Definition

Intuitively by ***a random variable** (r.v)* we mean a real number X connected with the outcome of a random experiment E.

Thus, to each outcome w of this experiment, there corresponds a real function $X(w)$ with domain S (the sample space) and range $(-\infty, \infty)$.

Notation:

Thus $X(w)$ represents the real number, with the random variable X which is associates with the outcome w.

Random variable(s) are denoted by capital letters X, Y, Z, ... etc. and typical outcome(s) of the random experiment (i.e., a typical element of the "sample space") are denoted by x, y, z, ... etc.

Types of random variables

Discreate vs. Continuous

If a random variable takes ***at most countable number of values***, it is called a discrete otherwise it is said to be continuous.

Example:

"In tossing of a coin, if the coin is tossed twice (say) the sample space is $S = \{Head, Tail\}$ the possible outcomes are $\{HH, HT, TH, TT\}$."

If $X(w)$ is the possible number of heads then,

X is said to be a random variable and the values w are 0, 1 or 2

Properties (theorems) of Random variables

- If X_1 and X_2 are random variables and C is a constant then CX_1 , $X_1 \pm X_2$, X_1X_2 are also random variables
- It will follow that $C_1X_1 + C_2X_2$ is also a random variable for constants C_1 and C_2 .

- If X_1 and X_2 are random variables then (i) max $\{X_1, X_2\}$ and (ii) min $\{X_1, X_2\}$ are also random variables. These are common in order statistics
- If X is a r.v and $f(\cdot)$ is a continuous or an increasing function then $f(X)$ is a random variable

Notation

If x is a-real number, then the set of all w in S such that $X(w) = x$ is denoted briefly by writing $X = x$.

- $P(X = a \text{ or } X = b) = P\{(X = a) u (X = b)\}$

- $P(X = a \text{ and } X = b) = P\{(X = a) n (X = b)\}$

Probability and distribution functions

Discrete Random Variable

If a random, variable takes at most a countable number of values, it is called a discrete random variable.

In other words, a real valued function defined on a discrete sample space is called a discrete random variable.

Probability Mass Function

Suppose X is a discrete random variable taking at most a countably infinite number of values $x_1, x_2, \ldots$

With each possible outcome x_i, we associate a number $p_i = P(X = x_i) = p(x_i)$ which is termed as probability of x_i.

Then numbers $p(x_i); i = 1, 2, \ldots$. Satisfying the following conditions:

a. $p_i \geq 0, \forall i$, and

b. $\sum p_i = 1, i = 1, 2, \ldots$

Is called the *probability mass function* of the random variable X

and the set (x_1, p_1) is called the *probability distribution (p.d.)* of the random variable X.

Discrete Distribution Function.

In this case there are a countable number of points

$x_1, x_2, \ldots$ and numbers p_i and $\sum p_i = 1$ such that

$F(X) = \sum_{x_i \le x} p_i$.

The graph of $F(x)$ is a "step function"

Example

A random variable X has the following probability function:

Values of X, x :	-2	-1	0	1	2	3
p(x) :	0·1	k	0·2	2k	0·3	k

a. Find the value of k

b. Construct the distribution Function

c. Draw the graph of p_i and $F(X)$

Examples of Discrete Distributions

Special Distributions

- Binomial
- Multinomial
- Negative binomial
- Poisson
- Geometric
- Hypergeometric

Phenomenon

Each of the above distribution has certain phenomenon in which it can be applied and eachh can be defined by its own probability mass function in such phenomenon

Binomial

Experimental conditions

Each trial results in two mutually disjoint outcomes termed as success and failure. The number of trials "n" is finite. The' trials are independent of each other. The probability of success 'p_1' is constant for-each trial

Poisson

Experimental conditions:

This occurs when there are events which do not occur as outcomes of a definite number of trials (unlike that in binomial) of an experiment;

but which occur at random points of time and space wherein our interest lies only in the number of occurrences of the event, not in its non-occurrences.

