CHAPTER 5

PROBABILITY AND STATISTICS

FUNDAMENTALS OF PROBABILITY

Probability is used to measure the degree of certainty or uncertainty of the occurrence of events.

Hence, if y is the total number of outcomes and x is the favorable number of outcomes then probability of occurrence of an event A is denoted by $P(A)$ and given by

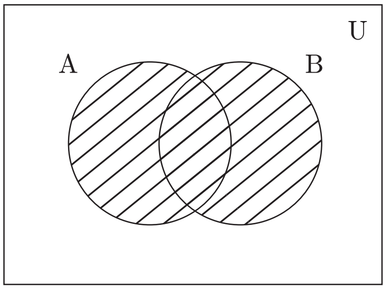
$$P(A) = \frac{x}{y}$$

A process to observe and measure the probability of an event is called an experiment. The result obtained through that experiment is the outcome of that experiment. An experiment which when repeated under identical conditions do not produce the same outcome every time is called a random experiment. The set of all the possible outcomes of a random experiment is called sample space.

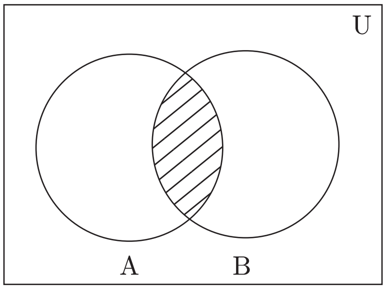
A set of the sample space associated with a random experiment is called an event.

Suppose we have two events A and B, then some of the operations performed on these events are as follows:

1. **Union of sets:** $A \cup B$

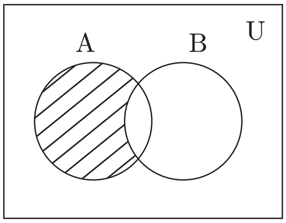


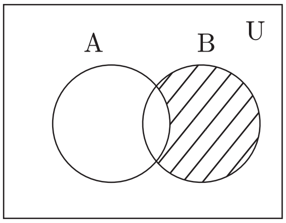
2. **Intersection of sets:** $A \cap B$



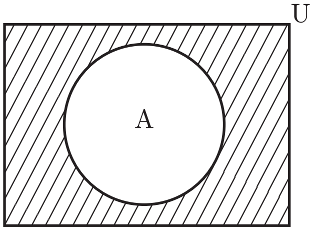
3. **Difference of sets:**

$A - B$ $B - A$

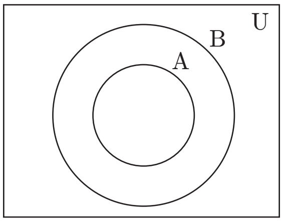




4. **Complement of a set:** A^c



5. **A is subset of B:** $A \subset B$



Types of Events

1. Each outcome of a random experiment is called an elementary event.
2. An event associated with a random experiment that always occurs whenever the experiment is performed is called a certain event.
3. An event associated with a random experiment that never occurs whenever the experiment is performed is called an impossible event.
4. If the occurrence of any one of two or more events, associated with a random experiment, presents the occurrence of all others, then the events are called mutually exclusive events.
5. If the union of two or more events associated with a random experiment includes all possible outcomes, then the events are called exhaustive events.
6. If the occurrence or non-occurrence of one event does not affect the probability of the occurrence or non-occurrence of the other, then the events are independent.
7. Two events are equally likely events if the probability of their occurrence is same.
8. An event which has a probability of occurrence equal to $1 - P$, where P is the probability of occurrence of an event A, is called the complementary event of A.

Approaches to Probability

These are two basic approaches of quantifying probability of an event.

1. **Classical approach:** Probability of an event E is calculated by the ratio of number of ways an event can occur to the number of ways a sample space can occur. This approach assumes that occurrence of all outcomes is equally probable or likely

$$P(E) = \frac{n(E)}{n(S)}$$

where $n(E)$ is the number of favorable outcomes and $n(S)$ is the number of total outcomes or sample space.

2. **Frequency approach:** Probability of an event E is defined as the relative frequency of occurrence of E. This approach is used when all outcomes are not equally probable or likely

$$P(E) = \lim_{N \to \infty} \frac{n(E)}{N}$$

where N is the number of times an experiment is performed and $n(E)$ is the number of times an event occurs.

