Statistical Analysis of Storms Data

By Md. Zubayer

Table of Contents

Problem Statement

Estimate confidence intervals for the difference between means, proportions, and the ratio of variances in the Storms dataset to explore changes in mean latitude, proportions of storm classifications, and variability in pressure data over different years.

Data Description

The Storms dataset, a subset of the NOAA (National Oceanic and Atmospheric Administration) Atlantic hurricane database best track data, encompasses information about tropical storms measured at different time points over the years. The dataset contains 13 variables, including:

- name: The name of the tropical storm.
- year: The year in which the storm occurred.
- month: The month in which the storm occurred.
- day: The day on which the storm occurred.
- hour: The hour at which the storm was recorded.
- lat: Latitude coordinates of the storm.
- long: Longitude coordinates of the storm.
- status: The status of the storm (e.g., tropical depression, tropical storm, hurricane).
- category: The category of the storm.
- wind: Wind speed associated with the storm.
- pressure: Atmospheric pressure associated with the storm.
- tropicalstorm_force_diameter: Diameter of tropical storm force winds.
- hurricane_force_diameter: Diameter of hurricane force winds.

The dataset consists of a total of 19,066 records, each providing detailed information about a specific observation in the Atlantic hurricane best track data.

Objectives

To explore and quantify variations in the Storms dataset by estimating confidence intervals for key parameters.

1. To Estimate the Confidence Interval (CI) of the Difference Between Two Population Means:
    - Investigate the difference in the mean latitude of storms between the years 1975 and 1991.
    - Utilize statistical methods to estimate the confidence interval for this difference.
    - Interpret the results to determine if there is a significant change in mean latitude between the two years.
2. To Estimate the Confidence Interval (CI) of the Difference Between Two Population Proportions:
    - Examine the difference in the proportion of storms classified as tropical depression in 2020 and tropical storm in 2021.
    - Apply statistical techniques to estimate the confidence interval for this difference.
    - Interpret the results to identify if there is a significant variation in the proportion tropical depression in 2020 and tropical storm in 2021.
3. To Estimate the Confidence Interval (CI) of the Ratio of Two Population Variances:
    - Explore the variation in the pressure data of storms between the years 1999 and 2000.
    - Employ statistical tools to estimate the confidence interval for the ratio of variances.
    - Interpret the results to understand if there is a significant difference in the variability of pressure between the two years.

Task 1

Estimating the confidence interval (CI) for the difference between two population means of the Latitude of the storm between the years 1975 and 1991.

For task 1 solution, I have written the below code in **R-studio**,

```
install.packages("dplyr")
library(dplyr)
#installing BSDA package for the Z test
install.packages("BSDA")
library(BSDA)

# Load the Storms dataset
data(storms)

# Subset the data for the years 1975 and 1991
storms_1975 <- filter(storms, year == 1975)
storms_1991 <- filter(storms, year == 1991)

# Calculating the population variances for both years
var_1975 <- var(storms_1975$lat)
var_1991 <- var(storms_1991$lat)

#Defining the variables for Z test
s_1975 <-storms_1975$lat
s_1991 <-storms_1991$lat

#Conducticting the Z test
z_test_result <- z.test(s_1975,s_1991,sigma.x = sqrt(var_1975),sigma.y =
sqrt(var_1991),conf.level = 0.95)
print(z_test_result)
```

Output of the Code (Task-1):

Two-sample z-Test	
Data	s_1975 and s_1991
z	-3.6728
p-value	0.0002399
Alternative Hypothesis	True difference in means is not equal to 0
95% Confidence Interval	-4.967200 to -1.510421
Sample estimates	
Mean of x	28.31975
Mean of y	31.55856

Output Breakdown (Task-1)

- Two-sample z-Test: This line indicates that a two-sample z-test is being conducted.
- data: s_1975 and s_1991: Specifies the two samples being compared that are latitude data for the years 1975 and 1991.
- z = -3.6728, p-value = 0.0002399: The calculated z-statistic is -3.6728. The p-value associated with this z-statistic is 0.0002399. The z-statistic measures how many standard deviations the sample means are from each other.
- Alternative hypothesis: True difference in means is not equal to 0 that indicates that there is a true difference in means.
- 95% Confidence Interval of -4.967200 to -1.510421: The confidence interval provides a range of plausible values for the true difference between the means. In this case, the 95% confidence interval for the difference in means is from -4.967200 to -1.510421.
- Sample Estimates:

  mean of x: The mean latitude for the year 1975 is estimated to be 28.31975.

  mean of y: The mean latitude for the year 1991 is estimated to be 31.55856.

