Probability Distributions

By: Alemakef Wagnew (Bsc, MPH)

Email: alemkelem3@gmail.com

University of Gondar, April 2019

Probability Distributions

- ***Random variable***: is a variable, which can take more than one value with given probability.

- A random variable is said to be discrete if there are always gaps between possible values of the random variable (often the random variable take only integer values).

- A random variable is continuous if it can take any value between any two of its possible values.

- ***Probability distribution*** of a random variable is a table, graph, or formula that gives the probabilities with which the random variable takes different values or ranges of values.

Discrete random variable:

- Are variables which can assume only a specific number of values. They have values that can be counted.

Examples:

- Toss a coin n time and count the number of heads.
- Number of industries in a country..
- Number of car accidents per week.
- Number of defective items in a given company.
- People coming to a theater on Monday.
- Number of bacteria per two cubic centimeter of water.

Continuous random variable

are variables that can assume all values between any two given values. It is a random variable whose values are not countable.

Examples:

- Height of students at certain college.
- Length of garment
- Life time of light bulbs.
- The price of a house
- Length of time required to complete a given training.

Discrete probability distribution

is a formula, a table, a graph or other devices used to specify all possible values of the discrete random variable (R.V) X along with their respective probabilities.

Probability function : A function that for each possible value of a discreet random variable takes on the probability of that value occurring.

Example: Consider the experiment of tossing a coin three times. Let X is the number of heads. Construct the probability distribution of X.

X=x	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

Properties of discrete probability distribution

1. $$\sum_{i=1}^{n} P(X = x_i) = 1$$

2. $$P(X = x_i) \geq 0 \text{ or } 0 \leq P(X = x_i) \leq 1$$

Properties of continuous probability distribution

1. The total area under the curve is one i.e. $\int_{-\infty}^{\infty} f(x) = 1$
2. $P(a \le X \le b) =$ the area under the curve between the point a and b.
3. $P(X) \ge 0$
4. $P(X = a) = 0$
5. $P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$

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

Introduction to expectation

Definition: the expected value (also known as the mean) of a random variable is a measure of the center location for the random variable.

1. Discrete R.V

$$E(X) = X_1P(X_1) + X_2P(X_2) + \dots + X_nP(X_n) = \sum_{i=1}^{n} X_i.P(X_i)$$

2. Continuous R.V

$$E(X) = \int_a^b X \cdot f(x) d(x)$$

Probability distribution

- The expected value of X is its mean

  Mean of X= E(X)

- The variance of X is given by:

  Variance of X=Var(x) = $E\left(X^2\right)-(E(X))^2$

$$E(X^2)=\sum_{i=1}^{n} X_i{}^2.P\left(X_i\right) \quad if\ X\ is\ discrete$$

$$=\int_{x} X^2 f(x)d(x) \quad if\ X\ is\ continuous$$

Example

Let X be a continuous R.V with distribution

$$f(x)=\begin{cases}\frac{1}{2}x & 0\le x\le 2\\ 0, & otherwise\end{cases}$$

Then find

a) P (1<x<1.5)

b) E(x)

c) Var(x)

d) $E(3x^2-2x)$

Common Types of Probability Distributions (Discrete)

a) Binomial distribution

- Lies on Bernoulli-trials. Conditions are:

a) The experiment should be repeated n number of times, where n>1.

b) Each trial should result in only two possible out comes named as success (p) or failure (1-p) arbitrarily.

c) The probability of success (p) remains constant from trial to trial.

d) All the n trials are independent.

Examples of binomial experiments

- Tossing a coin 20 times to see how many tails occur.
- Asking 200 people if they watch BBC news.
- Registering a newly produced product as defective or non defective.
- Asking 100 people if they favour the ruling party.
- Rolling a die to see if a 5 appears.

