JSS Mahavidyapeetha

JSS Science and Technology University

JSS Technical Institution Campus,

Mysore-570006



**“TEMPERATURE ANALYSIS OF A MONTH”**

Problem solving approach in Engineering Mathematics

**Under the guidance of:** **Name of the candidates**:

Dr.T.D.Roopamala 1. Ankita Ashok Baleri

Head of Department of Mathematics. USN : 01JST17CS020

2. Disha Bhat

USN : 01JST17CS049

3. Nanda Krishna K.S

USN : 01JST17CS095

CS 4th sem, ‘A’ section

**Contents**

1. Abstract
2. Introduction
3. Algorithm
4. Program with Output
5. Applications
6. Conclusion
7. Reference

**Problem statement:**

* Find the mean, median, variance and standard deviation of the collected data of the temperature.
* Forecast weather of current weekday by Logistic Regression Classification (Machine Learning Model).

**Abstract:**

Aim of this project is a general definition of probability, of its main mathematical features and the features it presents under particular circumstances. The behaviour of probability is linked to the features of the phenomenon we would predict. This link can be defined probability distribution. Given the characteristics of phenomena (that we can also define variables), there are defined probability distribution. For categorical (or discrete) variables, the probability can be described by a binomial or Poisson distribution in the majority of cases. For continuous variables, the probability can be described by the most important distribution in statistics, the normal distribution. Distributions of probability are briefly described together with some examples for their possible application.

Statistical discrete processes – for example, the number of accidents per driver, the number of insects per leaf in an orchard, the number of thunderstorms per year, the number of earthquakes per year, the number of patients visit emergency room in a certain hospital per day - often occur in real life. To approximate (or fit) a process, statistical probabilistic distributions are often used.

**INTRODUCTION**

# PROBABILITY

Probability theory, a branch of mathematics concerned with the analysis of random phenomena. The outcome of a random event cannot be determined before it occurs, but it may be any one of several possible outcomes. The actual outcome is considered to be determined by chance.

Probability is the measure of the likelihood that an event will occur. See glossary of probability and statistics. Probability quantifies as a number between 0 and 1, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur. A simple example is the tossing of a fair (unbiased) coin. Since the coin is fair, the two outcomes ("heads" and "tails") are both equally probable; the probability of "heads" equals the probability of "tails"; and since no other outcomes are possible, the probability of either "heads" or "tails" is 1/2 (which could also be written as 0.5 or 50%).

These concepts have been given an axiomatic mathematical formalization in probability theory, which is used widely in such areas of study as mathematics, statistics, finance, gambling, science (in particular physics), artificial intelligence/machine learning, computer science, game theory, and philosophy to, for example, draw inferences about the expected frequency of events. Probability theory is also used to describe the underlying mechanics and regularities of complex systems.

# RANDOM VARIABLE

In probability and statistics, a random variable or stochastic variable is a variable whose possible values are outcomes of a random phenomenon. More specifically, a random variable is defined as a function that maps the outcomes of an unpredictable process to numerical quantities, typically real numbers. It is a variable (specifically a dependent variable), in the sense that it depends on the outcome of an underlying process providing the input to this function, and it is random in the sense that the underlying process is assumed to be random.

## DISCRETE RANDOM VARIABLE

A discrete random variable is one which may take on only a countable number of distinct values such as 0,1,2,3........ Discrete random variables are usually (but not necessarily) counts. If a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family, the Friday night attendance at a cinema, the number of patients in a doctor's surgery, the number of defective light bulbs in a box of ten.

The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also sometimes called the probability function or the probability mass function.

Suppose a random variable X may take k different values, with the probability that X = xi defined to be P(X = xi) = pi. The probabilities pi must satisfy the following.

1: 0 < pi < 1 for each i

2: p1 + p2 + ... + pk = 1.

## CONTINUOUS RANDOM VARIABLE

A continuous random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.

