MATH 2310 Hypothesis Testing

Jakob Balkovec

2024-05-22

This lab will deal with hypothesis tests. In this lab, you will use software to calculate summary statistics for data, and then use those summary statistics to conduct a hypothesis test by hand.

Goals for this assignment:

- Use R to generate summary statistics
- Practice calculations for hypothesis tests by hand
- Interpret results of hypothesis tests

Grading: there are four possible points for each skill objective and analysis objective.

```r
# Import libraries, and read the data set in a data-frame

suppressPackageStartupMessages({
  library(dplyr)
  library(readxl)
  library(knitr)
})

library(readxl)

path <- "clouds.xlsx"
clouds_df <- suppressMessages(read_excel(path, col_names = TRUE))

# Update data to match the R format (floats instead of tuples)
clouds_df$Rainfall <- as.numeric(sub(",", ".", clouds_df$Rainfall))
```

Activity One

In previous labs, we looked at a research paper from 1983 by Chambers, Cleveland, Kleiner, and Tukey examining the effectiveness of cloud seeding using silver nitrate. Total rainfall (in acre-feet) was measured for 26 seeded clouds and 26 unseeded clouds. We will again be examining this data set further.

We would like to use a hypothesis test to assess whether cloud seeing is effective. As we have small sample sizes, we need to be able to assume our data is approximately normally distributed. In a previous lab, you showed that the histograms for this data are extremely skewed. However, many of you noted that using a log transformation resulted in data that looked approximately normally distributed.

Skill Objective:

Use R to take a log transformation of the rainfall data. Then, use R to calculate the means and standard deviations for the log of rainfall for seeded clouds, and for unseeded clouds.

Analysis Objective:

Using these results, conduct a hypothesis test to determine whether there is evidence that cloud seeding increases rainfall.

```r
# Log transform the data
clouds_df$LogRainfall <- log(clouds_df$Rainfall)
```

```r
# Means and standard deviations

seeded_log <- subset(clouds_df, Treatment == "Seeded")$LogRainfall
unseeded_log <- subset(clouds_df, Treatment == "Unseeded")$LogRainfall

mean_seeded_log <- mean(seeded_log, na.rm = TRUE)
sd_seeded_log <- sd(seeded_log, na.rm = TRUE)

mean_unseeded_log <- mean(unseeded_log, na.rm = TRUE)
sd_unseeded_log <- sd(unseeded_log, na.rm = TRUE)

results <- data.frame(
  Group = c("Seeded [Log]", "Unseeded [Log]"),
  Mean = c(round(mean_seeded_log, digits = 2), round(mean_unseeded_log, digits = 2)),
  SD = c(round(sd_seeded_log, digits = 2), round(sd_unseeded_log, digits = 2))
)

kable(results, col.names = c("Group", "Mean", "Standard Deviation"), caption = "Summary Statistics for Log-Transf
ormed Rainfall Data")
```

Summary Statistics for Log-Transformed Rainfall Data

Group	Mean	Standard Deviation
Seeded [Log]	5.13	1.60
Unseeded [Log]	3.99	1.64

Conducting a `t-test`:

```r
n_seeded <- length(seeded_log)
n_unseeded <- length(unseeded_log)

# Return a reference so it can be used further down in the stack
std_dev <- function(n_seeded, n_unseeded) {
  return (pooled_sd <- sqrt(((n_seeded - 1) * sd_seeded_log^2 + (n_unseeded - 1) * sd_unseeded_log^2) / (n_seeded
+ n_unseeded - 2)))
}

std_err <- function(pooled_sd) {
  return(pooled_sd * sqrt(1/n_seeded + 1/n_unseeded))
}

pooled_sd <- std_dev(n_seeded, n_unseeded)
se_diff <- std_err(pooled_sd)

# calculate the t-statistic
t_stat <- (mean_seeded_log - mean_unseeded_log) / se_diff
deg_of_freedom <- n_seeded + n_unseeded - 2

# crit value for a one-tailed test at alpha = 0.05
alpha <- 0.05

critical_value <- qt(1 - alpha, deg_of_freedom)
p_value <- pt(t_stat, deg_of_freedom, lower.tail = FALSE)
```

Hypotheses:

Null Hypothesis H_0: Cloud seeding does not increase rainfall. **Alternative Hypothesis** H_1: Cloud seeding increases rainfall.

```r
# Alpha = 0.05

values <- data.frame(
  t_stat = round(t_stat, digits = 2),
  crit_value = round(critical_value, digits = 2),
  p_values = round(p_value, digits = 2)
)

kable(values, col.names = c("Test Statistic", "Critical Value", "P-Values"), caption = "Summary of the Hypothesis
Test Results")
```

Summary of the Hypothesis Test Results

Test Statistic	Critical Value	P-Values
2.54	1.68	0.01

Observtions:

- Using a **significance level** of $\alpha = 0.05$, I conducted a hypothesis test and obtained the following results. I calculated the **test statistic** to be $t_{\text{statistic}} = 2.54$. The **critical value** ($t_\alpha = 1.68$), represents the threshold for the test statistic ($t_{\text{statistic}}$) beyond which we reject the **null hypothesis** (H_0) for a one-tailed test at the chosen **significance level** ($\alpha = 0.05$).
- The **p-value** ($p\text{-value} = 0.01$) is less than the **significance level** ($\alpha = 0.05$), we can reject the **null hypothesis** (H_0) based on the p-value.
- Additionally, since the **test statistic** ($t_{\text{statistic}} = 2.54$) is greater than the **critical value** ($t_\alpha = 1.68$). Therefore, we can also reject the **null hypothesis** (H_0).

Conclusion:

- In conclusion, because the **test statistic** is greater than the critical value and the **p-value** is less than the **significance level**, there is sufficient evidence to reject the **null hypothesis** (H_0). Therefore, our alternative hypothesis (H_1) holds. We can conclude that there is evidence that cloud seeding does increase rainfall.

Checking the result using R's functionality

We can use R's built-in function t_test to compare our observations/findings.

```r
t_result <- t.test(seeded_log, unseeded_log, alternative = "greater")

print(t_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  seeded_log and unseeded_log
## t = 2.5444, df = 49.966, p-value = 0.007042
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3903948       Inf
## sample estimates:
## mean of x mean of y
##  5.134187  3.990406
```

Observations:

- The results from the Welch's Two Sample t-test match our findings obtained "by hand"
- Since the **p-value** ($p\text{-value} = 0.007042$) is less than the chosen **significance level** ($\alpha = 0.05$), we reject the **null hypothesis** (H_0).
- The **test statistic** ($t_{\text{statistic}} = 2.54$) indicates that the difference in mean log-transformed rainfall between the seeded and unseeded groups is statistically significant.
- With a 95% confidence level, we can say that the mean log-transformed rainfall in the seeded group is significantly greater than the mean log-transformed rainfall in the unseeded group.
- Again, in conclusion, there is evidence to suggest that cloud seeding does increase rainfall, based on the results of the Welch's two-sample t-test.