Module 4: Probability Distribution and Sampling Theory

Sr. No.	Topics	Hours
4.1	Probability Distribution: Poisson and Normal distribution	7 hr
4.2	Sampling distribution, Test of Hypothesis, Level of Significance, Critical region, One-tailed, and two-tailed test, Degree of freedom.	
4.3	Students' t-distribution (Small sample). Test the significance of mean and Difference between the means of two samples. Chi-Square Test: Test of goodness of fit and independence of attributes, Contingency table.	
	Self-learning Topics: Test significance for Large samples, Estimate parameters of a population, Yate's Correction.	

Poisson Distribution

Many experimental situation occur in which we observe the counts of events within a set unit of time, area, volume, length etc. For example,

- The number of cases of a disease in different towns;
- The number of customer visit in medical shop during night;
- Number of new cases of SARS that occur in women in given region in a month;
- The number of births per hour during a given day.

In such situations we are often interested in whether the events occur randomly in time or space, or not.

Assumptions of Poisson Distribution

Assume that an interval is divided into a very large number of subintervals so that the probability of the occurrence of an event in any subinterval is very small.

- The probability of an occurrence of an event is constant for all subintervals: independent events;
- You are counting the number times a particular event occurs in a unit; and
- As the unit gets smaller, the probability that two or more events will occur in that unit approaches zero.

Poisson Probability Distribution

The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

If we let X = The number of events in a given interval.

Then, if the mean number of events per interval is

The probability of observing x events in a given interval is given by

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, 2, 3, 4, \ldots$$

Note:

- P(X=x) is the probability of x successes over a given period of time or space, given λ
- e is a mathematical constant. e 2:718282.
- λ is the parameter of the distribution. We say X follows a Poisson distribution with parameter λ
- The mean and variance of the Poisson probability distribution are same i.e. 'λ'.

Example: The number of births per hour during a given day

If Births in a hospital occur randomly at an average rate of 1.8 births per hour. What is the probability of observing 4 births in a given hour at the hospital?

Solution: Let X = No. of births in a given hour

(i) Events occur randomly

(ii) Mean rate λ = 1:8

Thus, X follows Poisson distribution

We can now use the formula to calculate the probability of observing exactly 4 births in a given hour

$$P(X = 4) = \mathrm{e}^{-1.8}\frac{1.8^4}{4!} = 0.0723$$

Example: The number of births per hour during a given day

What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

We want P(X $\geq$ 2) = P(X = 2) + P(X = 3) + : : :

i.e. an infinite number of probabilities to calculate

but

$$
\begin{aligned}
P(X \geq 2) &= P(X = 2) + P(X = 3) + \ldots \\
&= 1 - P(X < 2) \\
&= 1 - (P(X = 0) + P(X = 1)) \\
&= 1 - \left(e^{-1.8} \frac{1.8^0}{0!} + e^{-1.8} \frac{1.8^1}{1!} \right) \\
&= 1 - (0.16529 + 0.29753) \\
&= 0.537
\end{aligned}
$$

Example: Disease incidence

Suppose there is a disease, whose average incidence is 2 per million people. What is the probability that a city of 1 million people has at least twice the average incidence?

Solution:

Let X=number of cases in 1 million people has Poisson distribution with parameter 2

Twice the average incidence would be 4 cases.

Then, We can now use the formula to calculate the probability of observing greater then 4 cases

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \left(e^{-2}\frac{2^0}{0!} + e^{-2}\frac{2^1}{1!} + e^{-2}\frac{2^2}{2!} + e^{-3}\frac{2^3}{3!} \right)$$
$$= 0.143.$$

Example: Changing the size of the interval

Suppose we know that births in a hospital occur randomly at an average rate of 1.8 births per hour. What is the probability that we observe 5 births in a given 2 hour interval?

Solution: if births occur randomly at a rate of 1.8 births per 1 hour interval

Then births occur randomly at a rate of 3.6 births per 2 hour interval

- Let Y = No. of births in a 2 hour period

$$P(Y = 5) = \mathrm{e}^{-3.6}\frac{3.6^5}{5!} = 0.13768$$

Example: Changing the size of the interval

Suppose if new cases of Covid-19 in India are occurring at a rate of about 2 per month, then what's the probability that exactly 4 cases will occur in the next 3 months?

Solution: if new cases occur randomly at a rate of 2 per month. Then new cases occur randomly at a rate of 6 in the next 3 month.

- Let X = No. of new cases in a 3 month period

$$P(X = 4) = \frac{6^4 e^{-(6)}}{4!} = 13.4\%$$

Example: Fitting of Poisson Distribution

Number Of deaths	Frequencies
0	224
1	102
2	23
3	5
4	1
5+	0

The expected (mean) number of monthly deaths is np, and that can be estimated from the observed mean number of deaths. If we approximate the binomial distribution by Po(λ), where = np , then we don't have to worry about the size of the population.

λ = mean=(102+46+15+4)/(224+102+23+5+1)

λ = 167/355 =0.47

No. of deaths	Frequency	Expected frequency N X P(X=x)	Probability (Poisson = 0:47)
0	224	221.9	0.625
1	102	104.3	0.294
2	23	24.5	0.069
3	5	3.8	0.011
4	1	0.45	0.001
5+	0	0.04	0.0001

Practice Problems

Qu.1 A life insurance company insures the lives of 5,000 men of age 42. If Actuarial studies show the probability of any 42-year-old man dying in a given year to be 0.001, the probability that the company will have to pay 4 claims in a given year can be approximated by the Poisson distribution.