A continuous random variable is not defined at specific values. Instead, it is defined over an interval of values, and is represented by the area under a curve (in advanced mathematics, this is known as an integral). The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Suppose a random variable X may take all values over an interval of real numbers. Then the probability that X is in the set of outcomes A, P(A) is defined to be the area above A and under a curve. The curve, which represents a function p(x), must satisfy the following:

1: The curve has no negative values (p(x) > 0 for all x)

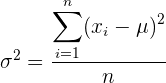
2: The total area under the curve is equal to 1.

# PROBABILITY DENSITY FUNCTION

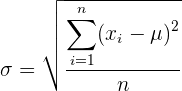
Probability distributions are generally divided into two classes. A discrete probability distribution (applicable to the scenarios where the set of possible outcomes is discrete, such as a coin toss or a roll of dice) can be encoded by a discrete list of the probabilities of the outcomes, known as a probability mass function. On the other hand, a continuous probability distribution (applicable to the scenarios where the set of possible outcomes can take on values in a continuous range (e.g. real numbers), such as the temperature on a given day) is typically described by probability density functions (with the probability of any individual outcome actually being 0). The normal distribution is a commonly encountered continuous probability distribution. More complex experiments, such as those involving stochastic processes defined in continuous time, may demand the use of more general probability measures.

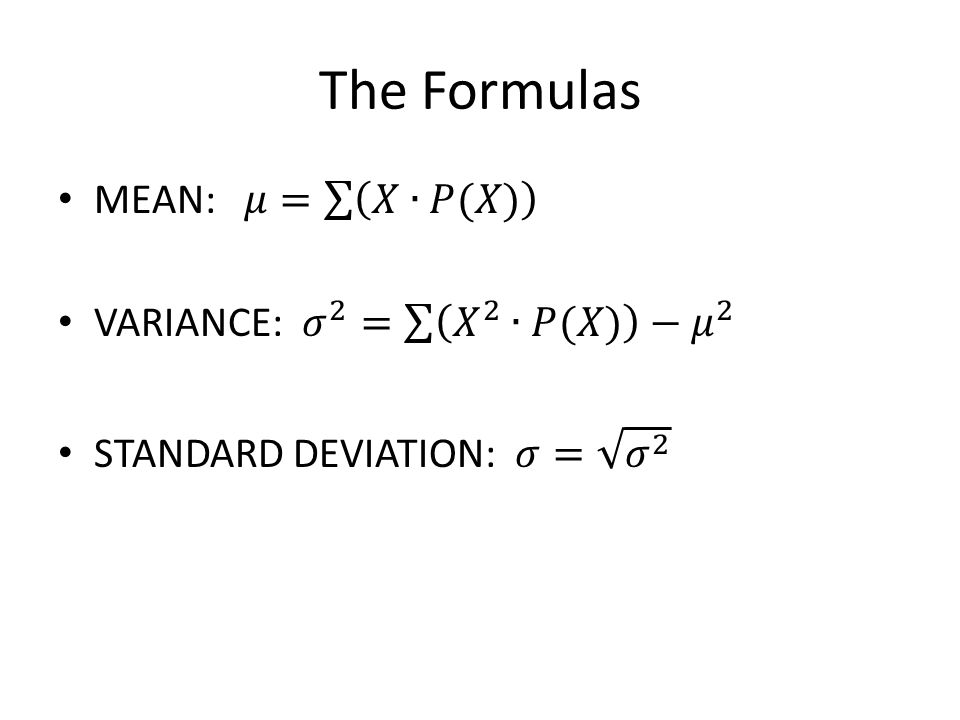
**Mean:** It is the sum of a list of numbers divided by the number of numbers in the list.

**Variance:** The average of the squared differences from the mean.



**Standard deviation:** Square root of the variance is the standard deviation.





**BRIEF INTRODUCTION TO OUR PROJECT**

Dataset is created by recording temperature of a month. On this data central measures of tendency (mean, median, mode) is calculated. In addition, variance and standard deviation of data is also analysed. Random Variable over the continuous data is obtained from user and probability function is determined. Sample space is all possible temperature of a city.

**Sample Space = {*Temp***: *Temp* being all possible temperature of a city in Celsius **}**

Machine Learning model (Logistic Regression) is trained on this dataset. This model obtains current weekday from operating system and forecast’s/predict the probability function of different weather, based on historical dataset.

