Statistics

Mid-Term Report

Ronak Gupta

23B2148

Mentor: Ashutosh Gandhe

Statistics is the art of learning from data. It is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of conclusions.

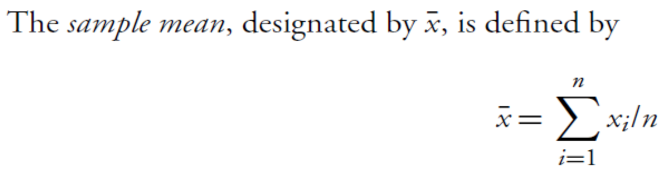
Statistics is of two types :- descriptive and inferential statistics

Descriptive Statistics is concerned with the description and summarization of data.

Inferential Statistics is concerned with the drawing of conclusions

Central Tendency

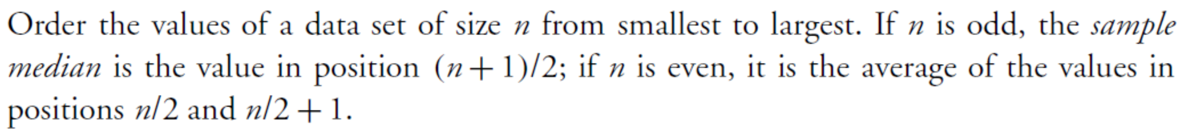
Mean *:-*



The *sample mean*, designated by $\bar{x}$, is defined by

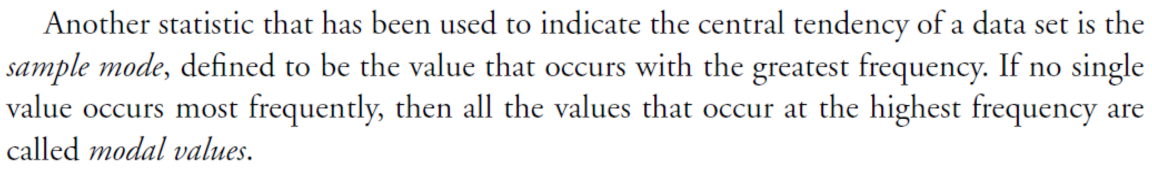
$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

Median *:-*



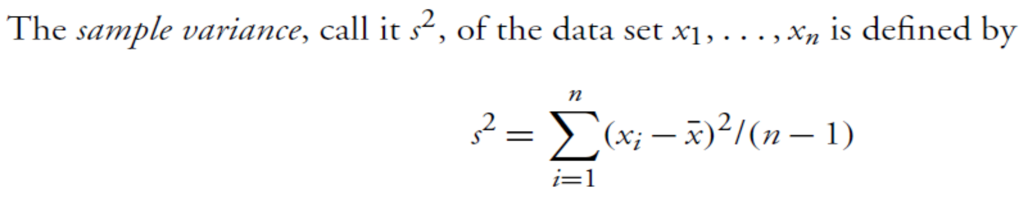
Order the values of a data set of size n from smallest to largest. If n is odd, the *sample median* is the value in position $(n+1)/2$; if n is even, it is the average of the values in positions $n/2$ and $n/2+1$.

Mode *:-*



Another statistic that has been used to indicate the central tendency of a data set is the *sample mode*, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*.

Variance*:-*



The *sample variance*, call it s^2, of the data set $x_1, \ldots, x_n$ is defined by

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

We divide by n-1 not n in case of sample as it corrects the bias and makes the sample variance closer to population variance.

We can use this identity to further simplify the variance calculation.

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

The computation of the sample variance can also be eased by noting that if

$$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

then $\bar{y} = a + b\bar{x}$, and so

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$

That is, if s_y^2 and s_x^2 are the respective sample variances, then

$$s_y^2 = b^2 s_x^2$$

Standard Deviation :-

The quantity s, defined by

$$s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$$

Sample Percentile :-

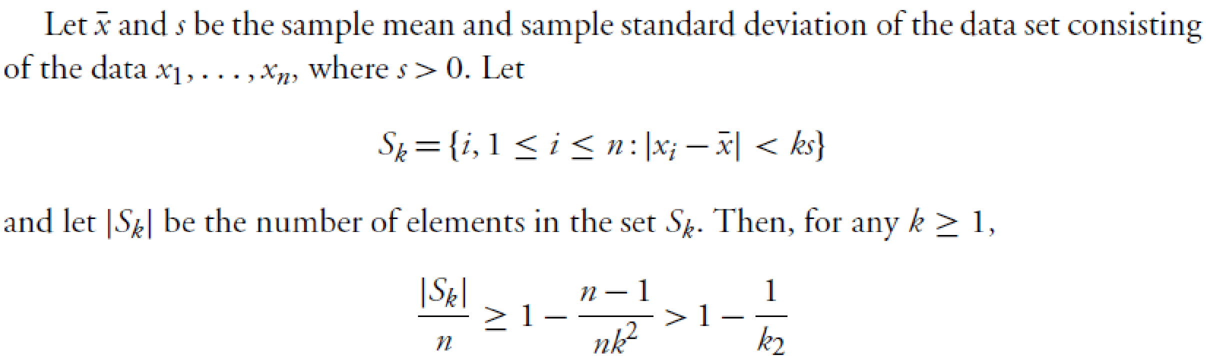
The *sample 100*p *percentile* is that data value such that at least 100p percent of the data are less than or equal to it and at least 100$(1-p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample 100p percentile is the arithmetic average of these two values.