P (X = 4 \ n = 5000, p = 0.001) = 0.1745

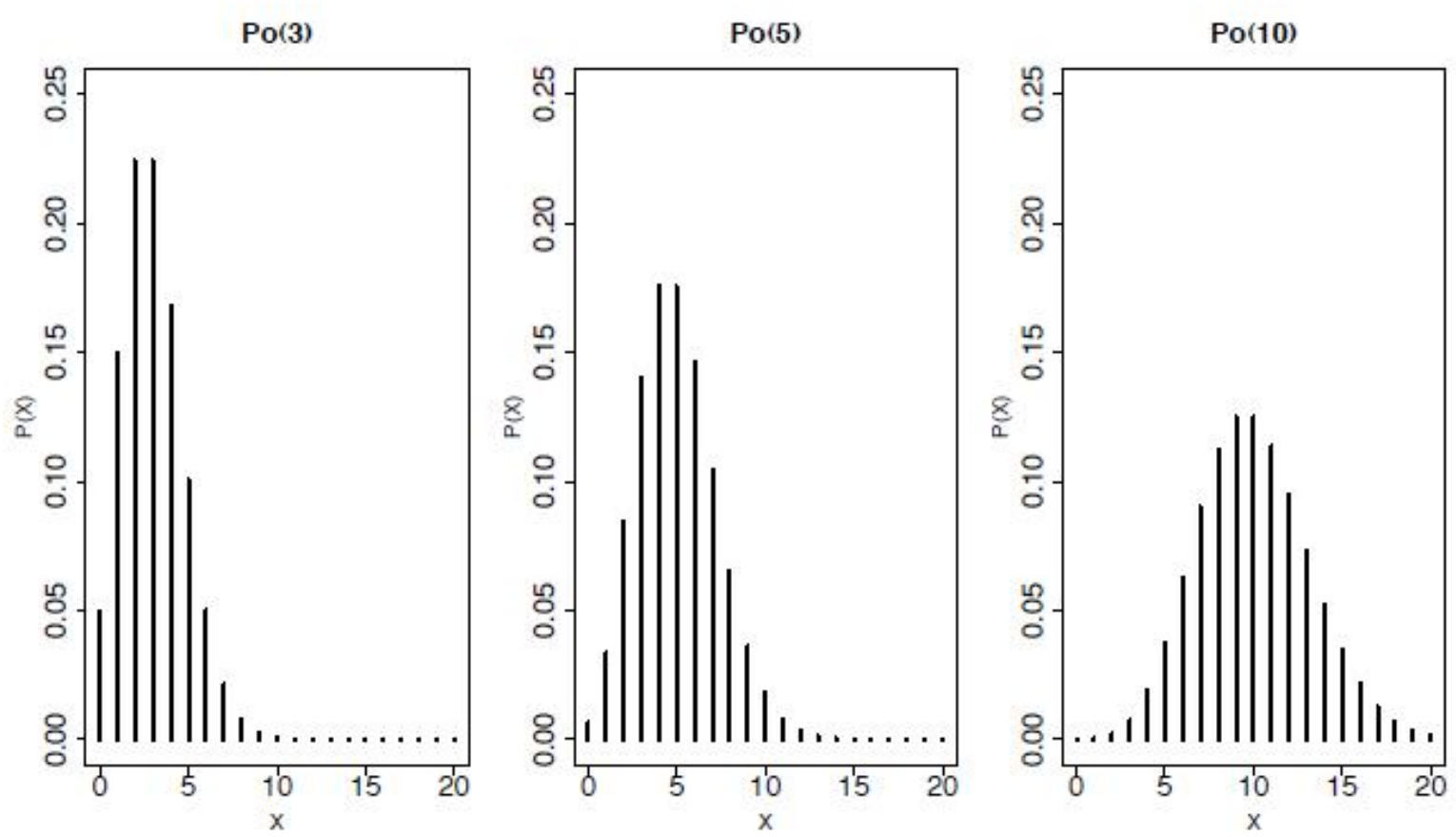
Qu.2 If calls to your cell phone are a Poisson process with a constant rate $\lambda=2$ calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour movie, your phone rings during that time?

