Section 6 | Hypothesis Testing

Mohammad Saqib Ansari

2023-12-01

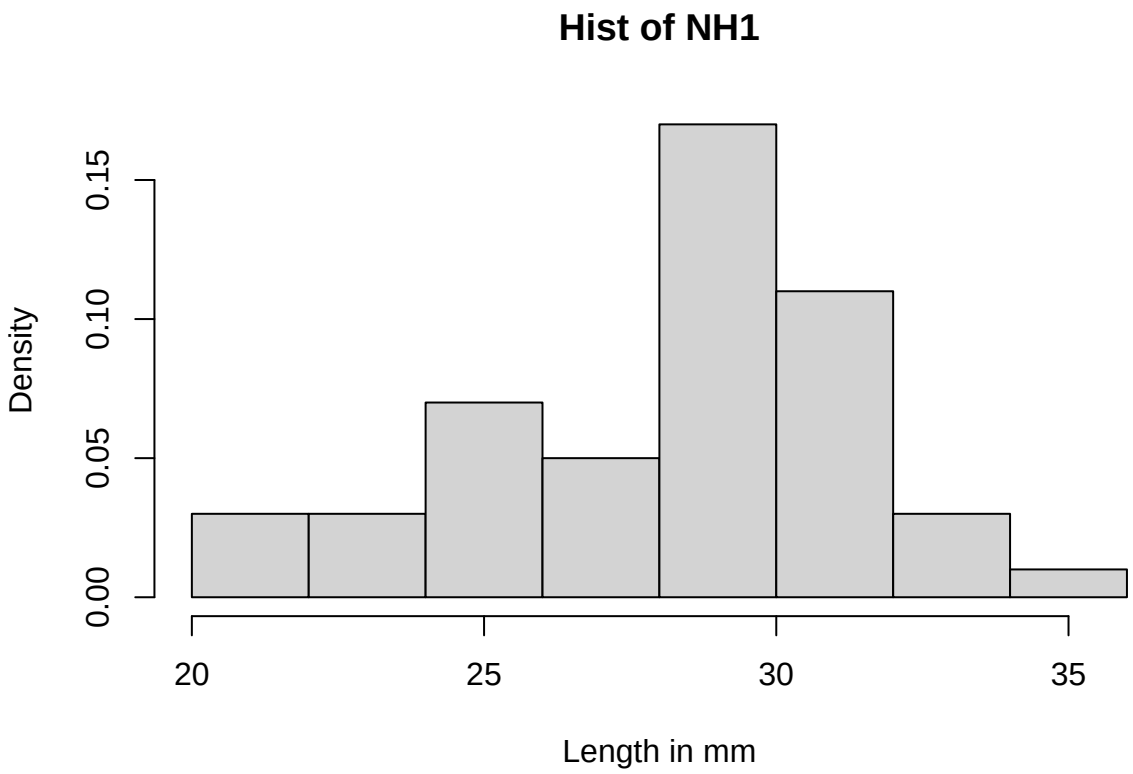
Confidence Interval for the Mean

Read Data and Descriptive Statistics

```
library("readxl")

# Reading data for NH1 and NH2
s1 <- read_excel("C:/Users/hp/OneDrive/Desktop/R For Statistics/New Data.xlsx", 1)
s2 <- read_excel("C:/Users/hp/OneDrive/Desktop/R For Statistics/New Data.xlsx", 2)

# Summary statistics for Nasal Height of 1 year
summary(s1$`Nasal Height of 1 years, mm`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    21.0    26.5    29.0    28.7    31.0    35.0
```

```
mean_nh1 <- mean(s1$`Nasal Height of 1 years, mm`)
sd_nh1 <- sd(s1$`Nasal Height of 1 years, mm`)

# Histogram for NH1
hist(s1$`Nasal Height of 1 years, mm`, main = "Hist of NH1", freq = FALSE,
     ylab = "Density", xlab = "Length in mm")
```




```
# Testing normality using Shapiro-Wilk test
shapiro.test(s1$`Nasal Height of 1 years, mm`)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  s1$‘Nasal Height of 1 years, mm‘
## W = 0.94469, p-value = 0.02073
```

The above code reads the data for nasal height of 1 year (`NH1`) from an Excel file, calculates summary statistics (mean and standard deviation), creates a histogram to visualize the distribution, and performs the Shapiro-Wilk test to assess the normality assumption. The output includes the summary statistics, histogram, and the Shapiro-Wilk test result.

