STUDENT PERFORMANCE ANALYSIS

Ryham Houari, Joudi J. G. Besaiso, Leen Alabudi, Abrar Al-harazi

Abstract This study aims to uncover the factors that influence students' exam scores by using various statistical methods, such as hypothesis testing, ANOVA, and chi-square tests. The variables we looked at include reading, writing, math, parental education level, gender, and whether the students took a test preparation course. We collected data from three American high schools, with a sample of 150 students drawn from a population of 1000 through systematic sampling. The findings shed light on how different factors impact student performance, providing valuable insights for improving educational outcomes.

Presentation link: https://youtu.be/x3DjcqR9hFU

1. INTRODUCTION

In this report, we delve into the factors that determine high school students' academic performance in subjects like mathematics, reading, and writing. By applying various statistical methods, we aim to identify significant factors affecting exam scores and provide actionable insights. These insights can help educators develop strategies to boost student performance and address disparities in education.

2. REQUIREMENTS

As shown in Figure 1. These are the important libraries.

3. DATA DESCRIPTION

The dataset, sourced from Kaggle, includes grades and demographic information on high school students' mathematical performance. Data were collected from three American high schools.

```
library(tidyverse)
library(janitor)
library(DataExplorer)
library(lubridate)
library(ggplot2)
library(gridExtra)
library(moments)
library(nortest)
library(car)
library(MASS)
library(corrplot)
library("BSDA")
```

Figure 1. Requirements

3.1 Data Collection

We used systematic sampling to select a sample of 150 students from a population of 1000. The dataset variables include gender, race/ethnicity, parental level of education, lunch type, participation in test preparation courses, and scores in math, reading, and writing.

3.2 Data Overview

We created a new column, 'average_score,' representing the average of the math, reading, and writing scores for each student.

3.2 Reading data

We started reading the data set using R, as shown in Figure 2.

3.3 First 5 instances of the population

As shown in Figure 3, we have 7 columns: gender, race/ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score.

```
# Data Collection and Cleaning
exams <- read_csv("exams.csv")
exams <- clean_names(exams)
View(exams)
```

Figure 2. Reading the data set

Filter

	gender	race_ethnicity	parental_level_of_education	lunch	test_preparation_course	math_score	reading_score	writing_score
1	female	group D	some college	standard	completed	59	70	78
2	male	group D	associate's degree	standard	none	96	93	87
3	female	group D	some college	free/reduced	none	57	76	77
4	male	group B	some college	free/reduced	none	70	70	63
5	female	group D	associate's degree	standard	none	83	85	86

Figure 3. First 5 instances of population

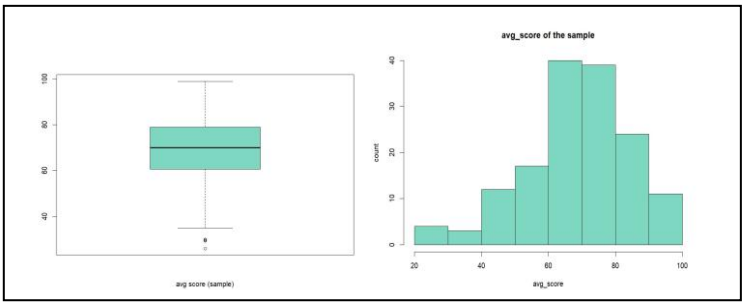

Figure 4. distribution of the average score of the sample

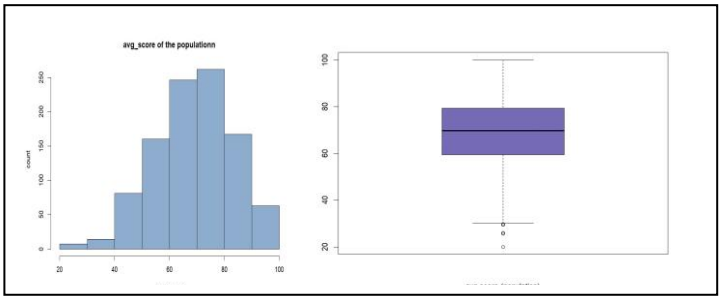

Figure 5. distribution of the average score of the pop

4. DATA VISUALIZATION

4.1 sample plots

We created various plots to visualize the sample data as shown in Figure 4, helping us understand the distribution and relationships between different variables. For example, histograms show the distribution of math scores, while scatter plots reveal correlations between different subjects. Box plots help us see differences in performance based on factors like gender and parental education.

4.2 Population Plots

Population plots give us a broader view of the entire dataset. These plots highlight key trends as shown in Figure 5, such as overall performance distribution and the impact of test preparation courses on student scores.

5. NORMALITY TESTS

To check if our data follows a normal distribution, we used QQ plots and the Pearson Coefficient of Skewness. The QQ plot compares our sample data to a normal distribution in Figure 6, while the skewness coefficient tells us if our data is skewed. If the skewness is greater than +1 or less than -1, it's considered significantly skewed. Our analysis found that some variables were approximately normal, but others were skewed, requiring non-parametric tests for accurate analysis.

