Statistical Analysis on Student Performance - Probability & Stats Project

Himanshu - Jaival - Shalini

2025-06-13

```r
# Set working directory and clear environment
setwd('/Users/shalini/Desktop/P&S Project')
rm(list=ls())

# Load the dataset
# The dataset is from kaggle.
# Link : https://www.kaggle.com/datasets/spscientist/students-performance-in-exams/
student_data <- read.csv('StudentsPerformance.csv', stringsAsFactors = TRUE)
head(student_data)
```

```
##   gender race.ethnicity parental.level.of.education        lunch
## 1 female        group B           bachelor's degree     standard
## 2 female        group C                some college     standard
## 3 female        group B              master's degree     standard
## 4   male        group A           associate's degree free/reduced
## 5   male        group C                some college     standard
## 6 female        group B           associate's degree     standard
##   test.preparation.course math.score reading.score writing.score
## 1                    none         72            72            74
## 2               completed         69            90            88
## 3                    none         90            95            93
## 4                    none         47            57            44
## 5                    none         76            78            75
## 6                    none         71            83            78
```

```r
# 1. Descriptive Statistics
# We have 5 categorical columns and 3 numercial columns. Some key insights are there are more women.
# Group C ethnicity has the highest count. Most parents have attended college.
# Most of the students use standard lunch. Most of them have not taken the test prep course.
# The mean score of math is 66.1, reading is 69.2 and writing is 68.1.
summary(student_data)
```

```
##     gender    race.ethnicity     parental.level.of.education          lunch
##  female:518   group A: 89    associate's degree:222          free/reduced:355
##  male  :482   group B:190    bachelor's degree :118          standard    :645
##               group C:319    high school       :196
##               group D:262    master's degree   : 59
##               group E:140    some college      :226
##                              some high school  :179
```

```
##  test.preparation.course   math.score     reading.score    writing.score
##  completed:358           Min.   :  0.00   Min.   : 17.00   Min.   : 10.00
##  none     :642           1st Qu.: 57.00   1st Qu.: 59.00   1st Qu.: 57.75
##                          Median : 66.00   Median : 70.00   Median : 69.00
##                          Mean   : 66.09   Mean   : 69.17   Mean   : 68.05
##                          3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
##                          Max.   :100.00   Max.   :100.00   Max.   :100.00
```

```r
# We can see from the correlation matrix that math and reading, math and writing has a high correlation
cor(student_data[, c(6, 7, 8)])
```

```
##               math.score reading.score writing.score
## math.score     1.0000000     0.8175797     0.8026420
## reading.score  0.8175797     1.0000000     0.9545981
## writing.score  0.8026420     0.9545981     1.0000000
```

```r
# 2. One Sample and Two Sample T-Test

# One-sample t-tests
# p is high → null will fly: Null hypothesis is accepted.
# sample mean is close to population mean
t.test(student_data$math.score, mu = 66.5)
```

```
##
##  One Sample t-test
##
## data:  student_data$math.score
## t = -0.85715, df = 999, p-value = 0.3916
## alternative hypothesis: true mean is not equal to 66.5
## 95 percent confidence interval:
##  65.14806 67.02994
## sample estimates:
## mean of x
##    66.089
```

```r
# p is low → null will go: Null hypothesis is rejected.
# sample mean is significantly different
t.test(student_data$math.score, mu = 70)
```

```
##
##  One Sample t-test
##
## data:  student_data$math.score
## t = -8.1564, df = 999, p-value = 1.029e-15
## alternative hypothesis: true mean is not equal to 70
## 95 percent confidence interval:
##  65.14806 67.02994
## sample estimates:
## mean of x
##    66.089
```

```r
# Two-sample t-test
# p is low → null will go: Null hypothesis is rejected.
# Means differ significantly between genders
t.test(math.score ~ gender, data = student_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  math.score by gender
## t = -5.398, df = 997.98, p-value = 8.421e-08
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -6.947209 -3.242813
## sample estimates:
## mean in group female   mean in group male
##             63.63320             68.72822
```

```r
# 3. One-Way ANOVA

# ANOVA: math score ~ parental education level
# We are comparing math scores across 6 levels of parental education (hence df = 5).
# The F-statistic is 6.522, and the p-value is 5.59e-06 (very small).
# Conclusion: The difference in mean math scores is statistically significant
# across different parental education levels.
aov_test1 <- aov(math.score ~ parental.level.of.education, data = student_data)
summary(aov_test1)
```

```
##                              Df Sum Sq Mean Sq F value   Pr(>F)
## parental.level.of.education   5   7296  1459.1   6.522 5.59e-06 ***
## Residuals                   994 222394   223.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA: math score ~ gender
