Unit 4:
Probability distributions and sampling theory

- Binomial distribution
- Poisson distribution
- Exponential distribution
- Normal distribution

Optional

- Central limit theorem
- Normal approximation to the binomial distribution

- Sampling distribution of means
- Sampling distribution of differences of means
- Sampling distribution of proportions

Discrete probability distribution

There are some distributions describe several real-life random phenomena.

For instance,

1) In a study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use the drug approximately follows a binomial distribution

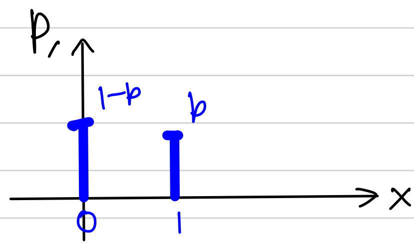
2) The number of white cells from a fixed amount of an individual's blood sample is usually random and may be described by a Poisson distribution

Bernoulli random variable; Parameter, $p \in [0,1]$

Bernoulli r.v, X takes the value 0 or 1.
i.e, $X = \{0,1\}$

Probability distribution (Bernoulli distribution)

$$P(X=1) = p$$
$$P(X=0) = 1-p, \quad \text{where } p \in [0,1].$$



A closed form is $P(X=x) = p^x (1-p)^{1-x}$.

It shows up

- When we consider a trial, or an experiment, whose outcome can be classified as either a success or a failure.

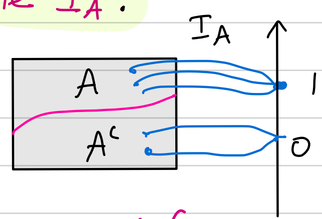
  We let $X=1$ when the outcome is a success and
  $X=0$ when it is a failure.

- When we make connection between events and r.v.

  Let A be an event such that the r.v takes the value 1 whenever the outcome lies in A and it takes 0 whenever the outcome lies in A^c.

  Such a r.v. is called an Indicator variable I_A.

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A^c \text{ occurs} \end{cases};$$



Ex: i) Toss a coin, let X be 1 if it is head (success) and 0 if it is tail (failure)

ii) Roll a dice, let X be 1 if outcome is 6, otherwise 0.

iii) Cricket match, Ind vs Aus, X be 1 if Ind wins and 0 if Ind loses.

Expected value : $E[X] = 0 \times (1-p) + 1 \times p = p$

Variance : $Var(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$

Binomial random variable; Parameters (n, p), where $n \in \mathbb{Z}^+$, $p \in [0,1]$

Suppose an experiment consist of n Bernoulli trials such that

1) The trials are independent.
2) Each trial results in only two possible outcomes, labelled as "success" and "failure"
3) The probability of success in each trial, denoted as p, remains constant.

If r. v. X represents the number of success that occur in the n trials, then X is said to be a binomial r. v. with parameters (n, p), 0< p < 1 and n = 1, 2, 3,....

Note: A Bernoulli r.v is a binomial r.v. with parameters $(1, p)$.

The probability mass function of X (Binomial distribution function) is

$$b(x : n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, 3, \dots$$

Ex: Consider n independent tosses of a coin with $P(H) = p$ and $P(T) = 1 - p$

Sample space is the set of sequences of H and T of length n. Let random variable X be the number of heads observed. Then X is called binomial random variable.

Suppose $n = 3$ the binomial distribution is

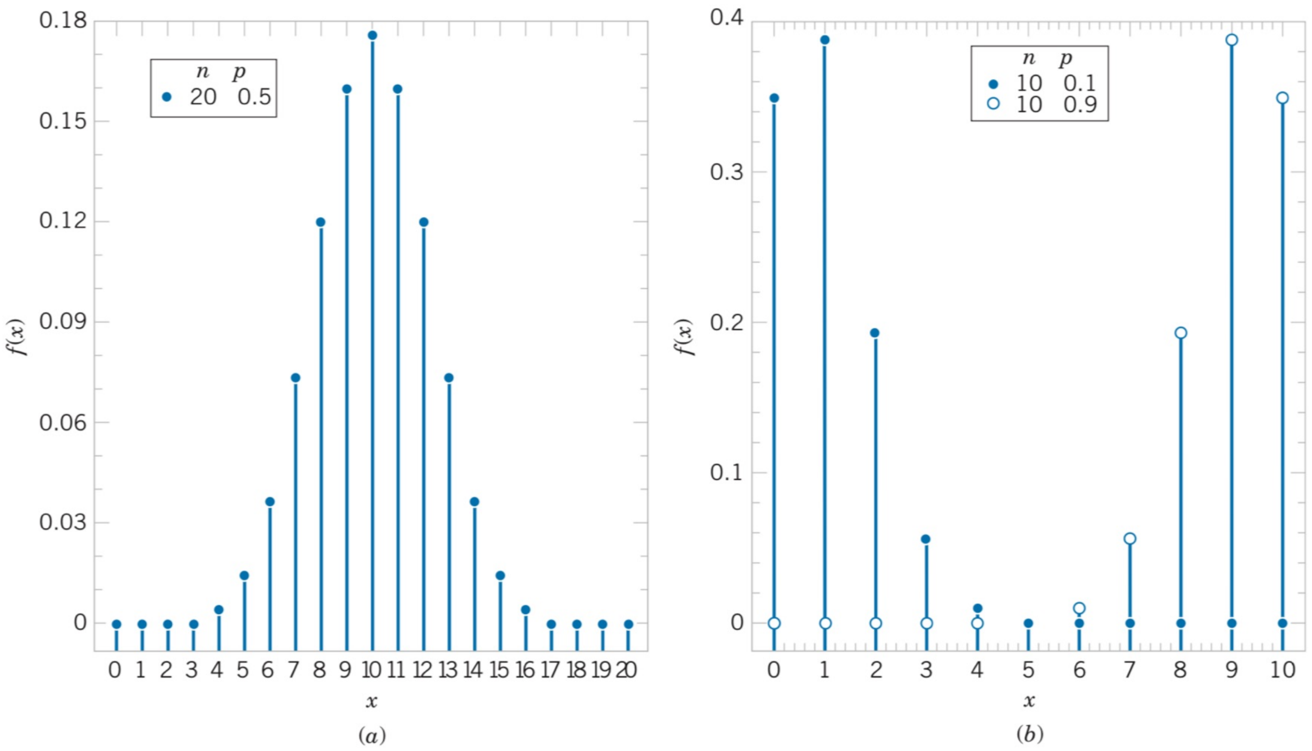
$$P(X=0) = P(TT) = (1-p)^3$$

$$P(X=1) = P(HTT) + P(THH) + P(TTH) = 3p(1-p)^2$$

$$P(X=2) = P(HHT) + P(THH) + P(THH) = 3p^2(1-p)$$

$$P(X=3) = P(HHH) = p^3$$

For any $x \in X$, $P(X = x) = \binom{3}{x} p^x (1-p)^{3-x}$



Binomial distribution for selected values of n and p.

Why the name binomial distribution?

The various values of $b(x; n, p) = \binom{n}{x} p^x (1-p)^x$ are terms in the binomial expansion of $(p+q)^n$ (where $q = 1-p$)

i.e

$$(p+q)^n = \binom{n}{0} q^n + \binom{n}{1} p q^{n-1} + \binom{n}{2} p^2 q^{n-2} + \cdots + \binom{n}{n} p^n$$

$$= b(0; n,p) + b(1; n,p) + b(2; n,p) + \cdots + b(n; n,p)$$

Since $p+q=1$, we see that

$$\sum_{x=0}^{n} b(x; n,p) = 1$$

Expectation and Variance of a binomial r.v.

Define new random variables

$$I_j = \begin{cases} 1 & \text{if } j\text{th trial is a success with prob. } p \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2, 3, \cdots, n.$

Therefore, in a binomial experiment the number of success can be written as the sum of the n independent indicator r.vs

Hence,

$$X = I_1 + I_2 + I_3 + \cdots + I_n$$

It is easy that

$$E[I_j] = 0(1-p) + 1\cdot p = p$$

and

$$Var(I_j) = E[I_j^2] - (E[I_j])^2 = p - p^2 = p(1-p)$$

The mean of the binomial distribution is

$$E[X] = E[I_1] + E[I_2] + \cdots + E[I_n]$$

$$\Rightarrow \quad \boxed{E[X] = np}$$

Variance of X is

$$Var(X) = Var(I_1) + Var(I_2) + \cdots + Var(I_n)$$

$$\Rightarrow \quad \boxed{Var(X) = np(1-p)}$$

Mean and Variance

If X is a binomial r.v. with parameter p and n,

$$\mu = E[X] = np \quad \text{and} \quad \sigma^2 = Var(X) = np(1-p)$$

Ex 1: Each sample of water has a 10% chance of containing a particular organic pollutant. Assume that the samples are independent with regard to the presence of the pollutant. Find the probability that in the next 18 samples, exactly 2 contain the pollutant.

a) Determine the probability that at least four samples contain the pollutant.

b) Determine the probability that more than or equal to 3 samples, but less than 7 contain the pollutant.

Soln: Let X = the no. of samples that contain the pollutant in the next 18 samples analyzed.

Then X is a binomial random variable with $p=0.1$ and $n=18$. Therefore,

$$P(X=2) = \binom{18}{2}(0.1)^2(0.9)^{18-2} = 0.284$$

a) The required probability is

$$P(X \geq 4) = \sum_{x=4}^{18} \binom{18}{x}(0.1)^x(0.9)^{18-x}$$

However it is easy to use the complementary event,

$$P(X \geq 4) = 1 - P(X < 4) = 1 - \sum_{x=0}^{3} \binom{18}{x}(0.1)^x(0.9)^x$$

$$= 1 - [0.15 + 0.3 + 0.284 + 0.168]$$

$$= 0.098$$

b) $$P(3 \leq X < 7) = \sum_{x=3}^{6} \binom{18}{x}(0.1)^x(0.9)^{18-x}$$

$$= 0.168 + 0.07 + 0.022 + 0.005$$

$$= 0.265$$

Ex 2: A large chain retailer purchases a certain kind of electronic device from a manufacturer. The manufacturer indicates that the defective rate of the device is 3%.

a) The inspector randomly picks 20 items from a shipment. What is the probability that there will be at least one defective item among these 20?

b) Suppose that the retailer receives 10 shipments in a month and the inspector randomly tests 20 devices per shipment. What is the probability that there will be exactly 3 shipments each containing at least one defective device among the 20 that are selected and tested from the shipment?

Soln: Let X denote the number of defective devices among the 20.

X is a binomial r.v with parameters $n=20$ and $p=0.03$

a) $P(X \geq 1) = 1 - P(X=0) = 1 - (0.03)^0 (1-0.03)^{20} = 0.4562.$

b) Testing of each shipment is a Bernoulli trial with success (containing atleast one defective item) with prob $p = 0.4562$.

Assuming independence from shipment to shipment and denoting by Y the no. of shipments containing at least one defective item,

Y is another binomial r.v. with parameters $n=10$ and $p = 0.4562$

$\therefore \quad P(Y=3) = \binom{10}{3} (0.4562)^3 (1-0.4562)^7 = 0.1602.$

Ex 3: If the mean and standard deviation of the number of correctly answered questions in a test given to 4096 students are 2.5 and $\sqrt{1.875}$. Find an estimate of the number of candidates answering correctly

i) 8 or more questions ii) 2 or less iii) 5 questions.

Soln: Let X denote correctly answered question.

X is binomial r.v. with parameters n and p.

Given, $E[X] = 2.5$ and $Var(X) = 1.875$

$\Rightarrow$ $np = 2.5$ and $np(1-p) = 1.875$

$\Rightarrow$ $1-p = \frac{1.875}{2.5} = 0.75$ or $p = 0.25$ and $n = 10$

prob. that x questions answered correctly is

$$P(X=x) = b(x;n,p) := \binom{n}{x} p^x (1-p)^{n-x}$$

$\therefore$ prob. that 4096 students answer x questions correctly is

$$4096 \times b(x;n,p)$$

i) No. of students answered 8 or more questions correctly is

$$4096 \times P(X \geq 8) = 4096 \times \sum_{x=8}^{10} b(x;n,p)$$

$$= 4096 \left(\binom{10}{8} (0.25)^8 (0.75)^2 + \binom{10}{9} (0.25)^9 (0.75) + 0.25^{10} \right)$$

$$\simeq 2$$

ii) $$4096 \times P(X \leq 2) = 4096 \times \sum_{x=0}^{2} b(x;n,p)$$

$$= 4096 \left((0.75)^{10} + \binom{10}{1} (0.25)(0.75)^9 + \binom{10}{2} (0.25)^2 (0.75)^8 \right)$$

$$\simeq 2153$$

iii) $$4096 \times P(X=5) = 4096 \times \binom{10}{5} (0.25)^5 (0.75)^5 \simeq 239$$

Ex 4: An air line know that 5% of the people making reservations on a certain flight will not turn up. Consequently their policy is to sell 52 tickets for a flight that can only hold 50 passengers. What is the prob. that there will be a seat for every passenger who turns up?

Soln: Let X be the no. of passengers who turn up.

X is a binomial r.v. with parameters $n = 52$ and $p = 0.95$

where p is prob that a passenger turns up.

Prob. that there will be a seat for every passenger who turns up is

$$P(X \le 50) = 1 - P(X > 50)$$

$$= 1 - \left(b(51; p, n) + b(52; p, n) \right)$$

$$= 1 - \binom{52}{51} (0.95)^{51} (0.05) - (0.95)^{52}$$

$$= 0.7405$$

Ex5: The probability of a shooter hitting a target is $1/3$. How many times he should shoot so that the probability of hitting the target atleast once is more than $3/4$.

Soln: Let X denote the no. of times shooter hits the target.

It is a r.v with parameters n and p.

Given, $p = 1/3$ (prob. of a shooter hitting a target)

and $P(X \ge 1) > 3/4$

$$\Rightarrow 1 - P(X < 1) > 3/4$$

$$\Rightarrow 1 - P(X = 0) > 3/4$$

$$\Rightarrow 1 - b(0; n, p) > 3/4$$

$$\Rightarrow 1 - \left(\tfrac{2}{3}\right)^n > 3/4 \Rightarrow \frac{1}{4} > \left(\tfrac{2}{3}\right)^n \text{ or } \log(1/4) < n \log(2/3)$$

$$n > \frac{\log(1/4)}{\log(2/3)}$$

$$n > 3.41$$

i.e No. of attempts, $n \geq 4$.

Ex6: A computer system uses passwords that are exactly six characters and each character is one of the 26 letters (a–z) or 10 integers (0–9). Suppose there are 10,000 users of the system with unique passwords. A hacker randomly selects (with replacement) one billion passwords from the potential set, and a match to a user's password is called a hit.
(a) What is the distribution of the number of hits?
(b) What is the probability of no hits?
(c) What are the mean and variance of the number of hits?

Soln: Let X denote number of hits with parameters p and n. where p is a prob. that a random select is a hit.

Given: $p = \frac{10^4}{36^6} = 0.0000045939$, $n = 10^9$, no. of random selection

a) Distribution of the r.v. X is binomial.

b) Prob. of no hits is $P(X=0) = (1-p)^{10^9} \approx 0$

c) mean, $E[X] = np = 4593.9$

Variance, $V(X) = np(1-p) \simeq np = 4593.9$ $(\because 1-p \simeq 0)$.

Poisson distribution ; Parameter $\lambda > 0$

The Poisson distribution is used to compute the probability of specific number of "events" during a particular period of time or span of space.

The process that models the occurence of events is called Poisson Process.

For instance, consider a model with characteristic of Poisson Process.

A model of customer arrivals in unit interval of time

Here customer arrival represents the Poisson event.

1) Let K denote the no. of customers arrive in the given unit time interval

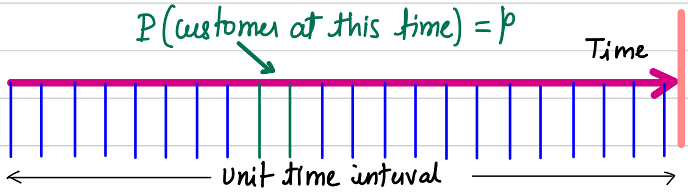
2) No two customers arrive simultaneously.

3) We make the assumption that the arrival times are independent from each other. i.e customer arriving at a specific time has no knowledge about the other customers nor their arrival times.

4) The average number of customers arrive in unit interval of time is constant, say λ i.e, $E[K] = \lambda$.

To obtain the probability distribution function, we consider the following construction which satisfy above properties:

- We divide time line into n subintervals each of length Δt.