Binomial distribution cont'd…

- In general, let the probability success be "p" and the trials be repeated n times.
- Then, the total number of successes (S or X) is a random variable and is said to have a binomial distribution, **bin (n, p)**. We can calculate each of the probabilities for the possible outcomes 0 to n by;

$$P(x\text{-successes}) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x} = {}^{n}C_{x}\, p^{x} (q)^{n-x};$$

where x= 0, 1, 2, …, n.

Binomial distribution cont'd…

The meaning of ${}_nC_x$ is the combination of n objects taken x at a time in an unordered subset of x of the n objects and it is calculated by

$$ {}_nC_x = \frac{n!}{x!(n-x)!} \quad \text{for } x=0,1,2,\ldots,n. $$

where n! is read as n factorial and $n! = n(n-1)(n-2)\ldots 2 \cdot 1$. $x! = x(x-1)\ldots(2)(1)$. $0! = 1$.

Here, n and p are the binomial parameters that specify the binomial distribution The populations mean of a binomial distribution is $\mu = np$ and its variance, $\sigma^2 = np(1-p) = npq$ and its standard deviation is $\sqrt{np(p-1)}$.

Binomial distribution cont'd…

Example 1: Suppose that in a certain population 52% of the all recorded births are males. If we randomly select five birth records from this population, what is the probability that exactly three of the records are male births?

Solution: n=5 x=3 (success), *p*=0.52

$$P(X=x) = f(3) = {}_5C_3\, p^3 q^{5-3} = \frac{5!}{3!(5-3)!}(0.52)^3(0.48)^2$$

$$= 10(.52)^3(0.48)^2 = 0.32$$

Exercise

What is the probability of getting three heads by tossing a fair coin four times?

Binomial distribution cont'd…

Example 3: Suppose that it is known that 30% of a certain population is immune to a given disease. If a random sample of size ten is selected from the population, what is the probability that it will contain exactly four immune persons?

Solution: n=10, X= 4, P (immune persons) = p = 0.3 and therefore, (1-p) =q= 0.7

$$P(X=4)=f(4)={}_{10}C_4 p^4 q^{10-4}=\frac{10!}{4!\,(10-4)!}(0.3)^4(0.7)^{10-4}$$

$$=0.2001$$

2. Poisson Distribution

- The Poisson distribution is used to model discrete events that occur infrequently in time and space i.e. rare events that occur in constant rate,
- The Poisson Distribution is a discrete distribution which takes on the values X = 0, 1, 2, 3, and so on.
- It is often used as a model for the number of events in a specific time period.
- It is determined by one parameter, lambda.

The Poisson distribution

- Suppose events happen randomly and independently in space or time at a constant rate in an interval.

- If events happen with rate λ (Greek lambda) events per unit time, the probability of x events happening in unit time is given as:

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

- λ is the parameter of the Poisson distribution and it is the average number of occurrences of the random event in the interval (or volume), and *e* is the constant and its value is 2.7183.

The Poisson distribution cont'd…

- Theoretically, an infinite number of occurrences of the event must be possible in the interval, and the probability of the single occurrence of the vent in a given interval is proportional to the length of the interval.

- Mean (μ) and variance (σ^2) are equal in Poisson distribution and are the same as λ. i.e. $E(x) = \lambda, Var(x) = \lambda$.
- Generally, for Poisson distribution:

$$X \sim \text{Poisson}(\lambda), \quad 0 < \lambda < \infty$$
$$\textbf{Sample space of } X\textbf{: } \{0, 1, \ldots\}$$

The Poisson distribution is used as a distribution of rare events, such as:

- The number of faults in 10 m lengths of cloth
- The number of end-breaks per hour on a spinning frame
- Number of misprints.
- Natural disasters like earth quake.
- Accidents.

Example 1

a factory contains a large number of similar machines which stops randomly at an average rate of 4.2 per day. What is the probability if number of stop pages per day will be 7?