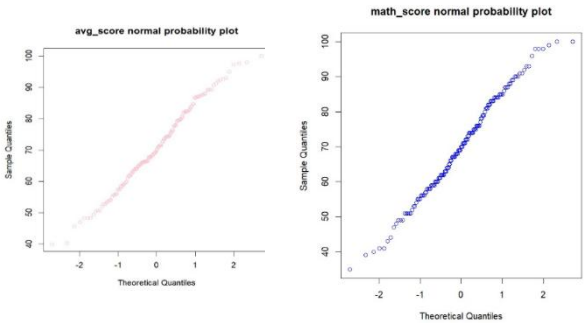

Figure 6. Q-Q plot

6. POINT ESTIMATIONS AND CONFIDENCE INTERVALS

6.1 Point eastimations

We calculated the mean and standard deviation for key variables and compared them with the overall population parameters as shown in Figure 7. This gave us initial insights into the central tendencies and variability of the data.
- *Math Score:* Mean = 67, SD = 15
- *Reading Score:* Mean = 70, SD = 14
- *Writing Score:* Mean = 69, SD = 16

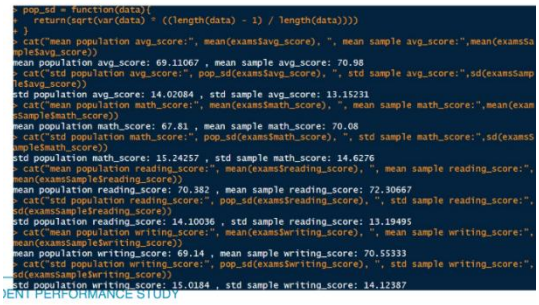

Figure 7. SD & mean for key variables

6.2 Confidence intervals

We used confidence intervals to estimate population parameters with a specified level of confidence.

- *T-test for average score mean:* With 95% confidence, the population mean of average scores is between 65 and 71 as shown in Figure 8.
- *Z-test for the proportion of male students:* With 95% confidence, the proportion of male students is between 45% and 55%, as shown in Figure 9.
- *Chi-square test for the standard deviation of math scores:* With 95% confidence, the population standard deviation is between 13 and 17, As shown in the Figure 10.

```
> #congidence intervals
> # avg_score confidence interval
> avg_score_sample_mean = mean(examsSample$avg_score)
> avg_score_sample_sd = sd(examsSample$avg_score)
> avg_score_sample_se = avg_score_sample_sd / sqrt(SAMPLE_SIZE)
> t_score = qt(p=1-alpha/2, df=SAMPLE_SIZE-1,)
> margin_error <- t_score * avg_score_sample_se
> lower_bound <- avg_score_sample_mean - margin_error
> upper_bound <- avg_score_sample_mean + margin_error
> cat('cl=',1-alpha,lower_bound,'< avg_score < ', upper_bound)
cl= 0.95 68.858 < avg_score <  73.102
```

Figure 8. The population mean of average scores

```
> num_males = sum(examsSample$gender == "male")
> p_hat = num_males / SAMPLE_SIZE
> q_hat = 1 - p_hat
> SE = sqrt(p_hat * q_hat / SAMPLE_SIZE)
> z_score = qnorm(1 - alpha / 2)
> moe = z_score * SE
> lower_bound = p_hat - moe
> upper_bound = p_hat + moe
> cat('cl =', 1 - alpha, lower_bound, '< proportion of male students <', upper_bound)
cl = 0.95 0.4873659 < proportion of male students < 0.6459674
```

Figure 9. the proportion of male students

```
> # Confidence Intervals for Variances and Standard Deviations
> df = SAMPLE_SIZE - 1
> chi_right = qchisq(1-alpha/2,df)
> chi_left = qchisq(alpha/2,df)
> cat('cl=',1-alpha,sqrt(df/chi_right) * sd(exams$math_score),' < sigma < ',sqrt(df/chi_left) * sd(exams$math_score))
cl= 0.95 13.69779  < sigma <  17.20259
```

Figure 10. SD of math scores

```
> # H_0: none__preparation_test and take__preparation_test student have the same avarage score
> # H_a: none__preparation_test and take__preparation_test student have different avarage score
> none__preparation_test = systematic_sampling(exams[exams$test_preparation_course == "none",], SAMPLE_SIZE)
> take__preparation_test= systematic_sampling(exams[exams$test_preparation_course == "completed",], SAMPLE_SIZE)
> t.test(none__preparation_test$avg_score, take__preparation_test$avg_score)

        Welch Two Sample t-test

data:  none__preparation_test$avg_score and take__preparation_test$avg_score
t = -3.5687, df = 291.91, p-value = 0.0004193
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.653925 -2.501630
sample estimates:
mean of x mean of y
 67.53556  73.11333
```

Figure 11. 1st hypothesis test

```
> #We assume the avg math score is 60 lets test that
> #with alpha=0.05
> #H0: the mean score of math scores equals 60
> #Ha: the mean score of math scores is less than 60
> t.test(examsSample$math_score, mu = 60, alternative = 'less')

        One Sample t-test

data:  examsSample$math_score
t = 8.4398, df = 149, p-value = 1
alternative hypothesis: true mean is less than 60
95 percent confidence interval:
    -Inf 72.0568
sample estimates:
mean of x
    70.08
```

Figure 12. 2nd hypothesis Test

7. HYPOTHESIS TESTING

We conducted several hypothesis tests to explore the relationships between different variables.

7.1 Test 1: Effect of Test Preparation Courses

- *H0:* No difference in average scores between students who completed the test preparation course and those who did not.
- *Ha:* There is a significant difference in average scores.
- *Result:* p-value = 0.0004193 (Reject H0)

This result shows that students who took the test preparation course had significantly different scores, suggesting the course's effectiveness. As shown in Figure 11

7.2 Test 2: Mean Number of Turns

- *H0:* The mean of the math scores equals 60.
- *Ha:* The mean of the math scores is less than 60.
- *Result:* p-value = 1 (Fail to reject H0)

hypothesis and resulted in a p-value of 1, as shown in Figure12. Which is not significant, so we fail to reject the null hypothesis.