$\therefore n \times \Delta t = 1$

\# of customers

$$p(k, \Delta t) = \begin{cases} \lambda \Delta t & \text{if } k = 1 \\ 1 - \lambda \Delta t & \text{if } k = 0 \\ 0 & \text{if } k > 1 \end{cases}$$

If the

- Probability that a new customer shows up on each subinterval is p
- Probability that more than one customer shows up on each subinterval is negligible. (tends to zero),

then $E[\# \text{ of customers}] = np = \lambda \Rightarrow p = \frac{\lambda}{n}$ $\left(\text{or } p = \lambda \Delta t\right)$

- It is clear that K is a binomial r.v. with parameters n and p.

$\therefore$ Prob. that k customers arrive in the given time interval is

$$P(K=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Since $E[K] = \lambda$, a constant and time is continuous we tend Δt to 0
$\Rightarrow n \to \infty$ and $p = \frac{\lambda}{n}$

$$\therefore \quad P(K=k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \left(\frac{\lambda}{n}\right)^k \frac{n(n-1)(n-2)\cdots n-(k-1)}{k!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}$$

$$= \left(\frac{\lambda}{\cancel{n}}\right)^k \frac{\cancel{n^k}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{(k-1)}{n}\right)}{k!} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}$$

As $n \to \infty$, $\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{(k-1)}{n}\right) \to 1$

$$\left(1-\frac{\lambda}{n}\right)^{-k} \to 1 \quad \text{and} \quad \left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$$

$$\therefore \quad P(K=k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, 3, \ldots$$

This distribution is called Poisson distribution and the entire model of customer arrivals is called Poisson Process.

Above ex can be generalized to include a broad array of random experiments.

The interval that was partitioned was a time interval. However, the same reasoning can be applied to any length, an area, or a volume.

For example, counts of

1) particles of contamination in semiconductor manufacturing
2) flaws in rolls of textiles
3) calls to a telephone exchange
4) atomic particles emitted from a specimen

In general, let K denote the no. of events in interval τ of real numbers

We partition the interval into subintervals of small length Δt and assume that as $\Delta t \to 0$,

1) the probability of more than one event in a subinterval tends to zero.

2) the prob. of occurrence of one event in a subinterval tends to $\lambda \Delta t$, λ is avg. no. of occurence of events in the unit interval

3) the event in each subinterval is independent of other subintervals,

A random experiment with these properties are called poisson process

The random variable K that equals the number of events in a time interval, τ is called Poisson random variable with a parameter $\lambda > 0$, λ being average no. of events in a unit interval

and the probability distribution function

$$p(k; \lambda\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!} \qquad k = 0, 1, 2, 3, \ldots$$

is called Poisson distribution.

Note that, i) $\sum_{k=0}^{\infty} p(k;\lambda\tau) = \sum_{k=0}^{\infty} \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!} = e^{-\lambda\tau} \cdot e^{\lambda\tau} = 1$

ii) $p(k,\lambda\tau) \geq 0$

Let us consider time interval $\tau = 1$. Then

Expectation : $E[K] = \lambda$

$$E[K] = \sum_{k=0}^{\infty} k\, p(k;\lambda) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!}$$

$$= \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!}$$

$$= \sum_{k=0}^{\infty} \lambda \frac{\lambda^k e^{-\lambda}}{k!} = \lambda$$

Variance : $Var(K) = \lambda$

$$Var(K) = \sum_{k=0}^{\infty} k^2 p(k;\lambda) - \lambda^2$$

$$= \sum_{k=1}^{\infty} \frac{k^2 \lambda^k e^{-\lambda}}{k!} - \lambda^2$$

$$= \sum_{k=1}^{\infty} \frac{k\,\lambda^k e^{-\lambda}}{(k-1)!} - \lambda^2$$

$$= \sum_{k=1}^{\infty} \frac{(k-1+1)\lambda^k e^{-\lambda}}{(k-1)!} - \lambda^2$$

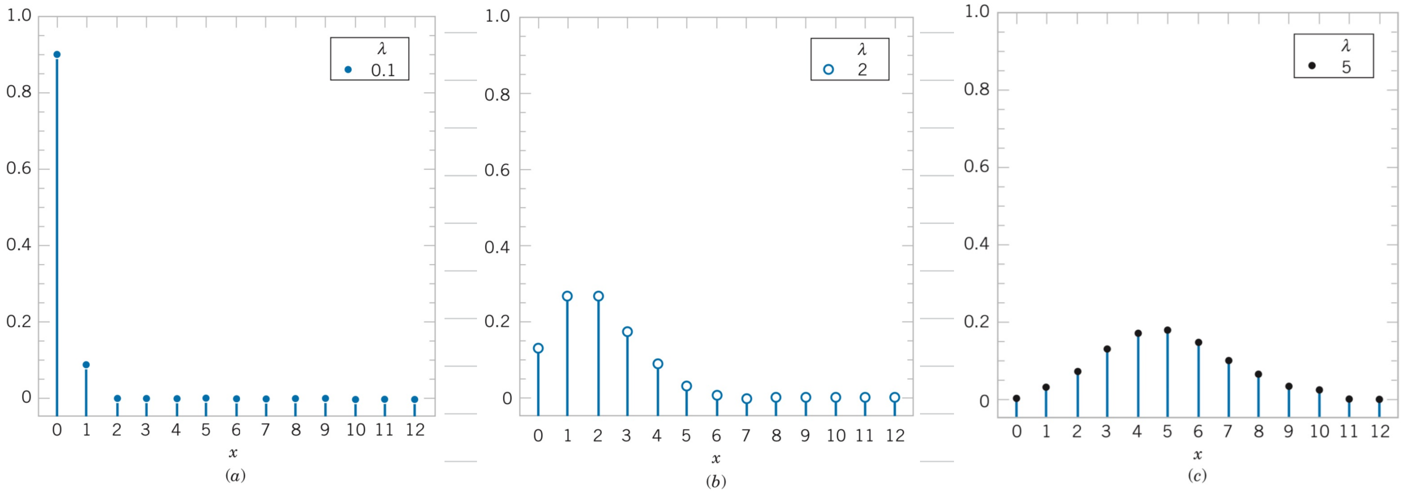
$$= \sum_{k=2}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} - \lambda^2$$

$$= \sum_{k=0}^{\infty} \lambda^2 \frac{\lambda^k e^{-\lambda}}{k!} + \sum_{k=0}^{\infty} \lambda \frac{\lambda^k e^{-\lambda}}{k!} - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

If the time interval is τ, then

$E[K] = \lambda\tau$ and $Var(K) = \lambda\tau$

Poisson distributions for selected values of the parameters.



Approximation of Binomial distribution by a Poisson distribution

Let X be a binomial r.v. with probability distribution $b(x;n,p)$. When $n \to \infty$, $p \to 0$, and $np = \lambda$ remains constant,

$$b(x;n,p) \xrightarrow{n \to \infty} p(x;\lambda)$$

Ex1: For the case of the thin copper wire, suppose that the number of flaws follows a Poisson distribution with a mean of 2.3 flaws per millimeter.

a) Determine the probability of exactly two flaws in 1 millimeter of wire.

b) Determine the probability of 10 flaws in 5 millimeters of wire.

c) Determine the probability of at least one flaw in 2 millimeters of wire.

Soln: a) Let K denote the no. of flaws in 1mm of wire.

Given $\lambda = 2.3$ flaw, (no. of flaws per unit mm)

$$P(K=2) = \frac{e^{-\lambda}\lambda^2}{2!} = \frac{e^{-2.3}(2.3)^2}{2!} = 0.265$$

b) Let K denote the no. of flaws in 5 mm of wire
The K has a Poisson distribution with

$$E[K] = \lambda 5 = 2.3 \text{ flaws/mm} \times 5\text{ mm} = 11.5 \text{ flaws}$$

$$\therefore P(K=10) = \frac{e^{-11.5}(11.5)^{10}}{10!} = 0.113$$

c) Let K denote the no. of flaws in 2mm of wire. Then K has a Poisson distribution with

$$E[K] = \lambda 2 = 2.3 \times 2 = 4.6 \text{ flaws}$$

$$\therefore P(K \geq 1) = 1 - P(K=0) = 1 - e^{-4.6} = 0.9899$$

Ex2: Contamination is a problem in the manufacture of optical storage disks (CDs). The number of particles of contamination that occur on an optical disk has a Poisson distribution, and the average number of particles per centimeter squared of media surface is 0.1. The area of a disk under study is 100 squared centimeters.

a) Find the probability that 12 particles occur in the area of a disk under study.

b) Determine the probability that 12 or fewer particles occur in the area of the disk under study.

Soln: Let K denote the no. of particles in the area of a disk under study

Given $\lambda = 0.1$ (avg. # of particles per cm^2)

$E[K] = \lambda 100 = 0.1 \text{ particles}/cm^2 \times 100 cm^2 = 10 \text{ particles}$

a) $\therefore P(K=12) = \dfrac{e^{-10} 10^{12}}{12!} = 0.095$

b) $P(K \le 12) = P(K=0) + P(K=1) + \cdots + P(K=12)$

$$= \sum_{k=0}^{12} \frac{e^{-10} 10^k}{k!} = 0.792$$

Ex 3: In a manufacturing process where glass products are made, defects or bubbles occur, occasionally rendering the piece undesirable for marketing. It is known that, on average, 1 in every 1000 of these items produced has one or more bubbles. What is the probability that a random sample of 8000 will yield fewer than 7 items possessing bubbles?

Soln: Let X denote the no. of products possessing defects or bubble.

X is a binomial r.v with parameters $n = 8000$, $p = 0.001$

Since n is quite large, and p is close to 0, we shall approx. with Poisson distribution.

$\lambda = 8000 \times 0.001 = 8$.

$$\therefore P(X < 7) \approx \sum_{x=0}^{6} p(x;8) = \sum_{x=0}^{6} \frac{e^{-8} 8^x}{x!} = 0.313$$

Ex 4: The number of particles emitted by a radioactive source is Poisson distributed. The source emits particles at a rate of 6 per minutes. Each emitted particle has a probability of 0.7 of being counted. Find the probability that 11 particles are counted in 4 minutes.

Soln: Let K denoted # of particles being counted in 4 minutes.

# of particles counted in one minute $\lambda = 6 \times 0.7 = 4.2$

$E[K] = \lambda 4 = 4.2 \times 4 = 16.8$

$$P(K=11) = \frac{e^{-\lambda 4} (\lambda 4)^{11}}{11!} = \frac{e^{-16.8} (16.8)^{11}}{11!} = 0.0381$$

Ex 5: If the probability that an individual suffers a bad reaction from an injection of a given serum is 0.001, determine the prob. that out of 2000 individuals

i) exactly 3; and ii) more than two individuals, will suffer a bad reaction.

Soln: Let X denote no. of individuals will suffer a bad reaction.
Clearly X is a binomial r.v with parameters:

$$p = 0.001 \text{ and } n = 2000$$

Since p is small and n is large, we can approx binomial to poisson distribution with parameter $\lambda = np = 2$

i) $$P(X=3) = \frac{e^{-\lambda}\lambda^3}{3!} = \frac{e^{-2}2^3}{6} = 0.1804$$

ii) $$P(X>2) = 1 - P(X \leq 1)$$

$$= 1 - P(X=1) - P(X=0)$$

$$= 1 - \frac{e^{-2}2}{1!} - e^{-2} = 1 - 3e^{-2} = 0.594$$

Exponential distribution ; parameter, $\lambda > 0$

The discussion of the Poisson distribution defined a random variable to be the number of events in an interval of length, $\tau > 0$.

Let K denote # of events in the interval τ and the Poisson distribution is

$$p(k, \lambda\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \qquad \text{———} \quad 0$$

where λ is # of events per unit interval.

The length of interval between events is another r.v.

The random variable X that equals the distance between successive events of a Poisson process with mean number of events $\lambda > 0$ per unit interval is an exponential random variable with parameter λ. The probability density function of X is

$$f(x, \lambda) = \lambda e^{-\lambda x}, \quad 0 \le x < \infty$$

The probability that a first event occur within a distance x from a starting point is given by

$$P(X \le x) = 1 - P(X > x) \qquad \text{———} \quad 1$$

is called the cumulative distribution function.

Where $P(X > x)$ the probability that the length until the first event will exceed x, but it is same as the probability that no Poisson event will occur in the distance x, $p(0; \lambda x)$

i.e $P(X > x) = p(0; \lambda x) = \dfrac{e^{-\lambda x}(\lambda x)^0}{0!}$ ——— 2 (from 0)

$\Rightarrow$ The Cumulative distribution function is

$$F(x) = P(X \le x) = 1 - e^{-\lambda x}$$

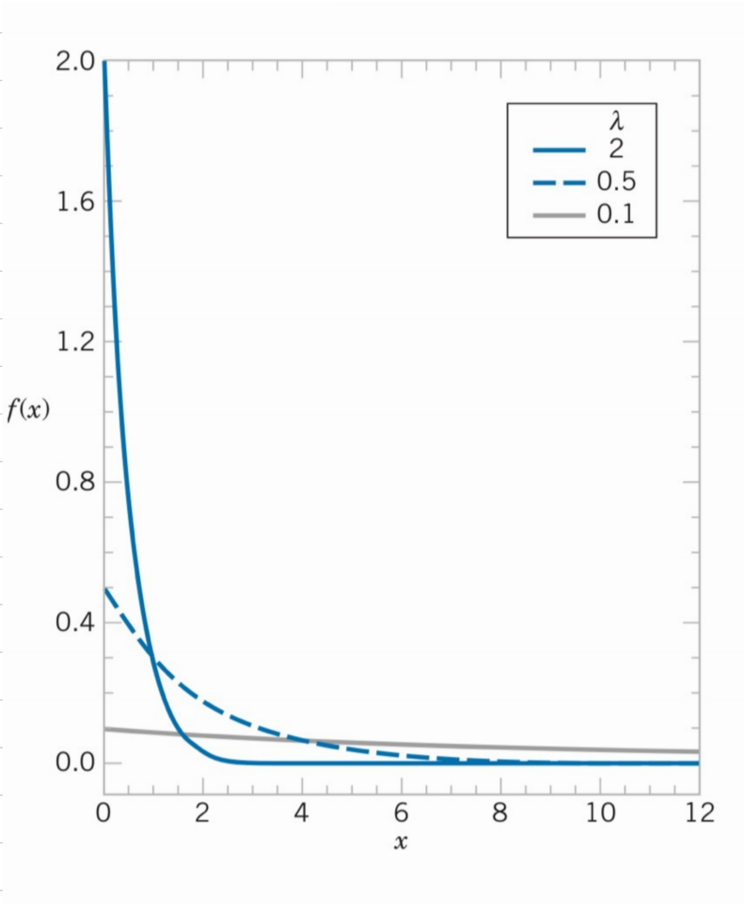
(from 1 and 2)

By differentiating $F(x)$, we get the probability density function of X and is calculated to be

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0.$$

is called exponential distribution.

Probability density function of exponential random variables for selected values of λ.



Expectation and Variance

If the r.v. X has an exponential distribution with parameter λ,

Expectation, $E[X] = \frac{1}{\lambda}$ and Variance, $Var(X) = \frac{1}{\lambda^2}$

$$E[X] = \int_0^\infty x \lambda e^{-\lambda x}\, dx$$

$$= \lambda \left[\frac{x e^{-\lambda x}}{-\lambda} - \frac{e^{-\lambda x}}{\lambda^2} \right]_{x=0}^{\infty} = \frac{1}{\lambda}$$

$$Var(X) = E[X^2] - (E[X])^2$$

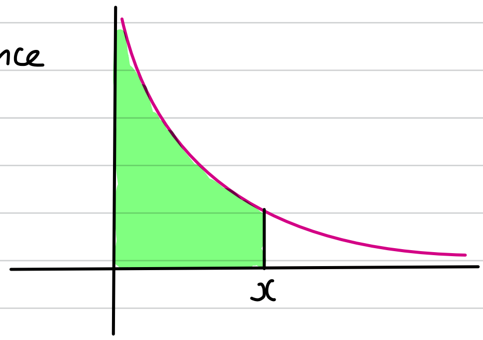
$$= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx - \frac{1}{\lambda^2}$$

$$= \lambda \left[x^2 \frac{e^{-\lambda x}}{-\lambda} - \frac{e^{-\lambda x}}{\lambda^2} 2x + \frac{e^{-\lambda x}}{-\lambda^3} \cdot 2 \right]_0^{\infty} - \frac{1}{\lambda^2}$$

$$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Remarks: Let $f(x; \lambda) = \lambda e^{-\lambda x}$, $x > 0$ be an exponential distributn

i) Prob. that an event occur within a distance x from the start is.

