FA 2: Estimation - The Influence of Internet Access on Math Performance: An Inferential Analysis

Madhumitha Sridhar

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)

# loading needed libraries
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(infer)
```

Introduction

Background

Education is a key driver in social mobility and economic growth and understanding the factors that influence student performance is vital for developing effective educational practices and interventions. In today's fast-paced, digital age, internet access has become an increasingly important for educational purposes, as it provides students with resources that go beyond traditional classroom materials.

Research Question

This report aims to investigate whether internet access at home has a significant impact on students' performance in mathematics. Specifically, we will examine:

1. Is there a significant difference in mean final math grades between students with and without internet access at home?
2. Is there a significant difference in the proportion of high-achieving students (those with grades above 15 out of 20) between students with and without internet access?

Population, Sample, and Data Source

Population: Secondary students studying mathematics **Sample**: Students from 2 Portugese secondary schools (abriel Pereira and Mousinho da Silveira) **Data Source**: This dataset was collected by Paulo Cortez and Alice Silva through both school reports and questionnaires. It contains various student attributes including demographics, social factors, and academic performance measurements.

Choice of Statistics

For this analysis, I will be using two different types of statistics:

1. **Difference in means**: Will be used to understand if internet access is associated with an overall differnce in avergae math performance
2. **Difference in proportions**: Will be used to examine if internet access is associated with a higher likelihood of being a high-achieving student.

These statistics will allow us to understand both the avergae effect and the effect of excellence rates, providing us with a more comprehensive picture than a standalone statistic can provide.

Data Preparation

First, we will start off by loading and examining the data:

```
student_data <- read.csv("student-mat.csv", sep=";")
```

```
# examining the structure with this code
glimpse(student_data)
```

```
Rows: 395
Columns: 33
$ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP",~
$ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "F",~
$ age        <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 15,~
$ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U",~
$ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", "LE~
$ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "T",~
$ Medu       <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4,~
$ Fedu       <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, 3,~
$ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "servic~
$ Fjob       <chr> "teacher", "other", "other", "services", "other", "other", ~
$ reason     <chr> "course", "course", "other", "home", "home", "reputation", ~
$ guardian   <chr> "mother", "father", "mother", "mother", "father", "mother",~
$ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1,~
$ studytime  <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1,~
$ failures   <int> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0,~
$ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no", "n~
$ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "yes",~
$ paid       <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes", ~
$ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "ye~
$ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "yes~
$ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "ye~
$ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes",~
$ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "no"~
$ famrel     <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3,~
$ freetime   <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, 1,~
$ goout      <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, 3,~
$ Dalc       <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,~
$ Walc       <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, 3,~
$ health     <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, 5,~
$ absences   <int> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, 16,~
$ G1         <int> 5, 5, 7, 15, 6, 15, 12, 6, 16, 14, 10, 10, 14, 10, 14, 14, ~
$ G2         <int> 6, 5, 8, 14, 10, 15, 12, 5, 18, 15, 8, 12, 14, 10, 16, 14, ~
$ G3         <int> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 14,~
```

```r
# next, let's look at the basic summary of this dataset
summary_stats <- student_data |>
  summarize(
    total_students = n(),
    with_internet = sum(internet == "yes"),
    without_internet = sum(internet == "no"),
```

```
    avg_final_grade = mean(G3),
    med_final_grade = median(G3)
  )

summary_stats
```

```
  total_students with_internet without_internet avg_final_grade med_final_grade
1            395           329                66        10.41519              11
```

The summary statistics above reveal that the dataset has 395 students, with 329 of them having internet access and 66 without, showing a significant imbalance between the two groups. The average final grade across all students is approximately 10.42/20, with a median of 11, which suggests a slightly left-skewed distribution of grades.

To make our analysis easier, we should create a binary variable for high-achieving students (those with G3 $>$ 15/20):

```
student_data <- student_data |>
  mutate(high_achiever = ifelse(G3 > 15, "yes", "no"))
```