$$P(X=0)=\frac{(2*1.5)^0 e^{-2(1.5)}}{0!}\frac{(3)^0 e^{-3}}{0!}=e^{-3}=.05$$

$\therefore$ P(X≥1)=1 – .05 = 95% chance

The shape of the Poisson distribution

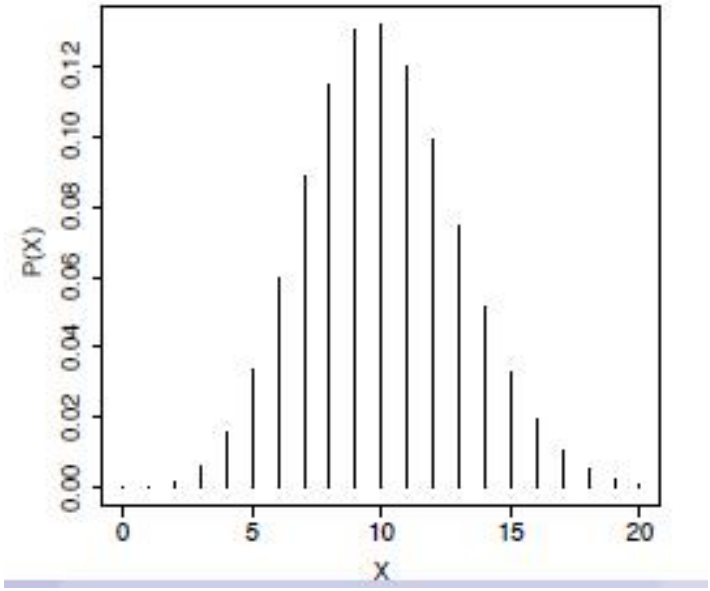
We can plot the theoretical distributions using MatLAB poisspdf(x, λ), e.g. poisspdf(0:20,3) gives



Discrete probability distribution

For the Binomial and Poisson distributions, the probability distributions were characterised by a formula for the probability of each possible discrete value.

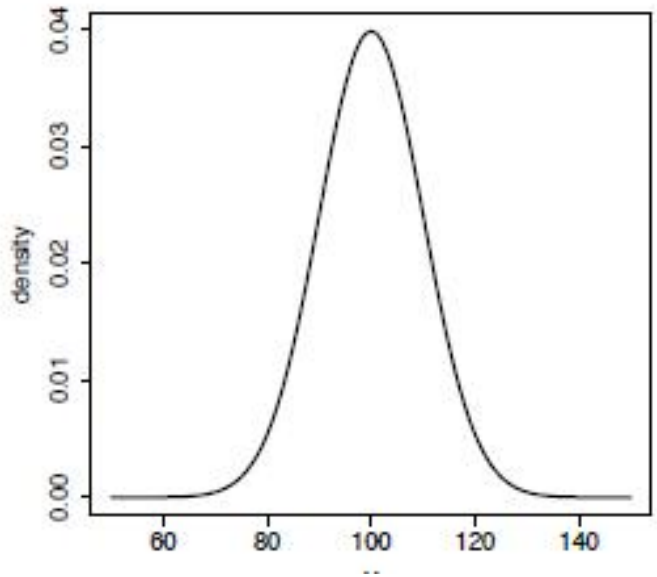
- All of the probabilities together sum up to 1.
- We can visualize the density by plotting the probabilities against the discrete values



Continuious probability distribution

For continuous data we don't have equally spaced discrete values. Instead we use a curve or function that describes the probability density over the range of the distribution.

- The curve is chosen so that the area under the curve is equal to 1.
- If we observe a sample of data from such a distribution we should see that the values occur in regions where the density is highest.



Normal Distribution

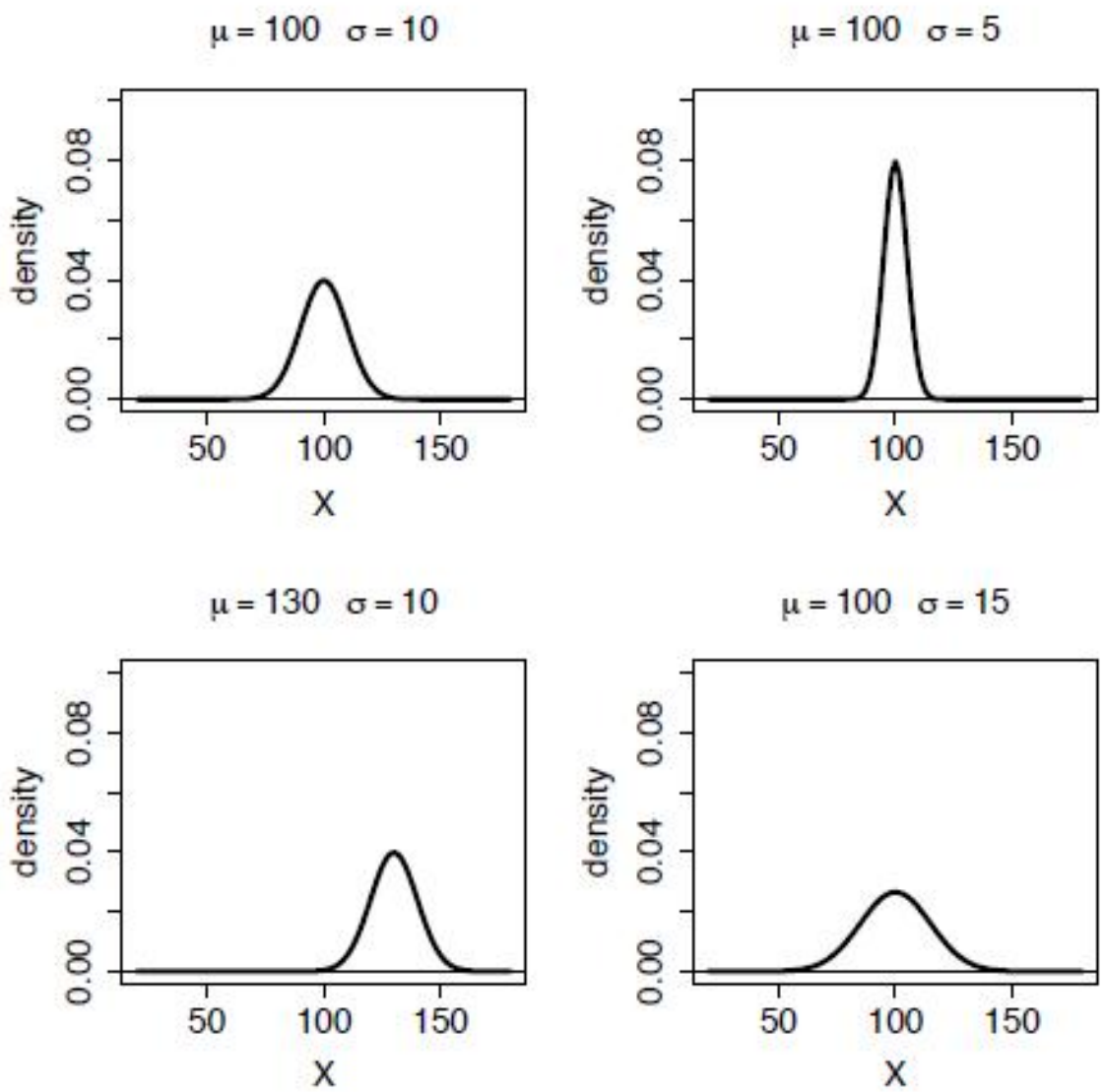
There are many, many possible probability density functions over a continuous range of values.