# We are comparing math scores between 2 gender groups (df = 1).
# The F-statistic is 28.98, and the p-value is 9.12e-08.
# Conclusion: There is a statistically significant difference in
# math scores between genders.
aov_test2 <- aov(math.score ~ gender, data = student_data)
summary(aov_test2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## gender        1   6481    6481   28.98 9.12e-08 ***
## Residuals   998 223208     224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA: math score ~ race/ethnicity
# We are comparing math scores across 5 racial/ethnic groups (df = 4).
# The F-statistic is 14.59, and the p-value is 1.37e-11.
# Conclusion: There is a highly significant difference in mean math scores
```

```r
# between at least some race/ethnicity groups.
aov_test3 <- aov(math.score ~ race.ethnicity, data = student_data)
summary(aov_test3)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## race.ethnicity   4  12729    3182   14.59 1.37e-11 ***
## Residuals      995 216960     218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 4. Linear Regression

# Simple Linear Regression
# We fitted a simple linear regression model to examine the effect of parental level of education on ma

#While the model is statistically significant overall (F(5, 994) = 6.52, p < 0.001),
#parental education alone is not a strong predictor of math performance.
#Only the lowest education categories show significant negative effects,
#suggesting that students with less-educated parents may be at a disadvantage,
#but other variables likely play a much larger role in explaining student math scores.


data.lm = lm(math.score ~ parental.level.of.education, data = student_data)
summary(data.lm)
```

```
##
## Call:
## lm(formula = math.score ~ parental.level.of.education, data = student_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.497  -9.138   0.186  10.503  36.862
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                                67.8829     1.0039  67.619
## parental.level.of.educationbachelor's degree   1.5069     1.7041   0.884
## parental.level.of.educationhigh school        -5.7451     1.4661  -3.919
## parental.level.of.educationmaster's degree     1.8629     2.1909   0.850
## parental.level.of.educationsome college       -0.7546     1.4134  -0.534
## parental.level.of.educationsome high school   -4.3857     1.5026  -2.919
##                                           Pr(>|t|)
## (Intercept)                                < 2e-16 ***
## parental.level.of.educationbachelor's degree  0.37674
## parental.level.of.educationhigh school       9.51e-05 ***
## parental.level.of.educationmaster's degree    0.39537
## parental.level.of.educationsome college       0.59356
## parental.level.of.educationsome high school   0.00359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.96 on 994 degrees of freedom
## Multiple R-squared:  0.03176,    Adjusted R-squared:  0.02689
```

```
## F-statistic: 6.522 on 5 and 994 DF,  p-value: 5.592e-06
```

```
# Multiple Linear Regression

# We then fitted a multiple linear regression model including parental education, gender, test preparat

#This model shows that gender and test preparation completion are strong predictors of math scores,
#while parental education continues to show a negative impact at lower levels.
#Although the explanatory power is still limited (Adjusted R² = 0.086), the model gives a more
#nuanced picture than using parental education alone. The p value is very low at <2e-16 and we reject t
#Parental.level.of.education, gender, test.preparation.course has an influence on math score.

data.mlm = lm(math.score ~ parental.level.of.education + gender + test.preparation.course, data = studer
summary(data.mlm)
```

```
##
## Call:
## lm(formula = math.score ~ parental.level.of.education + gender +
##     test.preparation.course, data = student_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.517  -9.803   0.261  10.033  41.203
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                 68.8034     1.2287  55.998
## parental.level.of.educationbachelor's degree   1.4552     1.6513   0.881
## parental.level.of.educationhigh school        -5.5143     1.4233  -3.874
## parental.level.of.educationmaster's degree     2.4965     2.1245   1.175
## parental.level.of.educationsome college       -0.5993     1.3698  -0.437
## parental.level.of.educationsome high school   -4.7949     1.4571  -3.291
## gendermale                                     5.3257     0.9188   5.796
## test.preparation.coursenone                   -5.4920     0.9606  -5.717
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## parental.level.of.educationbachelor's degree 0.378402
## parental.level.of.educationhigh school       0.000114 ***
## parental.level.of.educationmaster's degree   0.240230
## parental.level.of.educationsome college      0.661844
## parental.level.of.educationsome high school  0.001035 **
## gendermale                                   9.10e-09 ***
## test.preparation.coursenone                  1.43e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.49 on 992 degrees of freedom
## Multiple R-squared:  0.09284,    Adjusted R-squared:  0.08644
## F-statistic:  14.5 on 7 and 992 DF,  p-value: < 2.2e-16
```

```
# 5.Chi-Square test
#Check if two categorical varaiables are related to each other.
# Create contingency table
```

```r
data.chi <- table(student_data$gender, student_data$test.preparation.course)

# Perform Chi-Square Test of Independence
#Here we get the p value as 0.9. P value is high, null will fly.Null hypothesis will be accepted.
#These two categorical variables are no related to each other.
chisq.test(data.chi)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data.chi
## X-squared = 0.015529, df = 1, p-value = 0.9008
```