Negative Binomial

Experimental conditions

Suppose we have a succession of n Bernoulli trails. In which (i) the trials are independent, (ii) the probability of success 'p' in a trial remains constant from trial to trial in which there are x failures preceding the rth success in $x + r$ trials.

Geometric

<u>Experimental conditional</u>

Suppose we have a series of independent trials and on each trial the probability of success 'p' remains the same. Such that there are x *failures* preceding the *first success*.

Now, *the last trial result to a success event*

Definition: A random variable X is said to follow binomial distribution if it assumes only non-negative values and its probability mass function is given by,

$$P(X = x) = \binom{n}{x} p^x q^{n-x},$$

$$x = 0, 1, \dots n;$$

$$q = 1 - p$$

Negative Binomial

Experimental conditions

Suppose we have a succession of n Bernoulli trails. In which

(i) The trials are independent,

(ii) The probability of success 'p' remains constant from trial to trial

The main interest is to obtain r successes out of x failures such that there are x failures preceding the rth success.

Total number of trials in this distribution is

$$n = x + r .$$

Definition: A random variable X is said to follow a negative binomial distribution if its probability mass function is given by

$$P(X = x) = \binom{-r}{x}(-q)^x p^r,$$

$$x = 0, 1, 2 \ldots$$

The expected value, $E(X) = {}^{rq}/_{p}$

Variance, $Var\ (X) = {}^{rq}/_{p^2}$

Geometric

Experimental conditional

Suppose we have a series of independent trials and on each trial the probability of success 'p' remains the same.

Such that there are x *failures* preceding the *first success*.

The main interest in this distribution is the success, hence in this distribution *the last trial result to a success event.*

This can be considered as a special case of negative distribution in which $r = 1$

Definition. A random variable X is said to have a geometric distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = q^{x}p,$$

$$x = 0, 1, 2, \ldots \quad 0 < \mathrm{p} \leq 1$$

The expected value, $E(X) = {}^{q}/_{p}$

Variance, $Var\ (X) = {}^{q}/_{p^2}$

Poisson

Experimental conditions:

This occurs when there are events which do not occur as outcomes of a definite number of trials of an experiment;

But …

..... which occur at random *points of time and space* wherein our interest lies only in the number of occurrences of the event in that time interval.

Definition. A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}; x = 0,1,2, \ldots; \lambda > 0$$

- The expected value $= Variance = \lambda$

The notation $X \sim P(\lambda)$ denotes that X is a Poisson variate

Multinomial distribution

- This distribution can be regarded as a generalization of Binomial distribution.
- This distribution exists when there are more than two mutually exclusive outcomes in a trial.

Suppose $E_1, E_2, E_3, \ldots, E_k$ are k mutually exclusive and exhaustive outcomes of a trial with probabilities $p_1, p_2, p_3, \ldots, p_k$.

The probability that E_1 occurs x_1 times, E_2 occurs x_2 times ... E_k occurs x_k times in n independent trials;

Is given by

$$p(x_1, x_2, \ldots, x_k) = c_k \cdot p_1^{x_1} \cdot p_2^{x_2} \ldots p_k^{x_k}$$

Where $\sum x_i = n$, $c_k = \dfrac{n!}{x_1!\,x_2!\ldots x_k!}$ and $0 \le x_i \le n$

Expected values, $E\ (X_i) = n\, p_i$

Variances, $Var(X_i) = n\, p_i q_i$

$$i = 1,2,3, \ldots .k$$

Hypergeometric distribution

This distribution exists when the population is finite and the sampling is done without replacement, so that the events although being random are stochastically dependent.

Note: The population size here is N, of which is divided into two clusters of sizes M and $N - M$ respectively.

Then a random sample of size n is drawn from this population (without replacement).

The main interest here is the probability of getting k items out of this sample.

Definition: A discrete random variable X is said to follow the hypergeometric distribution if it assumes only non-negative values and its probability mass function is given by

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},$$

$$k = 0, 1, \ldots \min(n, M)$$

The expected value, $E(X) = {}^{nM}/_{N}$

Variance, $Var\ (X) = \dfrac{NM(N-M)(N-n).}{N^2\ (N-1}$

Definition

A random variable X is said to be continuous if it can take all possible values between certain limits.