Solution:

$$P(X = 7) = \frac{e^{-4.2}\, 4.2^{7}}{7!} = 0.0686$$

The Poisson distribution cont'd…

Example 2: In a study of suicides, the monthly distribution of adolescent suicides in an area for ten years interval closely followed a Poisson distribution with parameter $\lambda = 2.75$. Find the probability that a randomly selected month will be one in which three adolescent suicides occurred.

$$\textit{Solution: } P(X=3) = = \frac{\lambda^{-2.75} 2.75^{3}}{3!} = \underline{0.221584}$$

The Poisson distribution cont'd…

Example 3: $\lambda = 2$ which is the average number of items per sample and assuming that the number of items follows Poisson distribution, find the probability that the next sample taken will contain:

a. One or fewer items? Sol^{n} $p(X \leq 1) = 0.406$

b. Exactly three items? Sol^{n} $P(X=3) = 0.180$

c. More than five items? Sol^{n} $P(X>5) = 0.017$

Probability Distribution for continuous Variables

- If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.
- A continuous probability distribution differs from a discrete probability distribution in several ways by:
- Under different circumstances, the outcome of a random variable may not be limited to categories or counts.
- Because a continuous random variable X can take on an uncountable infinite number of values, the probability associated with any particular one value is almost equal to zero.
- As a result, a continuous probability distribution cannot be expressed in tabular form.

Probability distribution of continuous variables

- As a continuous variable can take an infinite number of values, it helps to **visualize the probability distribution as a curve and probabilities as 'area under the curve'.**
- The equation used to describe a continuous probability distribution is called a **probability density function** (pdf).
- A probability density function given over a rage $a \leq x \leq b$ satisfies the following:
- $f(x) \geq 0$ for all values of 'x' lying between a and b.

Probability distribution of continuous variables

- The total area covered under the curve of the function lying in the range -∞ and ∞ is equal to 1.
- So the probability of this continuous variable can now be found out by integrating this density function with respect to ‘x’ over the interval [a, b].

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

- Common examples of continuous probability distributions are: uniform distribution, normal distribution, chi squared distribution etc.

Characteristics of a distribution

- Features commonly used to describe a distribution are location, dispersion, modality and skewness.
  - ***Location*** tells us something about the average value of the variable.
  - ***Dispersion*** tells us something about how spread out, the values of the variable are.
  - ***Modality*** refers to the number of peaks in the distribution.
  - ***Skewness*** refers to whether or not the distribution is symmetric
- A distribution is said to be symmetric if it is symmetrically distribute about its mode.

The Normal Distribution

- The normal distribution is used extensively in the analyses of continuous variables and has an especially important role in statistics.
- It has been found to be a good approximation for many distributions that arise in practice.
- The normal distribution is a unimodal and symmetric.
- The normal distribution is completely described by two parameters, referred as the mean μ (read as 'mu') and standard deviation σ (read 'sigma').
- The mean μ can be any number (negative, positive or zero).
- The standard deviation σ must be a positive number.
- The mean μ defines the location of the distribution and the SD (standard deviation) σ defines the dispersion of the distribution about the mean.

$$f(x) = \frac{1}{\sqrt{2\prod\sigma}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}, \quad for -\infty < x < \infty$$