Axioms of Probability

1. The numerical value of probability lies between 0 and 1.
   Hence, for any event A of S, $0 \le P(A) \le 1$.
2. The sum of probabilities of all sample events is unity. Hence, $P(S) = 1$.
3. Probability of an event made of two or more sample events is the sum of their probabilities.

Conditional Probability

Let A and B be two events of a random experiment. The probability of occurrence of A if B has already occurred and $P(B) \neq 0$ is known as conditional probability. This is denoted by $P(A/B)$. Also, conditional probability can be defined as the probability of occurrence of B if A has already occurred and $P(A) \neq 0$. This is denoted by$P(B/A)$.

Geometric Probability

Due to the nature of the problem or the solution or both, random events that take place in continuous sample space may invoke geometric imagery. Some popular problems such as Buffon's needle, Birds on a wire, Bertrand's paradox etc. arise in a geometrical domain. Hence. geometric probabilities can be considered as non-negative quantities with maximum value of 1 being assigned to subregions of a given domain subject to certain rules. If P is an expression of this assignment defined on a domain S, then

$$0 < P(A) \le 1, A \subset S \text{ and } P(S) = 1$$

The subsets of S for which P is defined are the random events that form a particular sample spaces. P is defined by the ratio of the areas so that if $\sigma(A)$ is defined as the area of set A, then

$$P(A) = \frac{\sigma(A)}{\sigma(s)}$$

Rules of Probability

Some of the important rules of probability are given as follows:

1. **Inclusion–Exclusion principle of probability:**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive events, $P(A \cap B) = 0$ and then formula reduces to

$$P(A \cup B) = P(A) + P(B)$$

2. **Complementary probability:**

$$P(A) = 1 - P(A^{\mathrm{c}})$$

where $P(A^{\mathrm{c}})$ is the complementary probability of A.

3. $P(A \cap B) = P(A) * P(B/A) = P(B) * P(A/B)$ where $P(A/B)$ represents the conditional probability of A given B and $P(B/A)$ represents the conditional probability of B given A.

If $B_1, B_2, ..., B_n$ are pairwise disjoint events of positive probability, then

$$P(A)$$
$$= P\left(\frac{A}{B_1}\right)P(B_1) + P\left(\frac{A}{B_2}\right)P(B_2) + \cdots + P\left(\frac{A}{B_n}\right)P(B_n)$$

4. **Conditional probability rule:**

$$P(A \cap B) = P(B) * P(A/B)$$
$$\Rightarrow P(A / B) = \frac{P(A \cap B)}{P(B)}$$
$$\text{Or } P(B / A) = \frac{P(A \cap B)}{P(A)}$$

5. **Bayes' theorem:** Suppose we have an event A corresponding to a number of exhaustive events $B_1, B_2, ..., B_n$.

If $P(B_i)$ and $P(A/B_i)$ are given, then

$$P(B_i/A) = \frac{P(B_i)P(A/B_i)}{\sum P(B_i)P(A/B_i)}$$

6. **Rule of total probability:** Consider an event E which occurs via two different values A and B. Also, let A and B be mutually exhaustive and collectively exhaustive events.

Now, the probability of E is given as

$$P(E) = P(A \cap E) + P(B \cap E)$$
$$= P(A) * P(E/A) + P(B) * P(E/B)$$

This is called the rule of total probability.

STATISTICS

Statistics deals with the methods for collection, classification and analysis of numerical data for drawing valid conclusions and making reasonable decisions. The scope of statistics now includes collection of numerical data pertaining to almost every field, calculation of percentages, exports, imports, births—deaths, etc. Hence, it is useful in business, economics, sociology, biology, psychology, education, physics, chemistry and other related fields.

A value which is used to represent a given data is called central value and various methods of finding it are called measures of central tendency. Some measures of central tendency are arithmetic mean, median and mode. However, the central values are inadequate to give us a complete idea of the distribution as they do not tell us the extent to which the observations vary from the central value. Hence, to make better interpretation from the data, we should also have an idea how the observations are scattered around a central value.

The dispersion is the measure of variations in the values of variable. It measures the degree of scatterdness of the observations in a distribution around the central value. Some of the commonly used measures of dispersion are range, mean deviation, standard deviation and quartile deviation.