7.3 Test 3: Reading Scores by Gender

- *H0:* No difference in reading scores between male and female students.
- *Ha:* There is a significant difference in reading scores.
- *Result:* Small p-value, as shown in Figure 13 (Reject H0)

7.3 Test 4: Variance Comparison of Test Preparation Course Impact on Student Score

Fourth test: with alpha=0.05
-H0: the standard deviations of average scores of the students who take the test preparation course equal to the average score of student who haven't taken it
-Ha: the standard deviations of average scores of the students who take the test preparation course not equal to the average score of student who haven't taken it .
A F-test was used to evaluate the hypothesis and resulted in a p-value of 0.7786, as shown in Figure 14. Which is not significant, so we fail to reject the null hypothesis.

7.3 Test 5: Parental Degree Z-Test

Fifth test: with alpha=0.05
-H0: the proportion of parental bachelor's degree is equal to 0.5.
-Ha: the proportion of parental bachelor's degree is not equal to 0.5.
A z-test was used to evaluate the hypothesis and resulted in very small a p-value of, as shown in Figure 15. Which is significant to reject the null hypothesis in favor of the alternative hypothesis.

8. LINEAR REGRESSION

We used linear regression to predict student scores based on various factors. Our model, As shown in figure 17 identified significant predictors like parental education level and participation in test preparation courses.

- *Correlation between writing and math scores:* r = 0.65, As shown in figure 16

```
> cor(examsSample$writing_score,examsSample$math_score)
[1] 0.765073
```

Figure 16. correlation between writing and math scores

This regression analysis helps us understand how different factors contribute to academic success and allows us to make predictions about student performance.

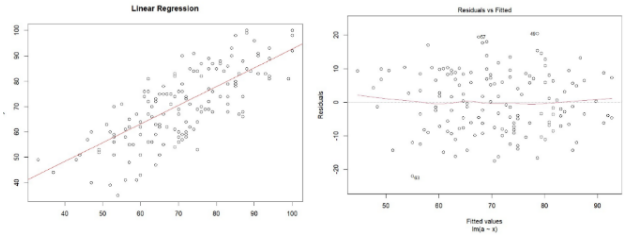

Figure 17. Model of linear regression

```
> #Third test: with alpha=0.05
> #H0: male and female students have the same reading score
> #Ha: male and female students have different reading scores
> female_sample = systematic_sampling(exams[exams$gender == 'female',], SAMPLE_SIZE)
> male_sample = systematic_sampling(exams[exams$gender == "male",], SAMPLE_SIZE)
> t.test(female_sample$reading_score, male_sample$reading_score)

        Welch Two Sample t-test

data:  female_sample$reading_score and male_sample$reading_score
t = 4.6057, df = 297.03, p-value = 6.104e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.142602 10.324064
sample estimates:
mean of x mean of y
 74.38000  67.14667
```

Figure 13. 3rd Hypothesis Test

```
> #Fourth test: with alpha=0.05
> #H0: the standard deviations of avg scores of the students who take the test preparation course
> #equal to the avg score of the students haven't take it
> #Ha: #H0: the standard deviations of avg scores of the students who take the test preparation course
> #not equal to the avg score of the students haven't take it
> none__preparation_test = systematic_sampling(exams[exams$test_preparation_course == 'none',], SAMPLE_SIZE)
> take__preparation_test = systematic_sampling(exams[exams$test_preparation_course == "completed",], SAMPLE_SIZE)
> var.test(none__preparation_test$avg_score, take__preparation_test$avg_score)

        F test to compare two variances

data:  none__preparation_test$avg_score and take__preparation_test$avg_score
F = 1.0472, num df = 149, denom df = 149, p-value = 0.7786
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7586413 1.4455701
sample estimates:
ratio of variances
           1.04722
```

Figure 14. 4th Hypothesis Test

```
> #Fifth test: with alpha=0.05
> #H0: the proportion of  parental bachelor's degree is equal to 0.5.
> #Ha: the proportion of  parentel bachelor's degree is not equal to 0.5
> # Count the number of parents with a bachelor's degree
> bachelors_degree = sum(examsSample$parental_level_of_education == "bachelor's degree")
> total_degrees = length(examsSample$parental_level_of_education)
> prop.test(x = bachelors_degree, n = total_degrees, p = 0.5)

        1-sample proportions test with continuity correction

data:  bachelors_degree out of total_degrees, null probability 0.5
X-squared = 100.86, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.04884108 0.14660455
sample estimates:
         p
0.08666667
```

Figure 15. 5th Hypothesis Test

```
> #chi square tests
> # Godness-of-fit
> #observed
> table(examsSample$lunch)

free/reduced     standard
          54           96
> r = table(examsSample$lunch)[[1]]
> s = table(examsSample$lunch)[[2]]
> o = c(r,s)
> #expected
> p = c(0.5,0.5)
> result = chisq.test(o,p=p)
> result

        Chi-squared test for given probabilities

data:  o
X-squared = 11.76, df = 1, p-value = 0.0006052

> #p-value = 0.0003274 reject the null
> result$expected
[1] 75 75
> result$observed
[1] 54 96
```

Figure 18. Chi-Square Goodness-of-Fit Test

9. CHI SQUARE TESTS

We used chi-square tests to examine relationships between categorical variables.