```
# count of high achievers overall
high_achievers_summary <- student_data |>
  group_by(high_achiever) |>
  summarize(count = n()) |>
  mutate(percentage = count / sum(count) * 100)

high_achievers_summary
```

```
# A tibble: 2 x 3
  high_achiever count percentage
  <chr>         <int>      <dbl>
1 no              355       89.9
2 yes              40       10.1
```

The results above show that only 40 students (10.13% of the total) qualify as high achievers, while the vast majority (355 students or 89.87%) fall below this threshold.

Exploratory Analysis

Internet Access Distribution

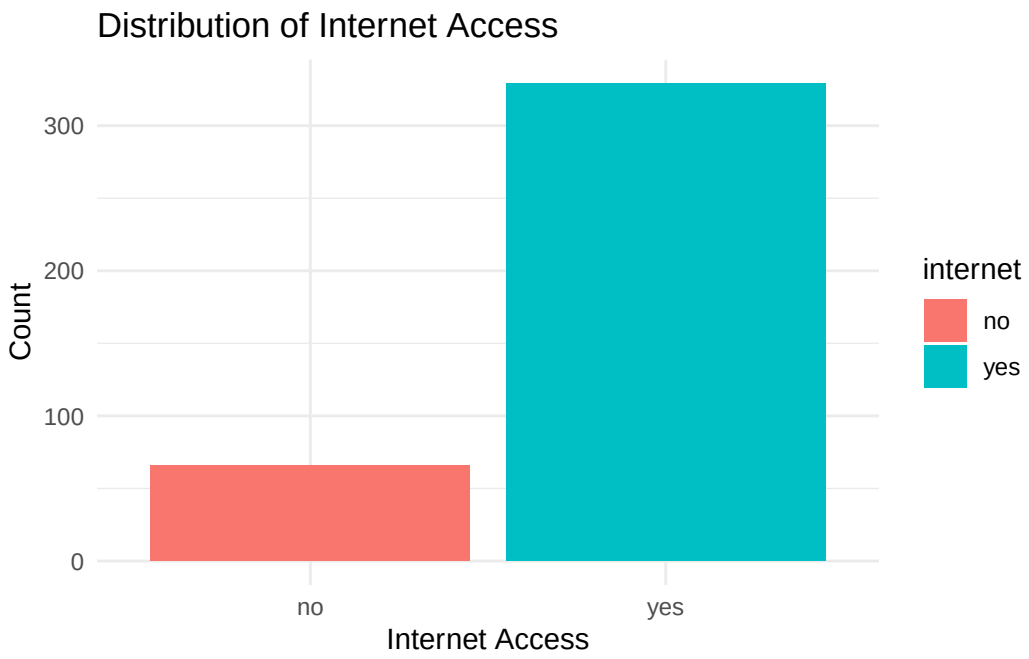
```
internet_dist <- student_data |>
  count(internet) |>
  mutate(percentage = n / sum(n) * 100)

