# **AIM**

To investigate how insurance firms use descriptive statistics and probability distributions to quantify and manage risks associated with different policy types.

## 

## **PREREQUISITES**

**Risk assessment by insurance firms**, focusing only on probability distributions and descriptive statistics are done using the following topics from the syllabus.

* **Descriptive Statistics:**

**-Mean, Median, and Mode:** To average claim amounts and typical claim sizes.

**-Standard Deviation and Variance:** To measure how much claim amounts vary, which is essential for assessing risk.

* **Probability Distributions**

**-Binomial Distribution:** Used for modelling risks with two outcomes like making a claim or not.

**-Poisson Distribution:** Models the frequency of claims for example number of accidents per year.

**-Normal Distribution:** This is assumption, it assumes that the claim or

other financial data follow a symmetric distribution around a mean,

useful for analyzing average costs.

### **OBJECTIVE of STUDY**

**Understanding Risk in few insurance policies**

* **Purpose of Risk Assessment:** Insurance companies use risk assessment to predicts potential future claims and to ensure they can cover these claims while remaining profitable.
* **Descriptive Statistics in Risk Analysis**

**. Role of Descriptive Statistics:** Descriptive Statistics helps insurers

summarize and analyze historical data on claims, which is used by the

insurance companies.

**. Key Measures** that are used are mentioned in the prerequisites.

* **Probability Distributions**

**. Importance in Risk Assessment:** Probability Distributions model the

likelihood of certain types of the occurring of claims. So due to this the insurers select a distribution that best fits the nature of risk.

**. Distributions** that are used are mentioned in the prerequisites.

#### **RESEARCH QUESTION**

**How do insurance firms utilize descriptive statistics and probability distributions to analyze data, predict potential risks, and determine fair premiums for policyholders?**

##### **METHODOLOGY**

**Identify Key Risk Factors**

- Determine the variables affecting risk (e.g., age, health conditions, frequency of claims, claim amounts).

- Group the data into categories based on these factors.

**Calculate Descriptive Statistics**

- Compute measures of central tendency (mean, median, mode) to identify typical values in the dataset, such as the average claim amount or age of policyholders.

- Calculate measures of dispersion (range, variance, standard deviation) to assess the variability in data, such as differences in claim amounts.

**Choose Appropriate Probability Distributions**

- Analyze the dataset to determine the suitable probability distribution:

- Binomial Distribution: For events with two outcomes (e.g., claim or no claim).

- Poisson Distribution: For the frequency of claims in a specific time period

- Normal Distribution: For continuous variables like claim amounts or ages of policy holders

**Estimate Probabilities**

- Calculate probabilities for specific events using the chosen distribution models

- Likelihood of high-value claims.

- Frequency of claims in a given period.



****

**Determine Expected Value**

- Use the formula

****

* E(X): Expected value
* P(xi): Probability of outcome xi
* Xi: Value of the outcome
* N: No of possible outcomes.

****

###### **EVIDENCE & ANALYSIS OF DATA**

**Dataset used in the project:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Policyholder ID** | **Age** | **Health Score (1-10)** | **Claim Frequency (per year)** | **Average Claim Amount (₹)** |
| **1** | **25** | **8** | **3** | **5000** |
| **2** | **30** | **7** | **1** | **4500** |
| **3** | **45** | **6** | **2** | **7000** |
| **4** | **50** | **5** | **4** | **10000** |
| **5** | **35** | **7** | **2** | **6000** |
| **6** | **60** | **4** | **5** | **12000** |
| **7** | **40** | **6** | **3** | **8000** |
| **8** | **55** | **5** | **4** | **9500** |
| **9** | **28** | **9** | **1** | **4000** |
| **10** | **32** | **8** | **1** | **4200** |
| **11** | **38** | **7** | **2** | **6100** |
| **12** | **46** | **6** | **3** | **7500** |
| **13** | **53** | **5** | **4** | **9800** |
| **14** | **29** | **9** | **1** | **3900** |
| **15** | **48** | **5** | **3** | **7400** |
| **16** | **42** | **6** | **2** | **8100** |
| **17** | **31** | **8** | **1** | **4300** |
| **18** | **34** | **7** | **2** | **5600** |
| **19** | **36** | **7** | **2** | **6200** |
| **20** | **39** | **8** | **1** | **5900** |
| **21** | **27** | **9** | **1** | **4000** |
| **22** | **33** | **8** | **2** | **6000** |
| **23** | **41** | **6** | **3** | **7400** |
| **24** | **37** | **7** | **2** | **8000** |
| **25** | **49** | **5** | **4** | **10500** |