9.1 Goodness of fit

- *H0:* The observed percentages match the expected values.
- *Ha:* The observed percentages do not match.
- *Result:* Significant p-value (Reject H0)

This test showed that the observed distribution of variables significantly differed from the expected distribution.
The plot of the different between the plot and observed, As shown in Figure 18.

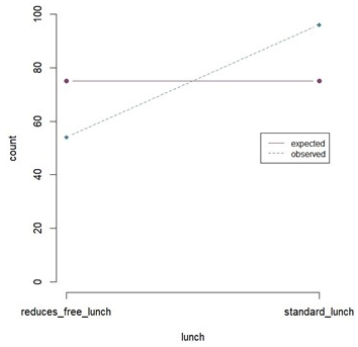

Figure 19. Different between expected and observed.

```
> #independence test
> #H0: The categorical variables gender and lunch are independent.
> #Ha: The categorical variables are dependent.
> table(examsSample$gender, examsSample$lunch)

        free/reduced standard
  female           27       38
  male             27       58
> chisq.test(table(examsSample$gender, examsSample$lunch))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(examsSample$gender, examsSample$lunch)
X-squared = 1.1324, df = 1, p-value = 0.2873
```

Figure 20. Independence test script

9.2 Independence test

- *H0:* Gender and lunch type are independent.

- *Ha:* Gender and lunch type are dependent.

- *Result:* Non-significant p-value, As shown in Figure 19 (Fail to reject H0)

This result suggests no significant relationship between gender and lunch type, indicating independence.

9.3 Homogeneity Test

- *H0:* Male and female students are homogeneous regarding lunch type.
- *Ha:* Differences exist in lunch type categories.
- *Result:* Non-significant p-value (Fail to reject H0)

This test indicated no significant differences in lunch type categories between male and female students.

```
> # homogeneity test
> #HO: The male and female are homogeneous when it comes to the categories of the lunch variable.
> #Ha: There is a difference when it comes to the categories of the variable lunch between male and female
> sample1 = systematic_sampling(exams[exams$gender == "male", ], SAMPLE_SIZE)
> # homogeneity test
> #HO: The male and female are homogeneous when it comes to the categories of the lunch variable.
> #Ha: There is a difference when it comes to the categories of the variable lunch between male and female
> sample1 = systematic_sampling(exams[exams$gender == "male", ], SAMPLE_SIZE)
> sample2 = systematic_sampling(exams[exams$gender == "female", ], SAMPLE_SIZE)
> sample3 = rbind(sample1, sample2)
> table(sample3$gender, sample3$lunch)

        free/reduced standard
  female            50      100
  male              44      106
> chisq.test(table(sample3$gender, sample3$lunch))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(sample3$gender, sample3$lunch)
X-squared = 0.38732, df = 1, p-value = 0.5337
```

Figure 21. Non-significant p-value

10. ANOVA

10.1 One-Way ANOVA

- *H0:* Reading score means are the same across all groups.
- *Ha:* Reading score means differ for at least one group.
- *Result:* Significant p-value (Reject H0) As shown in Figure 22.

```
> #one way ANOVA
> #HO: The reading score mean is the same for all group categories.
> #Ha: The reading score mean is the different for at least one group category.
> #unique(exams$race_ethnicity)
> examsSample$race_ethnicity = ordered(examsSample$race_ethnicity, levels = c("group A", "group B", "group C","group E", "group D"))
> anova_t1 = aov(reading_score~race_ethnicity, data=examsSample)
> summary(anova_t1)
              Df Sum Sq Mean Sq F value Pr(>F)
race_ethnicity   4   2105   526.2   3.201 0.0149 *
Residuals      145  23837   164.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 22. One way ANOVA.

```
> ScheffeTest(anova_t1)

  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$race_ethnicity
                      diff      lwr.ci    upr.ci   pval
group B-group A -4.734127 -16.8206555  7.352402 0.8273
group C-group A -2.512077 -13.6348420  8.610687 0.9736
group E-group A  6.685185  -5.4886118 18.858982 0.5702
group D-group A -1.442652 -13.2981250 10.412820 0.9975
group C-group B  2.222050  -7.3674481 11.811547 0.9710
group E-group B 11.419312   0.6283945 22.210230 0.0317 *
group D-group B  3.291475  -7.1390001 13.721949 0.9139
group E-group C  9.197262  -0.5019967 18.896522 0.0731 .
group D-group C  1.069425  -8.2271619 10.366012 0.9980
group D-group E -8.127838 -18.6593125  2.403637 0.2206

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 23. Scheffé Test

```
> anova_t2 = aov(reading_score~gender*test_preparation_course, data=examsSample)
> summary(anova_t2)
                                Df Sum Sq Mean Sq F value   Pr(>F)
gender                           1   1836  1836.2  11.507 0.000893 ***
test_preparation_course          1    441   441.0   2.764 0.098581 .
gender:test_preparation_course   1    366   366.2   2.295 0.131944
Residuals                      146  23298   159.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 24. Two way ANOVA

This analysis showed significant differences in reading scores across different groups, suggesting that group membership impacts reading performance. To see which category/categories were different the Scheffé test was applied As shown in Figure 23.

10.2 Two-Way ANOVA

- *Ho1:* No difference in reading scores by gender.
- *Ha1:* Difference in reading scores by gender.
- *Ho2:* No difference in reading scores by test preparation course.
- *Ha2:* Difference in reading scores by test preparation course.
- *Ho3:* No interaction effect between gender and test preparation course on reading scores.
- *Ha3:* Interaction effect exists.

- *Result:* Significant for Ho1 and Ho2 (Reject Ho1 and Ho2), not significant for Ho3 (Fail to reject Ho3) As shown in Figure 24.

The two-way ANOVA provided a deeper understanding of the individual and interaction effects of gender and test preparation courses on reading scores.

11. NON-PARAMETRIC TESTS

To account for non-normal data distributions, we conducted several non-parametric tests.

11.1 Sign Test

- *H0:* Median math score equals 90.
- *Ha:* Median math score is less than 90.
- *Result:* Significant p-value (Reject H0), as shown in Figure 25.

The sign test indicated that the median math score was significantly less than 90.