Arithmetic Mean

Arithmetic Mean for Raw data

Suppose we have values $x_1, x_2, ..., x_n$ and n are the total number of values, then arithmetic mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $\bar{x}$ = arithmetic mean.

Arithmetic Mean for Grouped Data (Frequency Distribution)

Suppose f_i is the frequency of x_i, then the arithmetic mean from frequency distribution can be calculated as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i$$

where $$N = \sum_{i=1}^{n} f_i$$

Median

Arithmetic mean is the central value of the distribution in the sense that positive and negative deviations from the arithmetic mean balance each other.

Median is defined as the central value of a set of observations. It divides the whole series of observation into two parts in the sense that the numbers of values less than the median is equal to the number of values greater than the median.

Median for Raw Data

Suppose we have n numbers of ungrouped/raw values and let values be $x_1, x_2, ..., x_n$. To calculate median, arrange all the values in ascending or descending order.

Now, if n is odd, then median $= \left[\dfrac{(n+1)}{2}\right]^{\text{th}}$ value

If n is even, then median

$$= \frac{\left[\left(\dfrac{n}{2}\right)^{\text{th}} \text{value} + \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{value}\right]}{2}$$

Median for Grouped Data

To calculate median of grouped values, identify the class containing the middle observation.

$$\text{Median} = L + \left[\frac{\dfrac{(N+1)}{2} - (F+1)}{f_m}\right] \times h$$

where $L =$ lower limit of median class

$N =$ total number of data items $= \sum f$

$F =$ cumulative frequency of class immediately preceding the median class

$f_m =$ frequency of median class

$h =$ width of class

Mode

Mode of raw values of data is the value with the highest frequency or simply the value which occurs the most number of times.

Mode of Raw Data

Mode of raw data is calculated by simply checking which value is repeated the most number of times.

Mode of Grouped Data

Mode of grouped values of data is calculated by first identifying the modal class, i.e. the class which has the target frequency. The mode can then be calculated using the following formula:

$$\text{Mode} = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times h$$

where

$L =$ lower limit of modal class

$f_m =$ frequency of modal class

$f_1 =$ frequency of class preceding modal class

$f_2 =$ frequency of class following modal class

$h =$ width of model class

Relation Between Mean, Median and Mode

Empirical mode = 3 Median − 2 Mean

When an approximate value of mode is required, the given empirical formula for mode may be used.

There are three types of frequency distribution:

1. **Symmetric distribution:** It has lower half equal to upper half of distribution. For symmetric distribution,

   $$\text{Mean} = \text{Median} = \text{Mode}$$

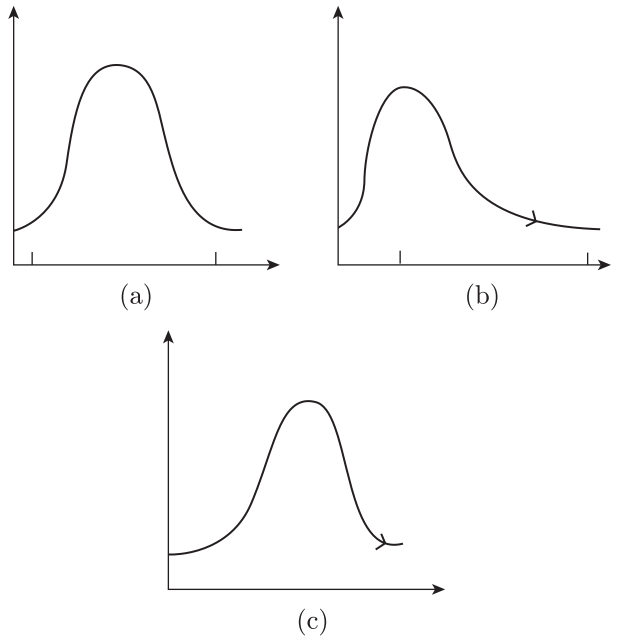
2. **Positively skewed distribution:** It has a longer tail on the right than on the left.

   $$\text{Mode} \le \text{Median} \le \text{Mean}$$

3. **Negatively skewed distribution:** It has a long tail on the left than on the right.

   $$\text{Mean} \le \text{Median} \le \text{Mode}$$

Figure 1 shows the three types of frequency distribution.


