2021101113_Hypothesis_Testing_Homework

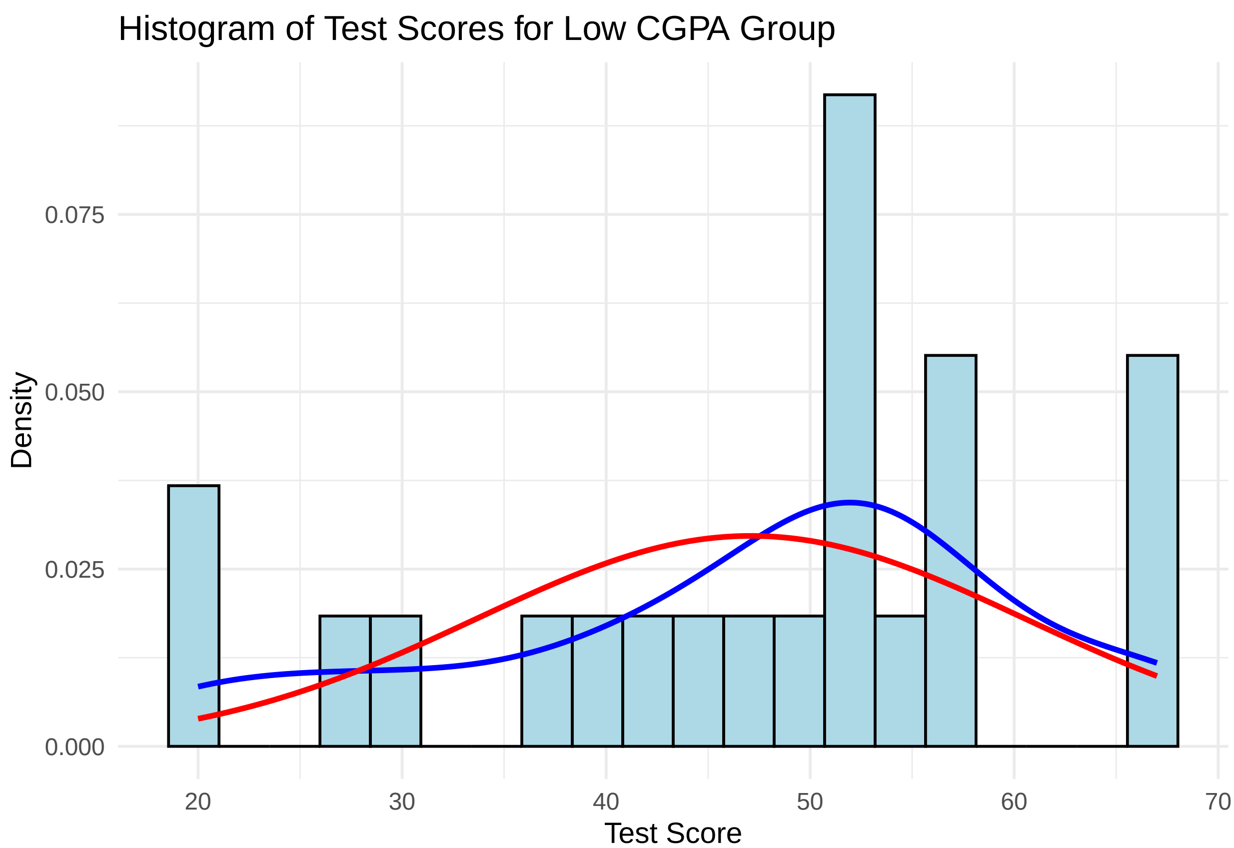
Gowlapalli Rohit

13/2/2024

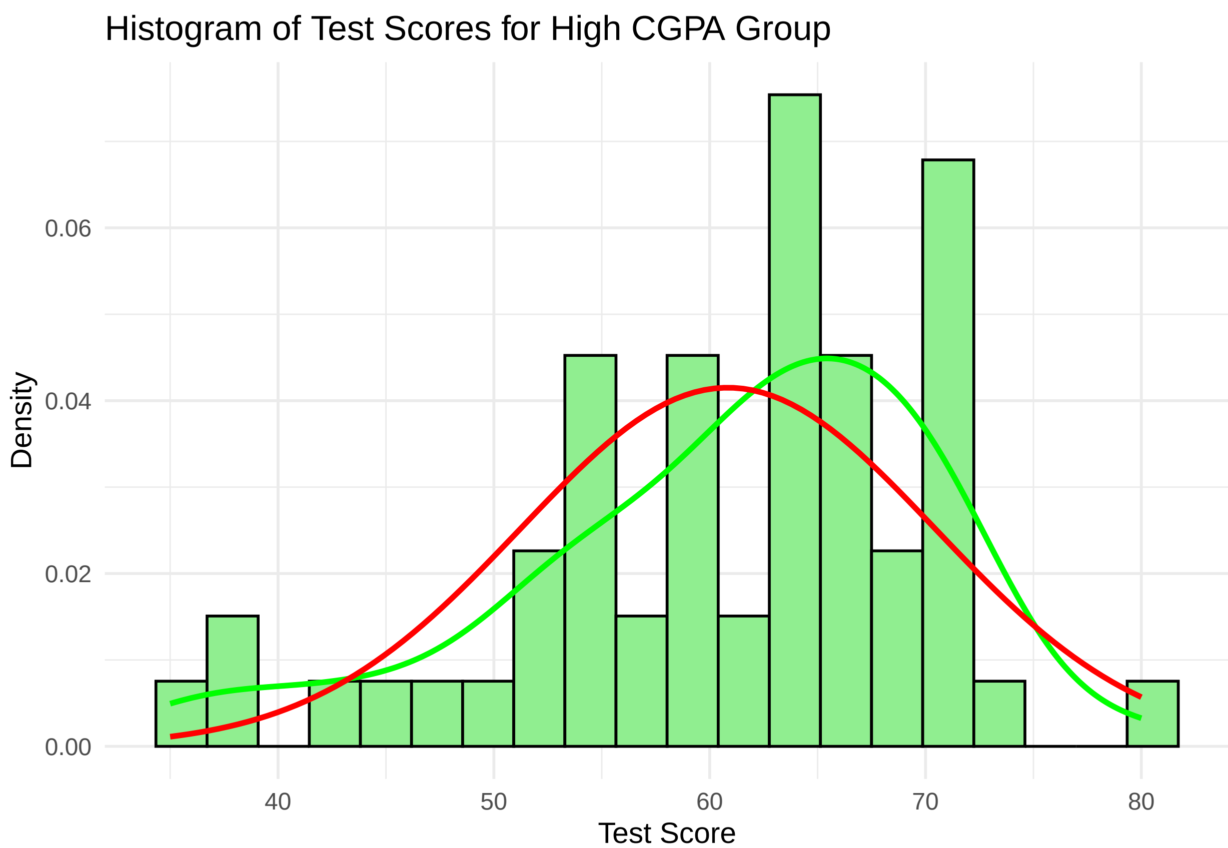
Contents


```
library(ggplot2)
data <- readxl::read_excel("data.xlsx")
low_cgpa <- data[data$GPA < 7, "TESTSCORE"]
high_cgpa <- data[data$GPA >= 7, "TESTSCORE"]
```

```
ggplot() +
  geom_histogram(data = low_cgpa, aes(x = TESTSCORE, y = ..density..), fill = "lightblue", color = "bla
  geom_density(data = low_cgpa, aes(x = TESTSCORE), color = "blue", size = 1) +
  stat_function(data = low_cgpa, aes(x = TESTSCORE), fun = dnorm, args = list(mean = mean(low_cgpa$TESTS
  labs(title = "Histogram of Test Scores for Low CGPA Group", x = "Test Score", y = "Density") +
  theme_minimal()
```




```
ggplot() +
  geom_histogram(data = high_cgpa, aes(x = TESTSCORE, y = ..density..), fill = "lightgreen", color = "bl
  geom_density(data = high_cgpa, aes(x = TESTSCORE), color = "green", size = 1) +
  stat_function(data = high_cgpa, aes(x = TESTSCORE), fun = dnorm, args = list(mean = mean(high_cgpa$TES
  labs(title = "Histogram of Test Scores for High CGPA Group", x = "Test Score", y = "Density") +
  theme_minimal()
```




```
test_scores_low_cgpa <- low_cgpa$TESTSCORE