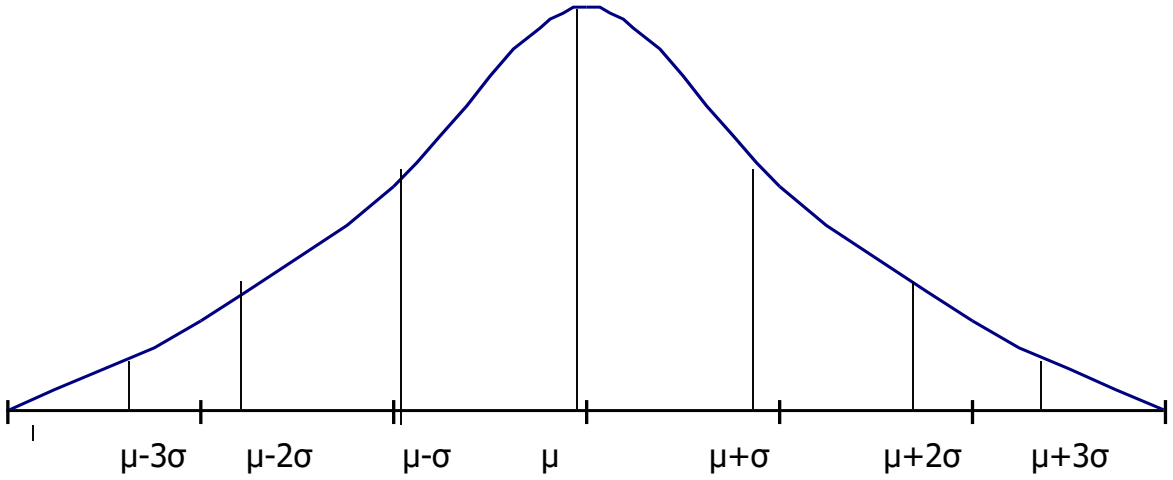


Fig.3. Percentage of area under a normal distribution with mean μ and standard deviation σ

For any normal distribution,

-about 68% (most) of the observations is contained within one SD of the mean.

-about 95% (majority) of the probability is contained within two SDs

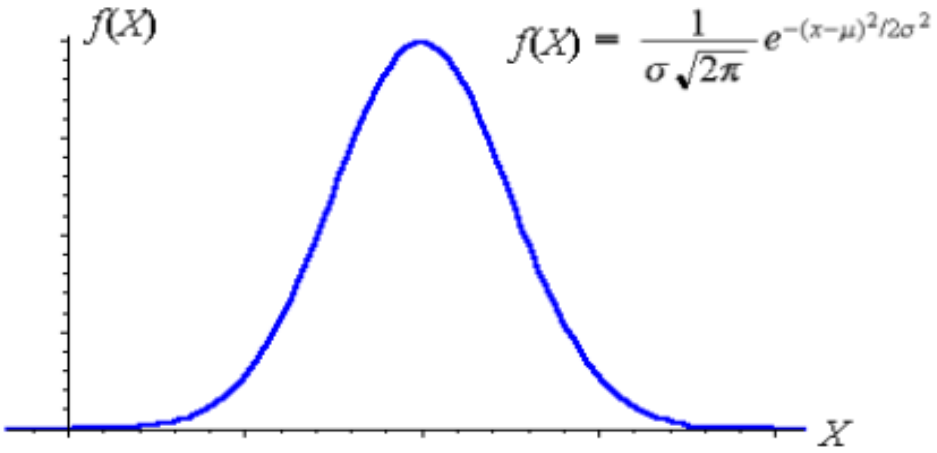
-and 99% (almost all) within three SDs of the mean.

Characteristics of Normal Distribution

- It links frequency distribution to probability distribution
- Has a Bell Shape Curve and is Symmetric
- It is Symmetric around the mean: Two halves of the curve are the same (mirror images)
- Hence Mean = Median = Mode
- The total area under the curve is 1 (or 100%)
- Normal Distribution has the same shape as Standard Normal Distribution.
- The curve never touches the x-axis.
- The distribution is completely determined by the parameters μ and σ^2

Normal Curve

- The graph of the normal distribution depends on two factors:
  - ✓ the mean and the standard deviation.
- The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph.
- When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow.
- All normal distributions look like a symmetric, bell-shaped curve.