(a)

(b)

(c)

Figure 1 (a) Symmetrical frequency distribution, (b) positively skewed frequency distribution and (c) negatively skewed frequency distribution.

Geometric Mean

Geometric mean (G.M.) is a type of mean or average which indicates the central tendency or typical value of a set of numbers by using the product of their values.

Geometric Mean of Raw Data

Geometric mean of n numbers $x_1, x_2, ..., x_n$ is given by

$$\left(\overset{n}{\underset{i=1}{\pi}}\right)^{1/n} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \cdots \cdot x_n}$$

Geometric Mean of Grouped Data

Geometric mean for a frequency distribution is given by

$$\log \text{G.M.} = \frac{1}{N}\sum_{i=1}^{n} f_i \log(x_i)$$

where $N = \sum_{i=1}^{n} f_i$.

Harmonic Mean

Harmonic mean (H.M.) is the special case of the power mean. As it tends strongly towards the least element of the list, it may (compared to the arithmetic mean) mitigate the influence of large outliers and increase the influence of small values.

Harmonic Mean of Raw Data

Harmonic mean of n numbers $x_1, x_2, x_3, ..., x_n$ is calculated as

$$\text{H.M.} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Harmonic Mean of Grouped Data

Harmonic mean for a frequency distribution is calculated as

$$\text{H.M.} = \frac{N}{\sum_{i=1}^{n} (f_i/x_i)}$$

where $N = \sum_{i=1}^{n} f_i$.

Range

Range is the difference between two extreme observations of the distribution. Hence, range is calculated by subtracting the largest value and the smallest value.

Mean Deviation

The mean deviation (M.D.) is the mean of absolute of the differences in values from the mean, median or mode.

Mean Deviation of Raw Data

Suppose we have a given set of n values $x_1, x_2, x_3, ..., x_n$, then the mean deviation is given by

$$\text{M.D.} = \frac{1}{n}\sum_{i=1}^{n} \left|x_i - \bar{X}\right|$$

where $\bar{x}$ = mean.

The following steps should be followed to calculate mean deviation of raw data:

1. Compute the central value or average 'A' about which mean deviation is to be calculated.
2. Take mod of the deviations of the observations about the central value 'A', i.e. $\left|x_i - \bar{x}\right|$.
3. Obtain the total of these deviations, i.e. $\sum_{i=1}^{n} \left|x_2 - \bar{X}\right|$.
4. Divide the total obtained in step 3 by the number of observations.

Mean Deviation of Discrete Frequency Distribution

For a frequency deviation, the mean deviation is given by

$$\text{M.D.} = \frac{1}{N}\sum_{i=1}^{n} f_i \left|x_i - \bar{x}\right|$$

where $N = \sum_{i=1}^{n} f_i$.

The following steps should be followed to calculate mean deviation of discrete frequency deviation:

1. Calculate the central value or average 'A' of the given frequency distribution about which mean deviation is to be calculated.
2. Take mod of the deviations of the observations from the central value, i.e. $\left|x_i - \bar{x}\right|$.
3. Multiply these deviations by respective frequencies and obtain the total $\sum f_i \left|x_i - \bar{x}\right|$.
4. Divide the total obtained by the number of observations, i.e. $N = \sum_{i=1}^{n} f_i$ to obtain the mean deviation.

Mean Deviation of Grouped Frequency Distribution

For calculating the mean deviation of grouped frequency distribution, the procedure is same as for a discrete frequency distribution. However, the only difference is that we have to obtain the mid-points of the various classes and take the deviations of these mid-points from the given central value.

Standard Deviation

The variance of X is the arithmetic mean of the squares of all deviation of X from arithmetic mean of the observations. It is denoted by σ^2. Standard deviation (or root men square deviation) is the positive square root of the variation of X. It is denoted by σ.

