CHAPTER 4: Probabilistic Features of Certain Data Distribution (Probability Distributions)

4.1 Introduction:

The concept of random variables is very important in Statistics. Some events can be defined using random variables.

There are two types of random variables:

$$\text{Random variables} \begin{cases} \textit{Discrete Random Variables} \\ \textit{Continuous Random Variables} \end{cases}$$

4.2 Probability Distributions of Discrete Random Variables:

Definition:

The probability distribution of a discrete random variable is a table, graph, formula, or other device used to specify all possible values of the random variable along with their respective probabilities.

Examples of discrete r v.'s

- The no. of patients visiting KKUH in a week.
- The no. of times a person had a cold in last year.

Example:

Consider the following discrete random variable.

X = The number of times a Saudi person had a cold in January 2010.

Suppose we are able to count the no. of Saudis which $X = x$:

x (no. of colds a Saudi person had in January 2010)	Frequency of x (no. of Saudi people who had a cold x times in January 2010)
0	10,000,000
1	3,000,000
2	2,000,000
3	1,000,000
Total	$N =$ 16,000,000

Note that the possible values of the random variable X are:

$$x = 0, 1, 2, 3$$

Experiment: Selecting a person at random
Define the event:

$(X = 0)$ = The event that the selected person had no cold.
$(X = 1)$ = The event that the selected person had 1 cold.
$(X = 2)$ = The event that the selected person had 2 colds.
$(X = 3)$ = The event that the selected person had 3 colds.

In general:

$(X = x)$ =The event that the selected person had x colds.

For this experiment, there are $n(\Omega) = 16{,}000{,}000$ equally likely outcomes.
The number of elements of the event (X = x) is:

n(X=x) = no. of Saudi people who had a cold x times in January 2010.
= frequency of x.

The probability of the event (X = x) is:

$$P(X = x) = \frac{n(X = x)}{n(\Omega)} = \frac{n(X = x)}{16000000} \quad \text{, for x=0, 1, 2, 3}$$

x	freq. of x $n(X = x)$	$P(X = x) = \frac{n(X = x)}{16000000}$ (Relative frequency)
0	10000000	0.6250
1	3000000	0.1875
2	2000000	0.1250
3	1000000	0.0625
Total	16000000	1.0000

Note:

$$P(X = x) = \frac{n(X = x)}{16000000} = \text{Re}\,lative\ Frequency = \frac{frequency}{16000000}$$

The probability distribution of the discrete random variable X is given by the following table:

x	$P(X=x)=f(x)$
0	0.6250
1	0.1874
2	0.1250
3	0.0625
Total	1.0000

Note: The table may contain a missing value.

Notes:

- The probability distribution of any discrete random variable X must satisfy the following two properties:

  (1) $0 \leq P(X=x) \leq 1$

  (2) $\sum_{x} P(X=x) = 1$

- Using the probability distribution of a discrete r.v. we can find the probability of any event expressed in term of the r.v. X.

Example:

Consider the discrete r.v. X in the previous example.

x	$P(X=x)$
0	0.6250
1	0.1875
2	0.1250
3	0.0625
Total	1.0000

(1) $P(X \geq 2) = P(X=2) + P(X=3) = 0.1250 + 0.0625 = 0.1875$

(2) $P(X > 2) = P(X=3) = 0.0625$ [note: $P(X>2) \neq P(X \geq 2)$]

(3) $P(1 \leq X < 3) = P(X=1) + P(X=2) = 0.1875 + 0.1250 = 0.3125$

(4) $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

$= 0.6250 + 0.1875 + 0.1250 = 0.9375$

another solution:

$P(X \leq 2) = 1 - P((\overline{X \leq 2}))$

$= 1 - P(X > 2) = 1 - P(X=3) = 1 - 0.625 = 0.9375$

(5) $P(-1 \leq X < 2) = P(X=0) + P(X=1)$

$= 0.6250 + 0.1875 = 0.8125$

(6) $P(-1.5 \le X < 1.3) = P(X = 0) + P(X = 1) = 0.6250 + 0.1875 = 0.8125$

(7) $P(X = 3.5) = P(\phi) = 0$

(8) $P(X \le 10) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = P(\Omega) = 1$