```
> #Sign test
> s2 = systematic_sampling(exams, 20)
> SIGN.test(s2$math_score ,md=90, alternative = 'less')

        One-sample Sign-Test

data:  s2$math_score
s = 1, p-value = 2.003e-05
alternative hypothesis: true median is less than 90
95 percent confidence interval:
     -Inf 75.45042
sample estimates:
median of x
      67.5

Achieved and Interpolated Confidence Intervals:

                 Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.9423   -Inf 74.0000
Interpolated CI       0.9500   -Inf 75.4504
Upper Achieved CI     0.9793   -Inf 81.0000
```

Figure 25. Sign Test

```
> cor.test(s2$math_score, s2$reading_score , method = "spearman", exact = FALSE)

        Spearman's rank correlation rho

data:  s2$math_score and s2$reading_score
S = 310.58, p-value = 8.1e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7664784
```

Figure 27. Spearman correlation Test

```
> wilcox.test(s2$math_score[s2$test_preparation_course == "completed"],s2$math_score[s2$test_preparation_course == "none"])

        Wilcoxon rank sum test with continuity correction

data:  s2$math_score[s2$test_preparation_course == "completed"] and s2$math_score[s2$test_preparation_course == "none"]
W = 87.5, p-value = 0.005109
alternative hypothesis: true location shift is not equal to 0
```

Figure 26. Wilcox Test

```
> runs_test_sample = exams$math_score[1:90] %%2
> runs_test_sample
 [1] 1 0 1 0 1 0 0 0 0 1 0 1 0 0 1 0 1 1 0 1 0 1 0 1 1 0 0 0 0 0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 1 0 0 1 0 1 0 0 1 1 0 1 1 1 1
[61] 0 1 1 1 1 1 1 1 0 0 1 0 1 1 1 1 0 1 1 1 0 1 1 1 1 0 0 0 0 0
> runs.test(runs_test_sample)

        Runs Test

data:  runs_test_sample
statistic = 0.42403, runs = 48, n1 = 45, n2 = 45, n = 90, p-value = 0.6715
alternative hypothesis: nonrandomness
```

Figure 28. Run Test

11.2 Wilcox Rank Sum Test

- *H0:* No difference in median math scores between students who took or did not take the test preparation course.
- *Ha:* Difference in median math scores exists.

- *Result:* Non-significant p-value (Fail to reject H0), as shown in Figure 26.

This test suggested no significant difference in median math scores based on participation in the test preparation course.

11.3 Spearman Correlation

- *H0:* No correlation between math and reading scores.
- *Ha:* Correlation exists between math and reading scores.
- *Result:* Significant p-value (Reject H0), as shown in Figure 27.

The Spearman correlation test showed a significant positive correlation between math and reading scores.

11.4 Runs Test

- *H0:* Sequence is random.
- *Ha:* Sequence is not random.
- *Result:* Non-significant p-value (Fail to reject H0), as shown in Figure 28.

The runs test indicated no significant departure from randomness in the sequence of scores.

12. R CODE

12.1 Install and download necessary packages

```r
# Install and load necessary packages
install.packages("tidyverse")
install.packages("janitor")
install.packages("DataExplorer")
install.packages("lubridate")
install.packages("ggplot2")
install.packages("gridExtra")
install.packages("moments")
install.packages("nortest")
install.packages("car")
install.packages("MASS")
install.packages("sampling")
install.packages("BSDA")
library(tidyverse)
library(janitor)
library(DataExplorer)
library(lubridate)
library(ggplot2)
library(gridExtra)
library(moments)
library(nortest)
library(car)
library(MASS)
library(corrplot)
library("BSDA")

# Data Collection and Cleaning
exams <- read_csv("exams.csv")
exams <- clean_names(exams)
View(exams)

# Remove NA values
exams <- drop_na(exams)



#add new column(avg_score) for the math and writing and reading scores
exams$avg_score = (exams$math_score + exams$reading_score+ exams$writing_score) / 3
head(exams)

View(exams)
```

12.2 systematic sampling

```r
#systematic sampling

SAMPLE_SIZE = 150
alpha = 0.05
systematic_sampling <- function(data, n) {

  total_obs <- nrow(data)  # Total number of observations

  sampling_interval <- ceiling(total_obs / n)  # Calculate the sampling interval

  start_point <- sample(1:sampling_interval, 1)  # Generate a random starting point

  # Generate indices while ensuring they stay within the bounds of the data
  sample_indices <- (start_point + sampling_interval * (0:(n-1))) %% total_obs
  sample_indices[sample_indices == 0] <- total_obs  # Replace 0 with total_obs to handle modulo wrapping correctly

  sample_data <- data[sample_indices, ]  # Select the samples

  return(sample_data)
}

set.seed(2004)
examsSample = systematic_sampling(exams, SAMPLE_SIZE)
View(examsSample)
```

12.3 sample plots