# displaying results from code above
internet_dist
```

```
  internet   n percentage
1       no  66   16.70886
2      yes 329   83.29114
```

Here, we can see that 329 students (83.29% of the sample) have internet access at home, while only 66 students (16.71%) do not. This substantial disparity in internet access highlights that having internet at home could be the norm among the studied student population.

Now, let's visualize the internet access distribution.

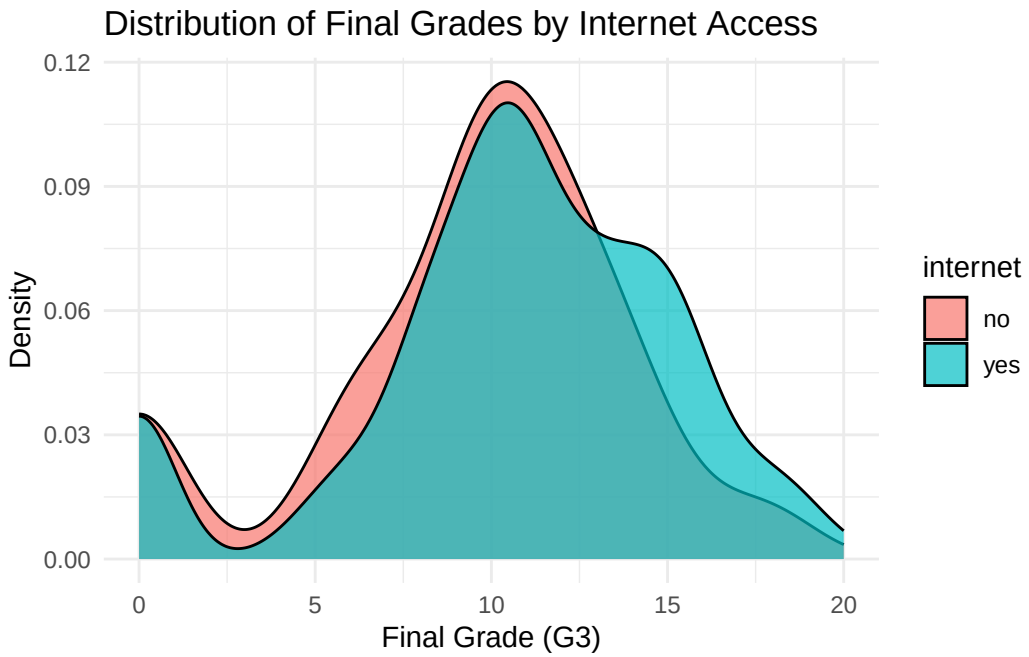
```
ggplot(student_data, aes(x = internet, fill = internet)) +
  geom_bar() +
  labs(title = "Distribution of Internet Access",
       x = "Internet Access",
       y = "Count") +
  theme_minimal()
```



This bar graph shows the distribution of internet access among students, showing that a significantly large number of students (about 325) have internet access compared to those without (about 65). This uneven distribution suggests that internet access is common among the student population, with roughly 5 times more students having access than not, which is important context in our analysis of the relationship between internet access and academic performance.

Final Grade Distribution by Internet Access

```
# visualizing grade distribution by internet access
ggplot(student_data, aes(x = G3, fill = internet)) +
  geom_density(alpha = 0.7) +
  labs(title = "Distribution of Final Grades by Internet Access",
       x = "Final Grade (G3)",
       y = "Density") +
  theme_minimal()
```



The density plot we generated above shows that students without internet access (red) have a slightly higher peak around grade 10, while the students with internet access (teal) have a more spread-out distribution with a secondary peak at higher grades (around 14-15). This suggests that while students without internet may cluster more tightly around average scores, those with internet access appear more likely to achieve scores in the higher range of the distribution.

```
# summary statistics by internet access
grade_by_internet <- student_data |>
  group_by(internet) |>
  summarize(
    mean_grade = mean(G3),
    median_grade = median(G3),
    sd_grade = sd(G3),
    n = n()
  )