Constructing Confidence Interval

```
# Confidence interval using t-test
ci_t_test <- t.test(s1$`Nasal Height of 1 years, mm`, conf.level = 0.95)$conf.int

# Confidence interval calculation by hand
n <- 900  # sample size
x_bar <- 3.4  # sample mean
sd <- 2.61  # sample standard deviation
```

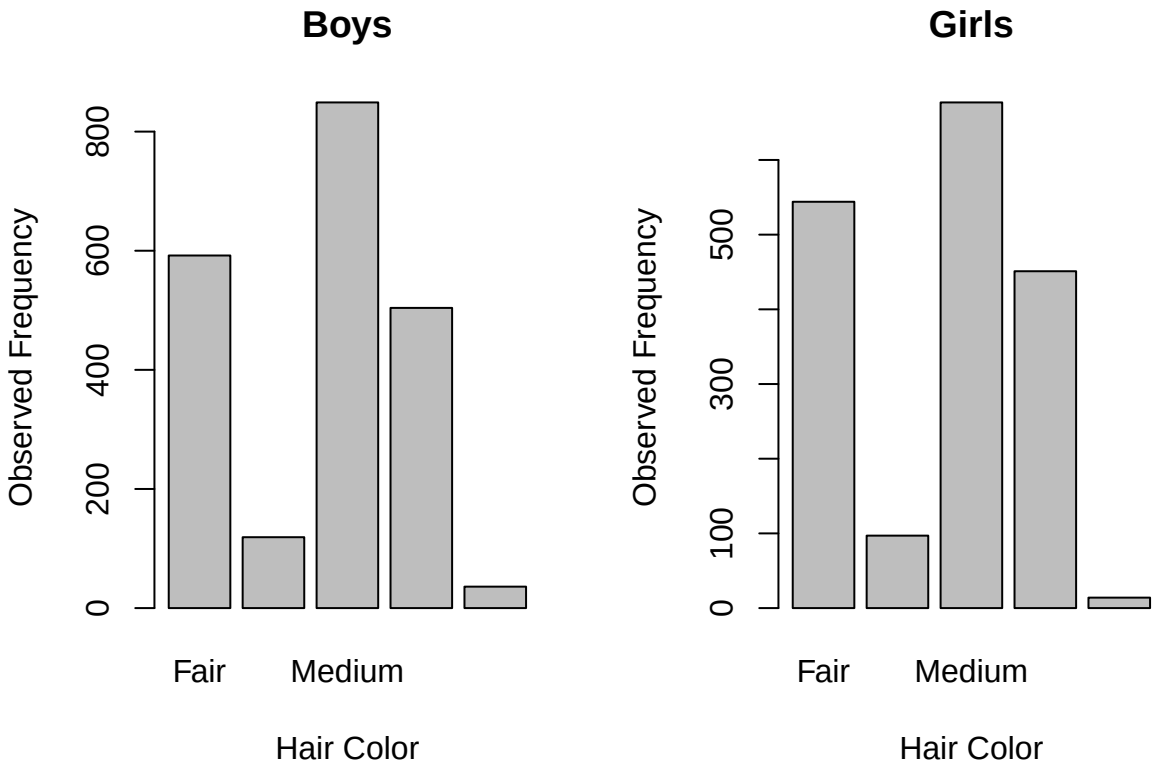
```r
p_mu <- 3.25  # population mean

# Calculating z-score
Z <- (x_bar - p_mu) / (sd / sqrt(n))

# 95% fiducial limits
upper_limit <- x_bar + qnorm(1 - 0.025) * (sd / sqrt(n))
lower_limit <- x_bar - qnorm(1 - 0.025) * (sd / sqrt(n))
```

This section constructs a confidence interval for the mean nasal height of 1 year using two methods: the `t.test` function and manual calculation using sample statistics.

Pearson's Chi-Square Test of Independence

Input Data and Visualize

```r
# Input data for Boys and Girls hair color
Boys <- c(592, 119, 849, 504, 36)
Girls <- c(544, 97, 677, 451, 14)
color <- rbind(Boys, Girls)
rownames(color) <- c("Boys", "Girls")
colnames(color) <- c("Fair", "Red", "Medium", "Dark", "Jet Black")

# Visualizing the data with barplots
par(mfrow = c(1, 2))
barplot(color[1, ], main = "Boys", xlab = "Hair Color", ylab = "Observed Frequency")
barplot(color[2, ], main = "Girls", xlab = "Hair Color", ylab = "Observed Frequency")
```




```
# Row and column profiles
round(100 * color / sum(color), 1)  # Joint frequency
```

```
##       Fair Red Medium Dark Jet Black
## Boys  15.2 3.1   21.9 13.0       0.9
## Girls 14.0 2.5   17.4 11.6       0.4
```

```
round(100 * prop.table(color, margin = 1), 1)  # Row profile
```

```
##       Fair Red Medium Dark Jet Black
## Boys  28.2 5.7   40.4 24.0       1.7
## Girls 30.5 5.4   38.0 25.3       0.8
```

```
round(100 * prop.table(color, margin = 2), 1)  # Column profile
```

```
##       Fair  Red Medium Dark Jet Black
## Boys  52.1 55.1   55.6 52.8        72
## Girls 47.9 44.9   44.4 47.2        28
```

```
# Chi-square test
results <- chisq.test(color)
```