In other words, a random variable is said to be continuous when its different values cannot be put in 1-1 correspondence with a set of positive integers.

A continuous random variable; is a random variable that (at least conceptually) can be measured to any desired degree of accuracy.

Examples of continuous random variables are age, height, weight etc

Probability Density Function

Concept and Definition

Consider the small interval $(x, x + dx)$ of length dx round the point x.

Let $f(x)$ be any continuous function of x so that $f(x)\,dx$ represents the probability that X falls in the infinitesimal interval $(x, x + dx)$.

Symbolically,

$$P\,(x \le X \le x + dx) = f_X(x)\,dx$$

The probability density function p.d.f. $f_X(x)$ of the r.v. X is defined as:

$$f_X(x) = \lim_{\partial x \to 0} \frac{P\,(x \le X \le x + \partial x)}{\partial x}$$

The probability for a variate value to lie in the interval dx is $f(x)\,dx$,

Hence the probability for a variate value to fall in the finite interval $[\alpha, \beta]$ is

$$P(\alpha \le X \le \beta) = \int_{\alpha}^{\beta} f(x)\,dx$$

a) $f(x) \geq 0, -\infty < x < \infty$

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

c) The probability P (E) given by

$$\int_{E} f(x)\, dx$$

is well defined for any event E.

Continuous Distribution Function

If X is a continuous random variable with the p.d.f. $f(x)$, then the function

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt\,, -\infty < x < \infty$$

is called the distribution function (dF.) or sometimes the cumulative distribution function (c.d.f.) of the random variable X.

note:

a) $0 \leq F_X(x) \leq 1, -\infty < x < \infty$

b) $\frac{d}{dx} F_X(x) = f_X(x)$ which is the p.d.f.

Normal Distribution

Historical note:

The normal distribution was first discovered in 1733 by English mathematician De-Movie. who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problems arising in the game of chance.

It was also known to Laplace. no later than 1774 but through a historical error it was credited to Gauss who first made reference to it in the beginning of 19th century (1809) as the distribution of errors in Astronomy.

Gauss used the normal curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies.

Definition

A random variable X is said to have a normal distribution with parameters μ (called "mean") and σ^2 (called "variance")....

.....if its density function is given by the probability law;

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

$$-\infty < x < \infty; \ -\infty < \mu < \infty; \ \sigma > 0$$

A random variable X following a normal with mean μ and variance σ^2, is expressed as $X \sim N\ (\mu, \sigma^2)$.

If $X \sim N\ (\mu, \sigma^2)$ then,

A random variable X following a normal with mean μ and variance σ^2, is expressed as $X \sim N(\mu, \sigma^2)$.

If $X \sim N(\mu, \sigma^2)$ then,

$$Z = \frac{x - \mu}{\sigma} \sim N(0, 1)$$

...Z is termed as standard normal variate and its p.d.f. is defined as

$$\varphi(Z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$$-\infty < z < \infty$$

Chief Characteristics

Of normal distribution:

a) Its curve is bell shaped and symmetrical about the line $x = \mu$.

b) Mean, median and mode of the distribution coincide.

c) As x increases numerically, f(x) decreases rapidly, the maximum probability occurring at the point $x = \mu$

d) $f(x)$ being the probability, can never be negative no portion of the curve lies below the $x - axis$

e) Linear combination of independent normal variates is also a normal variate.

Usually the probability of the normal question is obtained from statistical tables which are areas under the curve.

