TABLE OF CONTENTS

AIM 2
PREREQUISITES 4
OBJECTIVE of STUDY 5
RESEARCH QUESTION 5
METHODOLOGY 6
EVIDENCE & ANALYSIS OF DATA 8
LIMITATIONS OF MY PROJECT 18
FUTURE ENHANCEMENTS 19
CONCLUSION 20
BIBLIOGRAPHY 21
PLAGIARISM CHECK 22

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

Determine $$E(X) = \sum_{i=1}^{n} P(x_i) \cdot x_i$$ **Expected Value**

- Use the formula

$$E(X) = \sum_{i=1}^{n} P(x_i) \cdot x_i$$

- E(X): Expected value
- $P(x_i)$: Probability of outcome x_i
- X_i: 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**

<u>Age</u>

Mean of ages of the above policy holders

Mean= Sum Of All Ages/No of Datapoints

=983/25 = <u>**39.32**</u>

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= 13^{th} value i.e, <u>**38**</u>

Mode of ages of the policy holders

Mode= Highest frequency value

In the dataset the most recurring value is <u>**25**</u>

<u>**Health Score (1-10)**</u>

Mean of health scores of the above policy holders

Mean= Sum Of All Health Scores/No of Datapoints

=168/25 = <u>**6.72**</u>

Median of health scores of the above policy holders

Median

Sort the values:

Variable	Mean	Median	Mode
Age	39.32	38.0	25
Health Score (1-10)	6.72	7.0	7
Claim Frequency (per year)	2.36	2.0	2
Average Claim Amount (₹)	6836.00	6200.0	4000

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 = 13^{th} value i.e, **2**

Mode of claim frequencies of the above policy holders

Mode = Most frequent value

= **2(6times)**

H7 =VAR.S(C2:C26)

	A	B	C	D	E	F	G	H	I	J
1	Policyholder ID	Age	Health Score (1-10)	Claim Frequency (per year)	Average Claim Amount (₹)					
2	1	25	8	2	5000					
3	2	30	7	1	4500					
4	3	45	6	3	7000					
5	4	50	5	4	10000					
6	5	35	7	2	6000		Variance(Age)	Varaince(Health Score)	Variance(Claim Frequency)	Variance(Average Claim Amount)
7	6	60	4	5	12000		89.06	2.043333333	1.406666667	5127400
8	7	40	6	3	8000					
9	8	55	5	4	9500					
10	9	28	9	1	4000					
11	10	32	8	1	4200					

- D

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)

$$\frac{\lambda^x e^{-\lambda}}{x!}$$

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)=

For $x = 0, 1, 2, 3, 4$:

$$P(0) = \frac{1.76^0 \cdot e^{-1.76}}{0!} = \frac{1 \cdot 0.172044}{1} = 0.201897$$

Claim Frequency (x)	$P(x)$
0	0.201897
1	0.323034
2	0.258428
3	0.137828
4	0.055131

$$\lambda = \frac{\text{Sum of Claim Frequency}}{\text{Number of Rows}} = \frac{44}{25} = 1.76$$

$$\mu = 7308, \quad \sigma = 1557.89$$

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

Policy Holder Id	Average Claim Amount (x)	Z- value(z)	Probability (P(Z<z))
1	**5000**	-0.612088077	0.27
2	**4500**	-0.832899072	0.20
3	**7000**	0.271155901	0.61
4	**10000**	1.596021869	0.94
5	**6000**	-0.170466088	0.43
6	**12000**	2.479265848	0.99
7	**8000**	0.712777891	0.76
8	**9500**	1.375210874	0.92
9	**4000**	-1.053710066	0.15
10	**4200**	-0.965385668	0.17
11	**6100**	-0.126303889	0.45
12	**7500**	0.491966896	0.69
13	**9800**	1.507697471	0.93
14	**3900**	-1.097872265	0.14
15	**7400**	0.447804697	0.67
16	**8100**	0.75694009	0.78
17	**4300**	-0.92122347	0.18
18	**5600**	-0.347114884	0.36
19	**6200**	-0.08214169	0.47
20	**5900**	-0.214628287	0.42
21	**4000**	-1.053710066	0.15
22	**6000**	-0.170466088	0.43
23	**7400**	0.447804697	0.67
24	**8000**	0.712777891	0.76
25	**10500**	1.816832864	0.97