test_scores_high_cgpa <- high_cgpa$TESTSCORE

n1 <- length(test_scores_low_cgpa)
n2 <- length(test_scores_high_cgpa)

mean_low_cgpa <- mean(test_scores_low_cgpa)
mean_high_cgpa <- mean(test_scores_high_cgpa)

var1 <- var(test_scores_low_cgpa)
var2 <- var(test_scores_high_cgpa)

pooled_variance <- ((n1 - 1) * var1 + (n2 - 1) * var2) / (n1 + n2 - 2)

cat("Low CGPA Group:\n")
```

```
## Low CGPA Group:
```

```
cat("Observations Lower GPA:", n1, "\n")
```

```
## Observations Lower GPA: 22
```

```
cat("Mean:", mean_low_cgpa, "\n")
```

```
## Mean: 47.09091
```

```r
cat("Variance:", var1, "\n\n")
```

```
## Variance: 180.7532
```

```r
cat("High CGPA Group:\n")
```

```
## High CGPA Group:
```

```r
cat("Observations Higher GPA:", n2, "\n")
```

```
## Observations Higher GPA: 56
```

```r
cat("Mean:", mean_high_cgpa, "\n")
```

```
## Mean: 60.83929
```

```r
cat("Variance:", var2, "\n")
```

```
## Variance: 92.39188
```

```r
cat("\n")
```

```r
cat("Hypothesized Mean Difference:", mean_low_cgpa - mean_high_cgpa, "\n")
```

```
## Hypothesized Mean Difference: -13.74838
```

```r
cat("Pooled Variance:", pooled_variance, "\n")
```

```
## Pooled Variance: 116.8075
```