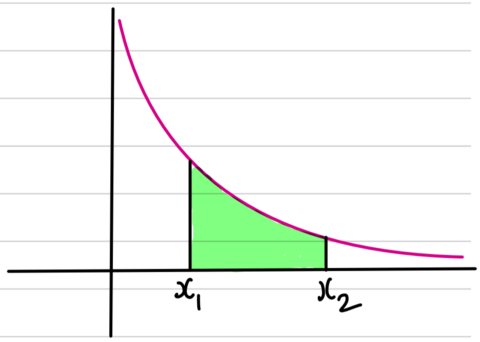


$$P(X \leq x) = F(x) = 1 - e^{-\lambda x}$$

or

$$P(X \leq x) = \int_0^{x} \lambda e^{-\lambda x} dx = 1 - e^{\lambda x}$$

ii) Prob. that an event occur between x_1 and x_2



$$P(x_1 < X < x_2) = F(x_2) - F(x_1)$$

$$= e^{\lambda x_1} - e^{\lambda x_2}$$

or

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \lambda e^{-\lambda x} dx = e^{\lambda x_1} - e^{\lambda x_2}$$

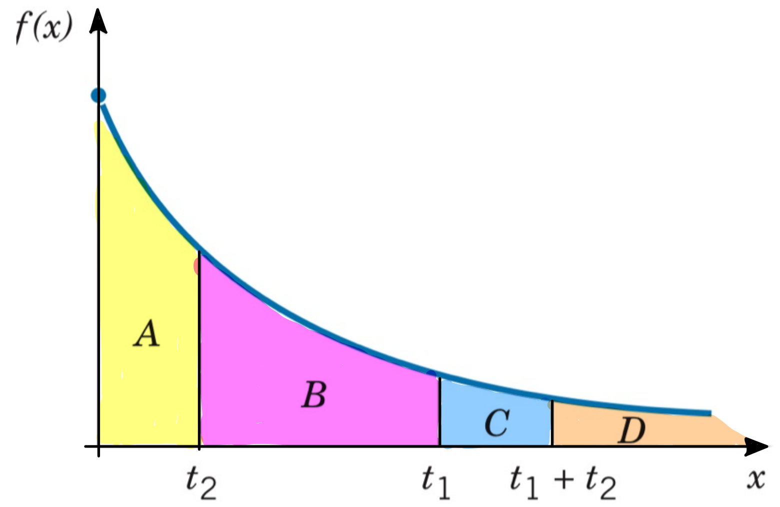
Memoryless property (lack of memory property)

A general statement of the property is as follows.

For an exponential random variable X,

$$P(X < t_1 + t_2 | X > t_1) = P(X < t_2)$$

Below figure graphically illustrates the lack of memory property.



Total area under PDF is

$A + B + C + D = 1$

$$P(X < t_2) = \frac{\text{area of region } A}{\text{total area}}$$

$$P(X < t_1 + t_2 \mid X > t_1) = \frac{P(t_1 < X < t_1 + t_2)}{P(X > t_1)} = \frac{\text{area of region } C}{\text{area of region } C+D}$$

property implies that

The proportion of the total area that is in A

= the proportion of the area in C and D that is in C

Interpretation:

After the length of interval t_1 without the occurence of an event, the probability of occurence of an event in the next length of interval t_2 is same as the probability of occurence in the interval $[0, t_2]$.

Pf: $P(X < t_1+t_2 \mid X > t_1) = \dfrac{P(t_1 < X < t_1+t_2)}{P(X > t_1)}$

$$= \frac{F(t_1+t_2) - F(t_1)}{1 - F(t_1)}$$

$$= \frac{1 - e^{-\lambda(t_1+t_2)} - (1 - e^{-\lambda t_1})}{1 - (1 - e^{-\lambda t_1})}$$

$$= \frac{e^{-\lambda t_1} - e^{-\lambda t_1} \cdot e^{-\lambda t_2}}{e^{-\lambda t_1}}$$

$$= 1 - e^{-\lambda t_2}$$

$$= F(t_2)$$

$$= P(X < t_2)$$

Note: The probability that the event occur after x_1 given that no event occured within distance x_2 $(x_2 < x_1)$ is the prob. that the event occur after $x_1 - x_2$

i.e $P(X > x_1 \mid X > x_2) = 1 - P(X \le x_1 \mid X > x_2)$

$$= 1 - P(X \le x_1 - x_2)$$

$$= P(X > x_1 - x_2) = e^{-\lambda(x_1 - x_2)}$$

By defn

$$P(X > x_1 \mid X > x_2) = \frac{P(X > x_1)}{P(X > x_2)} = e^{-\lambda(x_1 - x_2)}$$

Ex 1 In a large corporate computer network, user log-ons to the system can be modeled as a Poisson process with a mean of 25 log-ons per hour. What is the probability that there are no log-ons in an interval of 6 minutes?

a) What is the probability that the time until the next log-on is between 2 and 3 minutes? Upon converting all units to hours,

b) Determine the interval of time such that the probability that no log-on occurs in the interval is 0.90.

Soln: Let X denote the time hours from the start of the interval until the 1st log-on. Then

X represents exp. distribution with parameter $\lambda = 25$

The prob that there are no log-ons in an interval of 6 min

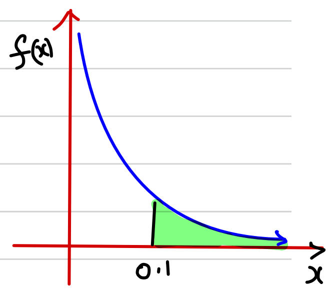
= The prob that there are log-ons after 6 mins

$$= P(X > 6\,min) = P(X > 0.1) \qquad (\because 6\,min = 0.1\,hr)$$

$$= 1 - P(X \leq 0.1)$$

$$= 1 - \left(1 - e^{-25(0.1)}\right)$$

$$= e^{-2.5} = 0.082$$



or

$$P(X > 0.1) = \int_{0.1}^{\infty} 25 e^{-25x}\, dx = e^{-2.5} = 0.085$$

a) $2\,min = \frac{2}{60} = 0.034\,hr$, $\quad 3\,min = \frac{3}{60} = 0.05\,hr$

$$P(2\,min < X < 3\,min) = P(0.034 < X < 0.05)$$

$$= F(0.05) - F(0.034)$$

$$= e^{-25(0.034)} - e^{-25(0.05)}$$

$$= 0.1409$$

b) Let the prob that no log-on occur in an interval $[0, x]$ be 0.9

i.e let the prob that log-on occur after the length x be 0.9

$$\Rightarrow \quad P(X > x) = 0.9$$

$$\Rightarrow \quad 1 - P(X < x) = 0.9$$

$$\Rightarrow \quad 1 - F(x) = 0.9$$

$$\Rightarrow \quad e^{-25x} = 0.9$$

$$\Rightarrow \quad -25x = Ln(0.9) \Rightarrow \quad x = 0.0042 \text{ hrs} \quad \text{or} \quad x = 0.253 \text{ min}$$

Ex2: Let X denote the time between detections of a particle with a Geiger counter and assume that X has an exponential distribution with E(X)=1.4 minutes.

a) What is the probability that we detect a particle within 30 seconds of starting the counter

b) Now, suppose we turn on the Geiger counter and wait 3 minutes without detecting a particle. What is the probability that a particle is detected in the next 30 seconds?

Soln: Given $\lambda = \frac{1}{1.4}$

a) prob. that a particle is detected within 30 sec of start

is $P(X \leq 30 \text{ sec}) = P(X \leq 0.5 \text{ min})$

$$= F(0.5) = 1 - e^{-\frac{1}{1.4}(0.5)}$$

$$= 0.3$$

b) Prob. that particle is detected with 3 min and 30 sec of starting the counter given that a particle is not detected in first 3 min is

$P(X \leq 3.5 \mid X > 3) = P(X \leq 0.5)$ (From memory less property)

$$= 0.3$$

Ex3: Suppose that a system contains a certain type of component whose time, in years, to failure is given by T . The random variable T is modeled nicely by the exponential distribution with mean time to failure $\beta = 5$. If 5 of these components are installed in different systems, what is the probability that at least 2 are still functioning at the end of 8 years?

Soln: Let T denote time, in years, a component fails. The r.v T is an exp. r.v.

mean, $\frac{1}{\lambda} = 5 \Rightarrow \lambda = \frac{1}{5}$

The prob. that a comp. still functioning after 8 years. is given by

$$P(T>8) = 1 - P(T \leq 8)$$
$$= 1-(1-e^{-\lambda 8})$$
$$= e^{-8/5} \simeq 0.2$$

Let X represents the no. of components working after 8 yrs. Then X is binomial r.v. with parameters $n=5$, $p=0.2$,

$$P(X \geq 2) = 1 - P(X<2)$$
$$= 1 - b(1;5,0.2) - b(0;5,0.2)$$
$$= 1 - \binom{5}{1} p^1 (1-p)^4 - \binom{5}{0} p^0 (1-p)^5$$
$$= 1 - 5(0.2)(0.8)^4 - (0.8)^5$$
$$= 1 - 0.7373 = 0.2627$$

Ex4: The life length (in months) of an electric component follows an exponential distribution with parameter $\lambda = 0.5$. What is the probability that the component survives at least 10 months given that already it had survived for more than 9 months?

Soln: Given $\lambda = 0.5$. The prob. that the component survives at least 10 months given that it has survived for more than 9 months is

$$P(X>10 \mid X>9) = P(X>1)$$

$$= 1 - P(X \leq 1)$$
$$= 1 - F(1)$$
$$= 1 - (1 - e^{-0.5(1)})$$
$$= e^{-0.5} = 0.6065$$

or

$$P(X > 10 \mid X > 9) = \frac{P(X > 10)}{P(X > 9)} = \frac{e^{-0.5(10)}}{e^{-0.5(9)}} = e^{-0.5}$$

Ex5: The sales per day in a shop is exponentially distributed with the average sale amounting to Rs 100 and net profit is 8%. Find the probability that the net profit exceeds Rs 30 on two consecutive days.

Soln: Let X denote sales per day, mean = Rs 100, $\therefore \lambda = \frac{1}{100}$

Net profit is 8%. If A is the amount for which profit is Rs 30, the $A \frac{8}{100} = 30 \Rightarrow A = 375$

prob. of profit exceeding Rs 30 on a single day.

= Prob. the sales exceed Rs 375

$$= P(X \geq 375)$$
$$= \int_{375}^{\infty} \lambda e^{-\lambda x} \, dx$$
$$= \int_{375}^{\infty} \frac{1}{100} e^{-x/100} \, dx = e^{-3.75} = 0.0235$$

$\therefore$ prob. that the profit exceeds on 2 consecutive days

$$= 0.0235 \times 0.0235 = 0.00055$$

Normal Distribution; Parameters μ and σ

The most important continuous probability distribution in the entire field of statistics is the normal distribution.

Its graph, called the normal curve, is the bell-shaped curve of Figure

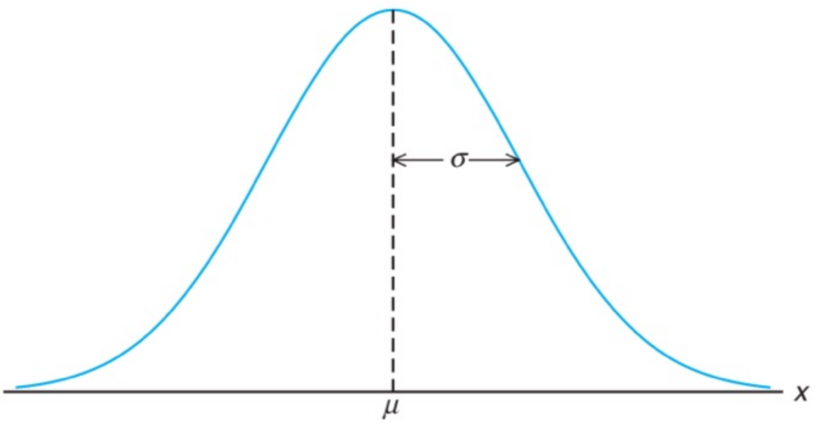


Figure: The normal curve.

In 1733, Abraham DeMoivre developed the mathematical equation of the normal curve.

The normal distribution is also referred as the Gaussian distribution, in honor of Karl Friedrich Gauss (1777-1855), who also derived its equation from a study of errors in repeated measurements of same quantity.

Whenever a random experiment is replicated, the random variable that equals the average (or total) result over the replicates tends to have a normal distribution as the number of replicates becomes large.

A continuous random variable X having the bell-shaped distribution of Figure is called a normal random variable.

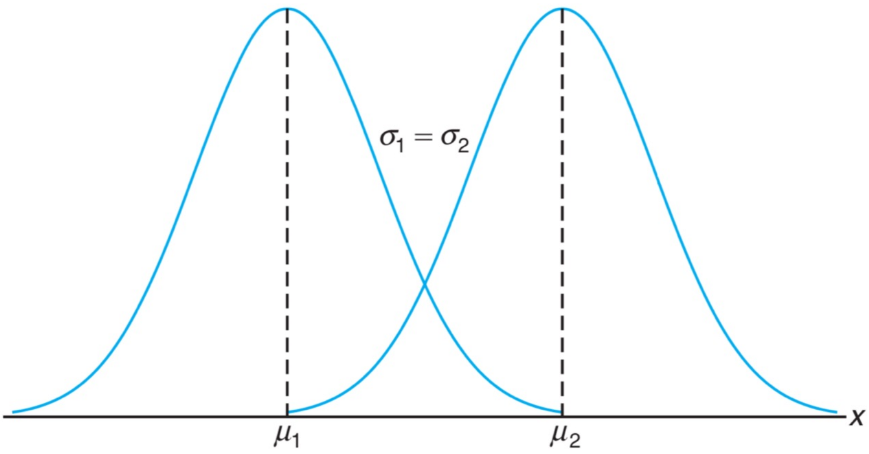
The density of the normal random variable X, with mean μ and variance σ^2, is denoted and given by

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

is called normal distribution.

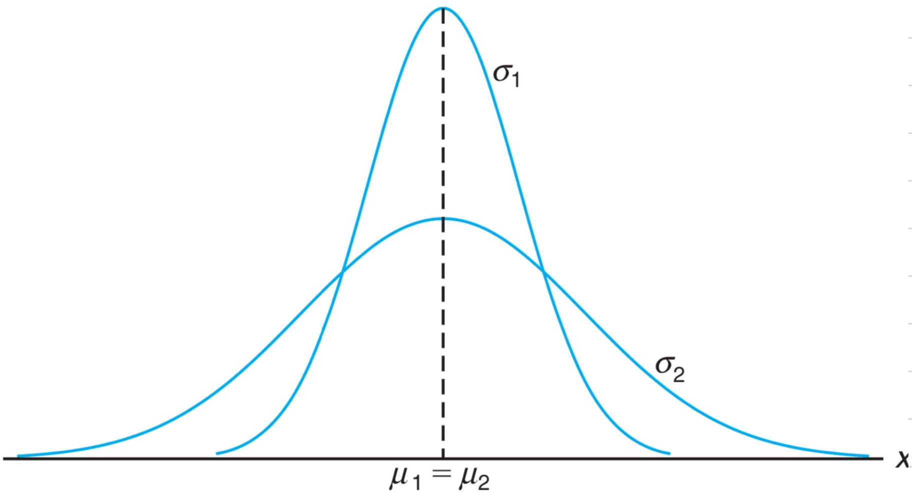
Graphs of normal curve for different μ and σ

i)



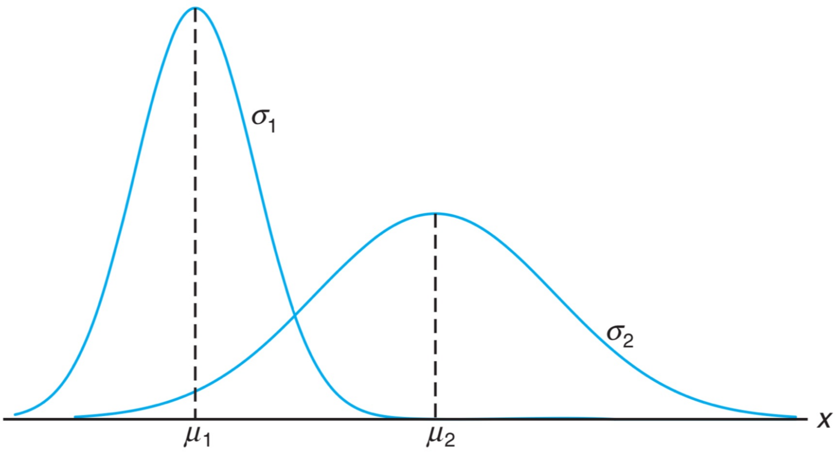
Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$.

ii)



Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$.

iii)



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$.