```r
#sample plots
hist(examsSample$avg_score, col='#7ED7C1', main='avg_score of the sample', xlab='avg_score', ylab = 'count')
hist(examsSample$math_score, col='#FF8080', main='math score of the sample', xlab='math score', ylab = 'count')
hist(examsSample$reading_score, col='#8EACCD', main='reading score of the sample', xlab='reading score', ylab = 'count')
hist(examsSample$writing_score, col='#79AC78', main='writing score of the sample', xlab='writing score', ylab = 'count')

hist(examsSample$avg_score[examsSample$test_preparation_course=="completed"], main='completed test preparation cource', xlab = 'avg_score', ylab = 'count', col='#BEADFA')

hist(examsSample$avg_score[examsSample$test_preparation_course=="none"], main='did not take the test preparation cource', xlab = 'avg_score', ylab = 'count', col='#756AB6')

hist(examsSample$avg_score[examsSample$math_score>80], main='math_score>80', xlab = 'avg_score', ylab = 'count', col='#7469B6')

clyinders1 = table(examsSample$test_preparation_course)
barplot(clyinders1, col=c('#BEADFA', '#756AB6'),xlab='test preparation cource of the sample', ylab = 'count')

clyinders2 = table(examsSample$gender)
barplot(clyinders2, col=c('#E1ACAC', '#BED7DC'),xlab='gender of the sample', ylab = 'count')

clyinders3 = table(examsSample$lunch)
barplot(clyinders3, col=c('#C5EBAA','#A5DD9B'),xlab='lunch price of the sample', ylab = 'count')

clyinders4 = table(examsSample$parental_level_of_education)
barplot(clyinders4, col=c('#FFCCCB', '#FFDAB9', '#FFE4E1', '#B0E0E6', '#98FB98', '#DDA0DD'),xlab='parental_level_of_education(sample)', ylab = 'count')

clyinders5 = table(examsSample$race_ethnicity)
barplot(clyinders5, col=c('#FFB6C1', '#ADD8E6', '#FFDAB9', '#98FB98', '#FFA07A')
,xlab='race_ethnicity', ylab = 'count')

boxplot(examsSample$avg_score, col='#7ED7C1', xlab='avg_score (sample)')
boxplot(examsSample$math_score, col='#FF8080', xlab='math score (sample)')
boxplot(examsSample$reading_score, col='#8EACCD', xlab='reading score (sample)')
boxplot(examsSample$writing_score, col='#79AC78', xlab='writing score (sample)')
```

12.4 Population plots

```r
#population plot
hist(exams$avg_score, col='#8EACCD', main='avg_score of the populationn', xlab='avg_score', ylab = 'count')
hist(exams$math_score, col='#FF8080', main='math score of the populationn', xlab='math score', ylab = 'count')
hist(exams$reading_score, col='#8EACCD', main='reading score of the populationn', xlab='reading score', ylab = 'count')
hist(exams$writing_score, col='#79AC78', main='writing score of the populationn', xlab='writing score', ylab = 'count')

hist(exams$avg_score[exams$test_preparation_course=="completed"], main='completed test preparation cource (pop)', xlab = 'avg_score', ylab = 'count', col='#BEADFA')

hist(exams$avg_score[exams$test_preparation_course=="none"], main='does not take the test preparation cource(pop)', xlab = 'avg_score', ylab = 'count', col='#756AB6')

clyinders6 = table(exams$test_preparation_course)
barplot(clyinders6, col=c("#BEADFA", "#756AB6"),xlab='test preparation cource of the populaton', ylab = 'count')

clyinders7 = table(exams$gender)
barplot(clyinders7, col=c("#8E7AB5", "#BED7DC"),xlab='gender of the populaton', ylab = 'count')

clyinders8 = table(exams$lunch)
barplot(clyinders8, col=c('#C5EBAA','#A5DD9B'),xlab='lunch price of the populaton', ylab = 'count')

clyinders9 = table(exams$parental_level_of_education)
barplot(clyinders9, col=c("#FFCCCB", "#FFDAB9", "#FFE4E1", "#B0E0E6", "#98FB98", "#DDA0DD"),xlab='parental_level_of_education_of_the_populaton', ylab = 'count')

clyinders10 = table(exams$race_ethnicity)
barplot(clyinders10, col=c("#FFB6C1", "#ADD8E6", "#FFDAB9", "#98FB98", "#FFA07A"),xlab='race_ethnicity_of_the_populaton', ylab = 'count')

boxplot(exams$avg_score, col='#756AB6', xlab='avg score (population)')
boxplot(exams$math_score, col='#FF8080', xlab='math score (population)')
boxplot(exams$reading_score, col='#8EACCD', xlab='reading score (population)')
boxplot(exams$writing_score, col='#79AC78', xlab='writing score (population)')
```

12.5 Normality Tests

```r
#normality tests
#Pearson's Coefficient (PC) of Skewness
#avg_score
qqnorm(examsSample$avg_score,col='pink', main='avg_score normal probability plot')
3*(mean(examsSample$avg_score) - median(examsSample$avg_score)) / sd(examsSample$avg_score)
#math_score
qqnorm(examsSample$math_score,col='blue', main='math_score normal probability plot')
3*(mean(examsSample$math_score) - median(examsSample$math_score)) / sd(examsSample$math_score)
#reading_score
qqnorm(examsSample$reading_score,col='purple', main='reading_score normal probability plot')
3*(mean(examsSample$reading_score) - median(examsSample$reading_score)) / sd(examsSample$reading_score)
#writing_score
qqnorm(examsSample$writing_score,col='green', main='writing_score normal probability plot')
3*(mean(examsSample$writing_score) - median(examsSample$writing_score)) / sd(examsSample$writing_score)
```

12.6 Point estimation