- The Normal distribution describes a special class of such distributions that are symmetric and can be described by two parameters

- μ = the mean of the distribution
- σ = the standard deviation of the distribution

- Changing the values of and alter the positions and shapes of the distributions.

Normal Distribution



Application of Normal Distribution

- This is the most important probability distribution in statistics and important tool in analysis of epidemiological data and management science.
- It's application goes beyond describing distributions
- It is used by researchers and modelers.
- The major use of normal distribution is the role it plays in statistical inference.
- The z score along with the t –score, chi-square and F-statistics is important in hypothesis testing.
- It helps managers/management make decisions.

Normal Distribution

If X is Normally distributed with mean and standard deviation , we write

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

μ And σ are the parameters of the distribution.

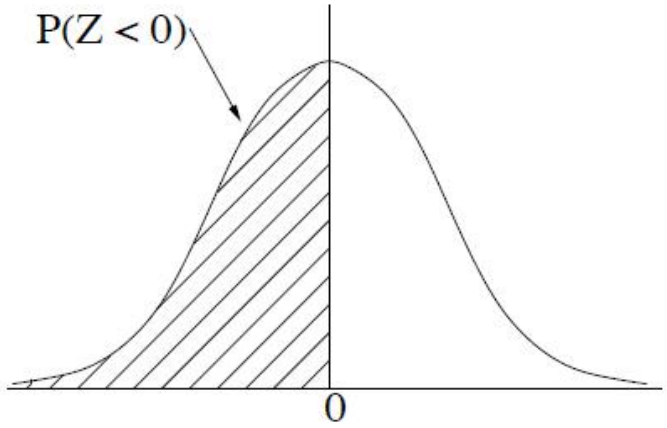
The probability density of the Normal distribution is given by

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

$f(X)$	=	frequency of random variable X
pi	=	3.14159; e = 2.71828
σ	=	population standard deviation
X	=	value of random variable (-ã< X < ã)
μ	=	population mean

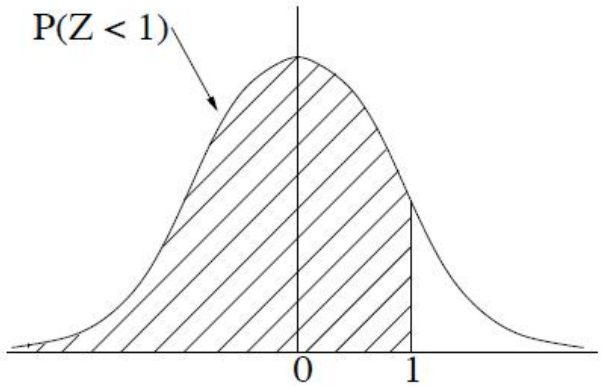
Calculating probability for continuous distributions

- For a discrete probability distribution we calculate the probability of being less than some value z, i.e. P(Z < z), by simply summing up the probabilities of the values less than z.
- For a continuous probability distribution we calculate the probability of being less than some value z, i.e. P(Z < z), by calculating the area under the curve to the left of z.
- For example, suppose Z is N(0, 1) and we want to calculate P(Z < 0) ?



- For this example we can calculate the required area as we know the distribution is symmetric and the total area under the curve is equal to 1, i.e. P(Z < 0) = 0:5.

What about P(Z < 1)?

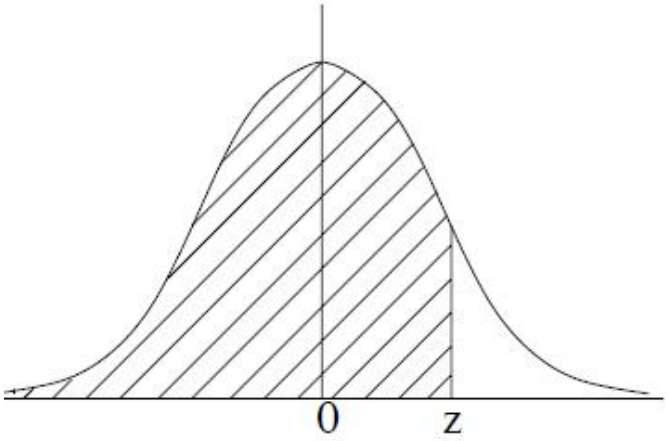


Calculating this area is not easy and so we use probability tables.