Properties of the normal curve.

1. The mean = median = mode, which is the point on the horizontal axis where the curve is a maximum.

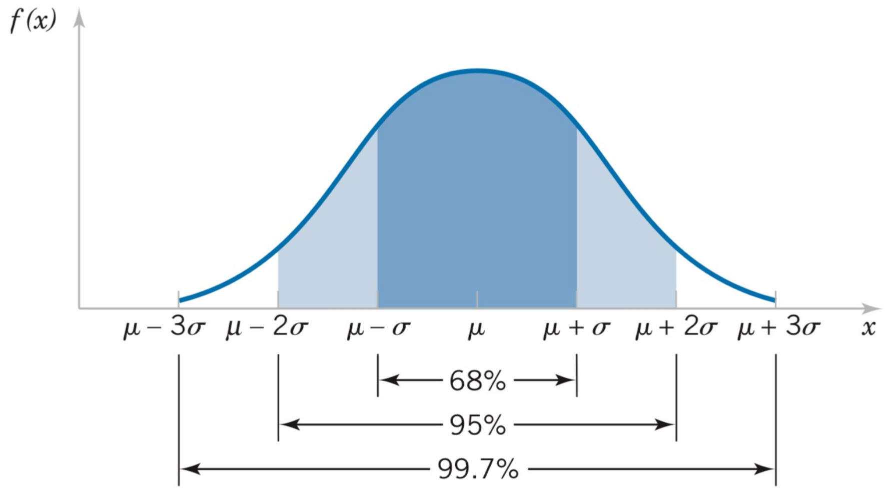
2. The curve is symmetric about a vertical axis through the mean μ.

3. The curve has its points of inflection at $X = \mu \pm \sigma$; it is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.

4. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.

5. The total area under the curve and above the horizontal axis is equal to 1.

6. $P(\mu - \sigma < X < \mu + \sigma) = 0.6827$
$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$
$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$



Mean of variance of the normal distribution

Let X be a normal r.v. Then $E[X] = \mu$ and $Var(X) = \sigma^2$

pf: Consider

$$E[X-\mu] = \int_{-\infty}^{\infty} \frac{x-\mu}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Sub. $z = \frac{x-\mu}{\sigma} \Rightarrow \sigma\, dz = dx$, we obtain

$$E[X-\mu] = \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} \sigma \, dz$$

$$= 0 \qquad \text{(Since integrand is odd fn)}$$

$$\Rightarrow E[X] - E[\mu] = 0$$

$$\Rightarrow E[X] = \mu \qquad (\because E[\mu] = \mu E[1] = \mu)$$

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Set $z = \dfrac{x-\mu}{\sigma} = \sigma dz = dx$, we obtain

$$E[(x-\mu)^2] = \int_{-\infty}^{\infty} \frac{\sigma^2 z^2}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left(-z e^{-\frac{1}{2}z^2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz \right)$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left(0 + \sqrt{2\pi} \right) = \sigma^2.$$

$$\left(\begin{array}{l} \because \int z \cdot z e^{-\frac{1}{2}z^2} dz = z\left(-e^{-\frac{1}{2}z^2}\right) - \int -e^{-\frac{1}{2}z^2} dz \\ \text{and } \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = \sqrt{2\pi} \end{array} \right)$$

Areas under the Normal Curves

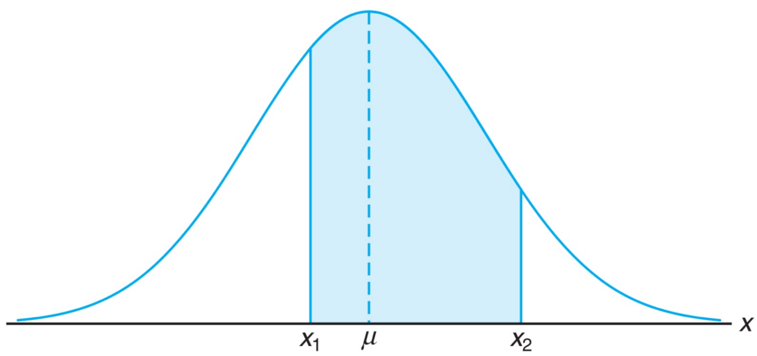
Let the normal distribution of a normal r.v. X with mean μ and variance σ^2 be

$$n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Then the probability that the r.v. assumes a value between $x=x_1$ and $x=x_2$ equals the area under the curve bounded by the ordinates $x=x_1$ and $x=x_2$.

ie $$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x;\mu,\sigma)\,dx = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{x_1}^{x_2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

= area of the shaded region in the below figure.



Since the normal curve depends on μ and σ, the area under the curve between any two ordinates also depends on μ and σ.

We can see that it is difficult to integrate normal curves, however, normal curve areas can be tabulated for quick reference using some statistical software.

It is a hopeless task to attempt to set up seperate tables for every value of μ and σ. Instead we transform any normal curve to a normal curve (called standard normal curve) with $\mu=0$ and $\sigma^2=1$.

Standard normal curve

A normal random variable with

$$\mu = 0 \quad \text{and} \quad \sigma^2 = 1$$

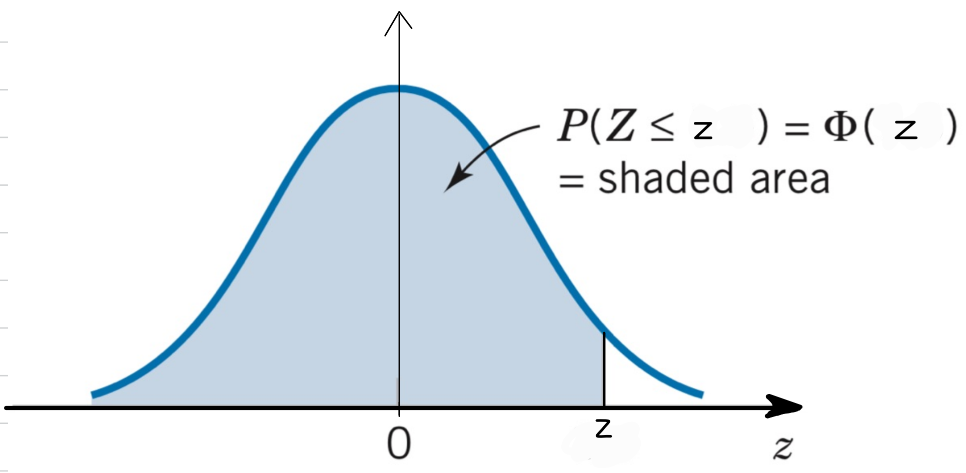
is called a standard normal r.v. and is denoted by Z. The probability density fn for Z is

$$n(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

It is called standard normal distribution or Z-distribution.

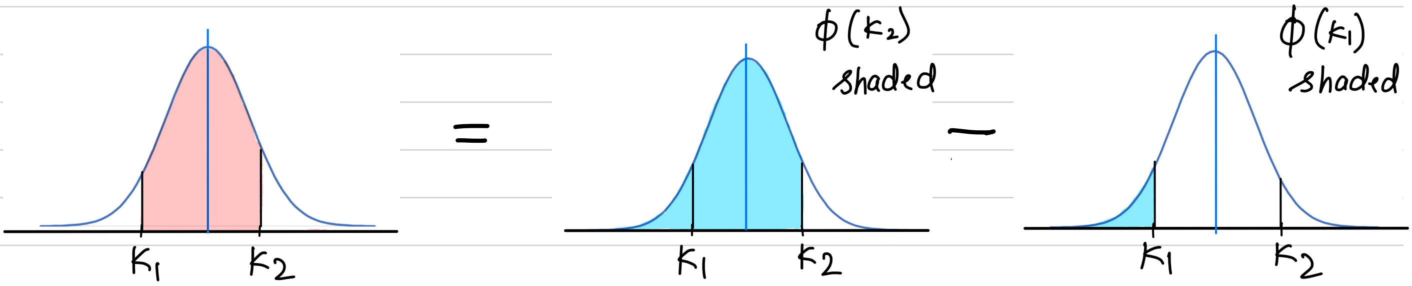
The cumulative distribution function of a standard normal r.v. is denoted by

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\, dt$$

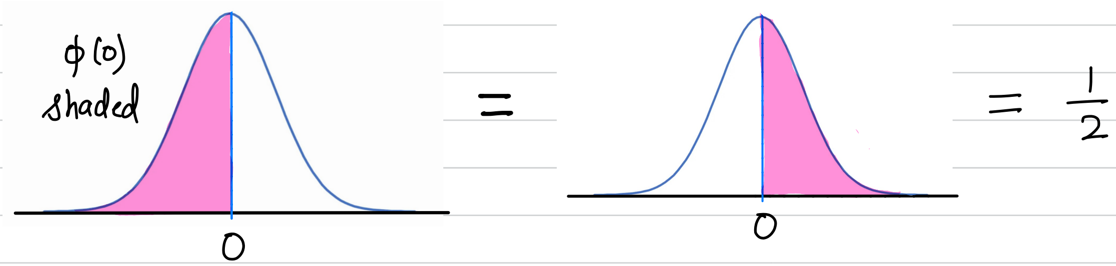


Note: 1) $P(k_1 < Z < k_2) = \Phi(k_2) - \Phi(k_1)$

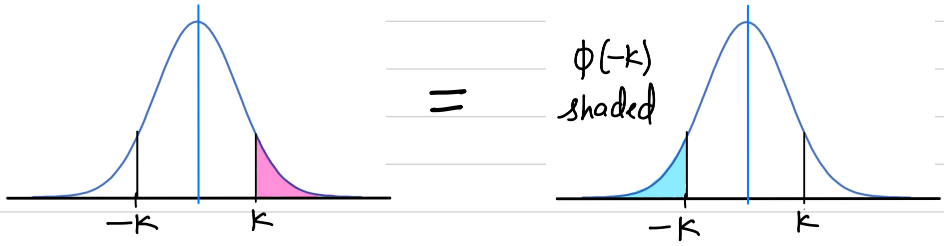
i.e



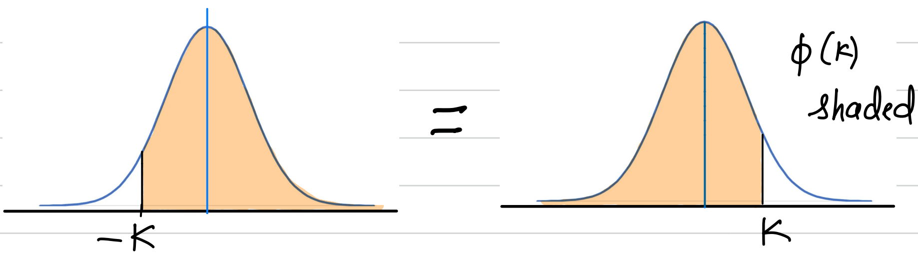
2) $P(Z \le 0) = P(Z \ge 0) = \phi(0) = \frac{1}{2}$ ($\because$ total area $= 1$)



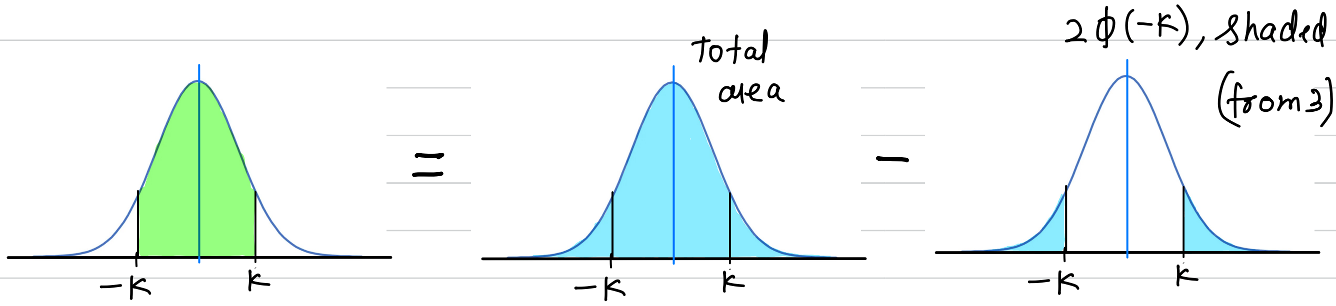
3) $P(Z \ge k) = P(Z \le -k) = \phi(-k)$



4) $P(Z \ge -k) = P(Z \le k) = \phi(k)$ for some $k \in \mathbb{R}^+$.



5) $P(-k \le Z \le k) = 1 - 2\phi(-k)$



Below table given cumulative probabilities for a standard normal r.v.

It gives the value of $\phi(z)$ for $-3.99 \le Z \le 3.99$

(note that r.v. Z is rounded to two decimals)

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

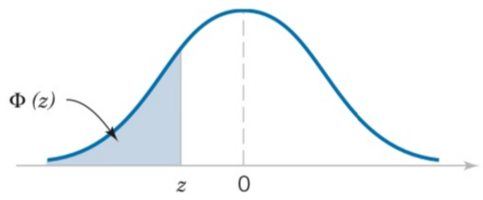


Table III Cumulative Standard Normal Distribution

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	−0.00
−3.9	0.000033	0.000034	0.000036	0.000037	0.000039	0.000041	0.000042	0.000044	0.000046	0.000048
−3.8	0.000050	0.000052	0.000054	0.000057	0.000059	0.000062	0.000064	0.000067	0.000069	0.000072
−3.7	0.000075	0.000078	0.000082	0.000085	0.000088	0.000092	0.000096	0.000100	0.000104	0.000108
−3.6	0.000112	0.000117	0.000121	0.000126	0.000131	0.000136	0.000142	0.000147	0.000153	0.000159
−3.5	0.000165	0.000172	0.000179	0.000185	0.000193	0.000200	0.000208	0.000216	0.000224	0.000233
−3.4	0.000242	0.000251	0.000260	0.000270	0.000280	0.000291	0.000302	0.000313	0.000325	0.000337
−3.3	0.000350	0.000362	0.000376	0.000390	0.000404	0.000419	0.000434	0.000450	0.000467	0.000483
−3.2	0.000501	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687
−3.1	0.000711	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968
−3.0	0.001001	0.001035	0.001070	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.001350
−2.9	0.001395	0.001441	0.001489	0.001538	0.001589	0.001641	0.001695	0.001750	0.001807	0.001866
−2.8	0.001926	0.001988	0.002052	0.002118	0.002186	0.002256	0.002327	0.002401	0.002477	0.002555
−2.7	0.002635	0.002718	0.002803	0.002890	0.002980	0.003072	0.003167	0.003264	0.003364	0.003467
−2.6	0.003573	0.003681	0.003793	0.003907	0.004025	0.004145	0.004269	0.004396	0.004527	0.004661
−2.5	0.004799	0.004940	0.005085	0.005234	0.005386	0.005543	0.005703	0.005868	0.006037	0.006210
−2.4	0.006387	0.006569	0.006756	0.006947	0.007143	0.007344	0.007549	0.007760	0.007976	0.008198
−2.3	0.008424	0.008656	0.008894	0.009137	0.009387	0.009642	0.009903	0.010170	0.010444	0.010724
−2.2	0.011011	0.011304	0.011604	0.011911	0.012224	0.012545	0.012874	0.013209	0.013553	0.013903
−2.1	0.014262	0.014629	0.015003	0.015386	0.015778	0.016177	0.016586	0.017003	0.017429	0.017864
−2.0	0.018309	0.018763	0.019226	0.019699	0.020182	0.020675	0.021178	0.021692	0.022216	0.022750
−1.9	0.023295	0.023852	0.024419	0.024998	0.025588	0.026190	0.026803	0.027429	0.028067	0.028717
−1.8	0.029379	0.030054	0.030742	0.031443	0.032157	0.032884	0.033625	0.034379	0.035148	0.035930
−1.7	0.036727	0.037538	0.038364	0.039204	0.040059	0.040929	0.041815	0.042716	0.043633	0.044565
−1.6	0.045514	0.046479	0.047460	0.048457	0.049471	0.050503	0.051551	0.052616	0.053699	0.054799
−1.5	0.055917	0.057053	0.058208	0.059380	0.060571	0.061780	0.063008	0.064256	0.065522	0.066807
−1.4	0.068112	0.069437	0.070781	0.072145	0.073529	0.074934	0.076359	0.077804	0.079270	0.080757
−1.3	0.082264	0.083793	0.085343	0.086915	0.088508	0.090123	0.091759	0.093418	0.095098	0.096801
−1.2	0.098525	0.100273	0.102042	0.103835	0.105650	0.107488	0.109349	0.111233	0.113140	0.115070
−1.1	0.117023	0.119000	0.121001	0.123024	0.125072	0.127143	0.129238	0.131357	0.133500	0.135666
−1.0	0.137857	0.140071	0.142310	0.144572	0.146859	0.149170	0.151505	0.153864	0.156248	0.158655
−0.9	0.161087	0.163543	0.166023	0.168528	0.171056	0.173609	0.176185	0.178786	0.181411	0.184060
−0.8	0.186733	0.189430	0.192150	0.194894	0.197662	0.200454	0.203269	0.206108	0.208970	0.211855
−0.7	0.214764	0.217695	0.220650	0.223627	0.226627	0.229650	0.232695	0.235762	0.238852	0.241964
−0.6	0.245097	0.248252	0.251429	0.254627	0.257846	0.261086	0.264347	0.267629	0.270931	0.274253
−0.5	0.277595	0.280957	0.284339	0.287740	0.291160	0.294599	0.298056	0.301532	0.305026	0.308538
−0.4	0.312067	0.315614	0.319178	0.322758	0.326355	0.329969	0.333598	0.337243	0.340903	0.344578
−0.3	0.348268	0.351973	0.355691	0.359424	0.363169	0.366928	0.370700	0.374484	0.378281	0.382089
−0.2	0.385908	0.389739	0.393580	0.397432	0.401294	0.405165	0.409046	0.412936	0.416834	0.420740
−0.1	0.424655	0.428576	0.432505	0.436441	0.440382	0.444330	0.448283	0.452242	0.456205	0.460172
0.0	0.464144	0.468119	0.472097	0.476078	0.480061	0.484047	0.488033	0.492022	0.496011	0.500000

