Hypothesis Testing

Max Johansson

Spring 2024

1. Introduction

In this project I perform two hypothesis tests on simulated data. In the first hypothesis test I consider the mean income of the individuals in the data, using the null hypothesis that the mean income of the population is equal to 40 income units against the alternative hypothesis that it is not equal to 40 income units. In the second hypothesis test I consider the mean income between two groups: unmarried and married individuals, testing the null hypothesis that the means of the groups are equal in the population against the alternative hypothesis that the means of the groups are not equal in the population.

2. Data

I use a ISLR2 data set called ‘Credit’. The variables I consider for this project is the continuous ‘income’ variable and the binary ‘married’ variable.

3. Results

3.1. One Sample

I consider the Income variable. I consider testing if the mean of this variable is significantly different than the mean under the null hypothesis, which I set to 40. I use a significance level of 5%, rejecting the null hypothesis if the p-value of the observed t-statistic is equal to or less than 5%.

$$H_0 : \mu = 40 H_1 : \mu \neq 40$$

The p-value associated with the t-statistic is smaller than 5%, leading me to reject the null hypothesis that the mean income in the population is equal to 40 in favour of the alternative hypothesis that the mean income of the population is not equal to 0.

```r
# The mean and the standard deviation of the variable in the data:
m_hat_inc <- mean(df$Income)
sd_hat_inc <- sd(df$Income)

# Setting the null hypothesis:
m0_inc <- 40

# Setting the significance level:
a <- 0.05
```

```r
# Setting the sample size:
n <- nrow(df)

# The degrees of freedom:
d_free <- n - 1

# Critical t value:
t_crit <- qt(1- a/2, df = d_free)

# The t-statistic is:
t_stat <- (m_hat_inc - m0_inc)/ (sd_hat_inc / sqrt(n))

# Upper and lower limits:
ul <- m0_inc + t_crit*(sd_hat_inc / sqrt(n))
ll <- m0_inc - t_crit*(sd_hat_inc/ sqrt(n))

# The p-value of the outcome:
p_val <- 2 * (1- pt(q = abs(t_stat), df = d_free))

# Print the results:
cat("The sample mean is:",
    round(m_hat_inc,3),
    "and a 95% confidence interval estimate of the mean income is:",
    round(ll,3),
    "to",
    round(ul,3), "\n")
```

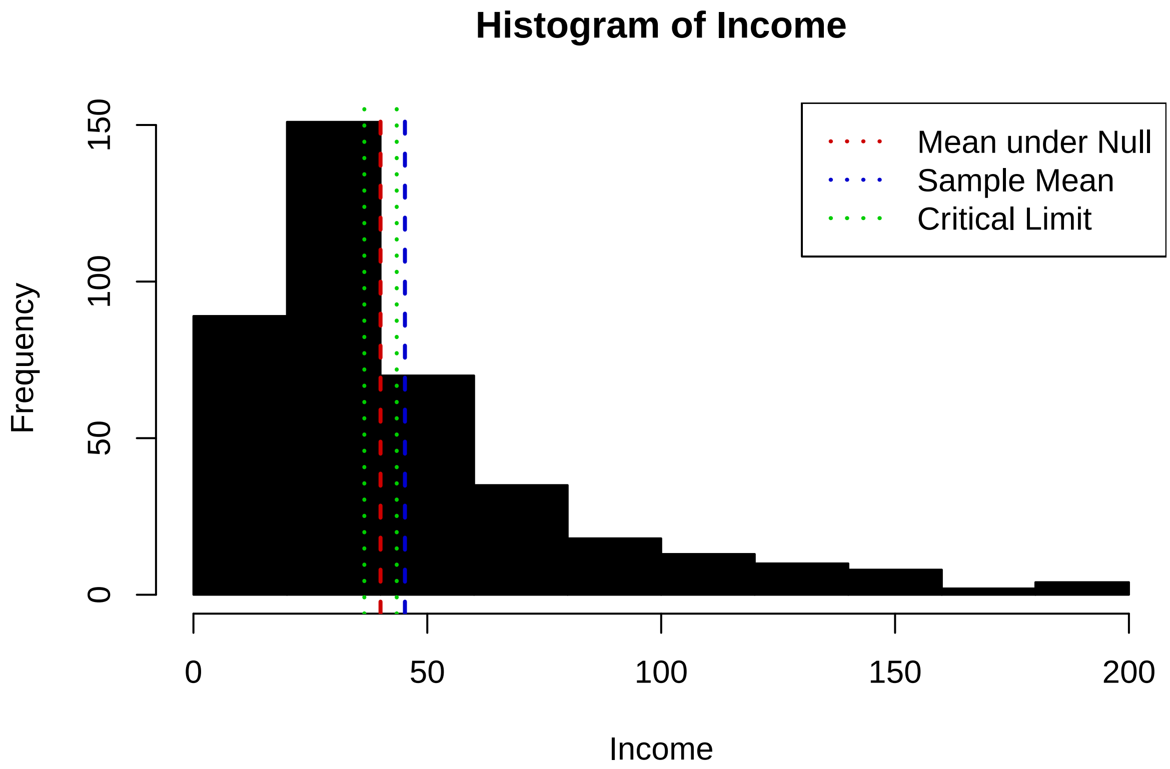
```
## The sample mean is: 45.219 and a 95% confidence interval estimate of the mean income is: 36.536 to 43
```

```r
cat("The t-statistic is:",
    round(t_crit,3),
    "and the associated p-value is:",
    round(p_val, 3),"\n")
```

```
## The t-statistic is: 1.966 and the associated p-value is: 0.003
```