****

**Calculations using DESCRIPTIVE STATISTICS**

* **Central Tendency Calculations**

**Age**

**Mean of ages of the above policy holders**

Mean= Sum Of All Ages/No of Datapoints

=983/25 = **39.32**

**Median of ages of the above policy holders**

Median

Sort the values:  
 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 35, 36, 37, 38, 39, 40, 41, 42, 45, 46, 48, 49, 50, 53,

55

Median= 13th value i.e, **38**

**Mode of ages of the policy holders**

Mode= Highest frequency value

In the dataset the most recurring value is **25**

**Health Score (1-10)**

**Mean of health scores of the above policy holders**

Mean= Sum Of All Health Scores/No of Datapoints

=168/25 = **6.72**

**Median of health scores of the above policy holders**

Median

Sort the values:

4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9

Median= 13th value i.e, **7**

**Mode of health scores of the above policy holders**

Mode = Most Frequent Health Score

The most recurring value is **7(5times)**

**Claim Frequency(per year)**

**Mean of claim frequencies of the above policy holders**

Mean = Sum of all claim frequencies/Number of Data Points

= 59/25 = **2.36**

**Median of claim frequencies of the above policy holders**

Sort the values:

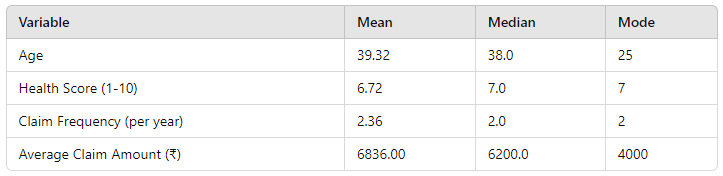
1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5

Median = 13th value i.e, **2**

**Mode of claim frequencies of the above policy holders**

Mode = Most frequent value

= **2(6times)**

**The Central Tendency Calculations for the above dataset**

- **Dispersion Calculations**

**Range**

**Range of age:**

Max(age)-Min(age)

= 55-25 = **30**

**Range of Health Score**

Max(Health score)-Min(Health score)

=9-5 = **4**

**Range of Claim Frequency**

Max(Claim Frequency)-Min(Claim Frequency)

= 5-1 = 4

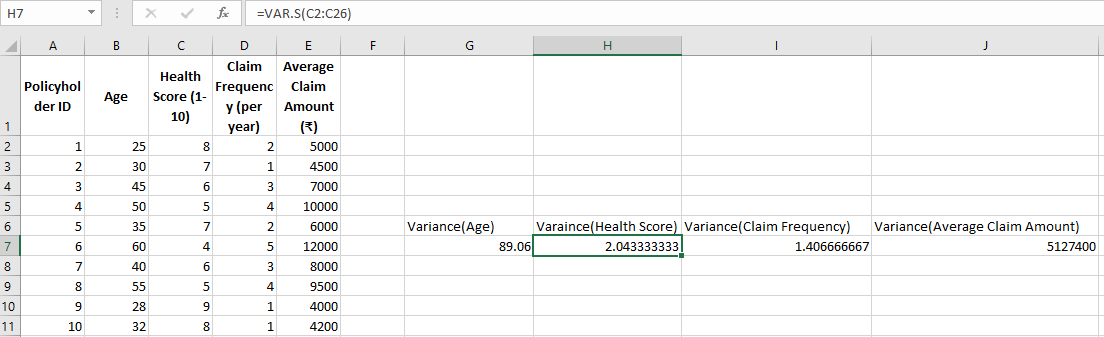
**Variance**

The calculation was done using excel the result along with screen shot is being

placed here.

