**Statistical Analysis of Traffic Incidents**

**and Weather Conditions**

A logo of a university of calgary

Description automatically generated

**Data 602 Group Project**

Anitha Joseph, Jincy Thomas, Megha Radhakrishnan Sanitha, Prichelle Lal

February 12th, 2025

**Introduction**

Transport Canada has undertaken to release a subset of its National Collision Database (NCDB) – a database containing all police-reported motor vehicle collisions on public roads in Canada. Selected variables (data elements) relating to fatal and injury collisions for the collisions from 2001 to 2021 data is collected from NCDB Online.

Highlights

1. Do weather variables, such as wind, rain, snow, speed limit, significantly influence road traffic incidents?
2. what are the specific effects of each weather parameter on traffic conditions? Additionally, to what extent are these impacts likely to be interconnected?
3. Gaining insight into the connection between weather and traffic could lead to more effective management of traffic and transit systems.

**Literature review**

<we have to some sentences here>

**Statistical Analysis**

**Statistical Analysis**

### **Question: Monte Carlo Simulation for Predicting the Probability of High Fatalities in Weather Conditions**

#### **Problem Statement:**

Given a dataset of weather-related incidents (including **injuries** and **fatalities**) for the years 2001–2020 across different weather conditions (such as **raining**, **snowing**, **freezing rain**, etc.), perform a Monte Carlo simulation to estimate the probability of exceeding a specific number of fatalities each year for each weather condition.

Specifically, you are tasked with simulating the total number of fatalities for each weather condition over multiple years, using Monte Carlo methods, and estimating the probability that the number of fatalities exceeds a threshold (e.g., 20 fatalities) under each weather condition.

#### **Steps:**

1. **Data Preparation:**
   1. Use the dataset containing **weather conditions**, **number of injuries**, **fatalities**, and other related information.
   2. For each weather condition (e.g., **snowing**, **raining**, etc.), determine the average rate of fatalities per year.
2. **Monte Carlo Simulation:**
   1. Simulate the number of fatalities for each weather condition over a given period (e.g., 1,000 simulations for each weather condition).
   2. Each simulation should randomly generate a number of fatalities based on the historical rate for that weather condition.
   3. For each simulated outcome, check if the number of fatalities exceeds the threshold (e.g., 20).
3. **Task in R:**
   1. Write an R script that:
      1. Simulates the number of fatalities for each weather condition using a Poisson distribution (since fatalities can be considered rare and independent events).
      2. Counts how many times the simulated fatalities exceed a threshold (e.g., 20 fatalities).
      3. Calculates the probability that the number of fatalities exceeds the threshold for each weather condition.
4. **Output:**
   1. R script should output the **probabilities** of exceeding the threshold (e.g., 20 fatalities) for each weather condition based on the Monte Carlo simulation. Compare these probabilities with the theoretical probabilities derived from the Poisson distribution using the formula for Poisson probabilities.  
        
        
        
        
      The output should show the probability of exceeding 20 fatalities for each weather condition, based on Monte Carlo simulations.

#### **Further Analysis:**

1. **Comparison with Poisson Distribution**:
   1. For each weather condition, you can also calculate the theoretical probability of exceeding 20 fatalities using the **Poisson cumulative distribution function** in R (ppois()).
   2. Compare the Monte Carlo estimated probabilities with the theoretical values to evaluate the accuracy of the simulation.
2. **Threshold Variations**:
   1. Vary the threshold (e.g., 10, 50, 100 fatalities) and repeat the simulation to explore how the probability of exceeding the threshold changes for each weather condition.

### **Question: Monte Carlo Simulation and Probability Estimation for Fatalities in Weather Conditions**

#### **Problem Statement:**

You are analyzing the number of **fatalities** that occurred under various weather conditions (e.g., **snowing**, **raining**, **freezing rain**) over the years 2001–2020. The number of fatalities in each weather condition can be modeled as a **Poisson process**, with different rates for each condition. You wish to estimate the probability that **at least 2 fatalities** occur in a given year for each weather condition.

Assume the following:

* You have N weather-related incidents in a year, where each incident can result in a certain number of fatalities (modeled using a Poisson distribution).
* You want to estimate the probability that **at least 2 fatalities** occur in a year for each weather condition, given the **Poisson rate parameter (λ)** for that condition.

Using Monte Carlo simulation, you will simulate the number of fatalities for each weather condition multiple times and estimate the probability that at least 2 fatalities occur.