```r
# Point estimation

pop_sd = function(data){
  return(sqrt(var(data) * ((length(data) - 1) / length(data))))
}

cat("mean population avg_score:", mean(exams$avg_score), ", mean sample avg_score:",mean(examsSample$avg_score))
cat("std population avg_score:", pop_sd(exams$avg_score), ", std sample avg_score:",sd(examsSample$avg_score))

cat("mean population math_score:", mean(exams$math_score), ", mean sample math_score:",mean(examsSample$math_score))
cat("std population math_score:", pop_sd(exams$math_score), ", std sample math_score:",sd(examsSample$math_score))

cat("mean population reading_score:", mean(exams$reading_score), ", mean sample reading_score:",mean(examsSample$reading_score))
cat("std population reading_score:", pop_sd(exams$reading_score), ", std sample reading_score:",sd(examsSample$reading_score))

cat("mean population writing_score:", mean(exams$writing_score), ", mean sample writing_score:",mean(examsSample$writing_score))
cat("std population writing_score:", pop_sd(exams$writing_score), ", std sample writing_score:",sd(examsSample$writing_score))
```

12.7 Confidence Intervals

```r
#congidence intervals
# avg_score confidence interval
avg_score_sample_mean = mean(examsSample$avg_score)
avg_score_sample_sd = sd(examsSample$avg_score)
avg_score_sample_se = avg_score_sample_sd / sqrt(SAMPLE_SIZE)

t_score = qt(p=1-alpha/2, df=SAMPLE_SIZE-1,)
margin_error <- t_score * avg_score_sample_se
lower_bound <- avg_score_sample_mean - margin_error
upper_bound <- avg_score_sample_mean + margin_error
cat('cl=',1-alpha,lower_bound,'< avg_score < ', upper_bound)
mean(exams$avg_score)

# Confidence Intervals for Proportions of Male Students

num_males = sum(examsSample$gender == "male")
p_hat = num_males / SAMPLE_SIZE
q_hat = 1 - p_hat
SE = sqrt(p_hat * q_hat / SAMPLE_SIZE)
z_score = qnorm(1 - alpha / 2)
moe = z_score * SE
lower_bound = p_hat - moe
upper_bound = p_hat + moe
cat('cl =', 1 - alpha, lower_bound, '< proportion of male students <', upper_bound)

# Confidence Intervals for Variances and Standard Deviations
df = SAMPLE_SIZE - 1
chi_right = qchisq(1-alpha/2,df)
chi_left = qchisq(alpha/2,df)
cat('cl=',1-alpha,sqrt(df/chi_right) * sd(exams$math_score),' < sigma < ',sqrt(df/chi_left) * sd(exams$math_score))
```

12.8 Hypothesis Tests

```r
# Hypotheses tests
# H_0: none__preparation_test and take__preparation_test student have the same avarage score
# H_a: none__preparation_test and take__preparation_test student have different avarage score
none__preparation_test = systematic_sampling(exams[exams$test_preparation_course == "none",], SAMPLE_SIZE)
take__preparation_test= systematic_sampling(exams[exams$test_preparation_course == "completed",], SAMPLE_SIZE)
t.test(none__preparation_test$avg_score, take__preparation_test$avg_score)
#p-value = 0.0004193 - reject
#The p-value is extremely small, indicating strong evidence against the null hypothesis.
#p-value = 4.422e-07

#We assume the avg math score is 60 lets test that
#with alpha=0.05
#H0: the mean score of math scores equals 60
#Ha: the mean score of math scores is less than 60
t.test(examsSample$math_score, mu = 60, alternative = 'less')
#p-value = 1 fail to reject

#Third test: with alpha=0.05
#H0: male and female students have the same reading score
#Ha: male and female students have different reading scores
female_sample = systematic_sampling(exams[exams$gender == 'female',], SAMPLE_SIZE)
male_sample = systematic_sampling(exams[exams$gender == "male",], SAMPLE_SIZE)
t.test(female_sample$reading_score, male_sample$reading_score)
#p-value = 0.000158 reject the null Hypothesis

#Fourth test: with alpha=0.05
#H0: the standard deviations of avg scores of the students who take the test preparation course
#equal to the avg score of the students haven't take it
#Ha: #H0: the standard deviations of avg scores of the students who take the test preparation course
#not equal to the avg score of the students haven't take it
none__preparation_test = systematic_sampling(exams[exams$test_preparation_course == 'none',], SAMPLE_SIZE)
take__preparation_test = systematic_sampling(exams[exams$test_preparation_course == "completed",], SAMPLE_SIZE)
var.test(none__preparation_test$avg_score, take__preparation_test$avg_score)
#p-value = 0.2895 fail to reject the null hypothesis

#Fifth test: with alpha=0.05
#H0: the proportion of  parental bachelor's degree is equal to 0.5.
#Ha: the proportion of  parentel bachelor's degree is not equal to 0.5
# Count the number of parents with a bachelor's degree
bachelors_degree = sum(examsSample$parental_level_of_education == "bachelor's degree")
total_degrees = length(examsSample$parental_level_of_education)
prop.test(x = bachelors_degree, n = total_degrees, p = 0.5)
#p-value < 2.2e-16 reject the null hypothesis
cot_v <- unlist(lapply(exams, is.numeric), use.names = FALSE)
corrplot(cor(exams[,cot_v]),method = "circle", type = "upper", order = "AOE")
```

12.9 The model of linear regression

