correlation between participation in extracurricular activities and overall academic performance in secondary school?

- What non-academic benefits do extracurricular activities offer, and how do they contribute to a student's overall development?
- Are certain types of extracurricular activities more influential on students' academic performance than others?

## 4. References


Caetano, L. (2012). Abandono escolar: Repercussões sócio-económicas na região centro. algumas reflexões. Finisterra, 40(79). https://doi.org/10.18055/Finis1503 (https://doi.org/10.18055/Finis1503)

Mukesh, H. V., Acharya, V., & Pillai, R. (2023). Are extracurricular activities stress busters to enhance students' well-being and academic performance? evidence from a natural experiment. Journal of Applied Research in Higher Education, 15(1), 152-168. https://doi.org/10.1108/JARHE-06-2021-0240 (https://doi.org/10.1108/JARHE-06-2021-0240)