**Part – A Each question carries 2 marks: 15 x 2 = 30 Marks**

1. **List the three axioms of Probability**
2. The first axiom of probability is that the probability of any event is a non-negative real number.
   * 1. P(A) >= 0 for all A ⊂ S or
     2. 0<=P(A)<=1
3. The second axiom of probability is that the probability of the entire sample space is one.

P(S) = 1

1. Additivity: The third axiom of probability deals with mutually exclusive events. If E1 and E2 are [mutually exclusive](https://www.thoughtco.com/mutually-exclusive-3126557), meaning that they have an empty intersection and we use U to denote the union, then

P(E1 U E2 ) = P(E1) + P(E2).

1. **Define conditional probability**



1. **Define Random Variable**

A random variable is a variable whose value is **unknown,** usually written as *X*, or a function that assigns values to each of an experiment's outcomes. Random variables are often designated by letters and can be classified as discrete, which are variables that have specific values, or continuous, which are variables that can have any values within a continuous range.

***1:****0 < pi < 1 for each i*

***2:****p1 + p2 + ... + pk = 1*

1. **What is a biased sample?**

In [statistics](https://en.wikipedia.org/wiki/Statistics), **sampling bias** is a [bias](https://en.wikipedia.org/wiki/Bias) in which a sample is collected in such a way that some members of the intended [population](https://en.wikipedia.org/wiki/Statistical_population) are less likely to be included than others.  It results in a **biased sample**, a non-random sample[[1]](https://en.wikipedia.org/wiki/Sampling_bias#cite_note-1) of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected.

1. **In a bivariate data, write the equations for determining the value of the regression coefficients which will minimize the sum of the squared errors.**
2. **If X is a random variable uniformly distributed over an interval (a,b).**

**That is, f(x) = ; a≤x≤b**

**What is the expected value of x?**

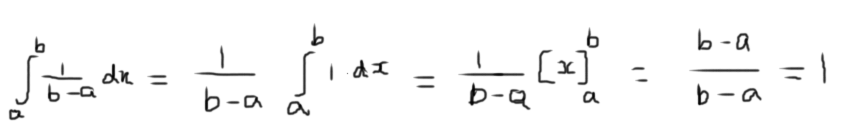
The expected value µ = E(X) is a measure of location or central tendency

**Definition**: Let X be a continuous random variable with range [a, b] and probability density function f(x). The expected value of X is defined by,

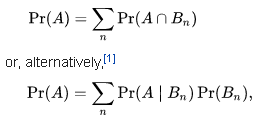
So f(x) dx represents the probability that X is in an infinitesimal range of width dx around x. Thus we can interpret the formula for E(X) as a weighted integral of the values x of X, where the weights are the probabilities f(x) dx.

The expected value is also called the **mean** or **average.**

f(x) = ; a≤x≤b

E(X) = 

1. **Write the expression for the total probability of an event.**



1. **X is a continuous random variable with density function given below.**

**f(x) = ; 1500≤x≤2500**

**= 0; elsewhere**

**Find the value of ‘a’**

1. **100 observations were made of two variables. Based on this sample, the correlation coefficient was found to be 0.275. Is this a significant value?**
2. **Write the condition to be satisfied for two random variables to be**
3. **Mutually exclusive**

P(ABC. . . ) = P(A)P(B)P(C)...

1. **Independent**

The variable which is not dependent on any other value(s)/variable(s) is called independent variable.

Let’s say, y = f(x) = 5x

Here x is independent variable

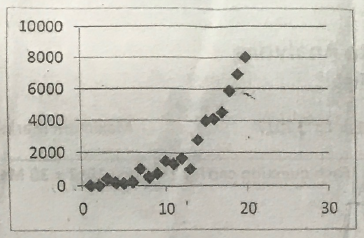
And y is dependent variable (dependent on x)

The value of y depends on the value chosen for x.

1. **A set of observed values of pulp production in metric tons and world pulp price in rupees was analyzed and the following statistics were calculated.**
   1. **Mean of pulp production**
   2. **Mean of world pulp price**
   3. **Standard deviation of pulp production**
   4. **Correlation coefficient between pulp production and world pulp price**

**What are the units of these statistics?**

1. **If a random variable X has a mean of , then what is the mean of the Random Variable Y = X - ?**
2. **The scatter plot of a set of observed values is shown below. Write your observation on the relation between the two variables.**



1. **10 observations were made of a random variable and the mean was found to be 10. A 11th observation of the variable yielding a value of 21. What is the new value of the mean including the 11th value?**
2. **Write the density functions of any two standard continuous distributions.**

**Continuous distributions:** A statistical distribution for which the variables may take on a continuous range of values.

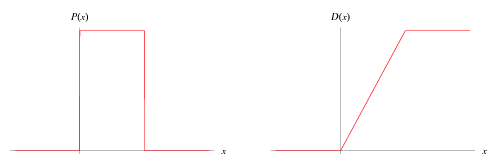
A continuous distribution describes the probabilities of the possible values of a continuous random variable. A continuous random variable is a random variable with a set of possible values (known as the range) that is infinite and uncountable.