grade_by_internet
```

```
# A tibble: 2 x 5
  internet mean_grade median_grade sd_grade     n
  <chr>         <dbl>        <dbl>    <dbl> <int>
1 no             9.41           10     4.49    66
```

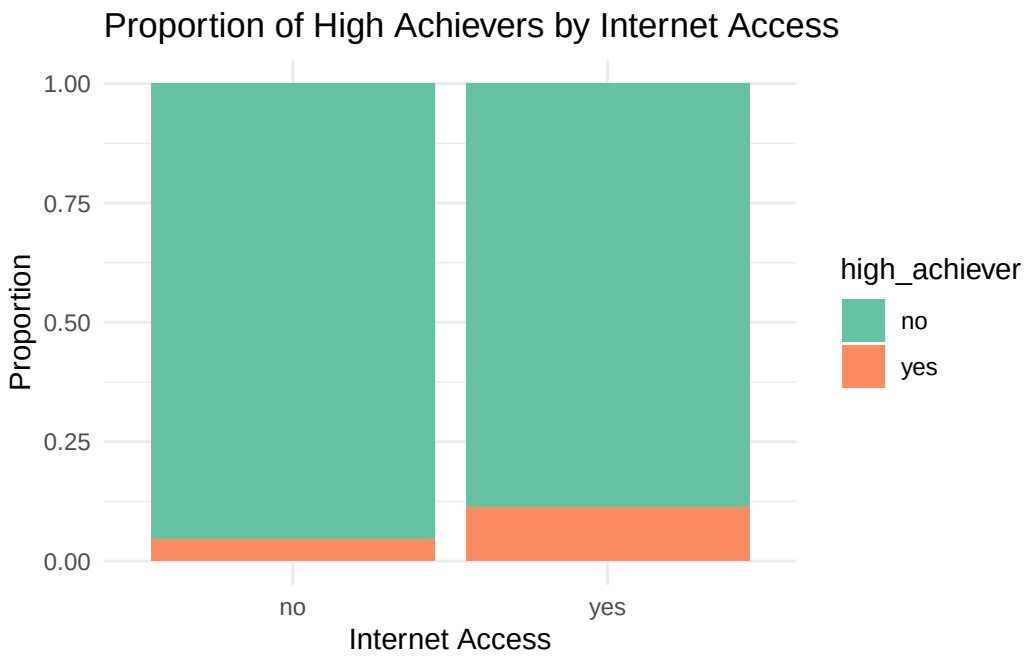
```
2 yes           10.6            11     4.58   329
```

The output above shows that students with internet access achieve a higher mean final grade (10.6 > 9.41) and higher final median grade than students without internet (11 > 10). Both groups also show similar standard deviations in performance (4.58 for students with internet and 4.49 for those without), suggesting comparable variability in grades within each group.

```
# high achievers by internet
high_achievers_by_internet <- student_data |>
  group_by(internet, high_achiever) |>
  summarize(count = n()) |>
  pivot_wider(names_from = high_achiever, values_from = count) |>
  mutate(
    total = yes + no,
    proportion_high_achievers = yes / total
  )

# displaying them
high_achievers_by_internet |>
  select(internet, yes, no, total, proportion_high_achievers)
```

```
# A tibble: 2 x 5
# Groups:   internet [2]
  internet   yes    no total proportion_high_achievers
  <chr>    <int> <int> <int>                     <dbl>
1 no           3    63    66                    0.0455
2 yes         37   292   329                    0.112
```

```
ggplot(student_data, aes(x = internet, fill = high_achiever)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of High Achievers by Internet Access",
       x = "Internet Access",
       y = "Proportion") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2")
```



Our stacked bar chart here shows that students with internet access (the "yes" bar) have a notably higher proportion of high acheivers (about 12-13%) compared to students without internet access (the "no" bar, about 3-4%). This reinforces the relationship between internet access and academic achievement, suggesting that students with internet at home are roughly 3 to 4 times more likely to be high achievers than those without.

Inferential Analysis

Difference in Means Final Grade

We can calculate the difference in means final grades between students with and without internet access and create a 95% confidence interval using bootstrapping:

```
# calculating point estimate (aka the observed difference in means)
point_estimate_diff_mean <- student_data |>
  specify(formula = G3 ~ internet) |>
  calculate(stat = "diff in means", order = c("yes", "no"))

point_estimate_diff_mean
```

```
Response: G3 (numeric)
Explanatory: internet (factor)
```

```
# A tibble: 1 x 1
   stat
  <dbl>
1  1.21
```

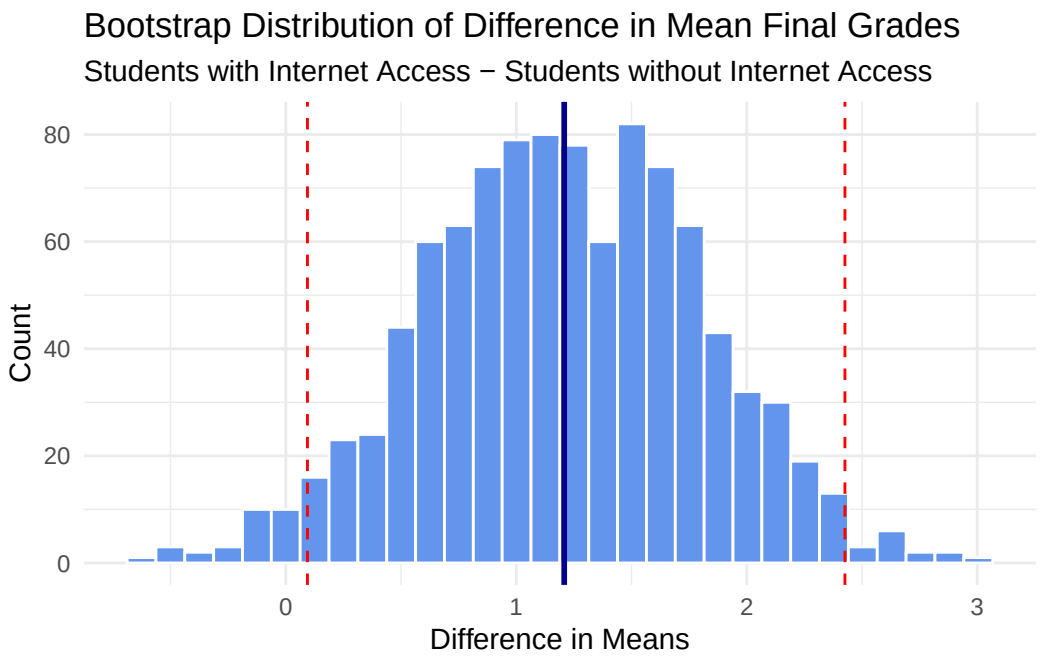
```
# calculating the 95% confidence interval using bootstrapping
ci_diff_mean <- student_data |>
  specify(formula = G3 ~ internet) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means", order = c("yes", "no")) |>
  get_confidence_interval(type = "percentile", level = 0.95)

ci_diff_mean
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1   0.0940     2.43
```

Let's visualize this:

```
# visualizing bootstrap distribution with confidence interval
# first, let's generate the bootstrap distribution to plot
bootstrap_dist_mean <- student_data |>
  specify(formula = G3 ~ internet) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means", order = c("yes", "no"))
```

```
# next, we can plot it
ggplot(bootstrap_dist_mean, aes(x = stat)) +
  geom_histogram(bins = 30, fill = "cornflowerblue", color = "white") +
  geom_vline(xintercept = ci_diff_mean |> pull(lower_ci), linetype = "dashed", color = "red"
  geom_vline(xintercept = ci_diff_mean |> pull(upper_ci), linetype = "dashed", color = "red"
  geom_vline(xintercept = point_estimate_diff_mean |> pull(stat), color = "darkblue", size =
  labs(title = "Bootstrap Distribution of Difference in Mean Final Grades",
       subtitle = "Students with Internet Access - Students without Internet Access",
       x = "Difference in Means",
       y = "Count") +
  theme_minimal()
```



Our outputs show that students with internet access score on average 1.21 points higher in their final math grades compared to those without internet (as shown by our point estimate). The bootstrap distribution histogram illustrates this difference with a solid vertical black line, while the red dashed lines represent the 95% confidence interval (0.11 to 2.46). This interval does not include 0, indicating a statistically significant difference in mean grades between the two groups, but this wide range suggests some uncertainty about the precise magnitude of this effect.

Difference in Proportion of High Achievers

Now, we can calculate the difference in the proportion of high achievers between students with and without internet access:

```
# calculating point estimate (aka observed difference in proportions)
point_estimate_diff_prop <- student_data |>
  specify(formula = high_achiever ~ internet, success = "yes") |>
  calculate(stat = "diff in props", order = c("yes", "no"))

point_estimate_diff_prop
```

```
Response: high_achiever (factor)
Explanatory: internet (factor)
```

```
# A tibble: 1 x 1
    stat
   <dbl>
1 0.0670
```

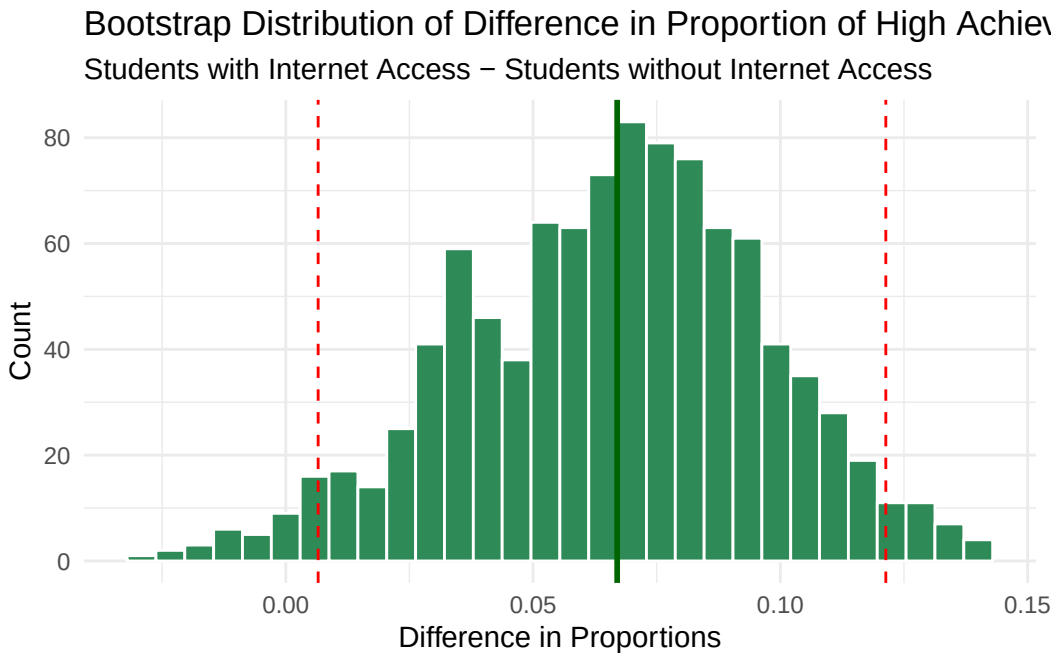
```r
# calculating 95% confidence interval using bootstrapping
ci_diff_prop <- student_data |>
  specify(formula = high_achiever ~ internet, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in props", order = c("yes", "no")) |>
  get_confidence_interval(type = "percentile", level = 0.95)

ci_diff_prop
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1  0.00655    0.121
```

Let's visualize the bootstrap distribution with confidence interval like we did earlier:

```r
# visualizing bootstrap distribution with confidence interval
# first, let's generate the bootstrap distribution to plot
bootstrap_dist_prop <- student_data |>
  specify(formula = high_achiever ~ internet, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in props", order = c("yes", "no"))
```