Since it is hard to integrate

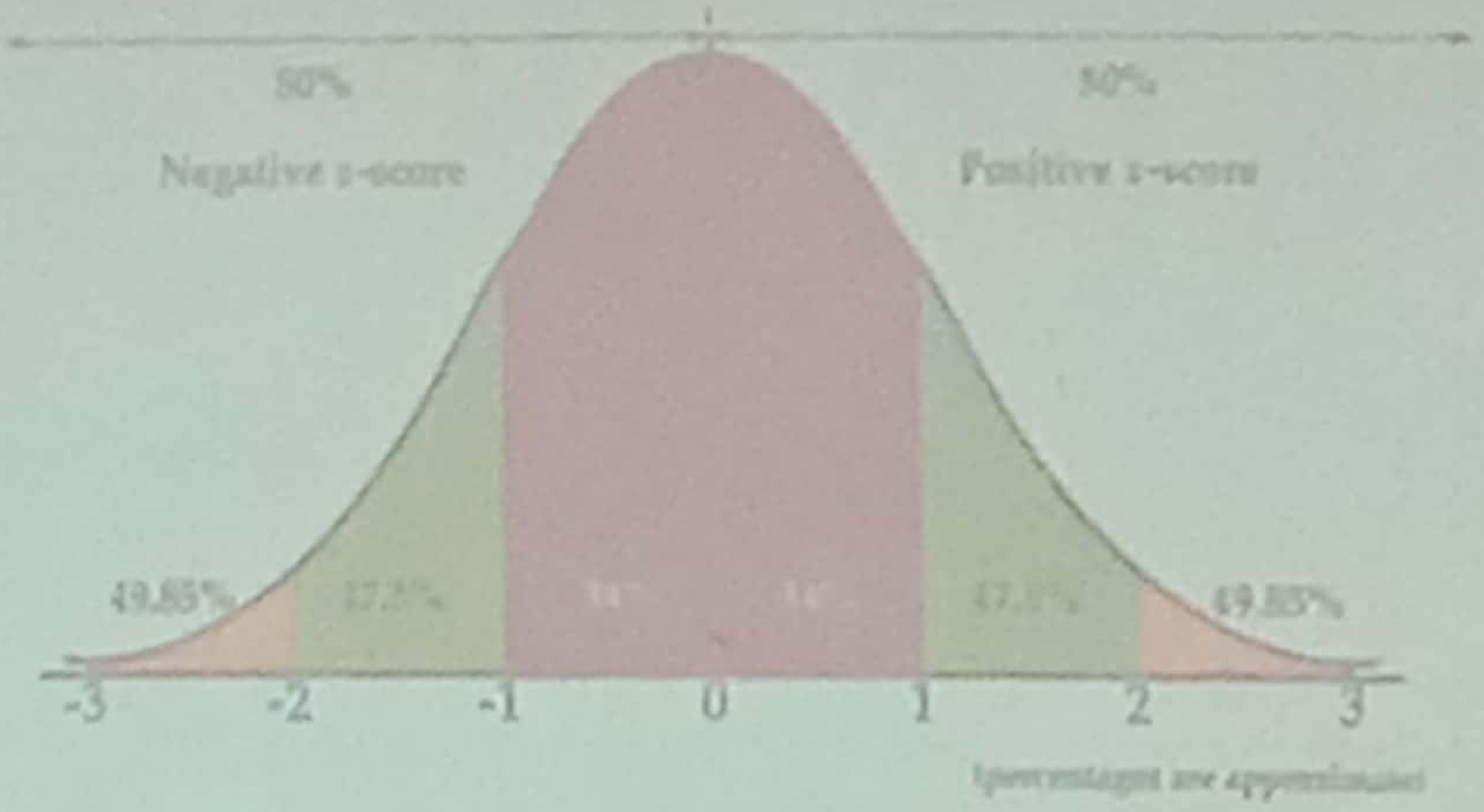
$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Usually the probability of the normal question is obtained from statistical tables which are areas under the curve.