To determine the sample 100p percentile of a data set of size n, we need to determine the data values such that

1. At least np of the values are less than or equal to it.
2. At least $n(1-p)$ of the values are greater than or equal to it.

The sample 25 percentile is called the *first quartile*; the sample 50 percentile is called the sample median or the *second quartile*; the sample 75 percentile is called the *third quartile*.

Chebyshev Inequality :-



Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, \ldots, x_n$, where $s > 0$. Let

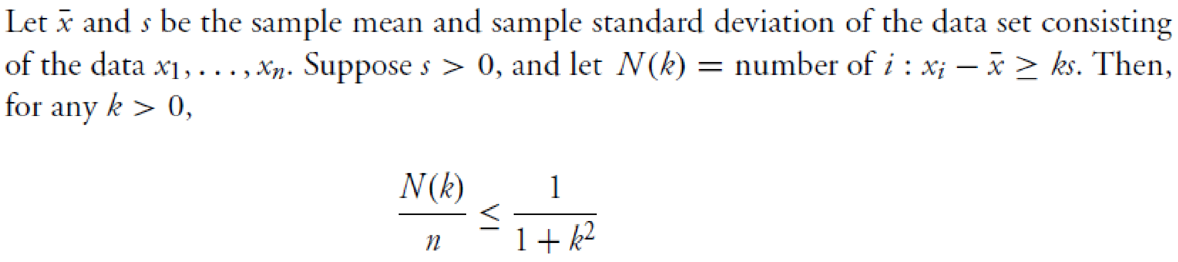
$$S_k = \{i, 1 \le i \le n : |x_i - \bar{x}| < ks\}$$

and let $|S_k|$ be the number of elements in the set S_k. Then, for any $k \ge 1$,

$$\frac{|S_k|}{n} \ge 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k_2}$$

But there is a problem, as Chebyshev's inequality holds universally, it might be expected for given data that the actual percentage of the data values that lie within the interval from $\overline{x}$ −ks to $\overline{x}$ +ks might be quite a bit larger than the bound given by the inequality.

So we often use one-sided Chebyshev Inequality

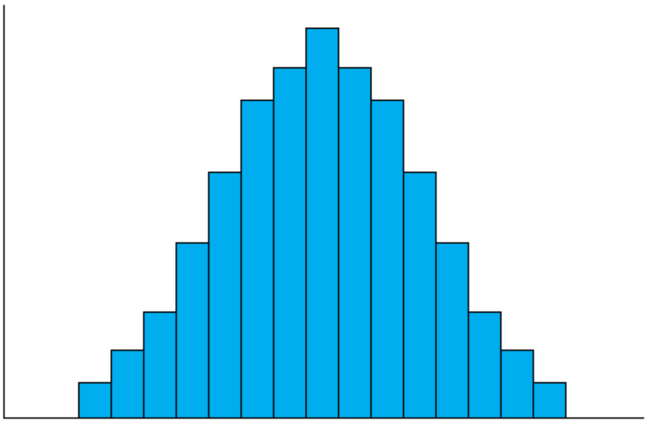


Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, \ldots, x_n$. Suppose $s > 0$, and let $N(k) =$ number of $i : x_i - \bar{x} \ge ks$. Then, for any $k > 0$,