**ALGORITHM**

* Mean, Median, Mode, Standard Variance:
* Basic Mathematical Functions
* Machine Learning Model:
* Logistic Regression

**PROGRAMMING LANGUAGE**

**Python**

**DEVELOPMENT EDITOR**

**Anaconda**

# DATASET

|  |  |  |
| --- | --- | --- |
| **Weekday** | **Climate** | **Temperature** |
| Sunday | Rainy | 20 |
| Monday | Rainy | 21 |
| Tuesday | Rainy | 21 |
| Wednesday | Rainy | 23 |
| Thursday | Rainy | 24 |
| Friday | Sunny | 23 |
| Saturday | Sunny | 23 |
| Sunday | Sunny | 24 |
| Monday | Rainy | 24 |
| Tuesday | Rainy | 25 |
| Wednesday | Sunny | 24 |
| Thursday | Rainy | 25 |
| Friday | Sunny | 26 |
| Saturday | Rainy | 25 |
| Sunday | Sunny | 26 |
| Monday | Rainy | 25 |
| Tuesday | Sunny | 26 |
| Wednesday | Rainy | 25 |
| Thursday | Rainy | 26 |
| Friday | Sunny | 25 |
| Saturday | Sunny | 25 |
| Sunday | Sunny | 25 |
| Monday | Sunny | 27 |
| Tuesday | Rainy | 27 |
| Wednesday | Sunny | 28 |
| Thursday | Sunny | 27 |
| Friday | Rainy | 28 |
| Saturday | Sunny | 29 |

**PROGRAM**

**MAIN MODULE:**

**#Importing Librabries**

import pandas as pd

import numpy as np

import math

import collections

import matplotlib.pyplot as plt

from sklearn.externals import joblib

**#importing dataset**

dataset=pd.read\_csv("Data.csv")

df=pd.DataFrame(dataset)

**#calculating frequency**

frequency=collections.Counter(df.iloc[:,-1])

**#to find mean and probability**

probability=[]

i=0

mean=0

for key,value in frequency.items():

probability.insert(i,float(value/28))

mean=mean+probability[i]\*key

i=i+1

**#Median**

d=dict(zip(frequency.keys(),probability))

median\_value=int(d.\_\_len\_\_()/2)

median=list(d.keys())[median\_value]

**#Mode**

mode=list(d.keys())[list(d.values()).index(max(d.values()))]

**#Plotting Probability Distribution:**

T=np.array(list(frequency.keys()))

power=np.array(probability)

from scipy.interpolate import spline

xnew = np.linspace(T.min(),T.max(),300)

power\_smooth = spline(T,power,xnew)

plt.plot(xnew,power\_smooth)

plt.axvline(x=mean,color='red',label='Mean')

plt.title('Probability Distribution Graph')

plt.xlabel('Temperature')

plt.ylabel('Probability Distribution Function')

plt.show()

**#to find variance**

i=0

var=0

for key,value in frequency.items():

var=var+((key-mean)\*\*2)\*probability[i]

i=i+1

**#Standard deviation**

sd=math.sqrt(var)

**#Printing**

print("\n\n\nMeasure of Central Tendency:\n")

print("Mean of probability distribution is %.3f"%mean)

print("Median of probability distribution is %.3f"%median)

print("Mode of probability distribution is %.3f"%mode)

print("Standard deviation of probability distribution is %.3f"%sd)

print("Variance of probability distribution is %.3f"%var)

**#Probability Distribution Of Test Random Variable:**

print("\n\n\nRandom Variable Testing:")

rand\_temp=float(input("Enter the random temperature"))

prob\_rand\_temp=np.interp(float(rand\_temp),T,power)

print("Probability function of Random temperature is %.1f" %prob\_rand\_temp)

**#Probability Application in Machine Learning:**

print("\n\n\nApplication of probability function to analyze weather using Logisstic Regression Classifiation(ML Model) of each weekday and predict the weather of current weekday\n")

from datetime import datetime

from datetime import timedelta

week = ['Sunday','Monday','Tuesday','Wednesday','Thursday','Friday','Saturday']

sample=np.array([week[datetime.today().weekday()]])