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$

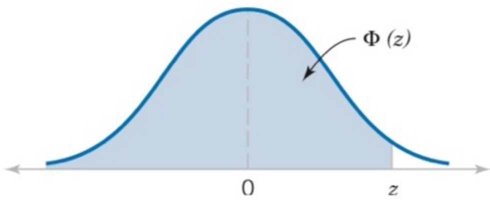


Table III Cumulative Standard Normal Distribution (*continued*)

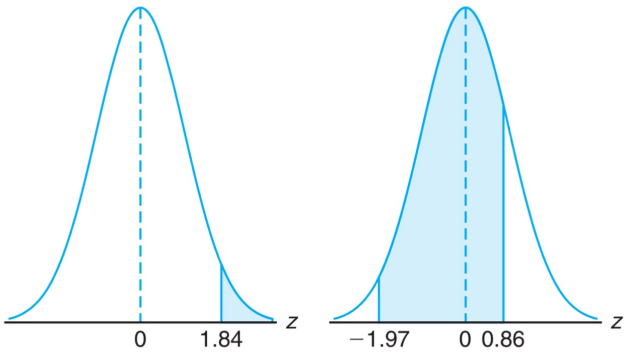
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967

Ex1: Given a standard normal distribution, find the area under the curve that lies

a) to the right of $z = 1.84$ and

b) between $z = -1.97$ and $z = 0.86$

Soln:



(a) (b)

a) Total area $= 1$.

Area to the right of 1.84 $= 1 -$ area to the left of 1.84

$= 1 - \phi(1.84)$

$= 1 - 0.9671 = 0.0329$

b) Area between $z = -1.97$ and $z = 0.86$

$=$ area to the left of 0.86 $-$ area to the left of -1.97

$= \phi(0.86) - \phi(-1.97)$

$= 0.8051 - 0.0244 = 0.7807$.

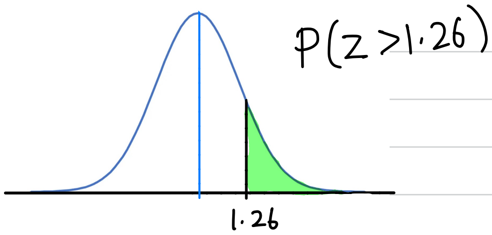
Ex2: Given the std normal distribution, find the foll.

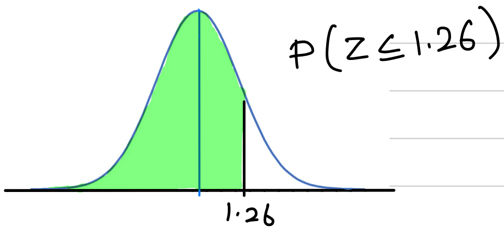
a) $P(Z > 1.26)$ b) $P(Z < -0.86)$

c) $P(Z > -1.37)$ d) $P(-1.25 < Z < 0.37)$

Soln: a) $P(Z > 1.26) = 1 - P(Z \le 1.26)$

$= 1 - \phi(1.26) = 1 - 0.8961 = 0.1038$





b) $P(Z<-0.86) = \phi(-0.86) = 0.1949$

c) $P(Z>-1.37) = 1-P(Z\leq -1.37) = 1-0.0853$

$= 0.9147$

$(\text{Also} = P(Z<1.37))$

d) $P(-1.25<Z<0.37) = \phi(0.37)-\phi(-1.25)$

$= 0.6443 - 0.1056 = 0.5387$

Ex3: Given a standard normal distribution, find the value of k such that

a) $P(Z>k) = 0.3015$ and

b) $P(k<Z<-0.18) = 0.4197$

Soln a) $P(Z>k) = 0.3015$

$\Rightarrow 1-P(Z\leq k) = 0.3015$

$\Rightarrow 1-\phi(k) = 0.3015$

$\Rightarrow \phi(k) = 0.6985$

$\Rightarrow k = 0.52$

b) $P(k<Z<-0.18) = 0.4197$

$\Rightarrow \phi(-0.18)-\phi(k) = 0.4197$

$\Rightarrow \phi(k) = \phi(-0.18)-0.4197 = 0.4286-0.4197$

$\Rightarrow \phi(k) = 0.0089$

$\Rightarrow k = -2.37$

Standardizing a normal random variable

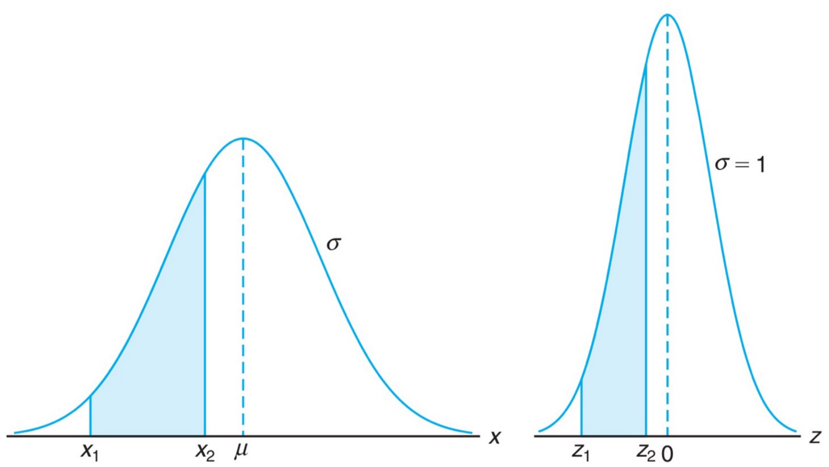
If X is a normal r.v. with $E[X]=\mu$ and $Var(X)=\sigma^2$, the random variable

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

is a normal r.v. with $E[Z]=0$ and $Var(Z)=1$.
That is, Z is a standard normal r.v.

- Whenever X assumes a value x, the corresponding value of Z is $z = \frac{x-\mu}{\sigma}$.
- If X falls between x_1 and x_2, the r.v. Z falls between $z_1 = \frac{x_1-\mu}{\sigma}$ and $z_2 = \frac{x_2-\mu}{\sigma}$.
- $$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \left(\text{put } z = \frac{x-\mu}{\sigma}\right)$$
$$= \int_{z_1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \qquad \left(= \int_{z_1}^{z_2} n(z;0,1)\,dz\right)$$
$$= P(z_1 < Z < z_2).$$

The original and transformed normal distributions.



Ex4. Given a r.v. X having a normal distribution with $\mu = 50$ and $\sigma = 10$, find the probability that X assumes a value between 45 and 62.

Soln: The z values corresponding to $x_1 = 45$ and $x_2 = 62$ are

$$z_1 = \frac{x_1 - \mu}{\sigma} = -0.5 \quad , \quad z_2 = \frac{x_2 - \mu}{\sigma} = 1.2$$

$$\therefore P(45 < X < 62) = P(-0.5 < Z < 1.2)$$

$$= \phi(1.2) - \phi(-0.5)$$

$$= 0.8849 - 0.3085 = 0.5764$$

Ex5: Given a normal distribution with $\mu = 40$ and $\sigma = 6$, find the value of x that has

a) 45% of the area to the left

b) 14% of the area to the right

Soln: a) find x such that

$$P(X \leq x) = 0.45$$

if z is corresponding value of the std normal r.v.,

$$P(Z \leq z) = 0.45$$

$$\Rightarrow \phi(z) = 0.45$$

from the above table $z = -0.13$.

Thus, $x = z\sigma + \mu = -0.13(6) + 4 = 39.22$

b) find x such that

$$P(X \geq x) = 0.14$$

if $z = \frac{x - \mu}{\sigma}$, then $P(Z \geq z) = 0.14$

$$\Rightarrow \quad 1 - P(Z < z) = 0.14$$

$$\Rightarrow \quad 1 - \phi(z) = 0.14$$

$$\Rightarrow \quad \phi(z) = 0.86$$

From the table, $z = 1.08$

$$\therefore \quad x = z\sigma + \mu = 1.08(6) + 4 = 46.48$$

Ex6: An electrical firm manufactures light bulbs that have a life, before burn-out, that is normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a bulb burns between 778 and 834 hours.

Soln: Let X denote life of light bulb, it is normal r.v.

Given $\mu = 800$ hrs and $\sigma = 40$ hrs

Prob. that a bulb burns b/w $x_1 = 778$ hrs and $x_2 = 834$ hrs is given by

$$P(x_1 \leq X \leq x_2)$$

The z values corresponding to $x_1 = 778$ and $x_2 = 834$ are

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{778 - 800}{40} = -0.55 \text{ and } z_2 = \frac{834 - 800}{40} = 0.85$$

Hence,

$$P(778 \leq X \leq 834) = P(-0.55 \leq Z \leq 0.85)$$

$$= \phi(0.85) - \phi(-0.55)$$

$$= 0.8023 - 0.2912 = 0.5111.$$

Ex7: In an industrial process, the diameter of a ball bearing is an important measurement. The buyer sets specifications for the diameter to be 3.0 ± 0.01 cm.

The implication is that no part falling outside these specifications will be accepted. It is known that in the process the diameter of a ball bearing has a normal distribution with mean μ = 3.0 and standard deviation σ = 0.005. On average, how many manufactured ball bearings will be scrapped?

Soln: Let $x_1 = 3 - 0.01 = 2.99\,cm$, $x_2 = 3 + 0.01 = 3.01\,cm$.

Given $\mu = 3$ and $\sigma = 0.005$

The corresponding z values are

$$z_1 = \frac{2.99-3}{0.005} = -2, \quad z_2 = \frac{3.01-3}{0.005} = 2$$

$$\therefore P(x_1 \le X \le x_2) = P(-2 \le Z \le 2) = \phi(2) - \phi(-2)$$
$$= 0.97725 - 0.02275$$
$$= 0.9545$$

Thus,

$$P(X \le x_1) + P(X \ge x_2) = 1 - 0.9545 = 0.0455$$

On average, 4.55% of manufactured ball bearings will be scrapped.

Ex8: Gauges are used to reject all components for which a certain dimension is not within the specification 1.50 ± d. It is known that this measurement is normally distributed with mean 1.50 and standard deviation 0.2. Determine the value d such that the specifications "cover" 95% of the measurements.

Solns: Let X denote measurements of components.

Given $\mu = 1.5$ and $\sigma = 0.2$.

Find d such that $P(1.50 - d < X < 1.50 + d) = 0.95$

Suppose $x_1 = 1.50 - d$ and $x_2 = 1.50 + d$. The corresponding z values are

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{\cancel{1.50} - d - \cancel{1.5}}{0.2}$$

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{\cancel{1.50} + d - \cancel{1.5}}{0.2}$$

Clearly $z_2 = -z_1 = \frac{d}{0.2}$

Thus

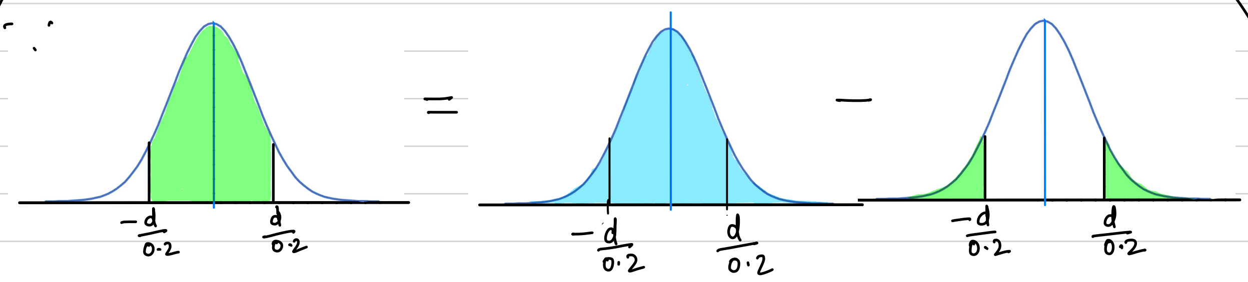
$$P(1.50 - d < X < 1.50 + d) = 0.95$$

$$\Rightarrow P\left(-\frac{d}{0.2} < Z < \frac{d}{0.2}\right) = 0.95$$

$$\Rightarrow 1 - 2\phi\left(-\frac{d}{0.2}\right) = 0.95$$

$$\Rightarrow \phi\left(-\frac{d}{0.2}\right) = 0.025 \Rightarrow -\frac{d}{0.2} = -1.96 \Rightarrow d = 0.392$$

$\because$



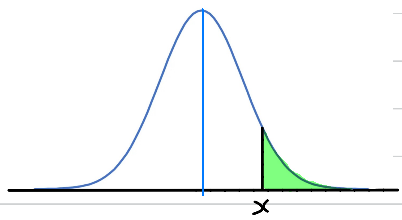
$$P\left(-\frac{d}{0.2} < Z < \frac{d}{0.2}\right) = 1 - 2\phi\left(-\frac{d}{0.2}\right)$$

Ex9: The average grade for an exam is 74, and the standard deviation is 7. If 12% of the class is given As, and the grades are curved to follow a normal distribution, what is the lowest possible A and the highest possible B? (Assume highest grade given is A)

Soln: Let X denote grade of an exam.

Given $\mu = 74$ and $\sigma = 7$.

Find x such that $P(X \geq x) = 0.12$.



The corresponding z value is

$$z = \frac{x - 74}{7} \Rightarrow x = 7z + 74$$

$$\therefore P(X \geq x) = 0.12$$

$$\Rightarrow P(Z \geq z) = 0.12$$

$$\Rightarrow 1 - \phi(z) = 0.12 \Rightarrow \phi(z) = 0.88$$

$$\Rightarrow z = 1.18$$

Thus, $x = 7(1.18) + 74 = 82.26$

$\therefore$ Lowest A is 83 and highest B is 82.

Ex 10: In a test on electric bulbs, it was found that the life time of a particular brand was distributed normally with an average life of 2000 hrs and std deviation of 60 hrs. If a firm purchases 2500 bulbs, find the no. of bulbs that are likely to last for

a) more than 2100 hrs b) between 1900 to 2100 hrs

Soln: Let X denote life time of an electric bulbs.

Given $\mu = 2000$ hrs and $\sigma = 60$ hrs

a) Let $x_1 = 2100$ hrs, the corresponding z value is

$$z = \frac{2100 - \mu}{\sigma} = \frac{2100 - 2000}{60} = 1.67$$

Prob. that a bulb last for more than 2100 hrs is

$$P(X > 2100) = P(Z > 1.6667)$$
$$= 1 - P(Z \leq 1.6667)$$
$$= 1 - \phi(1.67)$$
$$= 1 - 0.9525 = 0.0475$$

$\therefore$ # of bulbs lasts for more than 2100 hrs $= 0.0475 \times 2500$
$\simeq 119$

b) Let $x_1 = 1900$ and $x_2 = 2100$ hrs. The corres. z values are

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{1900 - 2000}{60} = -1.67 \text{ and } z_2 = \frac{2100 - 2000}{60} = 1.67$$

$$\therefore P(1900 < X < 2100) = P(-1.67 < Z < 1.67)$$
$$= \phi(1.67) - \phi(-1.67) = 0.905$$

$\therefore$ # of bulbs that are likely to last b/w 1900 and 2100 hrs

is $0.905 \times 2500 \simeq 2263$

Ex 11: In an examination 7% of students score less than 35% marks and 89% score less than 60% marks. Find the mean and std deviation if the marks are normally distributed.