Inferences:

The high CGPA group has a higher mean test score compared to the low CGPA group. Additionally, the variance in test scores is lower in the high CGPA group, suggesting less variability compared to the low CGPA group.

Normality Test

```r
perform_shapiro_test <- function(group) {
  shapiro_result <- shapiro.test(group)

  p_value <- shapiro_result$p.value
  test_statistic <- shapiro_result$statistic
  print(shapiro_result)
```

```r
  if (p_value > 0.05) {
    message <- "Data is normally distributed (p > 0.05)"
  } else {
    message <- "Data is not normally distributed (p <= 0.05)"
  }
  print(message)
  return(list(
    Test_Statistic = test_statistic,
    P_Value = p_value,
    Conclusion = message
  ))
}

cat("Shapiro-Wilk Test for Low CGPA Group:\n")
```

```
## Shapiro-Wilk Test for Low CGPA Group:
```

```r
shapiro_low_cgpa <- perform_shapiro_test(low_cgpa$TESTSCORE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.93498, p-value = 0.1558
##
## [1] "Data is normally distributed (p > 0.05)"
```

```r
cat("\nShapiro-Wilk Test for High CGPA Group:\n")
```

```
##
## Shapiro-Wilk Test for High CGPA Group:
```

```r
shapiro_high_cgpa <- perform_shapiro_test(high_cgpa$TESTSCORE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.94246, p-value = 0.009911
##
## [1] "Data is not normally distributed (p <= 0.05)"
```

Inferences:

Low CGPA Group:

- Shapiro-Wilk test for normality indicates that the data is normally distributed ($p > 0.05$) with $W = 0.93498$ and p-value $= 0.1558$.

High CGPA Group:

- Shapiro-Wilk test for normality indicates that the data is not normally distributed ($p <= 0.05$) with $W = 0.94246$ and p-value = 0.009911.

Since the Shapiro-Wilk test for the high CGPA group indicates that the data is not normally distributed ($p <= 0.05$), it's better to avoid parametric tests like the t-test, which assume normality. Instead, we should use non-parametric tests like the Wilcoxon rank-sum (Mann-Whitney U) test.

```
correlation_spearman <- cor.test(data$GPA, data$TESTSCORE, method = "spearman")
print(correlation_spearman)
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$GPA and data$TESTSCORE
## S = 34747, p-value = 9.409e-08
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5606084
```