**#LabelEncoder and OneHotEncoder**

labelencoder=joblib.load('LabelEncoderCategories')

sample=labelencoder.transform(sample)

onehotencoder=joblib.load('OneHotEncoderCategories')

sample=onehotencoder.transform(sample.reshape(1,-1)).toarray()

**#Predicting**

classifier=joblib.load("MLModelLogisticRegression")

prob=classifier.predict\_proba(sample)

print("The probability this weekday being Rainy=" + str(prob[0][0]) +"\nThe probability this weekday being Sunny=" +str(prob[0][1]))

**ML TRAINING MODULE:**

**#Importing Librabries**

import pandas as pd

import numpy as np

**#Importing DataSet**

dataset = pd.read\_csv("SunnyRainyDataset.csv")

X=dataset.iloc[:,:-2].values

y=dataset.iloc[:,-2].values

**#LabelEncoder and OneHotEncoder**

from sklearn.preprocessing import LabelEncoder,OneHotEncoder

from sklearn.externals import joblib

labelencoder=LabelEncoder()

X[:,0]=labelencoder.fit\_transform(X[:,0])

joblib.dump(labelencoder,'LabelEncoderCategories')

onehotencoder=OneHotEncoder(categorical\_features=[0])

X=onehotencoder.fit\_transform(X).toarray()

joblib.dump(onehotencoder,'OneHotEncoderCategories')

y=labelencoder.fit\_transform(y)

joblib.dump(labelencoder,'LabelEncoderCategoriesY')

#Spliting:

from sklearn.model\_selection import train\_test\_split

X\_train,X\_test,Y\_train,Y\_test=train\_test\_split(X,y,test\_size=0.10,random\_state=1348882)

**#ML Model:**

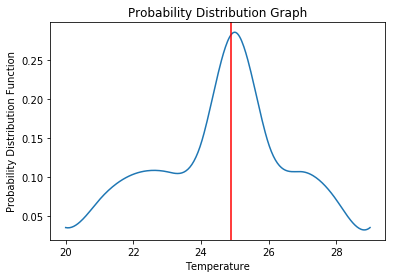
from sklearn.linear\_model import LogisticRegression

classifier = LogisticRegression(random\_state=0)

classifier.fit(X\_train,Y\_train)

joblib.dump(classifier,"MLModelLogisticRegression")

**OUTPUT**



***Measure of Central Tendency:***

Mean of probability distribution is 24.893

Median of probability distribution is 25.000

Mode of probability distribution is 25.000

Standard deviation of probability distribution is 2.093

Variance of probability distribution is 4.381

***Random Variable Testing:***

Enter the random temperature

->24.6

Probability function of Random temperature is 0.2

***Application of probability function to analyse weather using Logistic Regression Classification (ML Model) of each weekday and predict the weather of current weekday***

The probability of current weekday being Rainy=0.37

The probability of current weekday being Sunny=0.63

**APPLICATIONS**

In the statistical literature the large or small values assumed by a random variable from a finite set of measurements are termed extreme values. The variable of interest could be skin cancer, breast cancer brain tumour or any other chronic disease, etc. The largest and smallest value of the extremes are often of most interest. The extreme values are random variables. For example, number of people who are in the final stage of a certain cancer in a particular year exhibits random variations that are best described in terms of probabilities.

Biostatistics: application of statistical methods in biological sciences to deal with living things

Application of statistics in medicine

To compare efficacy of a drug

The % cured, relieved or died in experiment

To find an association B/W 2 attributes

In epidemiological studies

In addition, Probability Distribution function is extensively used in Machine Learning Algorithms.

**CONCLUSION**

A small dataset of temperature analysis of a month was created. The measure of central tendency (Mean, Median, Mode) was calculated on this dataset. Variance and standard deviation of the data were also analysed.

Machine Learning model was implemented on this small dataset. The ML model import’s the present weekday from operating system and forecasts different weather probability function on that day based on historical data.

We had a wonderful learning experience, while working on this project. This project took us through the various phases of project development and gave us the real insight into the applicability of probability function in various field of data science.