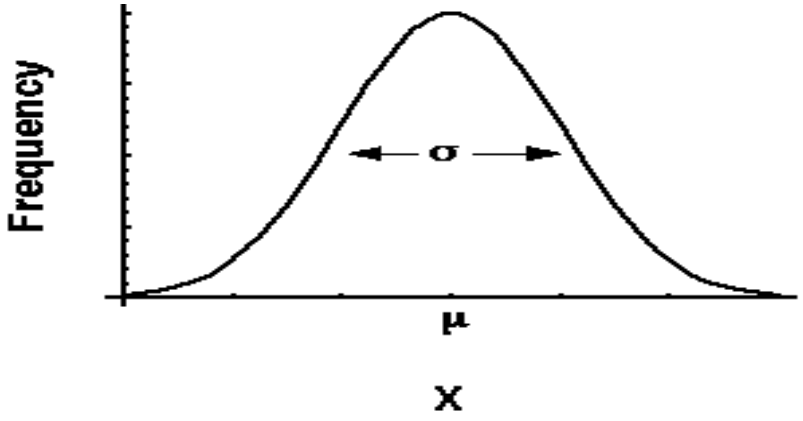


Standard Normal Distribution

- It makes life a lot easier for us if we standardize our normal curve, with a mean of zero and a standard deviation of 1 unit.

- We can transform all the observations of any normal random variable X with mean μ and variance σ to a new set of observations of another normal random variable Z with mean 0 and variance 1 using the following transformation:

$$Z = \frac{X - \mu}{\sigma}$$



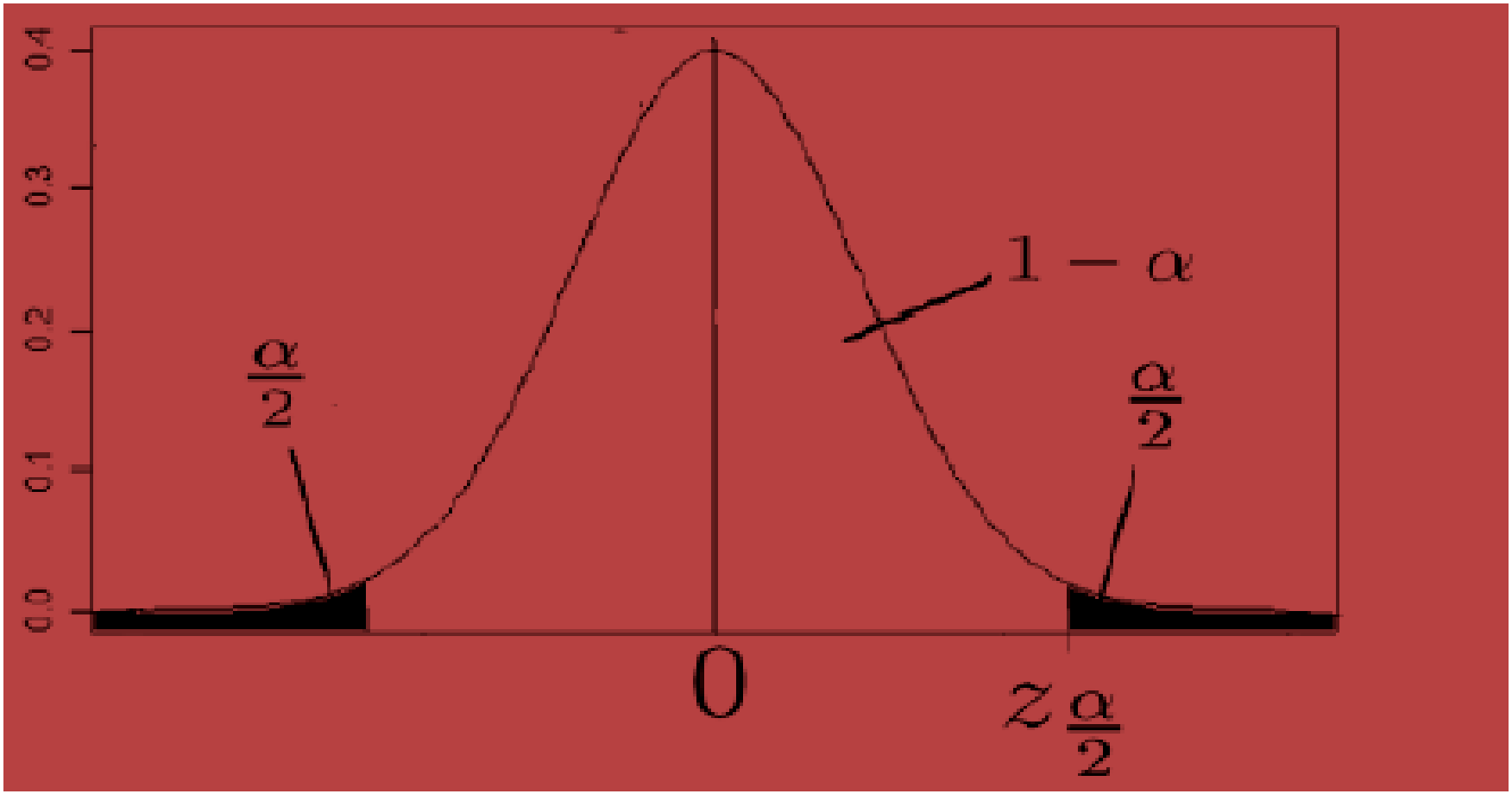
- Areas under the standard normal distribution curve have been tabulated in various ways. The most common ones are the areas between Z=0 and a positive value of Z.