**The formula used in excel sheet is**

**VAR.S(list of variables)**

****

 **Risk Assessment using Probability Distributions**

**In the data set chosen:**

Claim Frequency follows **Poisson Distribution**

Average Claim Amount follows **Normal Distribution**

**Why Claim Frequency Follows Poisson Distribution?**

* + **Discrete Nature**: Claim frequency is count data (e.g., 0, 1, 2), which Poisson is designed to model.
  + **Rare Events**: Insurance claims are infrequent for individual policyholders, aligning with Poisson's assumption.
  + **Constant Rate**: The average claim frequency (λ=1.76\lambda = 1.76λ=1.76) remains stable across the dataset.
  + **Independence**: The number of claims by one policyholder does not affect others.

**Why Average Claim Amount Follows Normal Distribution?**

* **Continuous Data**: Average claim amounts are continuous values (e.g., ₹5000, ₹7000).
* **Bell Curve**: Claim amounts often form a bell-shaped distribution with most values near the mean.
* **Central Limit Theorem (CLT)**: Averages of claim amounts approximate a Normal distribution over larger samples.
* **Symmetry**: The spread of values around the mean is fairly balanced.

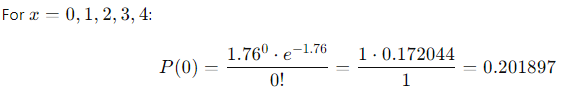
**Calculations:**

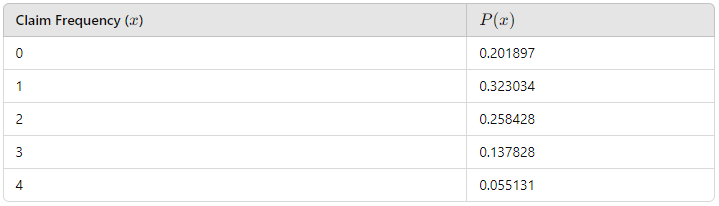
** Claim Frequency:**

**Poisson Distribution P(X)=**

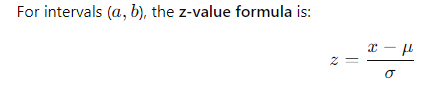
****

****

****Like this for all the other frequencies

****

For the average claim amount we group them in intervals for the ease of calculation

**Inferences from Probabilities and Descriptive Statistics**:

**From Descriptive Statistics:**

1. **Central Tendency (Mean, Median, Mode):**
   * The mean and median of claim frequencies and amounts indicate the average behavior of policyholders.
   * If the mean and median are close, the data is symmetrically distributed. If not, it may be skewed.
2. **Dispersion (Variance, Standard Deviation, Range):**
   * A high variance or standard deviation in claim amounts suggests a wider spread, indicating variability in customer claim behavior.
   * A small range or standard deviation in claim frequency shows consistency in the number of claims across policyholders.
3. **Skewness and Outliers:**
   * If there’s a significant difference between the mean and median, it indicates skewness.
   * Extreme values in claim amounts or frequencies might represent outliers, like unusually high claims.

**From Probability Distributions:**

1. **Poisson Distribution (Claim Frequency):**
   * The probabilities calculated show that most policyholders make 1-2 claims, with fewer making more or none. This aligns with the rare-event nature of claims.
   * A declining probability as 0.0555131 (claim frequency) increases is typical in Poisson, showing that multiple claims are less frequent.
2. **Normal Distribution (Claim Amount):**
   * Most claim amounts cluster near the mean (~₹7300), with lower probabilities for very high or very low claims.
   * The distribution suggests that claim amounts follow a predictable pattern, useful for setting insurance premiums.