Interpretation (Task 1)

1. The negative z-statistic (-3.6728) suggests that the mean latitude in 1975 is significantly lower than the mean latitude in 1991.
2. The very small p-value (0.0002399) indicates strong evidence against the null hypothesis of no difference in means.
3. The 95% confidence interval (-4.967200 to -1.510421) further supports the conclusion, as it does not include zero. This means that we can be 95% confident that the true difference in mean latitude is not zero.
4. The sample estimates indicate the mean latitudes for the years 1975 and 1991, respectively.

Comments (Task 1)

The 95% confidence interval for the difference between the two population means of latitude for storms between the years 1975 and 1991 is estimated to be between **-4.967200 and -1.510421**. This means –

We are 95% confident that the true difference in mean latitude is within -4.967200 and -1.510421, and it does not include zero, indicating a statistically significant difference in mean latitudes between the two years.

Task 2

Estimating the confidence interval (CI) for the difference between two population proportions of storms classified as tropical depression in 2020 and tropical storm in 2021.

For task 1 solution, I have written the below code in **R-studio**,

```
install.packages("dplyr")
library(dplyr)

#loading the dataset
data(storms)
# Subset the data for the years 2020 and 2021
s1 <-filter(storms, year ==2020)
```

```
s2 <-filter(storms, year ==2021)

# Sample Size for both cases
n1 <- nrow(s1)
n2 <- nrow(s2)

# Number of cases possess the attribute of interest
x1 <- sum(s1$status == "tropical depression")
x2 <-sum(s2$status == "tropical storm")

#For Prop_test variables(vector) Declaration
n <- c(n1, n2)
x <- c(x1, x2)
#Prop test
prop_test <- prop.test(x,n,conf.level = 0.95,correct = TRUE)

print(prop_test)
```

Output of the Code (Task 2)

2-sample test for equality of proportions with continuity correction	
Data	x out of n
X-squared	68.737
df	1
p-value	< 2.2e-16
Alternative Hypothesis	Two.Sided
95% Confidence Interval	-0.2361096 to -0.1423392
Sample estimates	
prop 1	0.1587486
prop 2	0.3479730

Output Breakdown (Task – 2)

- 2-sample test for equality of proportions with continuity correction: This line indicates that a two-sample test for equality of proportions is being conducted with continuity correction.
- Data is x out of n: Specifies the data used for the test, where 'x' represents the number of cases with the attribute of interest, and 'n' is the total sample size.
- X-squared = 68.737, df = 1, p-value < 2.2e-16: The chi-squared statistic (X-squared) is 68.737 with 1 degree of freedom (df). The extremely small p-value (< 2.2e-16) indicates strong evidence against the null hypothesis of equal proportions.
- Alternative Hypothesis: The alternative hypothesis suggests a two-sided test, indicating that the proportions are not equal.
- 95 percent confidence interval of -0.2361096 to -0.1423392: The confidence interval provides a range of plausible values for the true difference between proportions. In this case, the 95% confidence interval for the difference in proportions is from -0.2361096 to -0.1423392.
- Sample estimates:

  prop 1: The estimated proportion of storms classified as tropical depression in 2020 is 0.1587486; in other words, 15.87% of the storms in 2020 were classified as tropical depression.

  prop 2: The estimated proportion of storms classified as tropical storms in 2021 is 0.3479730; in other words, 34.797% of the storms in 2021 were classified as tropical storms.

Interpretation (Task – 2)

1. The statistical analysis indicates a significant difference in the proportions of storms classified as tropical depression in 2020 and tropical storm in 2021.
2. The extremely small p-value suggests strong evidence against the null hypothesis of equal proportions.
3. The 95% confidence interval, which does not include zero, further supports the conclusion that the proportion of tropical depression in 2020 is significantly lower than the proportion

in 2021. Therefore, there is a notable shift in the classification of storms between the two years.