Soln: Let X denote marks score by students.

Given, $P(X < 35) = 0.07$, $P(X < 60) = 0.89$

Suppose $x_1 = 35$ and $x_2 = 60$, the corres. z values are

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{35-\mu}{\sigma} \quad \text{——} (1) \quad \text{and} \quad z_2 = \frac{60-\mu}{\sigma} \quad \text{——} (2)$$

$$P(X < 35) = P(Z < z_1) = 0.07$$

$$\Rightarrow \phi(z_1) = 0.07$$

$$\Rightarrow z_1 = -1.47$$

$$P(X < 60) = P(Z < z_2) = 0.89$$

$$\Rightarrow \phi(z_2) = 0.89$$

$$\Rightarrow z_2 = 1.23$$

Sub. in (1) and (2)

$$\frac{35-\mu}{\sigma} = -1.47 \quad \Rightarrow \quad \mu - 1.47\sigma = 35 \quad \text{——} (3)$$

$$\frac{60-\mu}{\sigma} = 1.23 \quad \Rightarrow \quad \mu + 1.23\sigma = 60 \quad \text{——} (4)$$

Solving (3) and (4),

$$\sigma = 9.26 \quad \text{and} \quad \mu = 48.61$$

Extra information

Independent and identically distributed (i.i.d) random variables

Random variables $X_1, X_2, X_3, \cdots, X_n$ are said to be i.i.d if

a) All X_i's are independent from each other

b) Each X_i is drawn from same distribution.

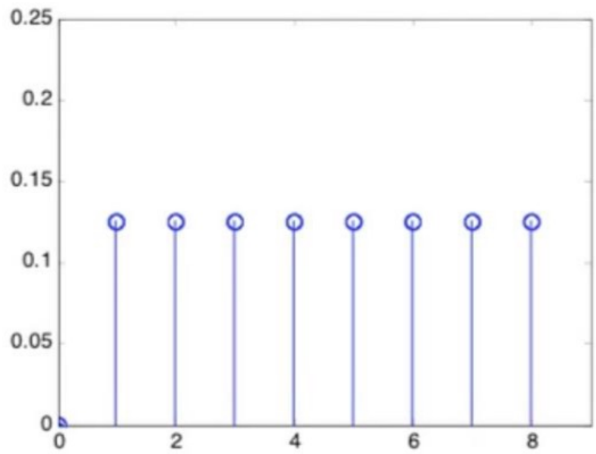
Probability distribution of the sum of i.i.d random variables

Roll an eight faced die.

Let $X = \{1, 2, 3, 4, 5, 6, 7, 8\}$ be a r.v. which represents possible outcomes.

If outcomes are equally likely, then PMF is

$P_X(x) = \frac{1}{8}$, $\forall x$. (Uniform distribution)



Two rolls of an eight faced die

Let $X_1 = \{1, 2, 3, 4, 5, 6, 7, 8\}$ be r.v. of 1st roll

$X_2 = \{1, 2, 3, 4, 5, 6, 7, 8\}$ be r.v. of 2nd roll.

Clearly, X_1 and X_2 are independent. The PMF of $Z = X_1 + X_2$,

We have $Z = \{2, 3, 4, 5, 6, 7, 8, \ldots, 15, 16\}$

$$P_Z(2) = P(X_1 = 1, X_2 = 1) = P(X_1 = 1)\, P(X_2 = 1) = \frac{1}{64}$$

$$p_Z(3) = P(X_1=1, X_2=2) + P(X_1=2, X_2=1) + P(X_1=3, X_2=0) + P(X_1=4, X_2=-1) + \cdots + P(X_1=16, X_2=-13)$$

$$= \sum_{x=1}^{16} P(X_1=x, X_2=3-x)$$

$$= \sum_{x=1}^{16} p_{X_1}(x)\, p_{X_2}(3-x)$$

$$= \frac{1}{64} + \frac{1}{64} + 0 + 0 + \cdots + 0 = \frac{1}{32}$$

$$p_Z(4) = \sum_{x=1}^{16} p_{X_1}(x)\, p_{X_2}(4-x) = \frac{3}{64}$$

$$p_Z(5) = \sum_{x=1}^{16} p_{X_1}(x)\, p_{X_2}(5-x) = \frac{4}{64}$$

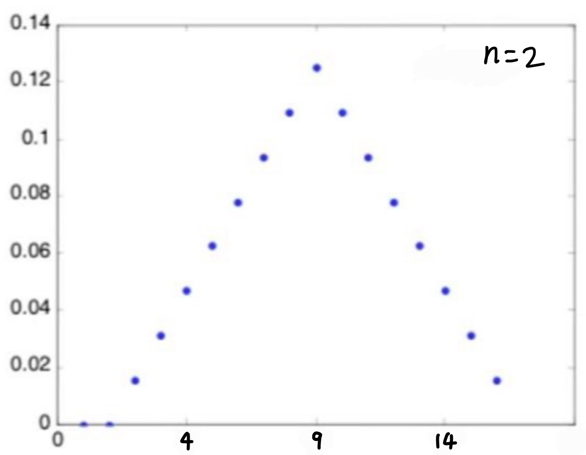
$$p_Z(6) = \sum_{x=1}^{16} p_{X_1}(x)\, p_{X_2}(6-x) = \frac{5}{64}$$

IIIly $p_Z(7) = \frac{6}{64}$, $p_Z(8) = \frac{7}{64}$, $p_Z(9) = \frac{8}{64}$, $p_Z(10) = \frac{7}{64}$)

$p_Z(11) = \frac{6}{64}$, $p_Z(12) = \frac{5}{64}$, $p_Z(13) = \frac{4}{64}$, $p_Z(14) = \frac{3}{64}$

$p_Z(15) = \frac{2}{64}$, $p_Z(16) = \frac{1}{64}$

Graph of $p_Z(z)$

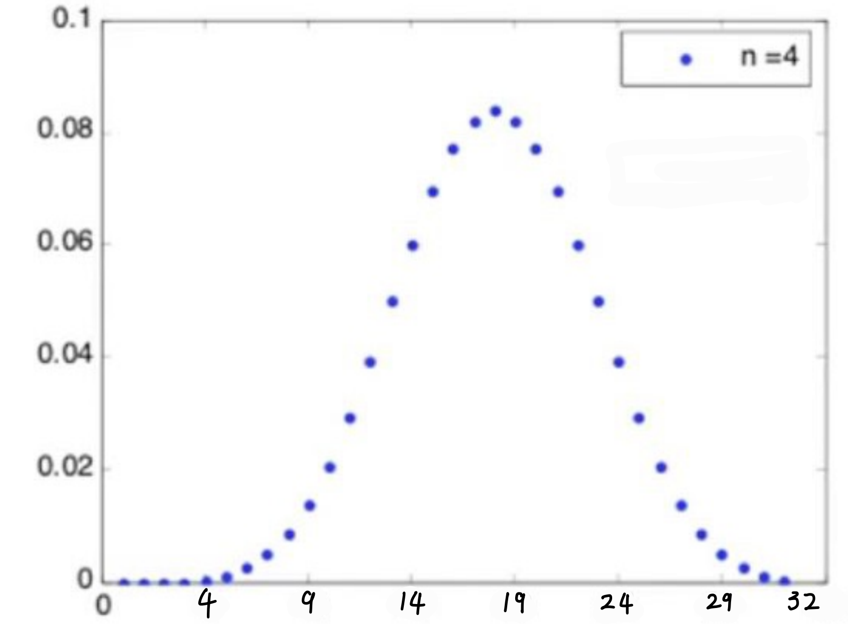


Four rolls of the die

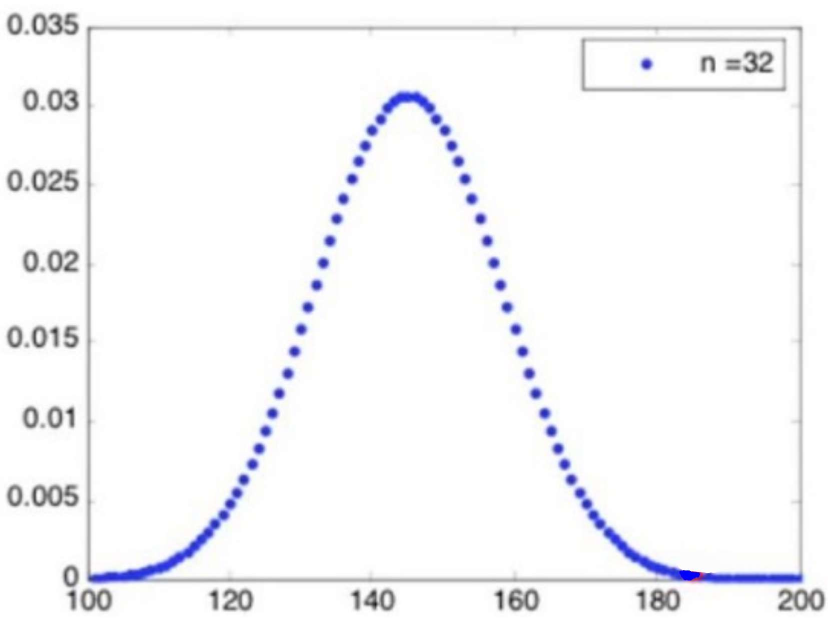
Let $X_i = \{1, 2, 3, 4, 5, 6, 7, 8\}$, $i = 1, 2, 3, 4$

For $Z = X_1 + X_2 + X_3 + X_4 = \{4, 5, 6, \cdots, 32\}$

The PMF graph is

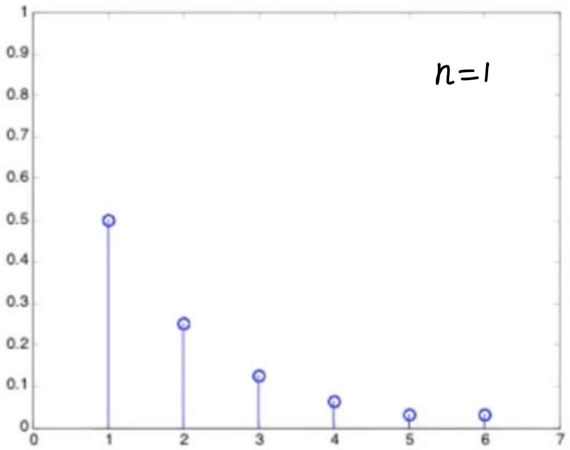


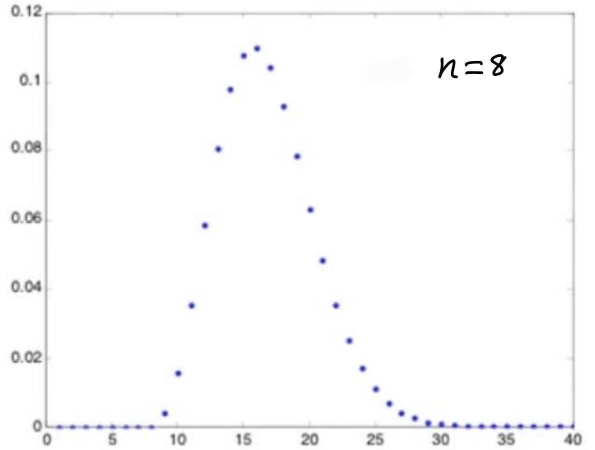
If $Z = X_1 + X_2 + \cdots + X_{32}$, then $p_Z(z)$ graph is

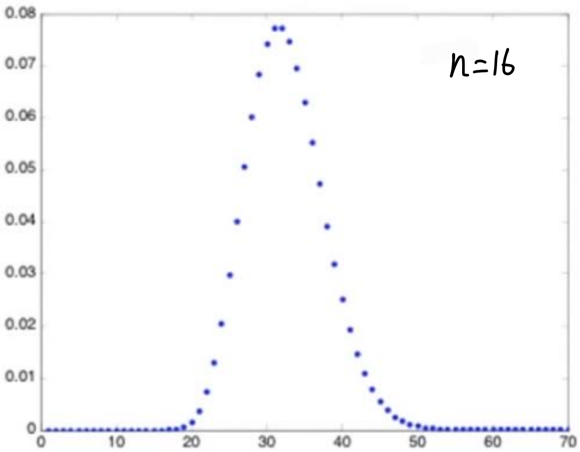


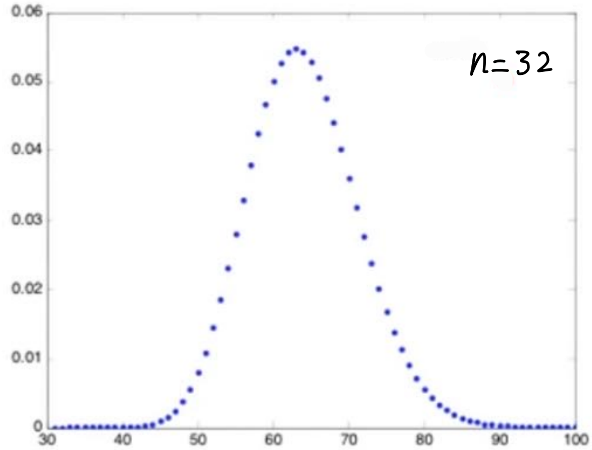
Consider distribution of the sum of geometric random variables

Let $Z = X_1 + X_2 + X_3 + \cdots + X_n$, where $P_{X_i}(x) = \frac{1}{2^x}$









Convolution

Let X and Y be i.i.d r.vs and $Z = X+Y$.

If X and Y are discrete : $p_Z(z) = \sum_{x \in X} p_X(x)\, p_Y(z-x)$

If X and Y are continuous : $f_Z(z) = \int_{x \in X} f_X(x)\, f_Y(z-x)\, dx$

Observations from above : If $X_1, X_2, \cdots, X_n$ are i.i.d r.vs, then the distribution of $X = \sum_{i=1}^{n} X_i$ is approximately normal distribution.

This is what we get from Central limit thm stated below.

Central limit theorem (CLT)

Let $X_1, X_2, X_3, \ldots, X_n$ be i.i.d random variables with finite mean μ and variance σ^2.

Consider the sum of the random variables

$$S_n = X_1 + X_2 + X_3 + \cdots + X_n$$

mean, $E[S_n] = n\mu$ and variance, $Var(S_n) = n\sigma^2$.

Then

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

is r.v. with $E[Z_n] = 0$ and $Var(Z_n) = 1$

Central limit Theorem states that

For every z: $\lim_{n\to\infty} P(Z_n \le z) = P(Z \le z)$

where Z is a standard normal r.v.

i.e. CDF of Z_n converges to standard normal CDF.

Note: a) From CLT we can say that

$S_n = \sum_{i=1}^{n} X_i$ is approximately a normal r.v.

b) If X_i's are normal, then S_n is normal for all n.

c) If X_i's are symmetric about μ, then S_n is normal for fairly small value of n.

d) In general, $S_n \sim$ normal when n is large. In practice $n \geq 30$ is considered to be large.

Normal approximation to the binomial and Poisson distribution

Let X be a binomial r.v. with parameters n and p. Then

$$X = X_1 + X_2 + X_3 + \cdots + X_n \quad \text{——— } (1)$$

where X_i's are Bernouli r.vs.

$$X_i = \{0, 1\}$$

Clearly, X_i's are independent and identically distributed,

$$p_{X_i}(x) = p^x (1-p)^{1-x}.$$

In View of CLT,

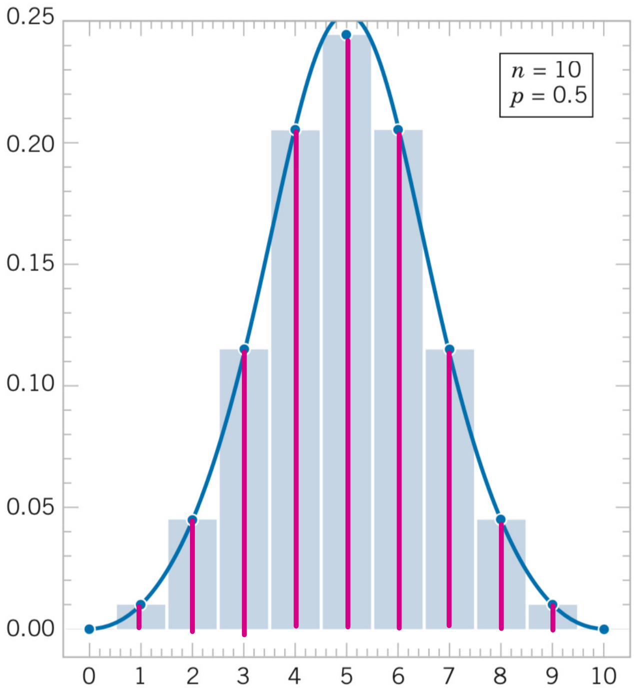
The binomial distribution of X with $E[X] = np$ and $Var(X) = np(1-p)$ approximates a normal distribution when n is large.