Standard Deviation of Raw Data

If we have n values $x_1, x_2, ..., x_n$ of X, then

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\Rightarrow \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The following steps should be followed to calculate standard deviation for discrete data:

1. Calculate mean $(\bar{X})$ for given observations.
2. Take deviations of observations from the mean, i.e. $(x_i - \bar{X})$.
3. Square the deviations obtained in the previous step and find
$$\sum_{i=1}^{n}(x_i - \bar{X})^2$$
4. Divide the sum by n to obtain the value of variance, i.e.
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$
5. Take out the square root of variance to obtain standard deviation,
$$\sigma^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2}$$

Standard Deviation of Discrete Frequency Distribution

If we have a discrete frequency distribution of X, then

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{n}(x_i - \bar{X})^2\right]$$

$$\sigma = \sqrt{\frac{1}{N}\left[\sum_{i=1}^{n}(x_i - \bar{X})^2\right]}$$

where $N = \sum_{i=1}^{n} f$.

The following steps should be followed to calculate variance if discrete distribution of data is given:

1. Obtain the given frequency distribution.
2. Calculate mean $(\bar{X})$ for given frequency distribution.
3. Take deviations of observations from the mean, i.e. $(x_i - \bar{X})$.
4. Square the deviations obtained in the previous step and multiply the squared deviations by respective frequencies, i.e. $f_i(x_i - \bar{X})$.
5. Calculate the total, i.e. $\sum_{i=1}^{n} f_i(x_i - \bar{X})^2$.
6. Divide the sum $\sum_{i=1}^{n} f_i(x_i - \bar{X})^2$ by N, where $N = \sum f_i$, to obtain the variance, σ^2.
7. Take out the square root of the variance to obtain standard deviation, $\sigma = \sqrt{\frac{1}{n}\left[\sum_{i=1}^{n} f_i(x_i - \bar{X})^2\right]}$.

Standard Deviation of Grouped Frequency Distribution

For calculating standard deviation of a grouped frequency distribution, the procedure is same as for a discrete frequency distribution. However, the only difference is that we have to obtain the mid-point of the various classes and take the deviations of these mid-points from the given central point.

Coefficient of Variation

To compare two or more series which are measured in different units, we cannot use measures of dispersion. Thus, we require those measures which are independent of units.

Coefficient of variation (C.V.) is a measure of variability which is independent of units and hence can be used to compare two data sets with different units.

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100$$

where σ represents standard deviation and $\bar{X}$ represents mean.

PROBABILITY DISTRIBUTIONS

Random Variable

Any variable whose value is subject to variations due to randomness is called a random variable. A random variable does not have a fixed single value. Conversely, it can take a different set of values from a sample space in which each value has an associated probability.

Suppose S is sample space associated with a given random experiment. Then, random variable is a real-valued function X which assigns to each event $n \in S$ a unique real number $X(n)$.

Random variable can be discrete and continuous.

Discrete random variable is a variable that can take a value from a continuous range of values.

Continuous random variable is a variable that can take a value from a continuous range of values.

If a random variable X takes $x_1, x_2, ..., x_n$ with respective probabilities $P_1, P_2, ..., P_n$, then

$$\begin{array}{rccccc} X: & x_1 & x_2 & x_3 & \ldots & x_n \\ P(X): & P_1 & P_2 & P_3 & \ldots & P_n \end{array}$$

is known as the probability distribution of X.

Properties of Discrete Distribution

1. $\sum P(x) = 1$
2. $E(x) = \sum x P(x)$
3. $V(x) = E(x^2) - [E(x)]^2 = \sum x^2 P(x) - [\sum x P(x)]^2$

where $E(x)$ denotes the expected value or average value of a random variable x and $V(x)$ denotes the variable of a random variable x.

Properties of Continuous Distribution

1. Cumulative distribution function is given by

$$F(x) = \int_{-\infty}^{x} f(x)\,dx$$

2. $E(x) = \int_{-\infty}^{\infty} x f(x)\,dx$
3. $V(x) = E(x^2) - [E(x)]^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left[\int_{-\infty}^{\infty} x f(x) dx\right]^2$
4. $P(a < x < b) = P(a \le x \le b) = P(a < x \le b)$

$$= P(a \le x \le b) = \int_a^b f(x)\,dx$$

Types of Discrete Distribution

General Discrete Distribution

Suppose a discrete variable X is the outcome of some random experiment and the probability of X taking the value x_i is P_i, then

$$P(X = x_i) = P_i \text{ for } i = 1, 2, ...$$

where, $P(x_i) \ge 0$ for all values of i and $\sum P(x_i) = 1$.