Choosing the right test

```
choose_test <- function(group1, group2) {
  if (shapiro_low_cgpa$P_Value > 0.05 && shapiro_high_cgpa$P_Value > 0.05) {
    message <- "Independent t-test can be used as both groups are normally distributed"
    test_result <- t.test(group1, group2, var.equal = TRUE)
  } else {
    message <- "Mann-Whitney U test can be used as at least one group is not normally distributed"
    test_result <- wilcox.test(group1, group2)
  }
  return(list(message, test_result))
}

cat("Choosing the appropriate test:\n")
```

```
## Choosing the appropriate test:
```

```
test <- choose_test(low_cgpa$TESTSCORE, high_cgpa$TESTSCORE)
print(test[1])
```

```
## [[1]]
## [1] "Mann-Whitney U test can be used as at least one group is not normally distributed"
```

```
print(test[2])
```

```
## [[1]]
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  group1 and group2
## W = 240.5, p-value = 3.084e-05
## alternative hypothesis: true location shift is not equal to 0
```

Inferences:

Choosing the Appropriate Test:

- Mann-Whitney U test is chosen as at least one group is not normally distributed.

Mann-Whitney U Test Results:

- The Wilcoxon rank sum test with continuity correction yielded a test statistic (W) of 240.5 and a p-value of 3.084e-05.
- The alternative hypothesis suggests that the true location shift is not equal to 0.

Calculating Effect-size

```
effect_size <- function(group1, group2) {
  mean_diff <- mean(group1) - mean(group2)
  pooled_sd <- sqrt(((length(group1) - 1) * var(group1) + (length(group2) - 1) * var(group2)) / (length
  cohen_d <- mean_diff / pooled_sd
  return(cohen_d)
}

cat("Effect Size:\n")
```

```
## Effect Size:
```

```
cat("Cohen's d:", abs(effect_size(low_cgpa$TESTSCORE, high_cgpa$TESTSCORE)),"\n")
```

```
## Cohen’s d: 1.272085
```

Inferences:

Effect Size: This indicates a large effect size, suggesting a substantial difference between the means of the two groups being compared.

Reporting Statistics

```r
report_statistics_mann_whitney <- function(test_result, alpha = 0.05) {
    cat("Mann-Whitney U Test Results:\n")
    cat("U-Statistic:", test_result$statistic, "\n")
    cat("P-Value:", test_result$p.value, "\n")
    cat("Conclusion:")
    if (test_result$p.value < alpha) {
        cat("Reject the null hypothesis and conclude that there is a significant difference between the
    } else {
        cat("Fail to reject the null hypothesis and conclude that there is no significant difference be
    }
}

report_statistics_t_test <- function(test_result, alpha = 0.05) {
    cat("Independent t-test Results:\n")
    cat("t-Statistic:", test_result$statistic, "\n")
    cat("Degrees of Freedom:", test_result$parameter, "\n")
    cat("P-Value:", test_result$p.value, "\n")
    cat("Confidence Interval:", test_result$conf.int, "\n")
    cat("Conclusion:")
    if (test_result$p.value < alpha) {
        cat("Reject the null hypothesis and conclude that there is a significant difference between the
    } else {
        cat("Fail to reject the null hypothesis and conclude that there is no significant difference be
    }
}

if (test[1] == "Independent t-test can be used as both groups are normally distributed") {
    report_statistics_t_test(test[[2]])
} else {
    report_statistics_mann_whitney(test[[2]])
}
```

```
## Mann-Whitney U Test Results:
## U-Statistic: 240.5
## P-Value: 3.084429e-05
## Conclusion:Reject the null hypothesis and conclude that there is a significant difference between th
```

Inferences:

- **U-statistic**: The U-statistic of 240.5 indicates the sum of ranks assigned to observations from the groups being compared. A higher U-statistic suggests greater difference between the distributions of the two groups.
- **p-value**: The very low p-value of 3.084429e-05 indicates strong evidence against the null hypothesis. It suggests that the observed difference between the two groups' distributions is unlikely to have occurred by random chance alone, leading to rejection of the null hypothesis.

Based on the Mann-Whitney U test results:

- **Question:** Do students with a GPA <= 7 have lower placement TESTSCORES than those with a GPA > 7?

- **Result:** The Mann-Whitney U test yielded a U-statistic of 240.5 and a very low p-value of 3.084429e-05.
- **Conclusion:** Reject the null hypothesis, indicating that there is a significant difference between the two groups.

Inference: Students with a GPA <= 7 are inferred to have lower placement test scores than those with a GPA > 7, based on the significant difference observed between the groups.