Further, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal distribution $n(z;0,1)$

Continuity correction:



Since a continuous normal distribution is used to approximate a discrete binomial distribution. The correction $+0.5$, called a continuity correction is used as follows.

$$P(X=x) = b(x;n,p)$$

$$\simeq \text{area under the normal curve from } x-0.5 \text{ to } x+0.5$$

$$= P\left(\frac{x-0.5-np}{\sqrt{np(1-p)}} \leq Z \leq \frac{x+0.5-np}{\sqrt{np(1-p)}}\right)$$

$$P(X\leq x) = \sum_{k=0}^{x} b(k;n,p)$$

$$\simeq \text{area under the normal curve to the left of } x+0.5$$

$$= P\left(Z \leq \frac{x+0.5-np}{\sqrt{np(1-p)}}\right)$$

$$P(a\leq X\leq b) = \sum_{k=a}^{b} b(k;n,p)$$

$$\simeq \text{area under the normal curve from } a-0.5 \text{ to } b+0.5$$

$$= P\left(\frac{a-0.5-np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b+0.5-np}{\sqrt{np(1-p)}}\right)$$

Note: 1) If p is close to 0 or 1, then binomial distribution is skewed so large n provide better normal approx.

2) If p is close to 0.5, the binomial distribution is almost symmetric so small n itself provided good normal approx.

Ex: Let X be a binomial r.v. with $n=50$ and $p=0.1$. Determine both exact and normal approx

i) $P(X\leq 2)$ ii) $P(X=5)$

Soln: Exact soln:

$$P(X\leq 2) = b(0;50,0.1) + b(1;50,0.1) + b(2;50,0.1)$$

$$= \binom{50}{0}(0.9)^{50} + \binom{50}{1}(0.1)(0.9)^{49} + \binom{50}{2}(0.1)^2(0.9)^{48}$$

$$= 0.112$$

$$P(X=5) = \binom{50}{5}(0.1)^5(0.9)^{45} = 0.1849$$

Normal approx:

$$P(X \le 2) = P\left(Z \le \frac{2.5 - np}{\sqrt{np(1-p)}}\right)$$

$$= p\left(Z \le \frac{2.5-5}{\sqrt{5 \times 0.9}}\right)$$

$$= P(Z \le -1.18)$$

$$= \phi(-1.18) = 0.119$$

$$P(X=5) = P(4.5 \le X \le 5.5)$$

$$= P\left(\frac{4.5-5}{\sqrt{5 \times 0.9}} \le Z \le \frac{5.5-5}{\sqrt{5 \times 0.9}}\right)$$

$$= P(-0.24 \le Z \le 0.24)$$

$$= \Phi(0.24) - \Phi(-0.24) = 0.19$$

Approx. is reasonable even for small $n=50$, when $p=0.1$.

Normal approximation to the Poisson distribution

If X is a poisson r.v. with $E[X]=\lambda$ and $Var(X)=\lambda$,

$$Z = \frac{X-\lambda}{\sqrt{\lambda}}$$

is approximately a standard normal r.v. The same continuity correction can be applied. The approx is good for $\lambda > 5$.

Sampling Theory

Defn: Population:

A population consists of the totality of the observations with which we are concerned.

- The number of observations in the population is defined to be the size of the population.
- Population size can be finite or infinite.

Ex: 1) If there are 600 students in a college and the type of blood is our interset, we say size of a population is 600.

2) The observations obtained by measuring the depth of a lake, from any conceivable position is an example of infinite size.

- Some finite population is so large that in theory we assume them to be infinite.

Mean and Variance of a population

If the size of a population is N and N observations of the population are denoted by $x_1, x_2, x_3, \ldots, x_N$, then

The population mean, $\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$

The population Variance, $\sigma^2 = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$

In the field of statistical inference, we are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. Therefore, we depend on a subset of observations from the population to help us make inference concerning the population.

Defn: Sample

A sample is a subset of a population.

In order to eliminate any possibility of bias in selecting a sample, it is desirable to choose random sample, i.e observations are made independently and at random.

Mean and Variance of a sample

Let $x_1, x_2, x_3, \ldots, x_n$ be a sample of n observations. Then

The sample mean, $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

The sample variance, $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Alternate form of s^2:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

We have,

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

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

Statistical methods are used to make decisions and draw conclusions about populations. This aspect of statistics is called statistical inference.

Statistical inference may be divided into two major areas.

i) Parameter estimation

ii) Hypothesis testing

Parameter and Statistic

A parameter is a numerical description of a population characteristic.

Some parameters of the population are

i) population mean, μ

ii) population standard deviation, σ

iii) proportion, p (items in a population that belong to a class of interest).

- Parameters are fixed values, usually, we don't know their exact values because it is impractical or impossible to measure the entire population.

- In general,

  the population parameter is denoted by θ.

A statistic is a numerical description of a sample characteristic.

Ex: i) sample mean, $\bar{x}$

ii) sample standard deviation, s

iii) proportion, $\hat{p}$ (items in a sample that belong to a class of interest).

- In general, a statistic for the parameter θ is denoted by $\hat{\theta}$.

- Statistics vary from sample to sample since they depend on the sample data collected.

Point estimate :

A number which provide a reasonable value (or good guess) for a true population parameter is called a Point estimate.

Statistics are used to estimate parameters when it is not feasible to measure the entire population.

For instance,

- A point estimate to the population mean μ is a statistic $\bar{x}$.
- Statistic $\hat{p}$ is a point estimate for the population proportion p.

This point estimate vary from sample to sample. Later, we use sampling distribution to find good and unbiased estimator to estimate the population parameter θ.

Now we discuss the notions of populations and samples in the context of the concept of random variables.

Recall that the totality of observations with which we are concerned constitute a population.

Let X denote a random variable, each observation in a population is a value of X having some probability distribution.

Ex: 1) In the blood-type experiment, the r.v. X represents the type of blood. There are 8 types of blood. So

$$X = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$$

Each student is given one of the values.

2) If our study is to determine the avg. life span of a certain brand of bulbs, the lifes of the bulbs are values assumed by continuous r.v. having perhaps a normal distribution.

Notation: When we refer population $f(x)$, we shall mean a population whose observations are values of r.v. having a distribution $f(x)$.

The mean and variance of a r.v. or probability distribution are also refered as the mean and variance of the corresponding population.

i.e, If X is a r.v. of population $f(x)$, then

$E[X] = \mu$, population mean and $Var(X) = \sigma^2$, population variance

Let us select a random sample of size n from a population $f(x)$. We define the r.v. X_i, $i = 1, 2, 3, \cdots, n$, to represent ith sample value that we observe.

The r.v.s $X_1, X_2, \cdots, X_n$ constitute a random sample with numerical values $x_1, x_2, \cdots x_n$ if the values are obtained independently under the same conditions.

Defn: Random Sample

The r.vs $X_1, X_2, \cdots, X_n$ are called a random sample of size n from the population $f(x)$ if

a) X_i's are independent r.vs, and

b) every X_i has the same probability distribution $f(x)$.

Note: The observed data are also referred to as a random sample, but the use of the same phrase should not cause any confusion.

Statistic

We discussed earlier that

- A sample mean $\bar{x}$ is a point estimate of the population parameter μ.
- We can say that $\bar{x}$ is a function of the observed values in the random sample.
- Because many samples are possible for the same population. We would expect $\bar{x}$ to vary from sample to sample.

That is, $\bar{x}$ is a value of a r.v. that we represent by $\bar{X}$.

IIIly s^2 and $\hat{p}$ are also values of a r.vs represented by S^2 and $\hat{P}$, respectively.

Such a random variable is called statistic.

Defn: Any function of the random variables constituting a random sample is called a statistic.

Defn: Sample mean and Sample Variance.

Let $X_1, X_2, X_3, \ldots, X_n$ be r.vs constituting a random sample of size n. Then

Sample mean, $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

Sample variance, $S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$

Sampling distribution

Since a statistic is a random variable, it has a probability distribution.

Defn: Sampling distribution

The probability distribution of a statistic is called a sampling distribution.

The sampling distribution of a statistic depends on

- The distribution of the population.
- The size of the samples.
- The method of choosing the samples.

It has applications to problems of Statistical inference.

Sampling distribution of the sample mean

The probability distribution of a sample mean $\bar{X}$ is called a sampling distribution of means.

Firstly, let us discuss the method to construct sampling distribution of mean from a finite population of size N.

1) Ramdomly draw all possible samples of size n.
2) Calculate mean for each sample
3) Summarize the mean obtained in step 2) in terms of relative frequency distribution | probability distribution.

Note: Random samples can be drawn with replacement or without replacement.

Ex1: Suppose we have a population of size $N=5$, consisting of the age of five children

6, 8, 10, 12 and 14.

Take sample of size 2

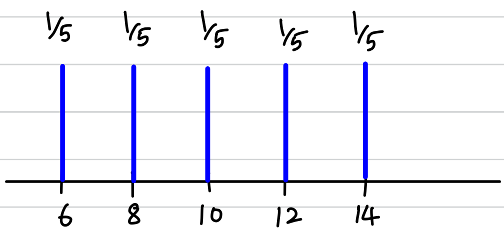
i) with replacement ii) without replacement

Construct sampling distribution of the sample mean.

Soln: Given, $X = \{6, 8, 10, 12, 14\}$ and population size, $N=5$

Population distribution: $p(x) = \frac{1}{5}, \ \forall x$

Graph:



We say that distribution is uniform

Population mean: $\mu = \frac{\sum x}{N} = \sum x\, p(x) = \frac{6+8+10+12+14}{5} = 10$

Population Variance: $\sigma^2 = \frac{\sum (x-\mu)^2}{N} = \sum (x-\mu)^2 p(x) = \frac{16+4+0+4+16}{5} = 8$

i) With replacement

We have total $N^n = 25$ possible samples

Sample values	Sample mean, $\bar{x}$
(6,6)	6
(6,8)	7
(6,10)	8
(6,12)	9
(6,14)	10
(8,6)	7
(8,8)	8
(8,10)	9
(8,12)	10

(8,14)	11
(10,6)	8
(10,8)	9
(10,10)	10
(10,12)	11
(10,14)	12
(12,6)	9
(12,8)	10
(12,10)	11
(12,12)	12
(12,14)	13
(14,6)	10
(14,8)	11
(14,10)	12
(14,12)	13
(14,14)	14

$\overline{X} = \{6, 7, 8, 9, 10, 11, 12, 13, 14\}$

$P(\overline{X}=6) = \frac{1}{25}$, $P(\overline{X}=7) = \frac{2}{25}$, $P(\overline{X}=8) = \frac{3}{25}$, $P(\overline{X}=9) = \frac{4}{25}$

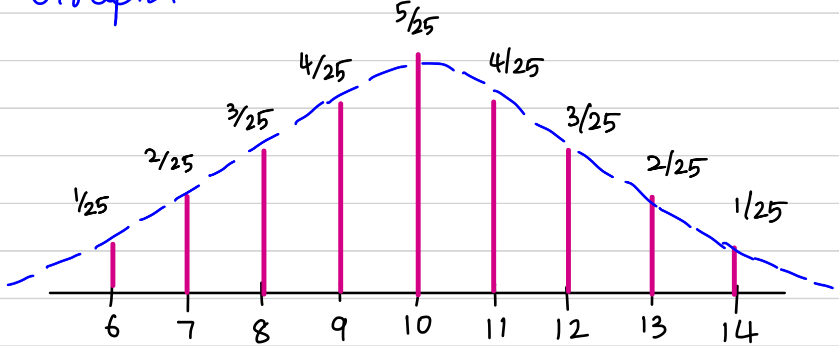
$P(\overline{X}=10) = \frac{5}{25}$, $P(\overline{X}=11) = \frac{4}{25}$, $P(\overline{X}=12) = \frac{3}{25}$, $P(\overline{X}=13) = \frac{2}{25}$

$P(\overline{X}=14) = \frac{1}{25}$

Sampling distribution of means:

$\bar{x}$	6	7	8	9	10	11	12	13	14
$p(\bar{x})$	1/25	2/25	3/25	4/25	5/25	4/25	3/25	2/25	1/25

Graph:



Mean of $\bar{X}$: $\mu_{\bar{X}} = \sum_{\bar{x}} \bar{x}\, p(\bar{x}) = 10 = \mu$ (Also, denoted by $E[\bar{X}]$)

Variance of $\bar{X}$: $\sigma_{\bar{X}}^2 = \sum (\bar{x} - \mu_{\bar{x}})^2\, p(\bar{x})$

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

$$= 104 - 100$$

$$= 4$$

$$= \frac{\sigma^2}{n}$$

ii) Without replacement:

Possible no. of samples : ${}^{N}C_n = {}^{5}C_2 = 10$

Sample values	Sample mean, $\bar{x}$
(6, 8)	7
(6, 10)	8
(6, 12)	9
(6, 14)	10
(8, 10)	9
(8, 12)	10
(8, 14)	11
(10, 12)	11
(10, 14)	12
(12, 14)	13

$\bar{X} = \{7, 8, 9, 10, 11, 12, 13\}$

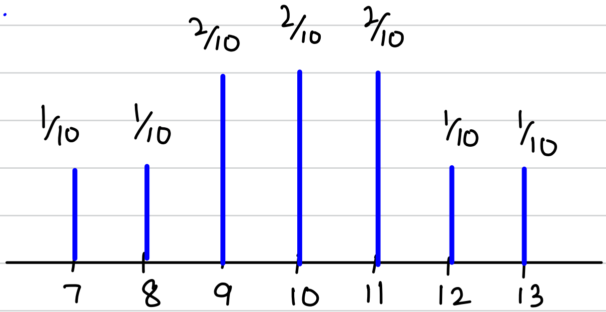
$P(\bar{X}=7) = \frac{1}{10}$, $P(\bar{X}=8) = \frac{1}{10}$, $P(\bar{X}=9) = \frac{2}{10}$, $P(\bar{X}=10) = \frac{2}{10}$

$P(\bar{X}=11) = \frac{2}{10}$, $P(\bar{X}=12) = \frac{1}{10}$, $P(\bar{X}=13) = \frac{1}{10}$

Sampling distribution of means

$\bar{x}$	7	8	9	10	11	12	13
$p(\bar{x})$	1/10	1/10	2/10	2/10	2/10	1/10	1/10

Graph:



Mean of $\bar{X}$: $\mu_{\bar{X}} = \sum_{\bar{x}} \bar{x}\, p(\bar{x}) = 10 = \mu$ (Also, denoted by $E[\bar{X}]$)

Variance of $\bar{X}$: $\sigma^2_{\bar{X}} = \sum (\bar{x} - \mu_{\bar{x}})^2 \, p(\bar{x})$

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

$$= 103 - 100$$

$$= 3$$

$$= \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

If sampling is with replacement:

Mean of $\bar{X}$, $\mu_{\bar{X}} = \mu \Rightarrow E[\bar{X}] = \mu$

Variance of $\bar{X}$, $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \Rightarrow Var(\bar{X}) = \frac{\sigma^2}{n}$

If sampling is without replacement:

Mean of $\bar{X}$, $\mu_{\bar{X}} = \mu \Rightarrow E[\bar{X}] = \mu$

Variance of $\bar{X}$, $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

when $N \to \infty$ or N is very large, then $\sigma^2_{\bar{X}} \approx \frac{\sigma^2}{n}$

Let $X_1, X_2, X_3, \cdots, X_n$ be r.vs constituting a random sample of size n taken from a population with mean μ and variance σ^2.