The set of values x_i with their probabilities P_i of a discrete variable X is called a discrete probability distribution. For example, the discrete probability distribution of X, the number which is selected by picking a card from a well-shuffled deck is given by the following table:

$X = x_i$:	Ace	2	3	4	5	6	7
P_i :	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$
$X = x_i$:	8	9	10	Jack	Queen	King	
P_i :	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	$\frac{1}{13}$	

The distribution function $F(x)$ of discrete variable X is defined by

$$F(x) = P(X \le x) = \sum_{i=1}^{n} P_i$$

where x is any integer.

The mean value ($\bar{x}$) of the probability distribution of a discrete variable X is known as its expectation and is denoted by $E(x)$. If $f(x)$ is the probability density function of X, then

$$E(x) = \sum_{i=1}^{n} x_i f(x_i)$$

Variable of a distribution is given by

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 f(x_i)$$

Binomial Distribution

Binomial distribution is concerned with trails of a respective nature whose outcome can be classified as either a success or a failure.

Suppose we have to perform n independent trails, each of which results in a success with probability P and in a failure with probability X which is equal to 1 − P. If X represents the number of successes that occur in the n trails, then X is said to be binomial random variable with parameters (n, p).

The binomial distribution occurs when the experiment performed satisfies the following four assumptions of Bernoulli's trails.

1. They are finite in number.
2. There are exactly two outcomes: success or failure.
3. The probability of success or failure remains same in each trail.
4. They are independent of each other.

The probability of obtaining x successes from n trails is given by the binomial distribution formula,

$$P(X) = {}^nC_x P^x (1-p)^{n-x}$$

where P is the probability of success in any trail and $(1 - p)$ is the probability of failure.

Poisson Distribution

Poisson distribution is a distribution related to the probabilities of events which are extremely rare but which have a large number of independent opportunities for occurrence.

A random variable X, taking on one of the values 0, 1, 2, …, n, is said to be a Poisson random variable with parameters m if for some $m > 0$,

$$P(x) = \frac{e^{-m} m^x}{x!}$$

For Poisson distribution,

$$\text{Mean} = E(x) = m$$

$$\text{Variance} = V(x) = m$$

Therefore, the expected value and the variable of a Poisson random variable are both equal to its parameters m.

Hypergeometric Distribution

If the probability changes from trail to trail, one of the assumptions of binomial distribution gets violated; hence, binomial distribution cannot be used. In such cases, hypergeometric distribution is used; say a bag contains m white and n black balls. If y balls are drawn one at a time (with replacement), then the probability that x of them will be white is

$$P(x) = \frac{{}^mC_x^n C_{y-x}}{{}^{m+n}C_x}, \quad x = 0,\ 1,\ \ldots,\ y, \quad y \leq m,\ n$$

This distribution is known as hypergeometric distribution.

For hypergeometric distribution,

$$\sum_{i=1}^{n} p(x) = 1, \text{ since } \sum_{i=1}^{n} {}^mC_x^n C_{y-x} = {}^{m+n}C_y$$

Geometric Distribution

Consider repeated trails of a Bernoulli experiment E with probability of success, P, and probability of failure, $q = 1 - p$. Let x denote the number of times E must be repeated until finally obtaining a success. The distribution is

$$p(x) = q^x p, \quad x = 0,\ 1,\ 2,\ \ldots,\ q = 1 - p$$

Also, $$\sum_{x=0}^{\infty} P(x) = P \sum_{x=0}^{\infty} q^x = p\frac{1}{1-q} = 1$$

The mean of geometric distribution $= q/p$.

The variance of geometric distribution $= q/p^2$.

Types of Continuous Distribution

General Continuous Distribution

When a random variable X takes all possible values in an interval, then the distribution is called continuous distribution of X.

A continuous distribution of X can be defined by a probability density function $f(x)$ which is given by

$$p(-\infty \leq x \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

The expectation for general continuous distribution is given by

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx$$

The variance for general continuous distribution is given by

$$V(x) = \sigma^2 = \int_{-\infty}^{\infty} \left(x - \bar{X}\right)^2 f(x)dx$$

Uniform Distribution

If density of a random variable X over the interval $-\infty < a < b < \infty$ is given by

$$f(x) = \frac{1}{b-a}, \quad a < x < b$$

Then the distribution is called uniform distribution.