$$\frac{N(k)}{n} \le \frac{1}{1+k^2}$$

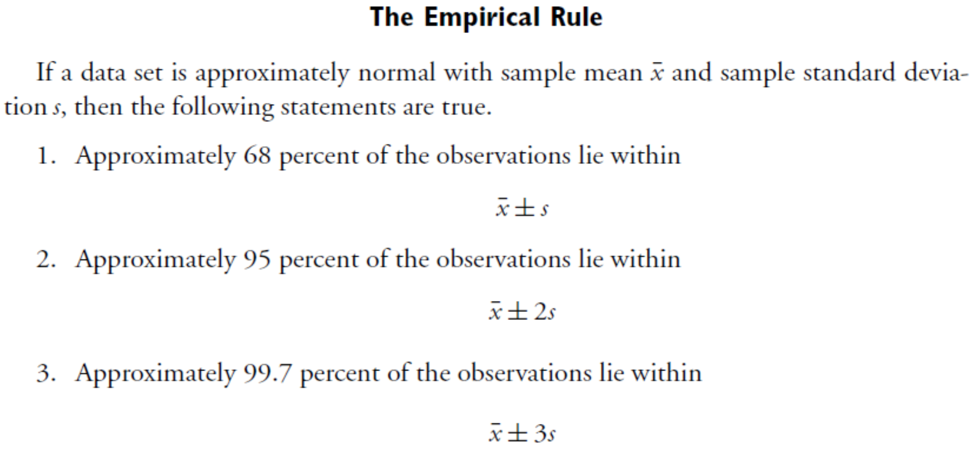
Normal Data Sets

Normal Data Sets often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion.



A data set whose histogram has two local peaks is said to be *bimodal*.

Normal Data Sets follow the Empirical Rule



The Empirical Rule

If a data set is approximately normal with sample mean $\bar{x}$ and sample standard deviation s, then the following statements are true.

1. Approximately 68 percent of the observations lie within

$$\bar{x} \pm s$$

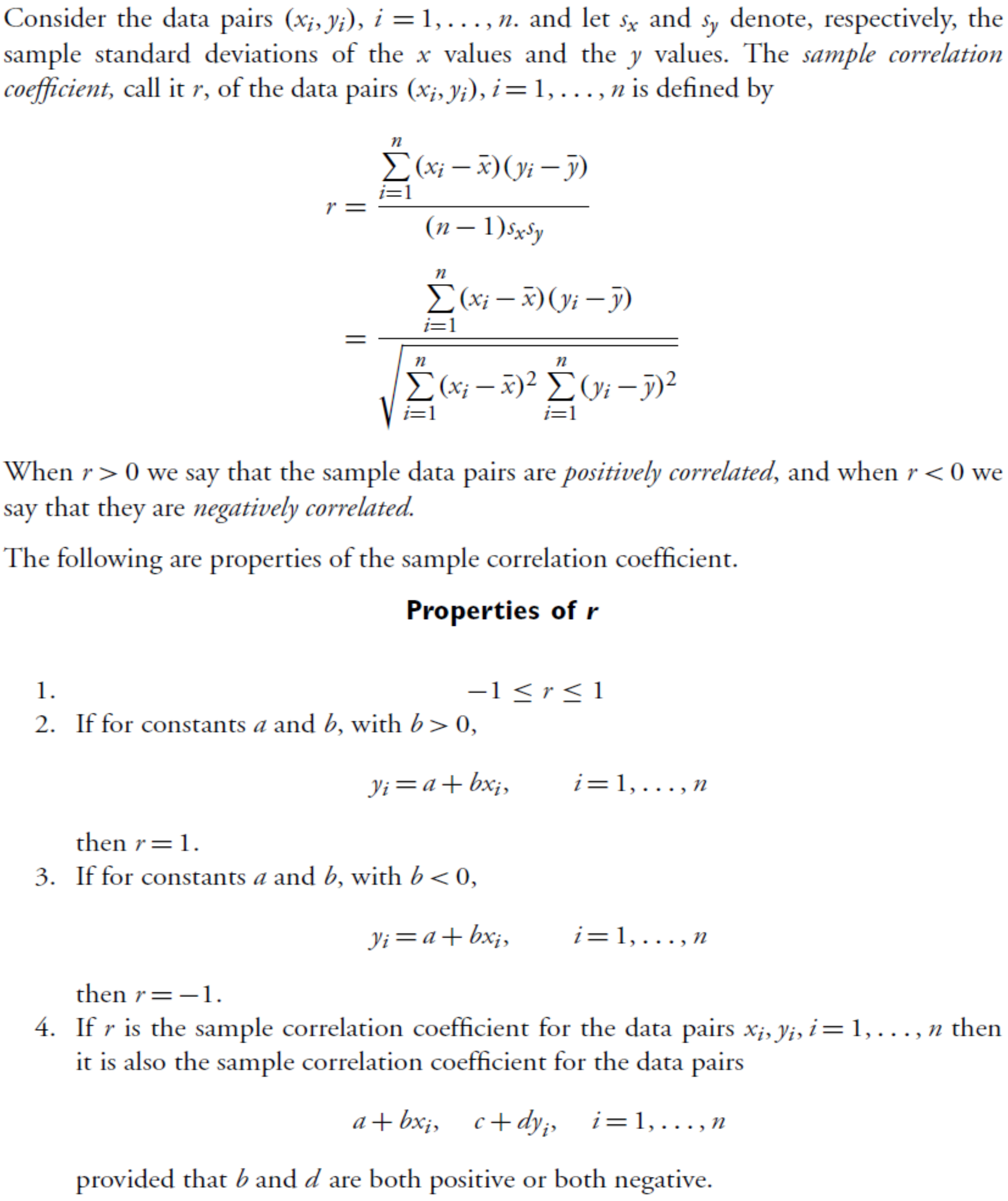
2. Approximately 95 percent of the observations lie within

$$\bar{x} \pm 2s$$

3. Approximately 99.7 percent of the observations lie within

$$\bar{x} \pm 3s$$

Correlation coefficient



Consider the data pairs (x_i, y_i), $i = 1, \dots, n$. and let s_x and s_y denote, respectively, the sample standard deviations of the x values and the y values. The *sample correlation coefficient,* call it r, of the data pairs (x_i, y_i), $i = 1, \dots, n$ is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