Obviously **it is impossible to tabulate** all possible probabilities for all possible Normal distributions so only one special Normal distribution,

N(0, 1), has been tabulated.

The tables allow us to read o probabilities of the form **P(Z < z).**



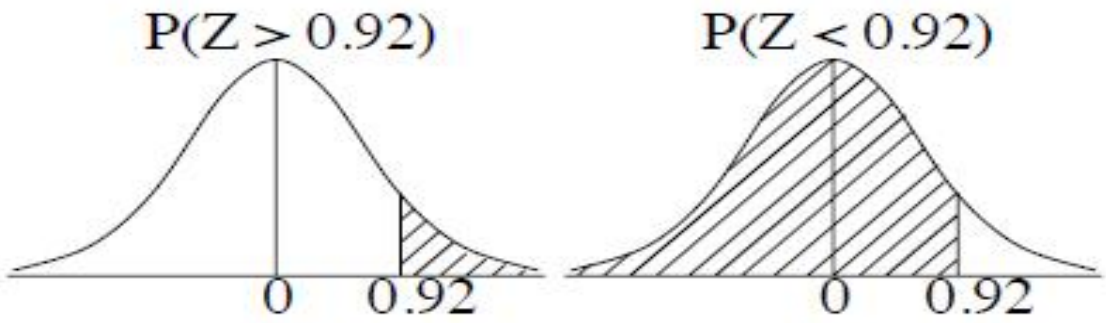
Z-Table

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.1	0.5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.2	0.5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0.3	0.6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.4	0.6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0.5	0.6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0.6	0.7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0.7	0.7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0.8	0.7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0.9	0.8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1.0	0.8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1.1	0.8643	8665	8686	8708	8729	8749	8770	8790	8810	8830

From this table we can identify that $P(Z < 1.0) = 0.8413$

Example-1

Suppose we want $P(Z > 0.92)$.



We know that

$$P(Z > 0.92) = 1 - P(Z < 0.92)$$

and we can calculate $P(Z < 0.92)$ from the tables.

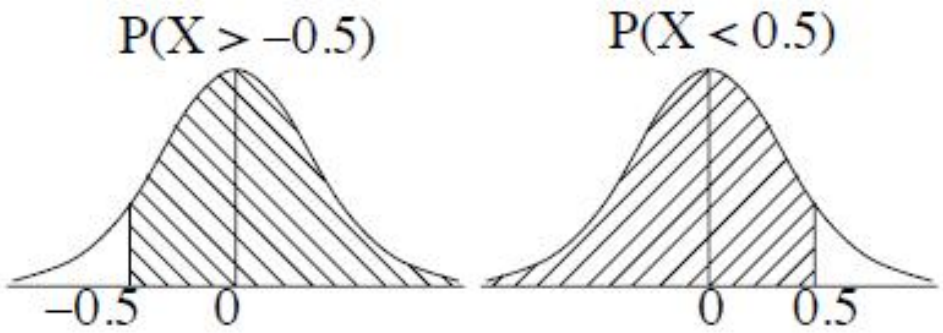
Thus, $P(Z > 0.92) = 1 - 0.8212 = 0.1788$.

Example-2

The table only includes positive values of z.

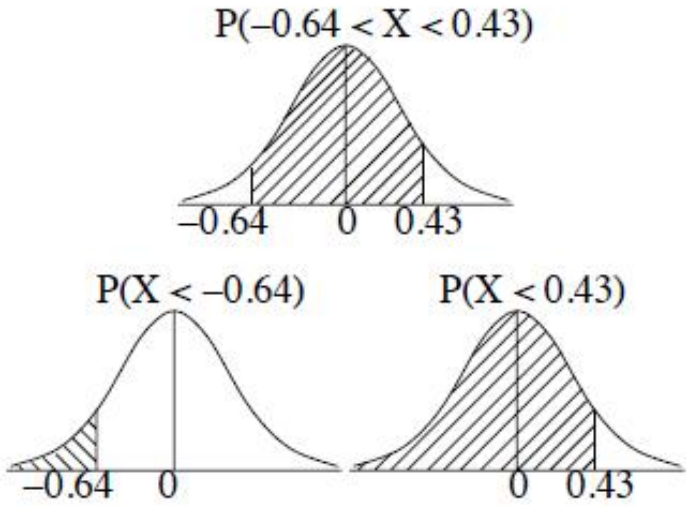
If we want to compute $P(Z > -0.5)$, we use the fact that the Normal distribution is symmetric:

$$P(Z > -0.5) = P(Z < 0.5) = 0.6915$$



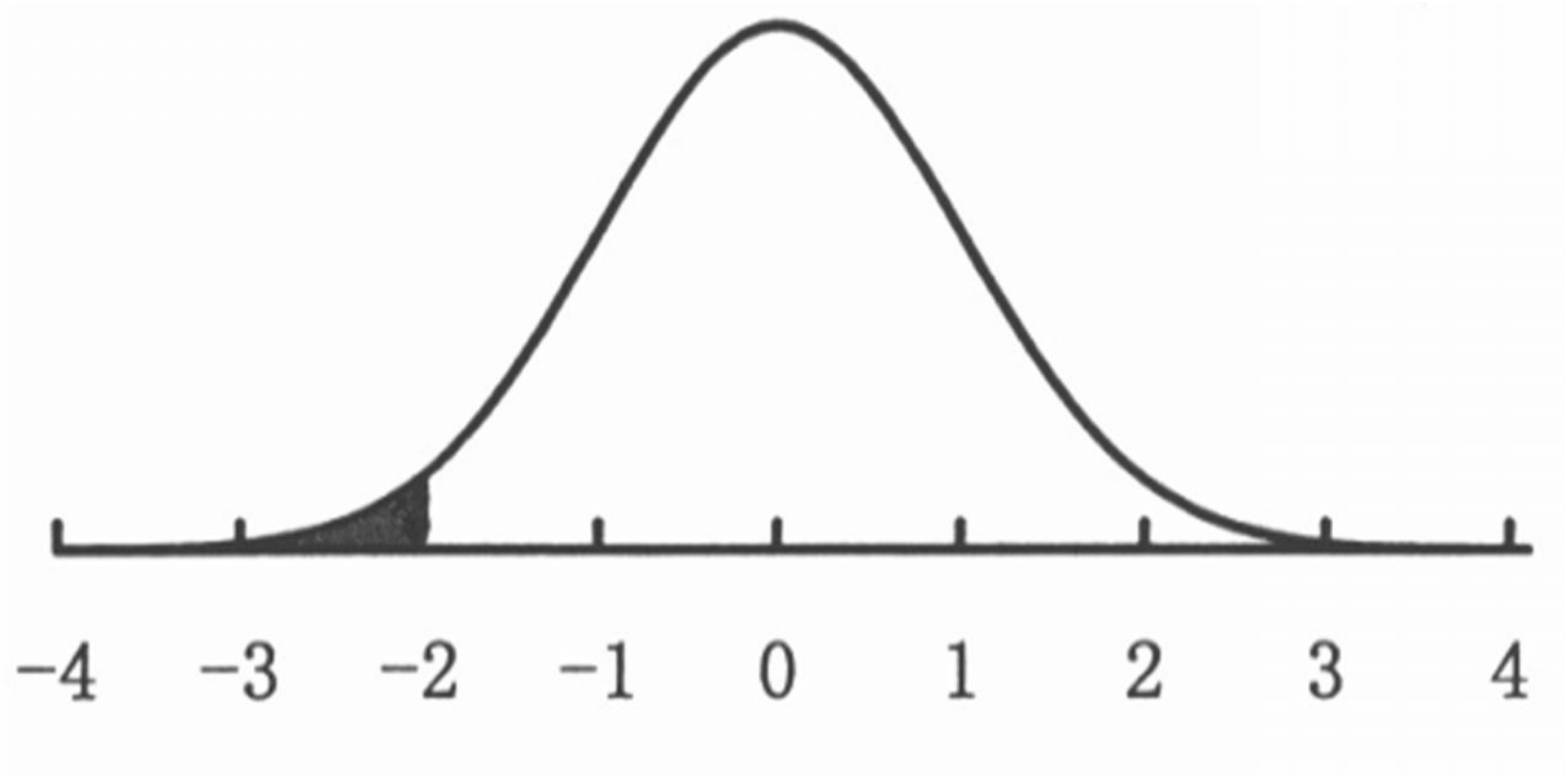
Example-3

How do we compute $P(-0.64 < Z < 0.43)$?



$$
\begin{aligned}
P(-0.64 < Z < 0.43) &= P(Z < 0.43) - P(Z < -0.64) \\
&= P(Z < 0.43) - P(Z > 0.64) \\
&= P(Z < 0.43) - (1 - P(Z < 0.64)) = 0.4053
\end{aligned}
$$