```r
#the model
#writing_score and math scores
cor(examsSample$writing_score,examsSample$math_score)

model = lm(examsSample$writing_score~examsSample$math_score)


cor(examsSample$writing_score,examsSample$math_score)
a <- examsSample$writing_score
x <- examsSample$math_score
model = lm(a~x)
#This makes predictions for writing_score based on given math_score
predict(model, data.frame(x=c(80, 90, 60)))

plot(examsSample$writing_score, examsSample$math_score, main = "Linear Regression", xlab = "x", ylab = "y")
abline(model, col = "red")

plot(model, which = 1)  # Residuals vs. Fitted values plot
```

12.10 Chi square Tests

```r
#chi square tests
# Godness-of-fit
#observed
table(examsSample$lunch)
r = table(examsSample$lunch)[[1]]
s = table(examsSample$lunch)[[2]]
o = c(r,s)
#expected
p = c(0.5,0.5)
#H0: the percentages in our data match the expected values.
#Ha: the percentages don't match the expected values.
result = chisq.test(o,p=p)
result
#p-value = 0.0003274 reject the null
result$expected
result$observed
#The difference between expected and observed is as plotted.
# Create a first line
plot(result$expected, type = "b", frame = FALSE, pch = 19,
     col = "#804674", xlab = "lunch", ylab = "count", ylim = c(0,100), xaxt = "n",)
axis(1,                         # Define x-axis manually
     at = 1:2,
     labels = c('reduces_free_lunch', 'standard_lunch'))
# Add a second line
lines(result$observed, pch = 18, col = "#51829B", type = "b", lty = 2)
# Add a legend to the plot
legend("right", legend=c("expected", "observed"),
       col=c("#804674", "#51829B"), lty = 1:2, cex=0.8)
#independence test
#H0: The categorical variables gender and lunch are independent.
#Ha: The categorical variables are dependent.
table(examsSample$gender, examsSample$lunch)
chisq.test(table(examsSample$gender, examsSample$lunch))
#p-value = 0.4332>0.05 fail to reject
```

12.11 Homogeneity Test

```
# homogeneity test
#H0: The male and female are homogeneous when it comes to the categories of the lunch variable.
#Ha: There is a difference when it comes to the categories of the variable lunch between male and female.
sample1 = systematic_sampling(exams[exams$gender == "male", ], SAMPLE_SIZE)
sample2 = systematic_sampling(exams[exams$gender == "female", ], SAMPLE_SIZE)
sample3 = rbind(sample1, sample2)
table(sample3$gender, sample3$lunch)
chisq.test(table(sample3$gender, sample3$lunch))
#p-value = 0.5337 fail to reject the null hypothesis
```

12.12 One way ANOVA

```
#one way ANOVA
#H0: The reading score mean is the same for all group categories.
#Ha: The reading score mean is the different for at least one group category.
#unique(exams$race_ethnicity)
examsSample$race_ethnicity = ordered(examsSample$race_ethnicity, levels = c("group A", "group B", "group C","group E", "group D"))
anova_t1 = aov(reading_score~race_ethnicity, data=examsSample)
summary(anova_t1)
#The p-value =  0.0755,fail to reject the null hypothesis
#perform Scheffe's test
ScheffeTest(anova_t1)
```

12.13 Two way ANOVA

```
##two-way
#Three hypotheses:
#H01: There is no difference between the means of Reading score for gender categories.
#Ha1: There is a difference between the means of Reading score for gender categories.

#H02: There is no difference between the means of Reading score for test preparation course categories.
#Ha2: There is a difference between the means of Reading score for test preparation course categories.

#H03: There is no interaction affect between the gender and the test preparation course
#categories on the Reading score means.
#Ha3: There is an interaction affect between the gender and the test preparation course categories on the Reading score means

anova_t2 = aov(reading_score~gender*test_preparation_course, data=examsSample)
summary(anova_t2)
```

12.14 Non parametric Tests

```
# non-parametric tests

#H0: The median of math_score equals 90
#Ha: The median of math_score is less than 90
#test
#Sign test
s2 = systematic_sampling(exams, 20)
SIGN.test(s2$math_score ,md=90, alternative = 'less')
#p-value is too small reject the null hypothesis


#Wilcox rank sum test.
#H0: The median of math_score for students who took or didn't take the test preparation course is the same.
#Ha: The median of math_score for tudents who took or didn't take the test preparation course is different.

wilcox.test(s2$math_score[s2$test_preparation_course == "completed"],s2$math_score[s2$test_preparation_course == "none"])
#fail to reject ,p-value = 0.3219
#Spearman correlation
#H0: There is no correlation between math_score and reading_score
#Ha: There is a correlation between math_score and reading_score
cor.test(s2$math_score, s2$reading_score , method = "spearman", exact = FALSE)
# reject the null hypothesis ,p-value = 5.467e-11

# runs test
#H0: The sequence is random.
#Ha: The sequence is not random.
runs_test_sample = exams$math_score[1:90] %%2
runs_test_sample
runs.test(runs_test_sample)
#fail to reject the null hypothesis
```

CONCLUSION

This study provides a comprehensive statistical analysis of factors affecting high school students' academic performance. The results highlight significant determinants such as gender, parental education level, and participation in test preparation courses. These insights can help educators and policymakers develop targeted strategies to improve student outcomes and address educational inequalities.

REFERENCES

- Kaggle Dataset: Student performance prediction (kaggle.com)