When $r > 0$ we say that the sample data pairs are *positively correlated*, and when $r < 0$ we say that they are *negatively correlated.*

The following are properties of the sample correlation coefficient.

Properties of *r*

1. $$-1 \le r \le 1$$
2. If for constants a and b, with $b > 0$,

$$y_i = a + bx_i, \qquad i = 1, \dots, n$$

then $r = 1$.

3. If for constants a and b, with $b < 0$,

$$y_i = a + bx_i, \qquad i = 1, \dots, n$$

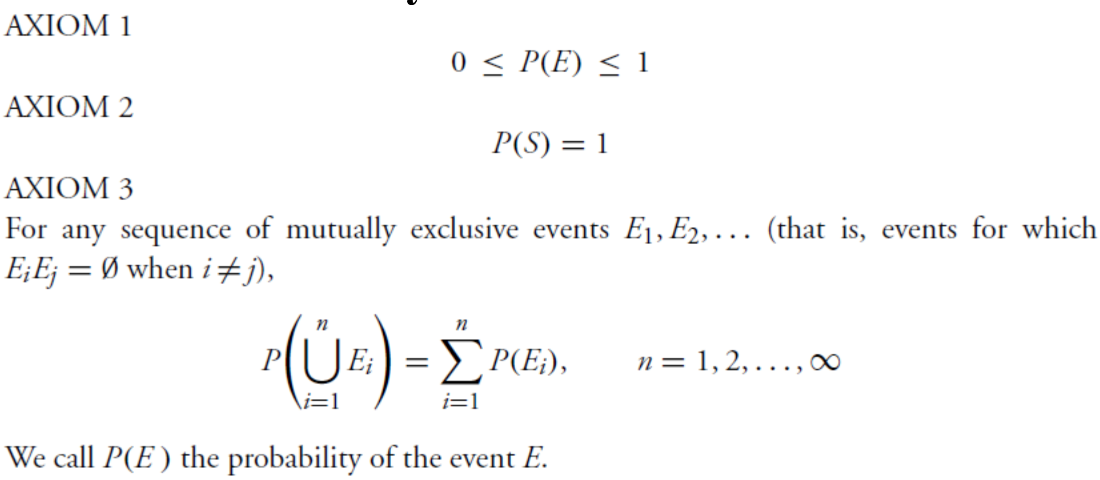
then $r = -1$.

4. If r is the sample correlation coefficient for the data pairs x_i, y_i, $i = 1, \dots, n$ then it is also the sample correlation coefficient for the data pairs

$$a + bx_i, \quad c + dy_i, \quad i = 1, \dots, n$$

provided that b and d are both positive or both negative.

Axioms of Probability



AXIOM 1

$$0 \leq P(E) \leq 1$$

AXIOM 2

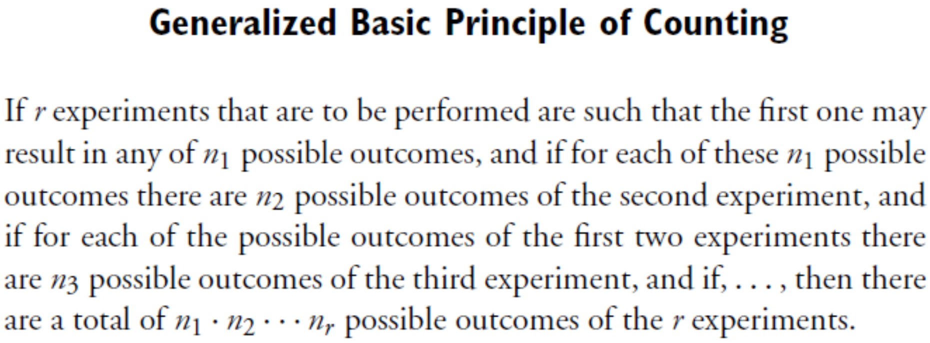
$$P(S) = 1$$

AXIOM 3

For any sequence of mutually exclusive events $E_1, E_2, \ldots$ (that is, events for which $E_iE_j = \emptyset$ when $i \neq j$),

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i), \qquad n = 1, 2, \ldots, \infty$$

We call $P(E)$ the probability of the event E.