Example-4



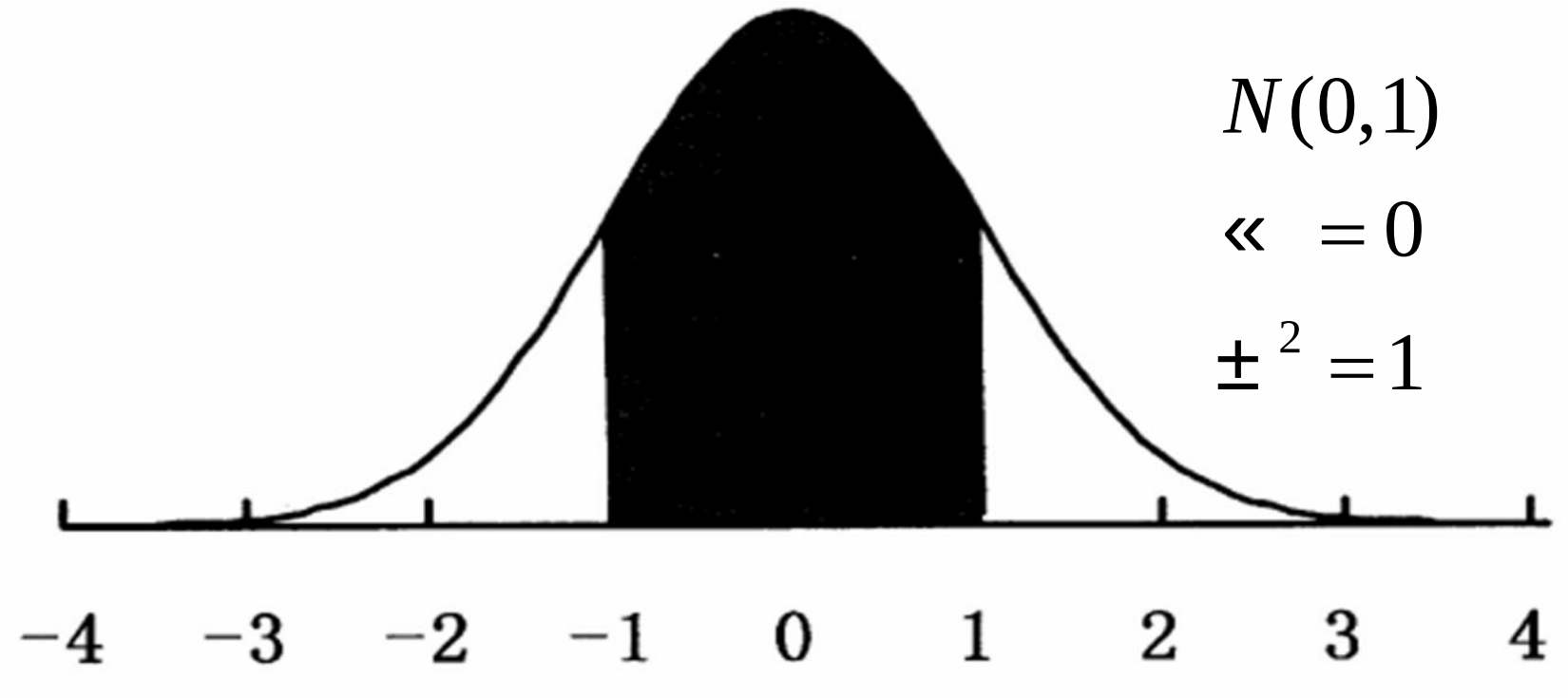
Area Below z = -2;

$P(z < -2) = P(z>2)=1-P(z<2)$

$=1-0.9772$

$=0.0228$

Example-5



(1) Area between -1, to +1;

P(-1 < z < +1) = P(z<1) - P(z<-1)=P(z<1)-P(z>1)=2P(z<1)-1

up to z = +1: .8413x2=1.6826

P(-1<z<1)=0.6826

Standardization

- All of the probabilities above were calculated for the standard Normal distribution N(0, 1).
- If we want to calculate probabilities from dierent Normal distributions we convert the probability to one involving the standard Normal distribution.
- This process is called standardisation.

Suppose $X \sim \mathrm{N}(3, 4)$ and we want to calculate $P(X < 6.2)$.

We convert this probability to one involving the N(0, 1) distribution by

1. Subtracting the mean μ
2. Dividing by the standard deviation σ

$$P(X < 6.2) = P\left(\frac{X-3}{2} < \frac{6.2-3}{2}\right) = P(Z < 1.6) = 0.9452$$

where $Z \sim \mathrm{N}(0,1)$

This process can be described by the following rule

If $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$ then $Z \sim \mathrm{N}(0, 1)$