- Given a normal distributed random variable X with Mean μ and standard deviation σ

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma} \right)$$

$$\Rightarrow P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma} \right)$$

Standard Normal Curve: Z ~ N (0, 1)



Standard Normal Distribution cont'd…

- To find the probability that z takes on a value between any two points on the z-axis, say z_0 and z_1, we must find the area bounded by the perpendiculars erected at these points, the curve, and the horizontal axis.

- We can find these areas from the standard normal (Z) table that contains the areas under the curve between and the specified values of z (say z_0) shown in the leftmost column of the table.

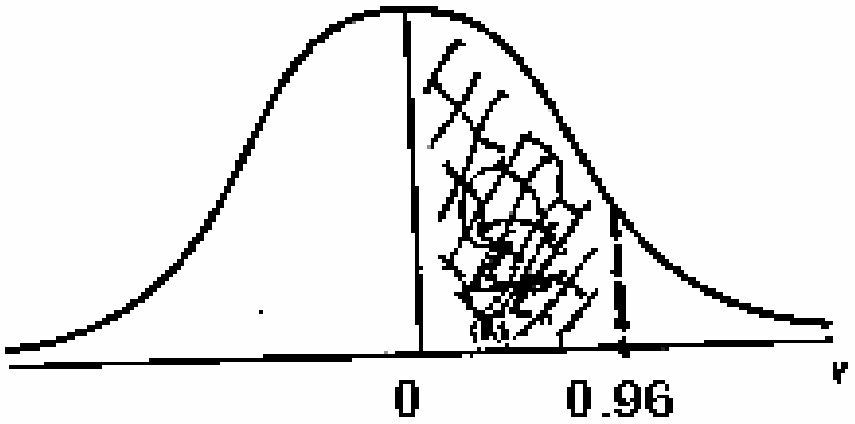
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890

Example 1:

Find the area under the standard normal distribution which lies

a) Between Z=0 and z= 0.96

Solution

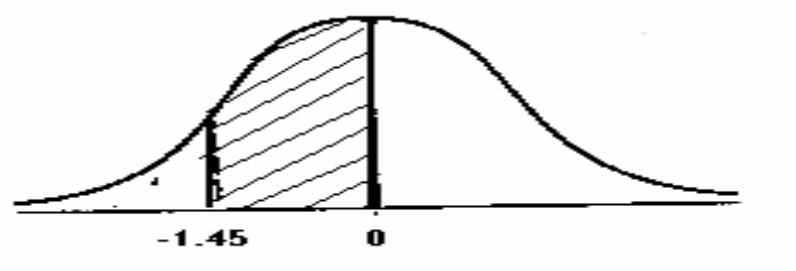


Area = P(0<Z<0.96) = 0.3315

Example 1

b) Z=-1.45 and Z=0

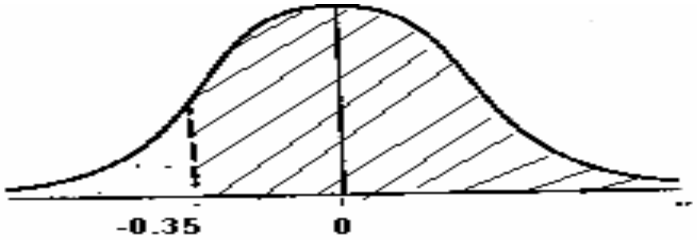
Solution:



Area = P(-1.45 < Z < 0) = P(0 < Z < 1.45) = 0.4265

C)The right of Z=-0.35

Solution:



Area = P(Z > - 0.35) = P(- 0.35 < Z < 0) + P(Z > 0)

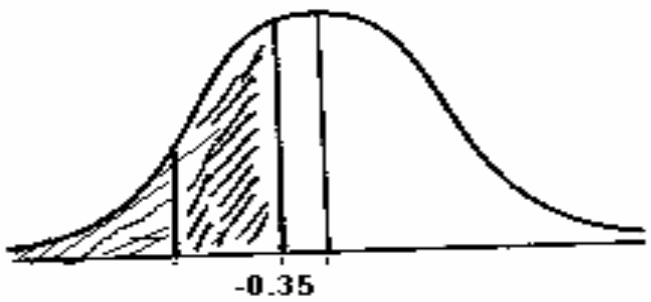
= P(0 < Z < 0.35) +P(Z > 0)

= 0.1368+0.5 = 0.6368

Example 1

d) To the left of Z=0.35

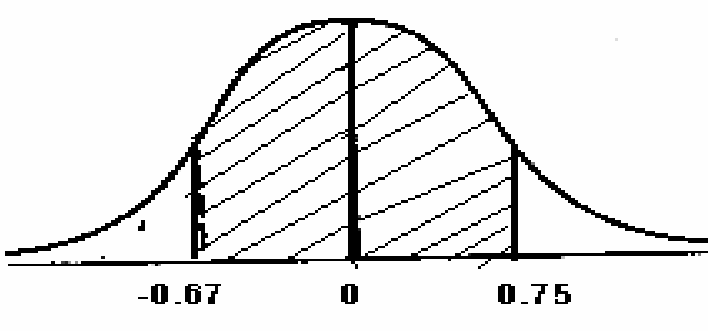
Solution:



Area = P(Z < -0.35) = 1- P(Z > - 0.35) = 1 – 0.6368 = 0.3632

e) Between Z=-0.67 and Z=0.75

Solution:



Area = P(- 0.67 < Z < 0.75) = P(-0.67 < Z < 0) + P(0 < Z < 0.75)

= P(0 < Z < 0.67) + P(0 < Z < 0.75) = 0.2486 + 0.2734 = 0.52

Example 2

Given the standard normal distribution, find P ($z \geq 2.71$)

Solution: *We obtain the area to the right of z=2.71 by subtracting the area between and +2.71 from 1. Thus, P ($z \geq 2.71$) = 1-P ($z \leq 2.71$) = 1-0.9966= 0.0034.*

Example 3: Suppose it is known that the heights of a certain population of individuals are approximately normally distributed with a mean of 70 inches and a standard deviation of 3 inches. What is the probability that a person picked at random from this group will be between 65 and 74 inches tall? I.e. P (65x74)?

Standard Normal Distribution cont'd…

Solution: *First use z- transformation to change this into standard normal distribution:*

_ The corresponding z-value for x=65 is z= 65-70/3 =

-1.67 and for x=74, z=74-70/3=1.33

P req = p(65≤x≤74) = P (-1.67≤z≤1.33)

Then, we find the area between -∞ and -1.67 =0.0475 and the area between - ∞ and 1.33 to be 0.9082.

Then, the area desired is the difference between these, 0.9082-0.0475= 0.8607.

Exercise

- The chest girths of a large sample of men were measured, and standard deviation of the measurements were found to be

Mean = 96m and standard deviation = 8cm

What is the probability that it will take a value?

a) Greater than 104cm

b) Less than 100cm

c) Less than 90cm

d) Between 100cm and 110cm