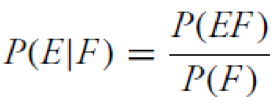
Generalized Basic Principle of Counting

If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes, and if for each of these n_1 possible outcomes there are n_2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments there are n_3 possible outcomes of the third experiment, and if, . . . , then there are a total of $n_1 \cdot n_2 \cdots n_r$ possible outcomes of the r experiments.

Conditional Probability

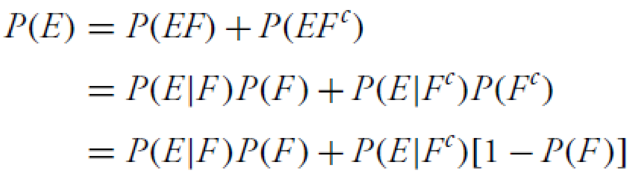
In conditional probability if the event F occurs, then in order for E to occur it is necessary that the actual occurrence be a point in both E and F; that is, it must be in EF.

We calc the conditional probability P(E|F) as



$$P(E|F) = \frac{P(EF)}{P(F)}$$

Bayes' Formula



$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)[1 - P(F)] \end{aligned}$$

Probability of the event E is a weighted average of the conditional probability of E given that F has occurred and the conditional probability of E given that F has not occurred, with each conditional probability being given as much weight as the event it is conditioned on has of occurring.

Suppose there are n events

$$P(E) = \sum_{i=1}^{n} P(EF_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$$

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

Independent Events :-

Independent events are those events whose occurrence is not dependent on any other event.

$$P(EF)=P(E)P(F)$$

Of course we may also extend the definition of independence to more than three events. The events $E_1, E_2, \ldots, E_n$ are said to be independent if for every subset $E_{1'}, E_{2'}, \ldots, E_{r'}$, $r \leq n$, of these events

$$P(E_{1'}E_{2'} \cdots E_{r'}) = P(E_{1'})P(E_{2'}) \cdots P(E_{r'})$$

Random Variables

These quantities of interest that are determined by the result of the experiment are known as random variables. If X is the random variable then

$$1 = P(S) = P\left(\bigcup_{i=2}^{12} \{X = i\}\right) = \sum_{i=2}^{12} P\{X = i\}$$

Random Variable are divided into two parts:- *Discrete* and *Continuous*

Random variables whose set of possible values can be written either as a finite sequence x1, . . . , xn, or as an infinite sequence x1, . . . are said to be **discrete**.

Random variables that take on a continuum of possible values are said to be **continuous**.

The *cumulative distribution function*, or more simply the *distribution function*, F of the random variable X is defined for any real number x by

$$F(x) = P\{X \leq x\}$$

For a discrete random variable X, we define the probability mass function p(a) of X by

$$p(a) = P\{X = a\}$$

The probability mass function $p(a)$ is positive for at most a countable number of values of a. That is, if X must assume one of the values $x_1, x_2, \ldots,$ then

$$\begin{aligned} p(x_i) &> 0, \qquad i = 1, 2, \ldots \\ p(x) &= 0, \qquad \text{all other values of } x \end{aligned}$$

Since X must take on one of the values x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

EXAMPLE 4.2a Consider a random variable X that is equal to 1, 2, or 3. If we know that

$$p(1) = \tfrac{1}{2} \qquad \text{and} \qquad p(2) = \tfrac{1}{3}$$

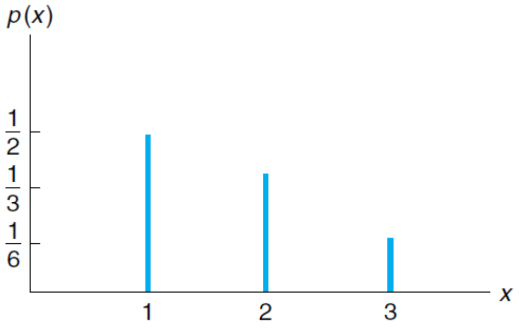
then it follows (since $p(1) + p(2) + p(3) = 1$) that

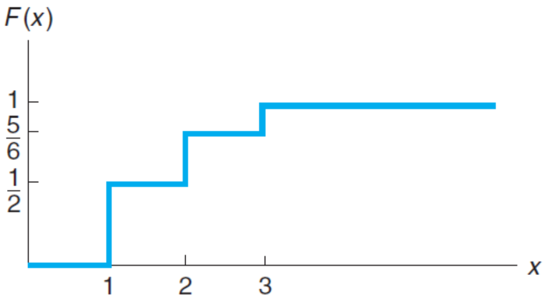
$$p(3) = \tfrac{1}{6}$$

A graph of $p(x)$ is presented in Figure 4.1. ■

The cumulative distribution function F can be expressed in terms of $p(x)$ by

$$F(a) = \sum_{\text{all } x \leq a} p(x)$$





We say that X is a continuous random variable if there exists a nonnegative function f (x), defined for all real x ∈ (−∞,∞), having the property that for any set B of real numbers

$$P\{X \in B\} = \int_B f(x)\, dx$$

The function f (x) is called the *probability density function* of the random variable X.