By CLT, the sample mean : $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ is approximately normal r.v. when the population is normal or n is sufficiently large with

mean : $\mu_{\overline{X}} = \frac{1}{n} \underbrace{(\mu + \mu + \cdots + \mu)}_{n \text{ times}} = \mu$

Variance : $\sigma_{\overline{X}}^2 = \frac{1}{n^2} \underbrace{\left(\sigma^2 + \sigma^2 + \cdots + \sigma^2\right)}_{n \text{ times}} = \dfrac{\sigma^2}{n}$

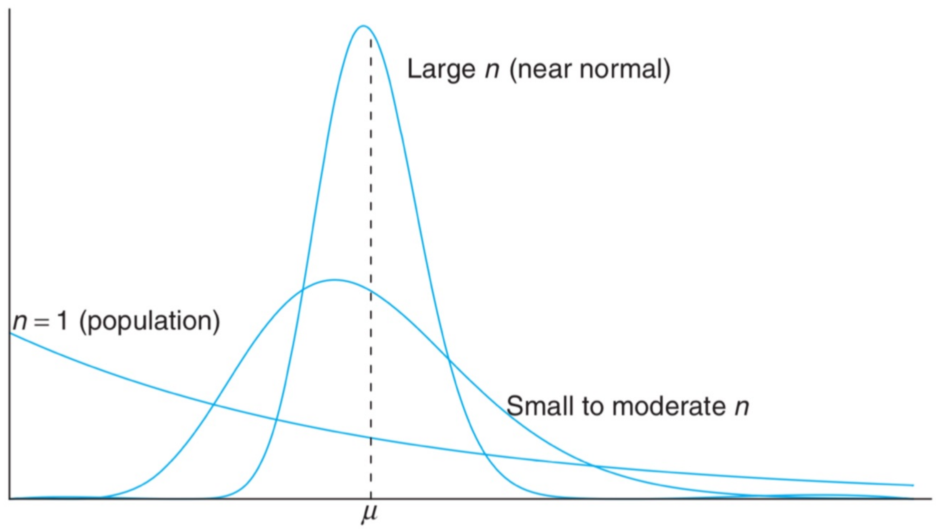
Standard deviation : $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ is also called Standard error of $\overline{X}$.

And the distribution of the r.v.

$$\boxed{Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}}$$

is standard normal distribution $n(z; 0, 1)$.

Distribution of $\overline{X}$ for $n=1$, moderate n, and large n



Ex2: An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of n = 25 resistors will have an average resistance less than 95 ohms.

Soln: Given, sample size $n=25$ and population distribution is normal with $\mu = 100\,\Omega$ and $\sigma = 10\,\Omega$

$\therefore$ The sampling distribution of $\bar{X}$ is normal, with

mean $\mu_{\bar{X}} = 100\,\Omega$

and variance $\sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n} = \dfrac{100}{25} = 4$

and standard deviation, $\sigma_{\bar{X}} = 2$

We have to find $P(\bar{X} < 95 \text{ ohms})$

Let $Z = \dfrac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \dfrac{\bar{X} - 100}{2}$

when $\bar{X} = 95$, then z-value is $z = \dfrac{95-100}{2} = -2.5$

$\therefore P(\bar{X} < 95) = P(Z < -2.5) = \Phi(-2.5) = 0.0062$

Practical Conclusion:

This example show that if the distribution of resistance is normal with mean 100 ohms and standard deviation of 10 ohms, then finding that a random sample of resistors with a sample mean smaller than 95 ohms is a rare event. If this actually happen, it caste doubt as to whether the true mean is really 100 ohms or if the true standard deviation is really 10 ohms.

Ex3: Suppose that a r.v X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{Otherwise} \end{cases}$$

Find the distribution of the sample mean of random sample of size $n = 40$.

Soln: Given distribution of X as

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$$

$$\therefore E[X] = \int_4^6 x\, f(x)\, dx = \int_4^6 x \cdot \tfrac{1}{2}\, dx = \frac{1}{4}(6^2 - 4^2)$$

$\Rightarrow$ mean, $\mu = 5$

$$Var(X) = E[X^2] - E[X]^2 = \int_4^6 x^2 \cdot \tfrac{1}{2}\, dx - 25$$

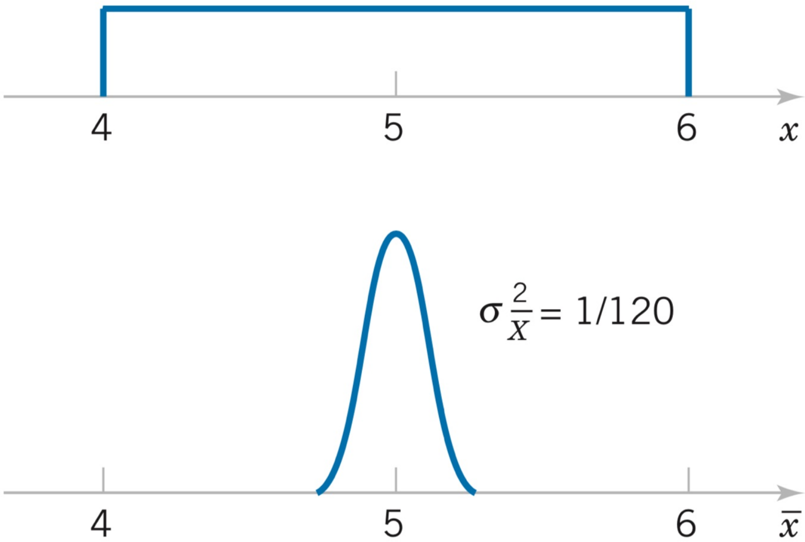
$\Rightarrow$ Variance, $\sigma^2 = \frac{1}{6}(6^3 - 4^3) - 25 = \frac{1}{3}$

Let $\overline{X}$ be the sample mean. Then by CLT distribution of $\overline{X}$ is approximately normal with

mean, $\mu_{\overline{X}} = 5$

Variance, $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n} = \dfrac{1}{3 \times 40} = \dfrac{1}{120}$ (given, $n = 40$)

Distributions of X and $\overline{X}$ are shown in the below fig.



Ex4 : An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Soln: Given that the population distribution is approximately normal with mean $\mu = 800$ hrs and standard deviation, $\sigma = 40$ hrs

$\therefore$ Sampling distribution of $\bar{X}$ will be approximately normal, with mean, $\mu_{\bar{X}} = 800$ hrs and Standard deviation, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{4} = 10$

(given $n = 16$)

The desired probability is

$P(\bar{X} < 775 \text{ hrs})$

The z-value with $\bar{x} = 775$ is $Z = \frac{775 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = -2.5$

$\therefore \quad P(\bar{X} < 775) = P(Z < -2.5) = \phi(-2.5) = 0.0062$

Ex5: The heights of 1000 students are approximately normally distributed with a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. Suppose 200 random samples of size 25 are drawn from this population. Determine

(a) the mean and standard deviation of the sampling distribution of $\bar{X}$;

(b) the number of sample means that fall between 172.5 and 175.8 centimeters inclusive;

(c) the number of sample means falling below 172.0 centimeters.

Soln: Given, the population mean $\mu = 174.5$ cm

population standard deviation $\sigma = 6.9$ cm

Sample size, $n = 25$

a) Let $\bar{X}$ be sample mean. The

mean of $\bar{X}$, $\mu_{\bar{X}} = \mu = 174.5$ cm

Variance of $\bar{X}$, $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{6.9^2}{25} = 1.9044\ cm^2$

SD of $\bar{X}$, $\sigma_{\bar{X}} = \sqrt{1.9044} = 1.38$ cm

b) No. of sample means that falls b/w 172.5 and 175.8

$= 200\ P(172.5 < \bar{X} < 175.8)$

$= 200\ P\left(\frac{172.5 - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < Z < \frac{175.8 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right)$

$= 200\ P\left(\frac{172.5 - 174.5}{1.38} < Z < \frac{175.8 - 174.5}{1.38}\right)$

$= 200\ P(-1.45 < Z < 0.94)$

$= 200\ (\phi(0.94) - \phi(-1.45))$

≈ 156

c) No. of sample means below 172 cm.

$$200\,P(\bar{X} < 172) = 200\,P\left(Z < \frac{172-174.5}{1.38}\right)$$

$$= 200\,P(Z < -1.81)$$

$$= 200 \times 0.035148$$

$$= 7$$

Sampling distribution of the differences of means

In some situations, we may be interested to draw the inference about the differences of two population means.
For example,

i) Two companies of bulbs are produced same type of bulbs and one may be interested to know which one is better.

ii) Two types of drugs, were tried on certain number of patients for controlling BP and one might be interested to know which one has better effect on controlling B.P

If we have two independent populations with means μ_1 and μ_2 and variance σ_1^2 and σ_2^2, and if $\overline{X}_1$ and $\overline{X}_2$ are the sample mean of two independent samples of sizes n_1 and n_2 from these populations, then the sampling distribution of the differences of means $\overline{X}_1 - \overline{X}_2$ is approximately normally distributed when n_1 and n_2 are sufficiently large, with

Mean : $\mu_{\overline{X}_1 - \overline{X}_2} = \mu_{\overline{X}_1} - \mu_{\overline{X}_2} = \mu_1 - \mu_2$

Variance: $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

Standard error of $\overline{X}_1 - \overline{X}_2 = \sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Hence,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

is approximately a standard normal variable.

Ex6 : The effective life of a component used in a jet-turbine aircraft engine is a random variable with mean 5000 hours and standard deviation 40 hours. The distribution of effective life is fairly close to a normal distribution.

The engine manufacturer introduces an improvement into the manufacturing process for this component that increases the mean life to 5050 hours and decreases the standard deviation to 30 hours. Suppose that a random sample of $n_1=16$ components is selected from the "old" process and a random sample of $n_2=25$ components is selected from the "improved" process. What is the probability that the difference in the two sample means $\bar{X}_2-\bar{X}_1$ is at least 25 hours? Assume that the old and improved processes can be regarded as independent populations.

Soln: Given:

Population 1	Population 2
$\mu_1 = 5000$	$\mu_2 = 5050$
$\sigma_1 = 40$	$\sigma_2 = 30$
$n_1 = 16$	$n_2 = 25$

Let $\bar{X}_1$ and $\bar{X}_2$ be sample mean of 1^{st} and 2^{nd} populatn respectively.

Then the sampling distribution of $\bar{X}_2-\bar{X}_1$ will be approx. normal with

$$\text{mean, } \mu_{\bar{X}_2-\bar{X}_1} = \mu_2-\mu_1 = 50$$

$$\text{Std error, } \sigma_{\bar{X}_2-\bar{X}_1} = \sqrt{\sigma^2_{\bar{X}_2}+\sigma^2_{\bar{X}_1}} = \sqrt{\frac{30^2}{25}+\frac{40^2}{16}} = \sqrt{136}$$

The desired prob.

$$P(\bar{X}_2-\bar{X}_1 \geq 25) = P\left(Z \geq \frac{25-\mu_{\bar{X}_2-\bar{X}_1}}{\sigma_{\bar{X}_2-\bar{X}_1}}\right)$$

$$= P\left(Z \geq \frac{25-50}{\sqrt{136}}\right)$$

$$= P(Z \geq -2.14)$$

$$= 1-\phi(-2.14) = 0.9838$$

Ex7 : The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Soln: Given:

Population A	Population B
$\mu_1 = 6.5$ yrs $\sigma_1 = 0.9$ yrs $n_1 = 36$	$\mu_1 = 6$ yrs $\sigma_1 = 0.8$ yrs $n_2 = 49$

To find $P(\bar{X}_1 - \bar{X}_2 \geq 1)$:

By CLT, distribution of $\bar{X}_1 - \bar{X}_2$ is approx. normal with

mean, $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 0.5$ yrs

Variance, $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{0.9^2}{36} + \frac{0.8^2}{49} = 0.03556$

Std error, $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{0.03556} = 0.189$

$\therefore P(\bar{X}_1 - \bar{X}_2 \geq 1) = P\left(Z \geq \frac{1 - \mu_{\bar{X}_1 - \bar{X}_2}}{\sigma_{\bar{X}_1 - \bar{X}_2}}\right)$ (Z is Std normal variable)

$= P(Z \geq 2.645)$

$= 1 - P(Z \leq 2.645)$

$= 1 - \phi(2.65)$

$= 1 - 0.996$

$= 0.004$

Sampling distribution of Sample proportion

When our interest is to study the proportion of population possess certain characteristics,

for ex:

i) the proportion of male in the population

ii) the proportion of heart patients admitted in hospitals.

In such situation, we study the inference about the population proportion.

Sampling distribution:

Let us consider the populalation $f(x)$ and let p be the proportion of the population possessing certain characteristic.

We define the r.V, $Y = \begin{cases} 1, & \text{if observation in the populatn possess certain charc.} \\ 0, & \text{otherwise} \end{cases}$

$E[Y] = p$ and $Var(Y) = p(1-p)$.

Consider a sample of size n.

Let X denote the r.v. which represents number of observations in the sample with the given characteristics.

Then $X = Y_1 + Y_2 + Y_3 + \dots + Y_n$, where $Y_i = \{0, 1\}$

and the proportion of sample with the given characteristics is $\hat{p} = \frac{X}{n}$.

mean of $\hat{p}$, $\mu_{\hat{p}} = \frac{E[X]}{n} = \frac{np}{n} = p$

Variance of $\hat{p}$, $\sigma^2_{\hat{p}} = \frac{Var(X)}{n^2} = \frac{p(1-p)}{n}$

Standard error of $\hat{p}$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

The distribution of $\hat{p}$ is approx. normal if n is large or $np \geq 5$ and $np(1-p) \geq 5$.

The distribution of

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is standard normal r.v.

Ex 8 : A machine produces a large number of items of which 15% are found to be defective. If a random sample of 200 items is taken from the population and sample proportion is calculated then find
(i) Mean and standard error of sampling distribution of proportion.
(ii) The probability that less than or equal to 12% defectives are found in the sample.

Soln : Given, population proportion, $p = 0.15$

Sample size, $n = 200$

Let $\hat{p}$ be the r.v which denote sample proportion. Then

i) mean of $\hat{p}$, $\mu_{\hat{p}} = p = 0.15$

Variance of $\hat{p}$, $\sigma^2_{\hat{p}} = \frac{p(1-p)}{n} = \frac{0.15\,(1-0.15)}{200} = 0.0006$

Standard error, $\sigma_{\hat{p}} = \sqrt{0.0006} = 0.025$

ii) The required probability:

$$P(\hat{p} \leq 0.12) = P\left(Z \leq \frac{0.12 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = P(Z \leq -1.2)$$

$$= \phi(-1.2)$$

$$= 0.1151$$

Ex 9 :

A recent study asked working adults if the'y worked most of their time remotely. The study found that 30% of employees spend the majority of their time working remotely. Suppose a sample of 150 working adults is taken.

1. What is the distribution of the sample proportion? Explain.
2. What is the mean and standard deviation of the sample proportion?
3. What is the probability that at most 27% of the workers in the sample work remotely most of the time?
4. What is the probability that at least 51 of the workers in the sample work remotely most of the time?
5. What is the probability that between 32% and 35% of the workers in the sample work remotely most of the time?

Soln: Population proportion $p = 0.3$

Sample size $n = 150$

1) Let $\hat{p}$ be the sample proportion. The distributn of $\hat{p}$ is approx. normal, since is large (also $np > 5$ and $n(1-p) > 5$)

2) mean, $\mu_{\hat{p}} = p = 0.3$

Variance, $\sigma^2_{\hat{p}} = \dfrac{p(1-p)}{n} = \dfrac{0.3 \times 0.7}{150} = 0.0014$

Standard error, $\sigma_{\hat{p}} = \sqrt{0.0014} = 0.0374$

3) $P(\hat{p} \le 0.27) = P\left(Z \le \dfrac{0.27 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right)$

$$= P\left(Z \le \frac{0.27 - 0.3}{0.0374}\right)$$

$$= P(Z \le -0.80)$$

$$= \phi(-0.80) = 0.2118$$

4) Proportion of worker work remotely is $\frac{51}{150} = 0.34$

$$P(\hat{P} \ge 0.34) = P\left(Z \ge \frac{0.34 - 0.3}{0.0374}\right)$$

$$= P(Z \ge 1.07)$$

$$= 1 - P(Z < 1.07)$$

$$= 1 - \phi(1.07) = 1 - 0.85769$$

$$= 0.1423$$

5) $P(0.32 < \hat{P} < 0.35) = P\left(\frac{0.32 - 0.3}{0.0374} < Z < \frac{0.35 - 0.3}{0.0374}\right)$

$$= P(0.53 < Z < 1.34)$$

$$= \phi(1.34) - \phi(0.53)$$

$$= 0.909877 - 0.701944$$

$$= 0.208$$