#### **Steps:**

1. **Data Preparation:**
   1. For each weather condition (e.g., **snowing**, **raining**, **freezing rain**), determine the **average number of fatalities** per year (λ). These values should come from the historical dataset.
2. **Monte Carlo Simulation:**
   1. Simulate the number of fatalities for each weather condition using the **Poisson distribution** for a set number of simulations (e.g., 10,000 simulations).
   2. For each simulation, check whether the number of fatalities is **at least 2**.
3. **R Script Implementation:**
   1. Write an R script that:
      1. Simulates the number of fatalities for each weather condition using the **Poisson distribution**.
      2. Counts how many times the simulated fatalities are greater than or equal to 2.
      3. Calculates the probability of having at least 2 fatalities in a year for each weather condition.
4. **Comparison with Exact Poisson Probability:**
   1. Compare the Monte Carlo approximated probabilities with the **exact Poisson probability** of having at least 2 fatalities. The exact probability for a Poisson distribution with parameter λ is given by:

*P(X≥2)=1−P(X=0)−P(X=1)P(X \geq 2) = 1 - P(X = 0) - P(X = 1)*P(X≥2)=1−P(X=0)−P(X=1)

Where *P(X=k)P(X = k)*P(X=k) for a Poisson distribution is:

*P(X=k)=e−λλkk!P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}*P(X=k)=k!e−λλk

1. **Output:**
   1. Present the **Monte Carlo estimated probabilities** and **exact Poisson probabilities** for **at least 2 fatalities** for each weather condition in a table and/or plot.

#### **Expected Results:**

The output should show a table with both the **Monte Carlo simulated probabilities** and the **exact Poisson probabilities** for each weather condition. For instance:

|  |  |  |
| --- | --- | --- |
| **Weather Condition** | **Monte Carlo Probability (≥ 2 Fatalities)** | **Exact Poisson Probability (≥ 2 Fatalities)** |
| Snowing | 0.95 | 0.93 |
| Raining | 0.55 | 0.56 |
| Freezing Rain | 0.65 | 0.63 |
| Clear and Sunny | 0.20 | 0.19 |
| Overcast | 0.35 | 0.33 |

The plot should display both the **Monte Carlo** and **exact Poisson probabilities** for easy comparison.

#### **Further Analysis:**

1. **Varying λ Values**:
   1. Run simulations with different values of λ (mean fatalities per year) and observe how the probability of **at least 2 fatalities** changes.
2. **Threshold Variations**:
   1. Vary the threshold (e.g., 1, 5, 10 fatalities) and see how the probabilities of exceeding the threshold change for each weather condition.

### **Poisson Regression for Analyzing the Impact of Weather Conditions on Fatalities**

**Problem Statement:** You want to model the number of **fatalities** in a given year based on weather conditions, using a **Poisson regression** model to account for the count nature of the data. You hypothesize that certain weather conditions (e.g., **snowing**, **raining**) may increase the expected number of fatalities.

**Null Hypothesis (H₀):** Weather conditions have no effect on the number of fatalities.

*H0:βsnowing=βraining=0H₀: \beta\_{\text{snowing}} = \beta\_{\text{raining}} = 0*H0 :βsnowing =βraining =0

**Alternative Hypothesis (H₁):** Weather conditions have a significant effect on the number of fatalities.

*H1:At least one β≠0H₁: \text{At least one } \beta \neq 0*H1 :At least one β=0

**Task:**

* Fit a **Poisson regression model** with fatalities as the dependent variable and weather conditions as independent variables.
* Test the significance of the weather conditions using **likelihood ratio tests**.
* Use a **significance level of 0.05** to determine which weather conditions significantly increase or decrease fatalities.

### **Comparing Fatalities Between Years Using a Chi-Square Test for Independence**

**Problem Statement:** You want to test if the **distribution of fatalities** is independent of the **year**. That is, you want to see if the number of fatalities in each weather condition is distributed similarly across different years from 2001 to 2020.

**Null Hypothesis (H₀):** The distribution of fatalities is independent of the year.

*H0:Fatalities are independent of the year.H₀: \text{Fatalities are independent of the year.}*H0 :Fatalities are independent of the year.

**Alternative Hypothesis (H₁):** The distribution of fatalities depends on the year.

*H1:Fatalities are dependent on the year.H₁: \text{Fatalities are dependent on the year.}*H1 :Fatalities are dependent on the year.

**Task:**

* Perform a **Chi-Square Test for Independence** to test whether the distribution of fatalities across different weather conditions differs by year.
* Use a **significance level of 0.05**.
* Interpret the results and assess whether there is a significant change in fatality distribution across the years.

### **Testing for Change in the Rate of Fatalities in Snowing Conditions Over Time**

**Problem Statement:** You want to test if the rate of **fatalities** during **snowing** conditions has increased or decreased significantly over the years from 2001 to 2020. The null hypothesis will assume no change, while the alternative hypothesis will assume a significant change in the rate of fatalities over time.

**Null Hypothesis (H₀):** There is no significant change in the rate of fatalities in snowing conditions over the years.

*H0:Rate of fatalities in snowing conditions is constant over time.H₀: \text{Rate of fatalities in snowing conditions is constant over time.}*H0 :Rate of fatalities in snowing conditions is constant over time.

**Alternative Hypothesis (H₁):** There is a significant change in the rate of fatalities in snowing conditions over the years.

*H1:Rate of fatalities in snowing conditions is different across the years.H₁: \text{Rate of fatalities in snowing conditions is different across the years.}*H1 :Rate of fatalities in snowing conditions is different across the years.

**Task:**

* Perform a **linear regression analysis** where the dependent variable is the number of fatalities in snowing conditions and the independent variable is the year. Test for the significance of the trend (the slope of the regression line).
* Use a **significance level of 0.05**.
* Interpret whether there is evidence for an increase or decrease in fatalities in snowing conditions over time.

Conclusion

Recommendations & Next Steps

References