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\, dx$$

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

If we let $a = b$ in the above, then

$$P\{X = a\} = \int_a^a f(x)\,dx = 0$$

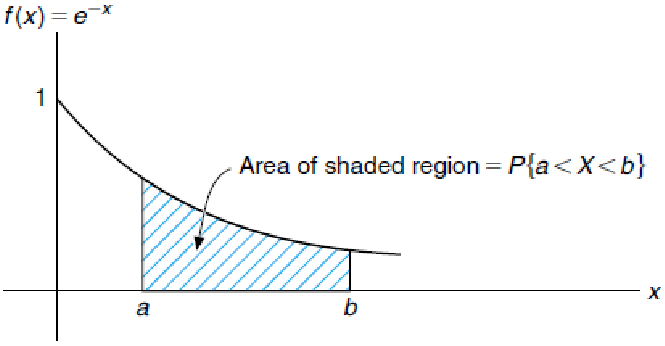
this equation states that the probability that a continuous random variable will assume any particular value is zero.

The relationship between the cumulative distribution $F(\cdot)$ and the probability density $f(\cdot)$ is expressed by

$$F(a) = P\{X \in (-\infty, a]\} = \int_{-\infty}^{a} f(x)\,dx$$

Differentiating both sides yields

$$\frac{d}{da}F(a) = f(a)$$



Jointly Distributed Random Variables

To specify the relationship between two random variables, we define the joint cumulative probability distribution function of X and Y by

$$F(x, y) = P\{X \le x, Y \le y\}$$

distribution function of X — call it FX — can be obtained from the joint distribution function F of X and Y as follows:

$$\begin{aligned} FX\ (x) &= P\{X \le x\} \\ &= P\{X \le x, Y < \infty\} \\ &= F(x,\infty) \end{aligned}$$

Similarly, the cumulative distribution function of Y is given by

$$FY\ (\ y) = F(\infty, y)$$

If X and Y are both discrete random variables them joint probability mass function of X and Y , p(xi, yj), by

$$p(xi, yj) = P\{X = xi\ ,\ Y = yj\}$$

$$\begin{aligned} P\{X = x_i\} &= P\left(\bigcup_j \{X = x_i, Y = y_j\}\right) \\ &= \sum_j P\{X = x_i, Y = y_j\} \\ &= \sum_j p(x_i, y_j) \end{aligned}$$

X and Y are jointly continuous if there exists a function f (x, y) defined for all real x and y, having the property that for every set C of pairs of real numbers (that is, C is a set in the two-dimensional plane).

$$P\{(X, Y) \in C\} = \iint\limits_{(x,y)\in C} f(x, y)\, dx\, dy$$

The function f (x, y) is called the joint probability density function of X and Y . If A and B are any sets of real numbers, then by defining C = {(x, y) : x ∈ A, y ∈ B}

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x, y)\, dx\, dy$$

If X and Y are jointly continuous, they are individually continuous, and their probability density functions can be obtained as follows:

$$\begin{aligned} P\{X \in A\} &= P\{X \in A, Y \in (-\infty, \infty)\} \\ &= \int_A \int_{-\infty}^{\infty} f(x, y)\, dy\, dx \\ &= \int_A f_X(x)\, dx \end{aligned}$$

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

Independent Random Variables

The random variables X and Y are said to be independent if for any two sets of real numbers A and B

$$P\{X \in A,\ Y \in B\} = P\{X \in A\}P\{Y \in B\}$$

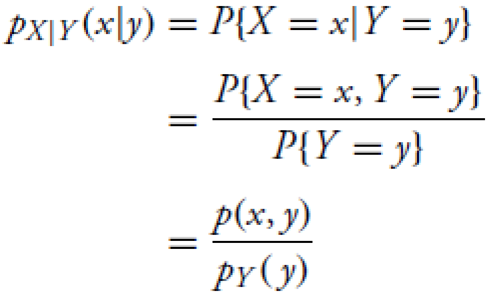
In other words, X and Y are independent if, for all A and B, the events EA = {X ∈ A} and FB = {Y ∈ B} are independent.
When X and Y are discrete independent random variables,

$$p(x, y) = pX\ (x)pY\ (\ y)$$

In the jointly continuous case, the condition of independence is equivalent to
f (x, y) = fX (x)fY (y) for all x, y.