The mean of uniform distribution is given by

$$E(x) = \int_a^b x \cdot F(x)dx$$

$$= \frac{1}{b-a}\left|\frac{x^2}{2}\right|_a^b$$

$$= \frac{a+b}{2}$$

In uniform distribution, x takes the values with the same probability.

The variance of uniform distribution is given by

$$V(x) = \sigma^2 = \frac{1}{12}(b-a)^2$$

Exponential Distribution

If density of a random variable x for $\lambda > 0$ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Then the distribution is called exponential distribution with parameter λ.

The cumulative distribution function $F(a)$ of an exponential random variable is given by

$$F(a) = P(x \leq a) = \int_0^a \lambda e^{-\lambda x} dx = (-e^{-\lambda x})^a$$
$$= 1 - e^{-\lambda a},\ a \geq 0$$

Mean for exponential distribution is given by

$$E(x) = 1/\lambda$$

Variance of exponential distribution is given by

$$V(x) = \frac{1}{\lambda^2}$$

Normal Distribution

A random variable X is a normal random variable with parameters μ and σ^2, if the probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty$$

where μ is mean for normal distribution and σ is standard deviation for normal distribution.

CORRELATION AND REGRESSION ANALYSES

Correlation

Until now we have discussed the analysis of observations on a single variable. In this topic, we discuss the cases where the changes in one variable are related to the changes in the other variable. Such simultaneous variation where the changes in one variable are associated with changes in the other is called *correlation*.

The correlation coefficient is a measure of linear association between two variables. The values of the correlation coefficient are always between −1 and 1. If an increase (or decrease) in the values of one variable corresponds to an increase (or decrease) in the other, the correlation is positive. Evidently, if an increase (or decrease) in the values of one variable corresponds to a decrease (or increase) in the other, the correlation is negative. A correlation coefficient of +1 indicates that the correlation is perfectly positive and a correlation coefficient of −1 indicates that the correlation is perfectly negative.

Coefficient of Correlation

Coefficient of correlation is defined as the numerical measure of correlation and can be calculated by the following relation:

$$r = \frac{\sum XY}{n\sigma_x \sigma_y}$$

where X is deviation from the mean $(x - \overline{x})$, Y is deviation from the mean $(y - \overline{y})$, σ_x is standard deviation of x-series, σ_y is standard deviation of y-series, and n is number of values of the two variables.

Coefficient of correlation for grouped data can be calculated using the following relation:

$$r = \frac{n\left(\sum fd_x d_y\right) - \left(\sum fd_x\right)\left(\sum fd_y\right)}{\sqrt{\left[\left\{n\sum fd_x^2 - \left(\sum fd_x^2\right)\right\} \times \left\{n\sum fd_y^2 - \left(\sum fd_y^2\right)\right\}\right]}}$$

where d_x is deviation of the central values from the assumed mean of x-series, d_y is deviation of the central values from the assumed mean of y-series, f is the frequency corresponding to the pair (x, y) and n is total number of frequencies $(= \sum f)$.

Lines of Regression

Sometimes, the dots of the scatter diagram tend to cluster along a well-defined direction which suggests a linear relationship between the variables x and y as shown in Fig. 2. Such a line giving the best fit for the given distribution of dots is known as line of regression.

The line giving the best possible mean values of y for each specified value of x is called the line of regression of y on x and the line giving the best possible mean values of x for each specified value of y is called the line of regression of x on y.

The regression coefficient of y on x is $r\dfrac{\sigma_y}{\sigma_x}$.

The regression coefficient of x on y is $r\dfrac{\sigma_x}{\sigma_y}$.

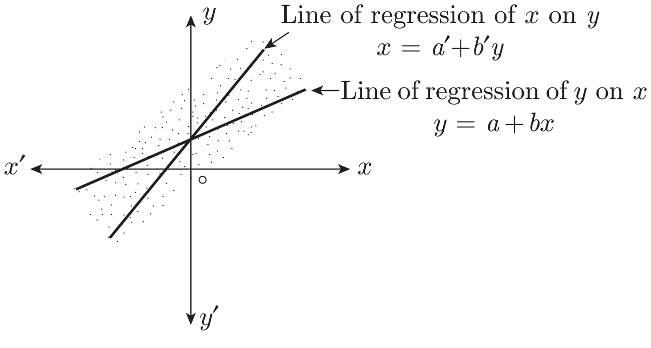

y
Line of regression of x on y
$x = a' + b'y$
Line of regression of y on x
$y = a + bx$
x'
x
O
y'

Figure 2 | Line of regression.