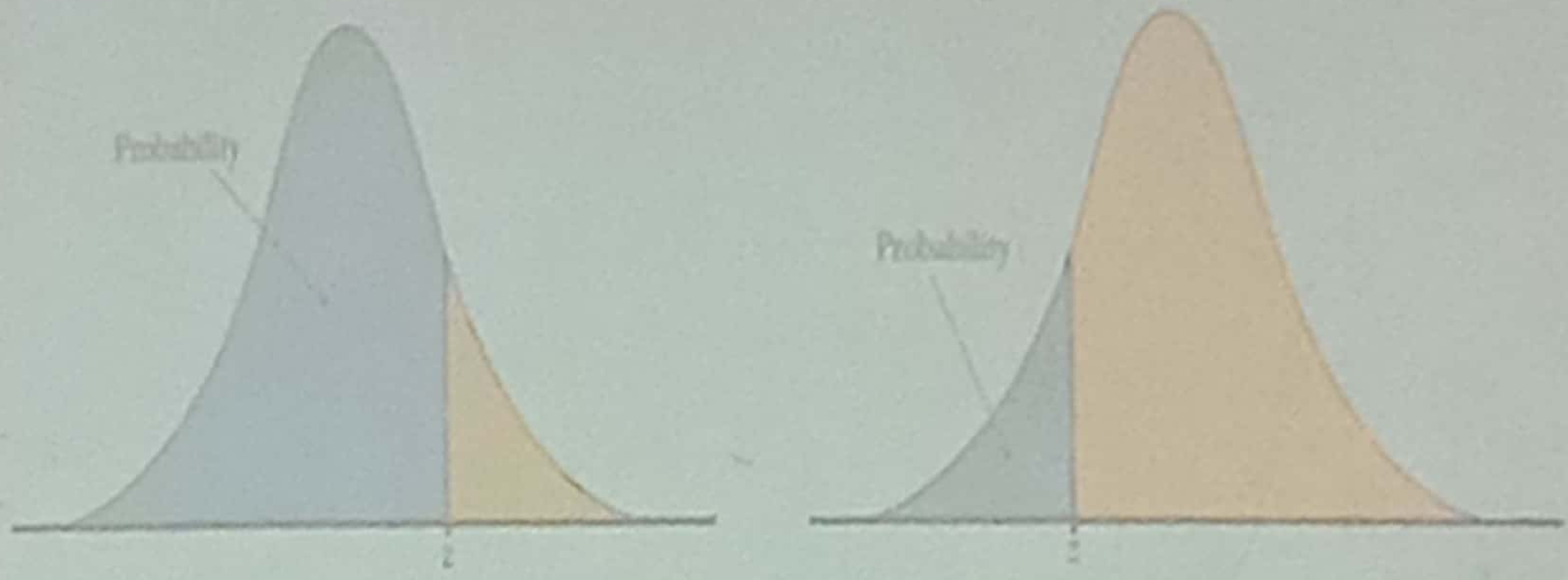
Since it is hard to integrate

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

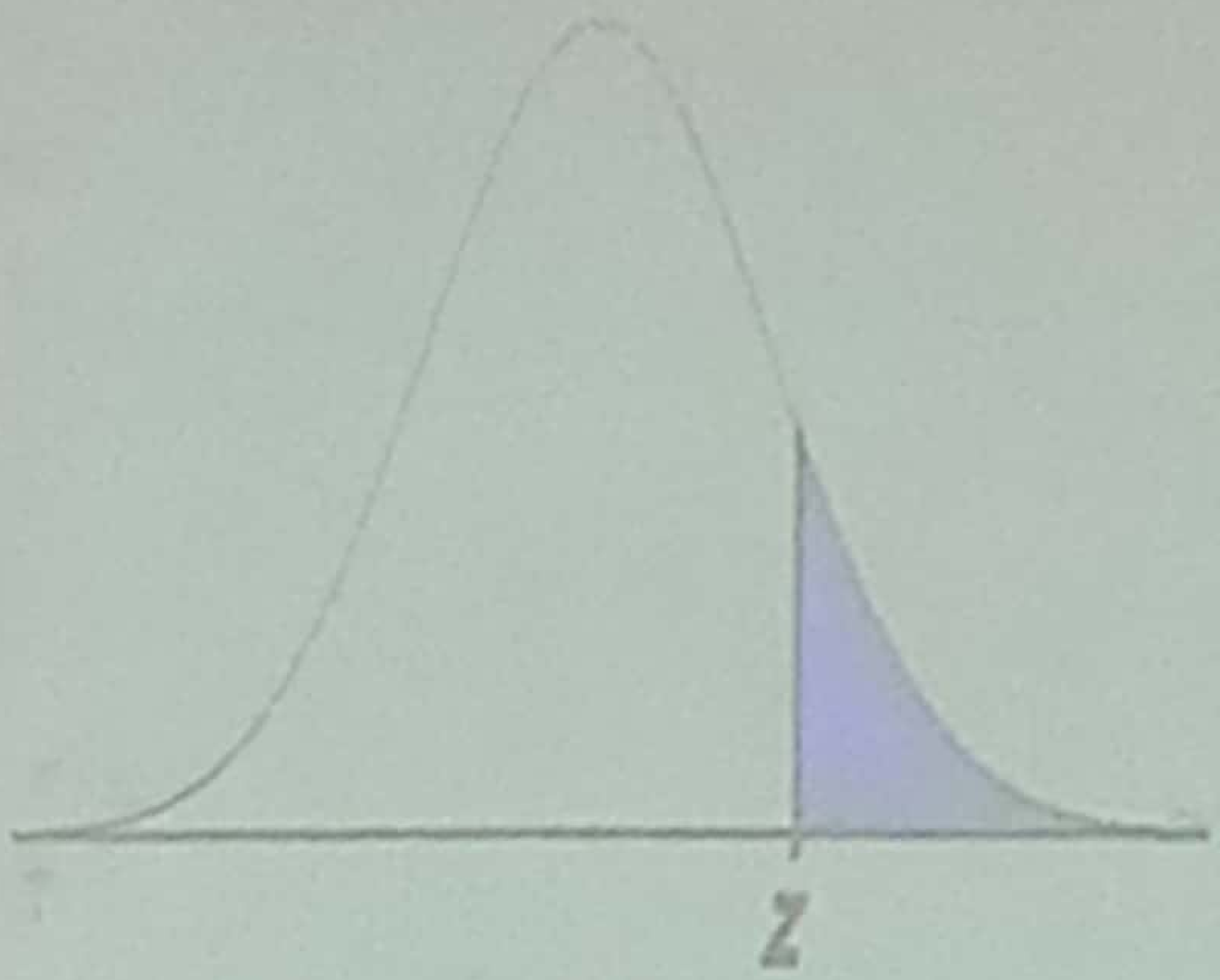
Normal curve

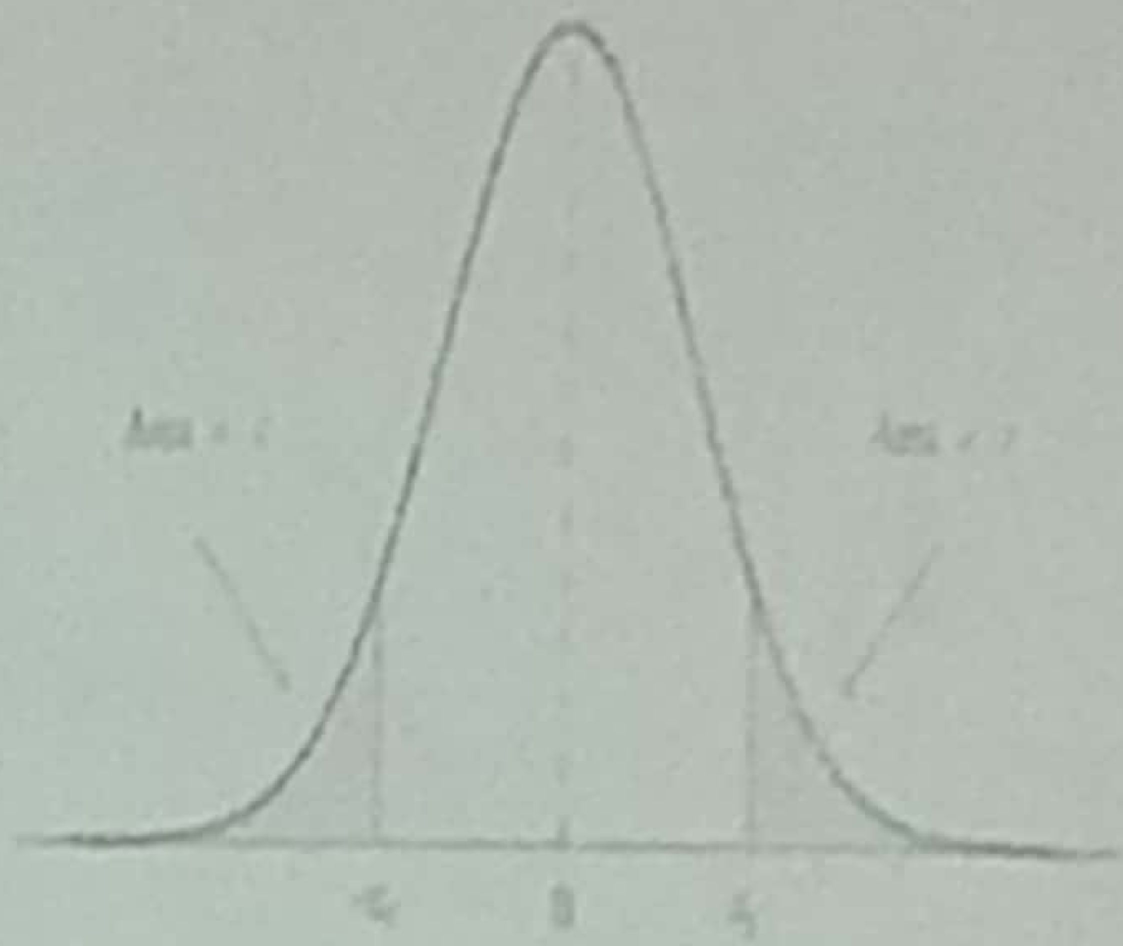


Possible curves



Scanned with CamScanner


Z


0

Importance of Normal distribution

Normal distribution plays a very important role in statistical theory because of the following reasons

- Most of the distributions occurring in practice, e.g., binomial, Poisson, Hypergeometric etc., can be approximated by normal distribution. Moreover, many of the sampling distributions. e.g., Student's 't', Snedecor's F, Chi-square distributions etc., tend to normality for large samples.

- Even if a variable is not normally distributed, it can sometimes be brought to normal form by simple transformation of that variable.