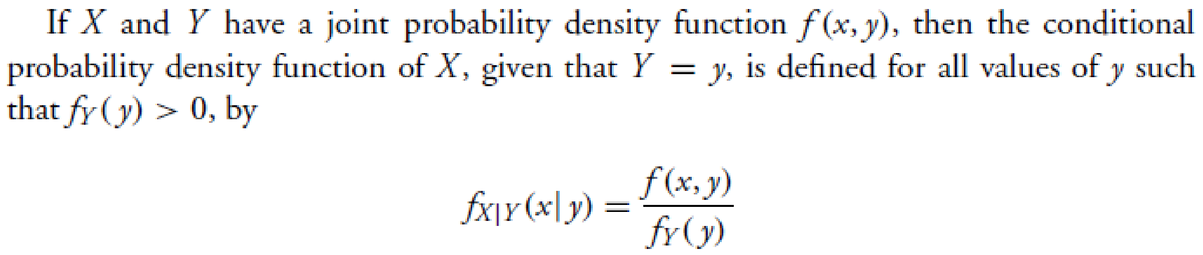
Conditional Distributions

if X and Y are discrete random variables, it is natural to define the conditional probability mass function of X given that Y = y, by



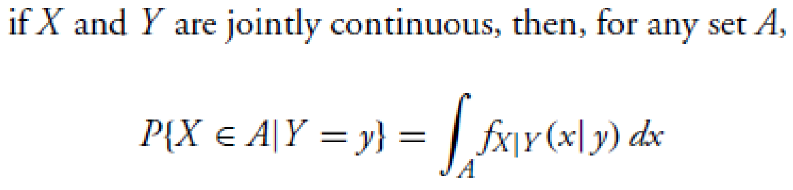
$$\begin{aligned} p_{X|Y}(x|y) &= P\{X = x|Y = y\} \\ &= \frac{P\{X = x, Y = y\}}{P\{Y = y\}} \\ &= \frac{p(x, y)}{p_Y(y)} \end{aligned}$$

for all values of y such that pY (y) > 0.



If X and Y have a joint probability density function $f(x, y)$, then the conditional probability density function of X, given that $Y = y$, is defined for all values of y such that $f_Y(y) > 0$, by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$



if X and Y are jointly continuous, then, for any set A,

$$P\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)\,dx$$

Expectation

It is the weighted average of the possible values that random variable can take,

For discrete random variables

$$\sum_{i=1}^{n} x_i p(x_i) = E[X]$$

For continuous random variables

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

Properties of expected value

(a) If X is a discrete random variable with probability mass function $p(x)$, then for any real-valued function g,

$$E[g(X)] = \sum_{x} g(x)p(x)$$

(b) If X is a continuous random variable with probability density function $f(x)$, then for any real-valued function g,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\, dx$$

If a and b are constants, then

$$E[aX + b] = aE[X] + b$$

The expected value of a random variable X, E[X], is also referred to as the mean or the first moment of X. The quantity E[Xn], n ≥ 1, is called the nth moment of X.

$$E[X^n] = \begin{cases} \sum_{x} x^n p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^n f(x)\, dx & \text{if } X \text{ is continuous} \end{cases}$$

Expected Value of Sums of Random Variables

If X and Y are random variables and g is a function of two variables

$$
\begin{aligned}
E[g(X, Y)] &= \sum_y \sum_x g(x, y)p(x, y) && \text{in the discrete case} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)\, dx\, dy && \text{in the continuous case}
\end{aligned}
$$

$$E[X_1 + X_2 \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

Variance

If X is a random variable with mean μ, then the *variance* of X, denoted by $\text{Var}(X)$, is defined by

$$\text{Var}(X) = E[(X - \mu)^2]$$

Another form is

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

Covariance and Variance of Sums of Random Variables

$$
\begin{aligned}
\text{Var}(X + X) &= \text{Var}(2X) \\
&= 2^2\ \text{Var}(X) \\
&= 4\ \text{Var}(X) \\
&\neq \text{Var}(X) + \text{Var}(X)
\end{aligned}
$$

There is, however, an important case in which the variance of a sum of random variables is equal to the sum of the variances; and this is when the random variables are independent.

The *covariance* of two random variables X and Y, written $\text{Cov}(X, Y)$, is defined by

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

where μ_x and μ_y are the means of X and Y, respectively.

- Cov(X, Y) =E[XY] − E[X]E[Y]
- Cov(X, Y) = Cov(Y , X)
- Cov(X, X) = Var(X)
- Cov(aX , Y) = a Cov(X, Y)
- Cov(X1 + X2, Y) = Cov(X1, Y) + Cov(X2, Y) (Covariance, like expectation, possesses an additive property.)