HYPOTHESIS TESTING

In statistics and probability theory, hypothesis testing is used for determining the probability that a given hypothesis is true. Let us illustrate the concept of hypothesis testing with the help of an example. Let us say that there are two schools of thoughts as regards how the presence of caffeine in coffee affects the alertness of people. The test is conducted on students of a class by dividing them into two groups; one that drank coffee with caffeine and the other that drank coffee without caffeine; their alertness (in terms of time spent sleeping in class) was tested and it was found that the group that drank coffee with caffeine slept less. One therefore tends to conclude that caffeine in coffee improves alertness. But the satistician may argue otherwise. He would say that difference in the average alertness of two groups was due to chance. There could be reasons other than the effect of caffeine, such as the students of caffeine-taker group having had a better sleep the previous night or being more interested in the class. He would argue that if the students were divided differently in two groups, the results would have been different and the conclusion drawn from the test might have been proven wrong. The satistician would therefore not conclude unless he performs a hypothesis test.

Hypothesis test allows us to make a rational decision between the hypothesis of real effects and chance explanations. Although the chance explanations cannot be completely eliminated, these may be unlikely if difference between the two groups is very large. A hypothesis would specify the required quantum of difference to conclude that the effects are real. The purpose of hypothesis testing is to eliminate false scientific conclusions as much as possible. The process of hypothesis testing consists of the following four major steps:

1. The first step is null hypothesis and alternative hypothesis. A null hypothesis (H_0) is a statistical hypothesis that is tested for possible rejection under the assumption that it is true. Rejection of null hypothesis would mean that the observations are not due to chance. The alternative hypothesis (H_a) is contrary to the null hypothesis. Alternative hypothesis believes that the observations are the result of a real effect with some amount of chance element included. The two hypotheses are stated in such a way that they are mutually exclusive. If one were true, other would be false and vice versa.
2. The second step is to identify a test statistic that can be used for assessing the truth of the null hypothesis. The analysis plan is to use sample data to either accept or reject the null hypothesis. The following elements need to be specified to complete the task of either accepting or rejecting null hypothesis. The first element is to choose a significance level. The significance level also denoted as α is the probability of rejecting the null hypothesis when it is true. A significance level of 0.02 indicates a 2% risk concluding that a difference exists when there is no actual difference. Researchers often choose significance level of 0.01, 0.05 or 0.1. Any value between 0 and 1 though can be used. The second and the important element is the test statistic and a sampling distribution. Computed from sample data, the test statistic might be a mean score, a proportion, difference between means, difference between proportions, etc. From the given sample distribution and test statistic, probabilities associated in the test statistic are computed. If the test statistic probability turns out to be less than the chosen significance level, the null hypothesis is rejected.
3. The third step is to analyze the sample data. It involves computing the P-value, which is the probability that a test statistic is at least as significant as the one observed, and would be obtained with the assumption that the null hypothesis was true. A smaller P-value signifies a stronger belief against the null hypothesis.
4. The P-value is next compared with the chosen (or an acceptable) significance level. If $P \leq \alpha$, then the observed effect is statistically significant. That is, null hypothesis is rejected and alternative hypothesis is valid.

HYPOTHESIS TESTING PROCEDURES FOR SOME COMMON STATISTICAL PROBLEMS

Hypothesis testing procedures for some common statistical problems are briefly described in the following paragraphs. The testing procedures discussed include hypothesis test procedures involving proportions, means, difference between proportions and difference between means.

Hypothesis Testing of a Proportion

The hypothesis testing of a proportion described as follows assumes that the following conditions are satisfied:

(a) The sampling method is a simple random sampling which implies that random sampling has a population of (N) objects; the sample consists of (n) objects and all possible samples of (n) objects are equally likely to occur.

(b) Each sample point can have two possible outcomes, namely, a success or a failure.

(c) The sample includes at least 10 successes and 10 failures.