- Many of the distributions of sample statistic (e.g., the distributions of sample mean, sample variance, etc.) tend to normality for large samples and as such they can best be studied with the help of the normal curves.

- The entire theory of small sample tests, viz., t, F, Chi-square tests-etc. is based on the fundamental assumption that the parent population from which the samples have been drawn follows normal distribution

- Theory of normal curves can be applied to the graduation of the curves which are not normal.

Examples

1. If the skulls are classified as A, B and C according as the length-breadth index is under 75, between 75 and 80, or over 80, find approximately (assuming that the distribution is 110rmal) the mean and standard deviation of a series in which A are 58%, B are 38% and C are 4%, being given that if

2. The number of units, x, needed of an item is discrete from 1 to 5. The probability, $p(x)$, is directly proportional to the number of units needed. The constant of proportionality is K.

i. Determine the pdf and CDF of x, and graph the resulting functions.

ii. Find the probability that x is an even value.

iii. If the numbers are grouped into even and odd. What are likely distributions involved in selection of odd? How?

25. In an examination it is laid down that a student passes if he secures 30 per cent or more marks. He is placed in the first, second or third division according as he secures 60% or more marks, between 45% and 60% marks and marks between 30% and 45% respectively. He gets distinction in case he secures 80% or more marks. It is noticed from the result that 10% of the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of students placed in the first, second and third divisions. (Assume normal distribution of marks.)