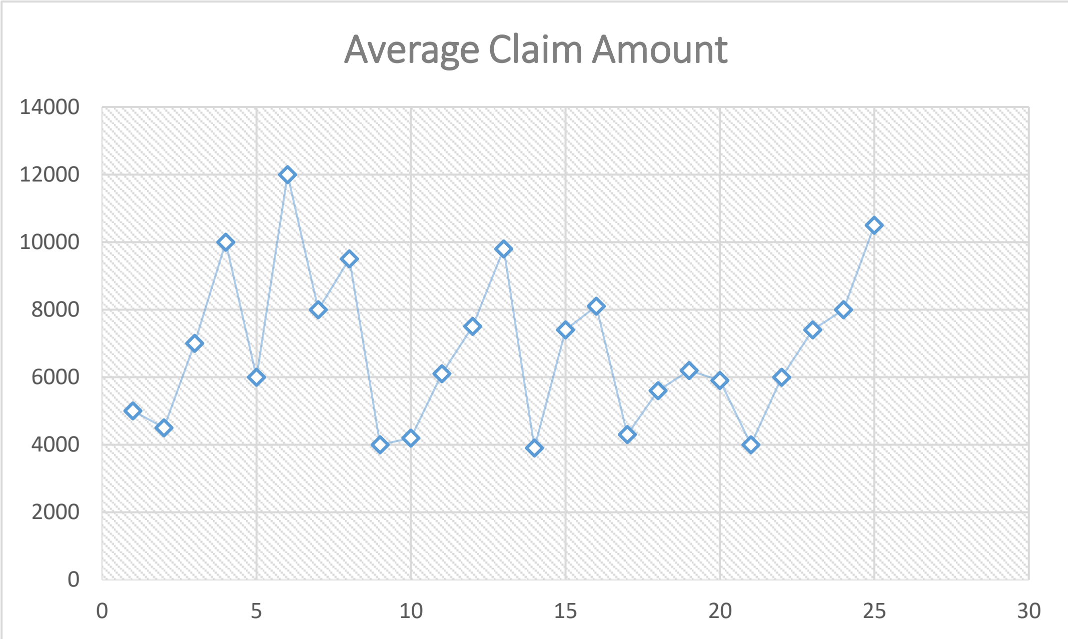
Probability Of Average Claim Amount Using Normal Distribution

The calculation was done using MS Excel using the formula

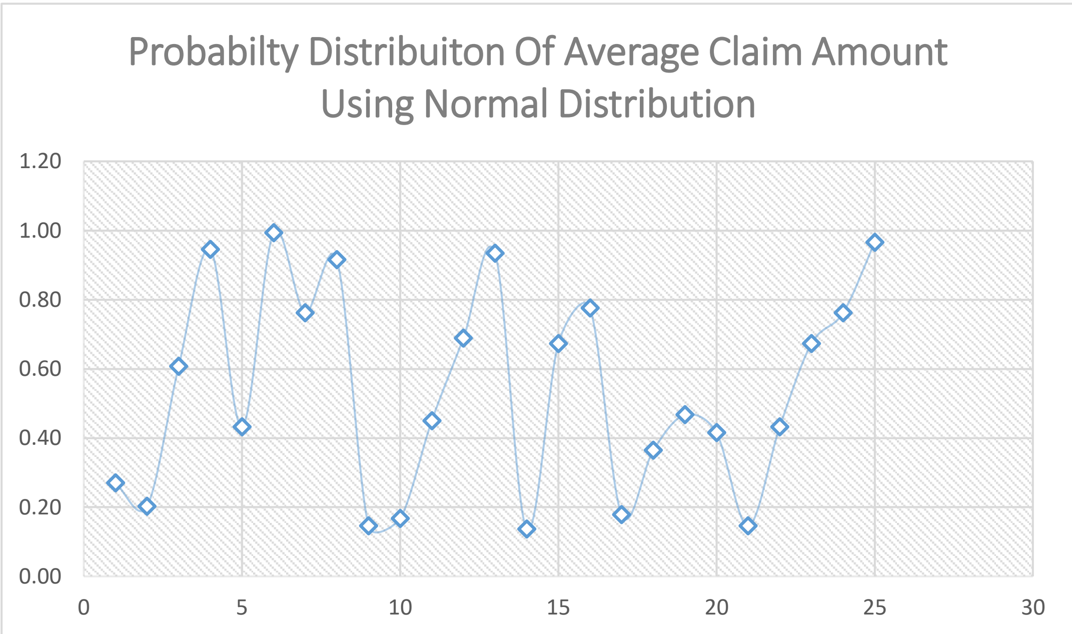
=NORM.S.DIST(z value,TRUE)