Probabilities of continuous random variables (X) are defined as the area under the curve of its PDF. Thus, only ranges of values can have a nonzero probability. The probability that a continuous random variable equals some value is always zero.

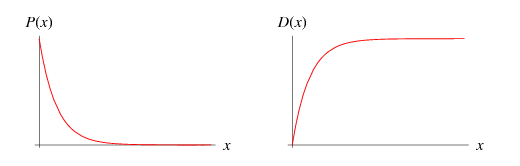
**Example 1: -** Uniform distribution

A uniform distribution, sometimes also known as a rectangular distribution, is a distribution that has constant probability.

|  |  |  |
| --- | --- | --- |
| P(x) | = | {0 for x<a; 1/(b-a) for a<=x<=b; 0 for x>b |
| D(x) | = | {0 for x<a; (x-a)/(b-a) for a<=x<=b; 1 for x>b. |



**Example 2: -** Exponential Distribution



Given a Poisson distribution with rate of change lambda, the distribution of waiting times between successive changes (with k=0) is

|  |  |  |  |
| --- | --- | --- | --- |
| D(x) | = | P(X<=x) |  |
| http://mathworld.wolfram.com/images/equations/ExponentialDistribution/Inline6.gif | = | 1-P(X>x) |  |
| http://mathworld.wolfram.com/images/equations/ExponentialDistribution/Inline9.gif | = | 1-e^(-lambdax), |  |

and the probability distribution function is

|  |
| --- |
| P(x)=D^'(x)=lambdae^(-lambdax). |

**Part B Each question carries 10 marks: 2 x 10 = 20 Marks**

1. **Given a sample of bivariate data, list the steps to be followed to build a prediction model.**

Scatter Plot

Identify Trend and Outliers

Transform data (if required)

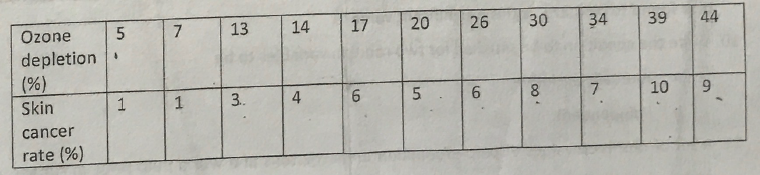
Validate the Trend

Fit a curve

Forecast

Verify the forecasted value

1. **Skin cancer rates have been steadily rising over recent years. It is thought that this may be due to ozone depletion. The following data are ozone depletion rates in various localities and the rates of skin cancer.**



* 1. **Fit a straight line regression model of the data**
  2. **What is the rate of skin cancer if ozone depletion is 40%**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Year** | **Ozone Depletion % (x)** | **Skin Cancer rate % (y)** | **x -͞x** | **(x -͞x)2** | **y-͞y** | **(y-͞y)2** | **(x -͞x)( y-͞y)** | **x²** | **xy** |
|  | 5 | 1 | -17.6364 | 311.0413 | -4.45455 | 19.84298 | 78.562 | 25 | 5 |
|  | 7 | 1 | -15.6364 | 244.4959 | -4.45455 | 19.84298 | 69.6529 | 49 | 7 |
|  | 13 | 3 | -9.63636 | 92.8595 | -2.45455 | 6.024793 | 23.6529 | 169 | 39 |
|  | 14 | 4 | -8.63636 | 74.58678 | -1.45455 | 2.115702 | 12.562 | 196 | 56 |
|  | 17 | 6 | -5.63636 | 31.7686 | 0.545455 | 0.297521 | -3.0744 | 289 | 102 |
|  | 20 | 5 | -2.63636 | 6.950413 | -0.45455 | 0.206612 | 1.19835 | 400 | 100 |
|  | 26 | 6 | 3.363636 | 11.31405 | 0.545455 | 0.297521 | 1.83471 | 676 | 156 |
|  | 30 | 8 | 7.363636 | 54.22314 | 2.545455 | 6.479339 | 18.7438 | 900 | 240 |
|  | 34 | 7 | 11.36364 | 129.1322 | 1.545455 | 2.38843 | 17.562 | 1156 | 238 |
|  | 39 | 10 | 16.36364 | 267.7686 | 4.545455 | 20.66116 | 74.3802 | 1521 | 390 |
|  | 44 | 9 | 21.36364 | 456.405 | 3.545455 | 12.57025 | 75.7438 | 1936 | 396 |
| **Sum** | 249 | 60 |  | 1680.545 |  | 90.72727 | 370.8182 |  |  |
|  | 22.63636 | 5.454545 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

**When x = 40%, then y = 9.25% (approximately)**

**Part C Optional 10 Marks**

**Prove that the value of the correlation coefficient can lie between -1 and +1 only.**