Calculating z-values

$Z \sim N(0,1)$

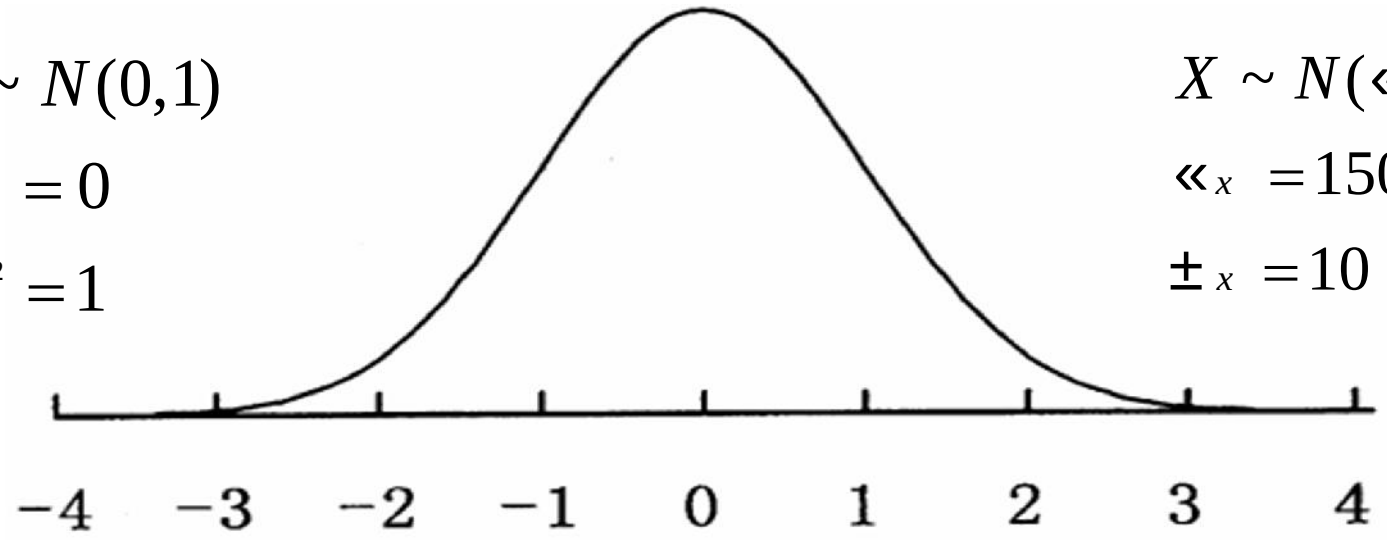
$\mu_z = 0$

$\sigma_z^2 = 1$

$X \sim N(\mu_x, \sigma_x)$

$\mu_x = 150$

$\sigma_x = 10$



$$z = \frac{x - \mu_x}{\sigma_x} \quad ; \quad \text{if } X \sim N(150,10) \text{ } i.e. \text{ } \mu_x = 150, \sigma_x = 10$$

$$\text{when } x = 150; \quad z = \frac{150 - 150}{10} = 0$$

$$\text{when } x = 170; \quad z = \frac{170 - 150}{10} = \frac{20}{10} = 2$$

Problem-01

X is a normally normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find

a) $P(x < 40)$

b) $P(x > 21)$

c) $P(30 < x < 35)$

Solution:

a) For $x = 40$, the z-value $z = (40 - 30) / 4 = 2.5$
Hence $P(x < 40) = P(z < 2.5) = 0.9938$

b) For $x = 21$, $z = (21 - 30) / 4 = -2.25$
Hence $P(x > 21) = P(z > -2.25) = P(z < 2.25) = 0.9878$

c) For $x = 30$, $z = (30 - 30) / 4 = 0$ and for $x = 35$, $z = (35 - 30) / 4 = 1.25$
Hence $P(30 < x < 35) = P(0 < z < 1.25) =$
$= 0.8944 - 0.5 = 0.3944$

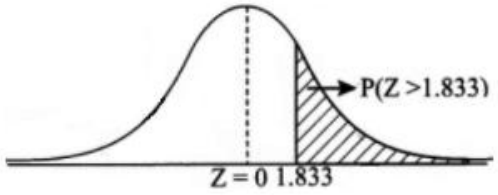
Problem-02

In a test on 2,000 electric bulbs, it was found that bulbs of a particular make, was normally distributed with an average life of 2,040 hours and a standard deviation of 60 hours. Estimate the number of bulbs likely to burn for

(i) more than 2,150 hours (ii) less than 1,950 hours

(iii) more 1,920 hours but less than 2,100 hours.

Solution: Let X be the number of hours for which the bulbs are in use. It is given that X is normally distributed with mean 2040 hours and a standard deviation of 60 hours, (i.e) X ~ N (2040, 602)



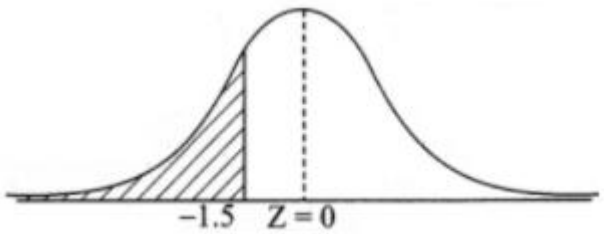
(i) P (X > 2150) We change to the standard normal.

The total area to the right of Z = 0 is 0.5.

The area between Z = 0 and 1.833 is 0.4664 (from tables)

So P (Z > 1.833) = 1- P(Z<1.833)=1-0.9664 = 0.0336 The number of bulbs likely to burn for more than 2150 hours is 2000 × 0.0336 = 67.2 ~ 67.

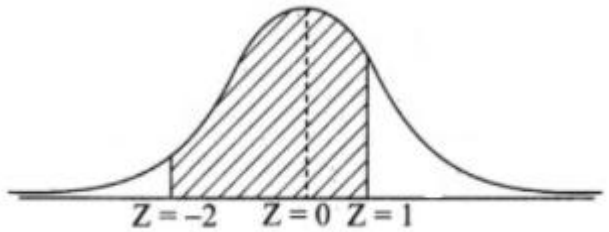
(ii) We want P (X < 1950) = P(X - μ)/σ = P(Z < -1.5)



P(Z < -1.5) = P(Z>1.5)=1- 0.9332 = 0.0668 Hence the number of bulbs likely to bum for less than 1950 hours is 2000 × 0.0668 = 133.6 ~ 134.

(iii) We want P(1920 < X < 2100) When X = 1920, Z = (X - μ)/σ = (1920 - 2040)/60 = -2 When X = 2100, Z = (2100 - 2040)/60 = 60/60 = 1

So P(1920 < X < 2100) = P(-2 < Z < 1) = P(Z<1) – P(Z< -2)=0.8413- P(Z>2)



0.8413- (1-P(Z<2))=0.8413- 0.0228 = 0.8185 Hence the number of bulbs likely to bum for more than 1920 hours but less than 2100 hours is 2000 × 0.8185 = 1637.

Problem-03

A certain number of articles manufactured in a batch were classified into three categories according to some particular characteristic, being less than 50, between 50 and 60 and greater than 60. If this characteristic is known to be normally distributed, determine the mean and standard deviation for this batch if 60%, 35% and 5% were found in these categories.

Problem-04

In a normal distribution 31% of the items are under 45 and 8% of the items are over 64. find the mean and standard deviation of the distribution.