C3 =NORM.S.DIST(B3,TRUE)

	A	B	C	D	E	F	G	H	I
1	x Average Claim amount	z value	normal distribution						
2	5000	-0.612088077	0.27						
3	4500	-0.832899072	0.20						
4	7000	0.271155901	0.61						
5	10000	1.596021869	0.94					Mean	Standard Deviation
6	6000	-0.170466088	0.43					6386	2264.38
7	12000	2.479265848	0.99						
8	8000	0.712777891	0.76						
9	9500	1.375210874	0.92						
10	4000	-1.053710066	0.15						
11	4200	-0.965385668	0.17						
12	6100	-0.126303889	0.45						
13	7500	0.491966896	0.69						
14	9800	1.507697471	0.93						
15	3900	-1.097872265	0.14						
16	7400	0.447804697	0.67						



Average Claim Amount
14000
12000
10000
8000
6000
4000
2000
0
0
5
10
15
20
25
30



Probabilty Distribuiton Of Average Claim Amount Using Normal Distribution
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0
5
10
15
20
25
30

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.

LIMITATIONS OF MY PROJECT

1. Limited Dataset

- The analysis is based on only 25 data points, which might not represent the entire population of claims.
- A larger dataset could provide more accurate and reliable results.

2. Assumption of Normality

- The project assumes that the claim amounts follow a normal distribution, which might not always hold true in real-world scenarios.
- If the actual distribution deviates significantly from normality, the results may be biased.

3. Simplified Risk Assessment

- The project uses basic descriptive statistics and probability distributions, which may oversimplify the complexities of real-world risk assessment.
- Real-world risk assessments often involve advanced statistical techniques, predictive models, and external factors.

4. Ignoring External Factors

- Factors like economic conditions, policy changes, or customer demographics are not included in the analysis, which could significantly affect insurance claim patterns.

5. Static Analysis

- The analysis is based on a single snapshot of data and does not consider temporal changes or trends over time.
- Claims data might vary seasonally or due to market conditions, which are not reflected here.

6. Rounding Errors

- Calculations involve rounding z-values and probabilities, which may introduce minor inaccuracies.

FUTURE ENHANCEMENTS

- **Larger Dataset**: Use more extensive data to improve the accuracy and representativeness of results.

- **Advanced Techniques**: Incorporate methods like regression or machine learning to analyze risks in-depth.

- **Dynamic Analysis**: Study trends over time using time-series data to identify seasonal or long-term patterns.

- **Inclusion of External Factors**: Analyze the impact of customer demographics, economic indicators, or policy types on claims.

- **Visualizations and Dashboards**: Create interactive dashboards for better data exploration and scenario simulations.

- **Validation with Real Data**: Test your findings with real-world insurance data to assess practical applicability

CONCLUSION

Based on the analysis, the following conclusions can be drawn:

1. **Risk Assessment Patterns**:
   - Most claim amounts are concentrated around the mean value (₹6386), with probabilities aligning with a normal distribution.
   - Claims closer to the mean have higher probabilities, indicating a consistent trend in claim patterns.

2. **Insights from Descriptive Statistics**:
   - The central tendency (mean, median, and mode) highlights the typical claim amount, providing a reliable estimate for policy pricing.
   - The spread (variance and standard deviation) reflects variability in claims, helping to identify the level of risk insurers face.

3. **Probability Analysis**:
   - Probabilities derived from the normal distribution help quantify the likelihood of claims falling within specific ranges.
   - For example, higher claims (e.g., above ₹9500) have lower probabilities, indicating they are less frequent but could represent significant risks.

BIBLIOGRAPHY

- www.stattrek.com
- www.investopedia.com
- www.thoughtco.com
- chatgpt.com

PLAGIARISM CHECK