```r
# plotting the bootstrap distribution
ggplot(bootstrap_dist_prop, aes(x = stat)) +
  geom_histogram(bins = 30, fill = "seagreen", color = "white") +
  geom_vline(xintercept = ci_diff_prop |> pull(lower_ci), linetype = "dashed", color = "red"
  geom_vline(xintercept = ci_diff_prop |> pull(upper_ci), linetype = "dashed", color = "red"
  geom_vline(xintercept = point_estimate_diff_prop |> pull(stat), color = "darkgreen", size =
  labs(title = "Bootstrap Distribution of Difference in Proportion of High Achievers",
       subtitle = "Students with Internet Access - Students without Internet Access",
       x = "Difference in Proportions",
       y = "Count") +
  theme_minimal()
```



Our outputs here show that students with internet access have a 6.7 percentage point higher proportion of high achievers compared to those without internet. The bootstrap distribution histogram shows this difference (through the solid black vertical line) with a 95% confidence interval ranging from 0.3% to 12.5% (shown by the red dashed lines). This confidence interval does not include zero, confirming a statistically significant difference in high achievement rates, though the wide interval indicates some uncertainty about the exact magnitude of this effect.

Conclusion

Summary of Findings

Based on the bootstrapped confidence interval, we can make the following inferences:

1. **Difference in Mean Final Grades**: The 95% confidence interval for the difference in mean final grades between students with and without internet access is [0.11, 2.46]. Since this interval does not contain zero, we can conclude that there is a statistically significant difference in mean final grades between these two groups, with students having internet access scoring on average 1.21 points higher than those without internet.
2. **Difference in Proportion of High Achievers**: Similarly, the 95% confidence interval for the difference in the proportion of high achievers between students with and without internet access is [0.003, 0.125]. Since this interval also does not contain zero, we can conclude that there is a statistically significant difference in the proportion of high

achievers between these two groups, with students having internet access being about 6.7 percentage points more likely to be high achievers than those without internet.

Assessment of Validity

Construct Validity: Our dataset measures internet access as a binary variable (yes/no) but it doesn't capture the quality, frequency, or purpose of internet use by these students with access to internet. Final grades (g3) are a direct measure of academic performance but this may not fully capture all aspects of learning or understanding. A more nuanced measurement of internet usage specifically for educational purposes will most probably better capture the relationship we are trying to understand here. However, despite these limitations, the measurements used are reasonable proxies for the constructs of interest.

External Validity: The sample consists of students from 2 Portugese secondary schools, which may limit generalizability to students in other countries, educational systems, or socioeconomic contexts. Additionally, the data was collected in 2008, and the role and the capabilities of the internet has evolved significantly since then. Urban or rural settings might also influence the relationship between internet access and academic performance differently. More diverse samples across various geographical and socioeconomic context would certaintly strengthen our external validity.

Internal Validity: This analysis is observational rather than experimental, which means that we **cannot** establish causality. Students with internet access may differ from those without in various ways (e.g., socioeconomic status, parental education) that could confound the relationship between internet access and academic performance. A more robust causal analysis would require controlling for potential confounders or randomized assignment of internet access.

Implications and Future Research

The results we've found suggest that internet access is positively associated with both higher average mathematics grades and an increased likelihood of being a high achiever, with statistically significant differences between the two groups for both measures. However, due to the limitations in validity discussed above, these findings should be interpreted cautiously, as the observational nature of the study prevents us from establishing causality, and potential confounding variables such as socioeconomic status may account for some or all of the observed differences.

Future research could address these limitations by: 1. Using more nuanced measures of internet usage specifically for educational purposes. 2. Including a more diverse sample of students across different regions and socioeconomic contexts. 3. Using experimental designs to better address causality. 4. Controlling for potential confounding variables like socioeconomic status and parental education.

Understanding the relationship between internet access and academic performance is increasingly important as education becomes more digital, especially with generative AI tools that can potentially just hand out answers, raising questions about how to balance digital access equity with increased academic performance, maintaining academic integrity and genuine learning outcomes. We can treat this analysis as a starting point for further investigation into how digital resources can be leveraged to improve educational outcomes.

References

Cortez, P. and A. M. Gonçalves Silva. "Using data mining to predict secondary school student performance." (2008).