Comments (Task – 2)

The 95% confidence interval for the difference between two population proportions of storms classified as tropical depression in 2020 and tropical storm in 2021 is estimated to be between -0.2361096 and -0.1423392. This indicates that we can be 95% confident that the true difference in proportions lies within -0.2361096 and -0.1423392, and it does not include zero. Therefore, there is a statistically significant difference in the proportions of storms classified as tropical depression between the two years. The negative interval suggests that the proportion in 2020 is significantly lower than the proportion in 2021.

Task 3

Estimating the confidence interval (CI) for the Ratio of Two Population Variances of the air pressure at the storm's center between the years 1999 and 2000.

For task 1 solution, I have written the below code in **R-studio**,

```
install.packages("dplyr")
library(dplyr)

#installing ConfIntVariance package for var.test
install.packages("ConfIntVariance")
library(ConfIntVariance)

#loading the dataset
data(storms)
# Subsetting the data for the years 2020 and 2021
s1999 <-filter(storms, year ==1999)
s2000 <-filter(storms, year ==2000)
```

```
#Conducting var.test
variance_ratio<-var.test(s1999$pressure,s2000$pressure)
print(variance_ratio)
```

Output of the Code (Task 3)

F test to compare two variances	
Data	s1999$pressure and s2000$pressure
F	1.9471
Numerator df	456
denominator df	448
p-value	2.413e-12
Alternative Hypothesis	True ratio of variances is not equal to 1
95% Confidence Interval	1.618741 to 2.341717
Sample estimates	
Ratio of Variances	1.947103

Output Breakdown (Task – 3)

- F test to compare two variances: Indicates that an F test is being conducted to compare the variances of two samples.
- Data is s1999$pressure and s2000$pressure: Specifies the data used for the test, which is the air pressure at the storm's center for the years 1999 and 2000.
- F = 1.9471, num df = 456, denom df = 448, p-value = 2.413e-12: The calculated F statistic is 1.9471. The numerator degrees of freedom are 456. The denominator degrees of freedom are 448. The extremely small p-value (2.413e-12) indicates strong evidence against the null hypothesis of equal variances.
- Alternative Hypothesis: The alternative hypothesis is that the true ratio of variances is not equal to 1.
- 95 percent confidence interval of 1.618741 to 2.341717: The confidence interval provides a range of plausible values for the true ratio of variances. In this case, the 95% confidence interval for the ratio of variances is from 1.618741 to 2.341717.

Sample Estimates: The estimated ratio of variances is 1.947103.

Interpretation (Task – 2)

1. The F test is used to compare the variances of air pressure at the storm's center between the years 1999 and 2000.
2. The p-value (2.413e-12) is extremely small, providing strong evidence against the null hypothesis of equal variances.
3. The 95% confidence interval (1.618741 to 2.341717) further supports the conclusion that the true ratio of variances is not equal to 1.
4. The estimated ratio of variances is 1.947103, indicating a substantial difference in variances between the two years.

Comments (Task – 3)

The 95% confidence interval for the ratio of two population variances of air pressure at the storm's center between the years 1999 and 2000 is estimated to be between 1.618741 and 2.341717. This means that we are 95% confident that the true ratio of variances lies within 1.618741 and 2.341717, and it does not include 1. Therefore, there is a statistically significant difference in the variances of air pressure between the two years.

Conclusion

This comprehensive analysis aimed to explore and identify meaningful insights from the Atlantic Storms dataset, focusing on three distinct tasks. Each task involved different statistical methodologies, providing valuable information about the characteristics of storms during specific years.

The statistical analysis carried out on the Atlantic Storms dataset showed patterns across different time periods. By closely examining latitude, storm classification proportions, and air pressure variances, I gained valuable insights into how Atlantic storms have changed over the specified years. The robust methodologies, such as Z-tests, chi-squared tests, Prop-test and F-tests, ensured the dependability of our findings.

In conclusion, the statistical analyses provide valuable insights into the trends and variations of storm attributes in the Atlantic. These discoveries contribute to our broader understanding of meteorological phenomena and hold potential implications for climate research and disaster preparedness.