I visualize the results by plotting the distribution of the income variable together with the null hypothesis mean, the sample mean and the critical limits. At a 5% significance level, the null hypothesis that the mean of income is equal to 40 is rejected in favor of the alternative hypothesis that the mean is not equal to 40.

```r
# Histogram
hist(df$Income, main = "Histogram of Income", xlab = "Income", ylab = "Frequency", col = "black")
abline(v = m0_inc, col = "red3", lwd = 2, lty = 2, label = "Mean under Null Hypothesis")
abline(v = m_hat_inc, col = "blue3", lwd = 2, lty = 2, label = "Sample Mean")
abline(v = c(ul, ll), col = "green3",lwd = 2, lty = 3)
legend("topright", legend = c("Mean under Null", "Sample Mean", "Critical Limit"), col = c("red3","blue
```



3.2. Two Samples

I now investigate if there a significant difference in income between unmarried people and married people. The hypotheses are:

$$H_0 : \mu_u = \mu_m H_1 : \mu_u \neq \mu_m$$

According to my results, the p-value associated with the t-statistic is greater than 5%, so the null hypothesis that the means of unmarried and married individuals is equal is not rejected.

```
# Income values per marital status:
vals_unmarried <- subset(df$Income, df$Married == "No")
vals_married <- subset(df$Income, df$Married == "Yes")

# Mean Income of non-married and married groups:
m_unmarried <- mean(vals_unmarried)
m_married <- mean(vals_married)

# The difference:
m_diff <- m_unmarried - m_married

# Standard deviation of the groups:
sd_unmarried <- sd(vals_unmarried)
sd_married <- sd(vals_married)
```

```r
# Number of observations in each group:
n_unmarried <- length(vals_unmarried)
n_married <- length(vals_married)

# Pooled standard deviation:
pooled_sd <- sqrt(((n_unmarried - 1) * sd_unmarried^2 + (n_married -1) * sd_unmarried^2) / (n_unmarried

# t statistic:
t <- m_diff / (pooled_sd * sqrt(1/n_unmarried + 1/ n_married))

# Degrees of freedom:
d_fr <- n_unmarried + n_married - 2

# P-value of t-statistic:
p <- 2 * (1- pt(abs(t), df = d_fr))

# Lower and Upper limit for the difference
ll <- m_diff - qt(0.975, d_fr) * pooled_sd * sqrt(1/n_unmarried + 1/n_married)
ul <- m_diff +  qt(0.975, d_fr)* pooled_sd * sqrt(1/n_unmarried + 1/n_married)

# Print the results:
cat("The difference in means is:",
    round(m_diff,3),
    "and a 95% confidence interval for the difference in means is between:",
    round(ll,3),
    "and",
    round(ul,3),
    "\n")
```

```
## The difference in means is: -2.576 and a 95% confidence interval for the difference in means is betw
```

```r
cat("The t-statistic is",
    round(t,3),
    "and the associated p-value is",
    round(p,3),
    "\n")
```

```
## The t-statistic is -0.775 and the associated p-value is 0.439
```

4. Findings

My results indicate that the mean of the income variable is not 40, and that the mean incomes of the unmarried and married subgroups are not significantly different from each other.

5. Libraries

ISLR2: https://cran.r-project.org/package=ISLR2