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} \text{Cov}(X_i, Y)$$

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$$

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \text{Cov}(X_i, X_j)$$

If X and Y are independent random variables, then

$$\text{Cov}(X, Y) = 0$$

and so for independent $X_1, \ldots, X_n$,

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Special Random Variables

Bernoulli and Binomial Random Variables

A random variable X is said to be a Bernoulli random variable if its probability mass function is given by

$$P\{X = 0\} = 1 - p$$
$$P\{X = 1\} = p$$

Where X=1 is when outcome is success and X=0 when outcome is failure.
Its expected value is

$E[X] = 1 \cdot P\{X = 1\} + 0 \cdot P\{X = 0\} = p$

If there are n independent trials, each with probability of success and failure is p and 1-p respectively. If X represents the number of successes that occur in n trails them X is said to be a binomial random variable with parameters (n,p). The probability mass function of a binomial random variable with parameters n and p is given by

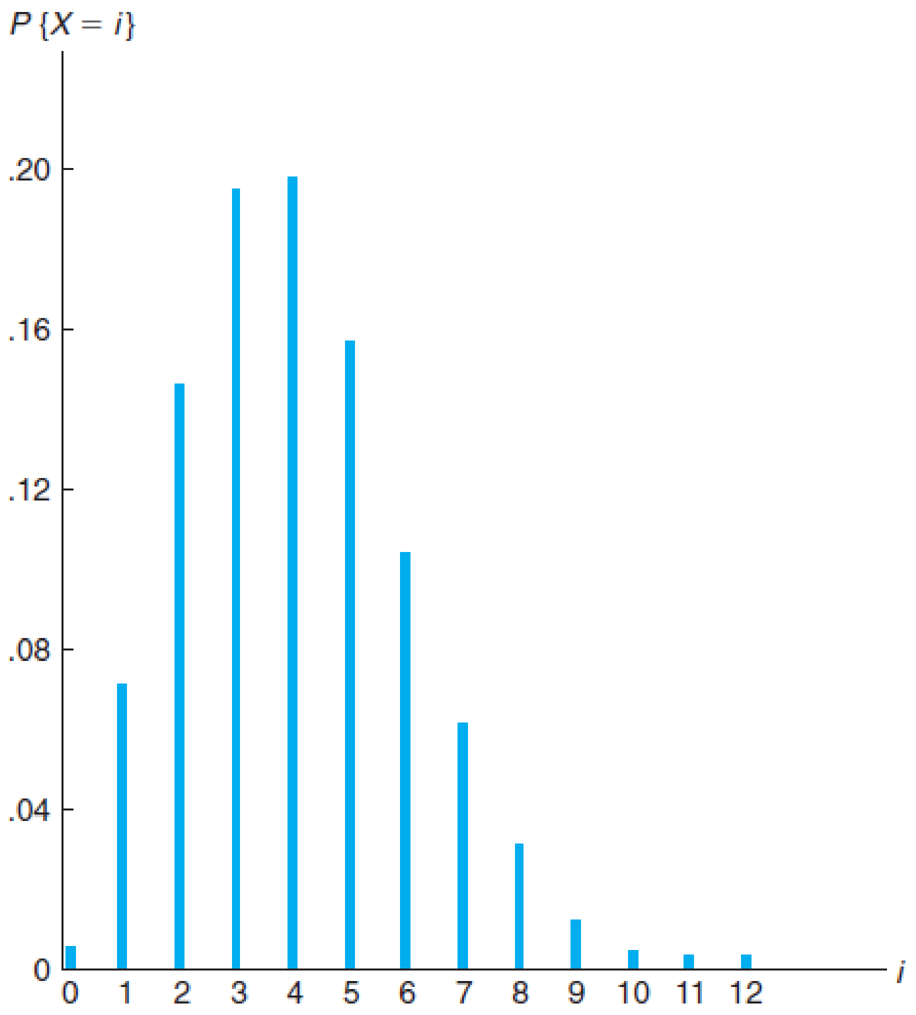
$$P\{X = i\} = \binom{n}{i} p^i (1 - p)^{n-i}, \quad i = 0, 1, \ldots, n$$

The Poisson Random Variable

A random variable X, taking on one of the values 0, 1, 2, . . . , is said to be a Poisson random variable with parameter λ, λ > 0, if its probability mass function is given by

$$P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}, \qquad i = 0, 1, \ldots$$

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda} \sum_{i=0}^{\infty} \lambda^i / i! = e^{-\lambda} e^{\lambda} = 1$$



Poisson probability mass function with λ = 4.

Both the mean and the variance of a Poisson random variable are equal to the parameter λ.

The Poisson random variable has a wide range of applications in a variety of areas because it may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small.

Updated POA

- Special Random Variables and Sampling Statistics : Week 5
- Sampling Statistics and Central Limit Theorem : Week 6
- Parameter Estimation : Week 7
- Hypothesis Testing : Week 7, 8
- Introduction to Regression- Simple, Multiple, Polynomial and Logistic : Week 9

End Term Report Submission