This section performs Pearson's Chi-Square Test of Independence to check if hair color is independent of gender among Boys and Girls. It starts with inputting data, visualizing it using barplots, and then calculates row and column profiles. Finally, it conducts the chi-square test using `chisq.test()`.

Results and Contributions

```
# Test results and contributions
results
```

```
##
##  Pearson's Chi-squared test
##
## data:  color
## X-squared = 10.467, df = 4, p-value = 0.03325
```

```
round(100 * results$residuals^2 / results$stat, 1)
```

```
##       Fair Red Medium Dark Jet Black
## Boys   7.8 0.4    6.5  2.9      28.4
## Girls  9.2 0.5    7.7  3.4      33.4
```

```
round(results$residuals, 3)
```

```
##         Fair    Red Medium   Dark Jet Black
## Boys  -0.903  0.202  0.825 -0.549     1.723
## Girls  0.979 -0.219 -0.896  0.596    -1.870
```

The output includes the chi-square test results showing the chi-square statistic, degrees of freedom, and p-value. Additionally, it presents the contributions of each cell to the chi-square statistic and the standardized residuals for each cell, indicating the deviation from expected frequencies.

```
#####################################
## Comparison of Two Means

# Step 1: Read the data
# Step 2: Compare two subpopulations graphically
# Step 3: Calculate descriptive statistics (mean, SD, quantiles)
# Step 4: (Optional) Test the normality of data in each subpopulation
# Step 5: Test the equality of variance
# Step 6: Test the equality of means

# Define data for Diet A and Diet B
dietA <- c(25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25)
dietB <- c(44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22)

# Graphical comparison using boxplot
label <- c("Diet A", "Diet B")
boxplot(dietA, dietB, xlab = "Diet", ylab = "Weight", names = label)

# Descriptive statistics for Diet A
mean_dietA <- mean(dietA)
sd_dietA <- sd(dietA)
quantile_dietA <- quantile(dietA)

# Descriptive statistics for Diet B
```

```r
mean_dietB <- mean(dietB)
sd_dietB <- sd(dietB)
quantile_dietB <- quantile(dietB)

# Testing normality for Diet A and Diet B
qqnorm(dietA)
qqline(dietA, col = "grey")
shapiro_test_dietA <- shapiro.test(dietA) # p_val = 0.1095

# Similar normality testing for Diet B

# Testing equality of variance
variance_test <- var.test(dietA, dietB, conf.level = 0.95) # p_value > 0.05

# Testing equality of means
t_test_result <- t.test(dietA, dietB, alternative = 'two.sided', conf.level = 0.95,
                        var.equal = TRUE) # p_value > 0.05

##########################################
## Testing Conformity of Proportion

# Step: Test the equality of the proportion to 50% with a type-one error rate 5%
binom_test_result <- binom.test(18, n = 20, p = 0.85, alternative = "greater") # p_value = 0.4049

# Comparing several proportions
Boys_col <- c(592, 119, 849, 504, 36)
total_pop <- c(1136, 216, 1526, 955, 50)

prop_test_result <- prop.test(Boys_col, total_pop) # p_value < 0.05
```

Explanation: - **Comparison of Two Means:** This section outlines steps to compare two datasets ('Diet A' and 'Diet B') regarding weight measurements. It includes graphical comparisons, descriptive statistics, normality testing, variance testing, and means testing between the two datasets. - **Testing Conformity of Proportion:** This part involves two tests. Firstly, it examines if a proportion differs significantly from 50% using a binomial test. Secondly, it compares multiple proportions using a proportion test.

Definitions:

- **Descriptive Statistics:** Measures that summarize and describe features of a dataset (e.g., mean, standard deviation, quantiles).
- **Normality Testing:** Checking if data follows a normal distribution.
- **Equality of Variance Test:** Testing if the variances of two datasets are equal.
- **Equality of Means Test:** Examining if the means of two datasets are significantly different.
- **Binomial Test:** A statistical test to assess if a proportion significantly differs from a specified value (e.g., 50%).
- **Proportion Test:** Comparing proportions across multiple groups to determine if they are significantly different.