(9) The probability that the selected person had at least 2 cold:

$$P(X \ge 2) = P(X = 2) + P(X = 3) = 0.1875$$

(10) The probability that the selected person had at most 2 colds:

$$P(X \le 2) = 0.9375$$

(11) The probability that the selected person had more than 2 colds:

$$P(X > 2) = P(X = 3) = 0.0625$$

(12) The probability that the selected person had less than 2 colds:

$$P(X < 2) = P(X = 0) + P(X = 1) = 0.8125$$

Graphical Presentation:

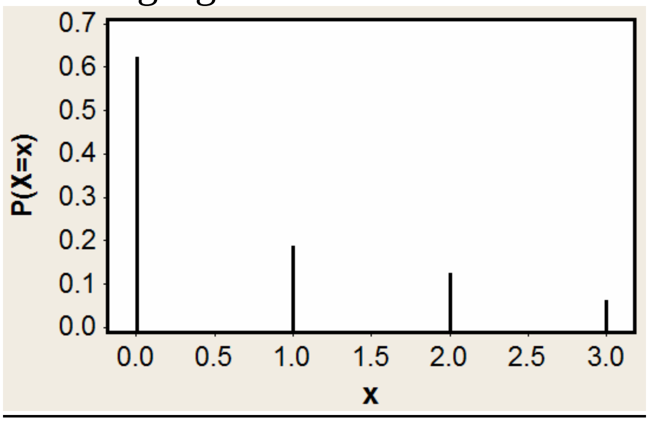
The probability distribution of a discrete r. v. X can be graphically represented.

Example:

The probability distribution of the random variable in the previous example is:

x	$P(X = x)$
0	0.6250
1	0.1875
2	0.1250
3	0.0625

The graphical presentation of this probability distribution is given by the following figure:



Find the Probability that the selected person had no cold in January 2010 ?

possible values of the random variable X are ....

Mean and Variance of a Discrete Random Variable

Mean: The mean (or expected value) of a discrete random variable X is denoted by μ or μ_X. It is defined by:

$$\mu = \sum_x x\, P(X = x)$$

Variance: The variance of a discrete random variable X is denoted by σ^2 or σ_X^2 . It is defined by:

$$\sigma^2 = \sum_x (x-\mu)^2 P(X = x)$$

Example:

We wish to calculate the mean μ and the variance of the discrete r. v. X whose probability distribution is given by the following table:

x	$P(X = x)$
0	0.05
1	0.25
2	0.45
3	0.25

Solution:

x	$P(X = x)$	$x\,P(X = x)$	~~$(x-\mu)$~~	~~$(x-\mu)^2$~~	~~$(x-\mu)^2 P(X = x)$~~
0	0.05	0	~~-1.9~~	~~3.61~~	~~0.1805~~
1	0.25	0.25	~~-0.9~~	~~0.81~~	~~0.2025~~
2	0.45	0.9	~~0.1~~	~~0.01~~	~~0.0045~~
3	0.25	0.75	~~1.1~~	~~1.21~~	~~0.3025~~
Total	1	$\mu =$ $\sum x\, P(X = x)$ $= 1.9$			~~$\sigma^2 =$~~ ~~$\sum (x-\mu)^2 P(X = x)$~~ ~~$= 0.69$~~

$$\mu = \sum_x x\ P(X = x) = (0)(0.05) + (1)(0.25) + (2)(0.45) + (3)(0.25) = 1.9$$

$$\sigma^2 = \sum_x (x - 1.9)^2\ P(X = x)$$

~~$= (0-1.9)^2(0.05) + (1-1.9)^2(0.25) + (2-1.9)^2(0.45) + (3-1.9)^2(0.25)$~~

$= 0.69$

Cumulative Distributions:

The cumulative distribution function of a discrete r. v. X is defined by:

$$P(X \le x) = \sum_{a \le x} P(X = a) \qquad \text{(Sum over all values } \le x\text{)}$$

Example:

Calculate the cumulative distribution of the discrete r. v. *X* whose probability distribution is given by the following table:

x	$P(X = x)$
0	0.05
1	0.25
2	0.45
3	0.25

Use the cumulative distribution to find:

P(X≤2), P(X<2), P(X≤1.5), P(X<1.5), P(X>1), P(X≥1)

Solution:

The cumulative distribution of *X* is:

x	$P(X \le x)$
0	0.05
1	0.30
2	0.75
3	1.0000

$$P(X \le 0) = P(X = 0)$$
$$P(X \le 1) = P(X = 0) + P(X = 1)$$
$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$P(X \le 3) = P(X = 0) + \cdots + P(X = 3)$$

Using the cumulative distribution,

P(X≤2) = 0.75

P(X<2) = P(X≤1) = 0.30

P(X≤1.5) = P(X≤1) = 0.30

P(X<1.5) = P(X≤1) = 0.30

P(X>1) = 1- P($\overline{(X > 1)}$) = 1-P(X≤1) = 1- 0.30 = 0.70

P(X≥1) = 1- P($\overline{(X \ge 1)}$) = 1-P(X<1) = 1- P(X≤0)

= 1- 0.05 = 0.95

Page 60

probability distribution $\longrightarrow$ cumulative distribution

x	$P(X = x)$
0	0.05
1	0.25
2	0.45
3	0.25

x	$P(X \le x)$
0	0.05
1	0.05+0.25= 0.30
2	0.05+0.25+0.45= 0.75
3	0.05+0.25+0.45+0.25=1

cumulative distribution $\longrightarrow$ probability distribution

x	$P(X \le x)$
0	0.05
1	0.30
2	0.75
3	1

x	$P(X = x)$
0	0.05
1	0.30-0.05= 0.25
2	0.75 -0.30 = 0.45
3	1 – 0.75 =0.25

Complement of probability:

- $P(X \le a) = 1 - P(X > a)$
- $P(X < a) = 1 - P(X \ge a)$

- $P(X \ge a) = 1 - P(X < a)$
- $P(X > a) = 1 - P(X \le a)$

In exercise

Example: (Reading Assignment)

Given the following probability distribution of a discrete random variable X representing the number of defective teeth of the patient visiting a certain dental clinic:

x	P(X = x)
1	0.25
2	0.35
3	0.20
4	0.15
5	K

a) Find the value of K.

b) Find the flowing probabilities:
   1. $P(X < 3)$
   2. $P(X \leq 3)$
   3. $P(X < 6)$
   4. $P(X < 1)$
   5. $P(X = 3.5)$

c) Find the probability that the patient has at least 4 defective teeth.

d) Find the probability that the patient has at most 2 defective teeth.

e) Find the expected number of defective teeth (mean of X).

f) Find the variance of X.

Solution:

a) $1=\sum P(X = x)=0.25+0.35+0.20+0.15+K$

$1=0.95+K$

$$K = 0.05$$

The probability distribution of X is:

x	P(X = x)
1	0.25
2	0.35
3	0.20
4	0.15
5	0.05
Total	1.00

b) Finding the probabilities:

1. $P(X < 3) = P(X=1)+P(X=2) = 0.25+0.35 = 0.60$
2. $P(X \leq 3) = P(X=1)+P(X=2)+P(X=3) = 0.8$
3. $P(X < 6) = P(X=1)+P(X=2)+P(X=3)+P(X=4)+P(X=5)= P(\Omega)=1$
4. $P(X < 1) = P(\phi)=0$
5. $P(X = 3.5) = P(\phi)=0$

c) The probability that the patient has at least 4 defective teeth

$$P(X\geq4) = P(X=4)+P(X=5) =0.15+0.05=0.2$$

d) The probability that the patient has at most 2 defective teeth

$$P(X\leq2) = P(X=1)+P(X=2) = 0.25+0.35=0.6$$

e) The expected number of defective teeth (mean of X)

x	P(X = x)	x P(X = x)
1	0.25	0.25
2	0.35	0.70
3	0.20	0.60
4	0.15	0.60
5	0.05	0.25
Total	$\sum P(X=x)=1$	$\mu=\sum xP(X=x)$=2.4

The expected number of defective teeth (mean of X) is

$$\mu=\sum xP(X=x)=(1)(0.25)+(2)(0.35)+(3)(0.2)+(4)(0.15)+(5)(0.05)=2.4$$

f) The variance of X:

x	$P(X=x)$	$(x-\mu)$	$(x-\mu)^2$	$(x-\mu)^2P(X=x)$
1	0.25	-1.4	1.96	0.49
2	0.35	-0.4	0.16	0.056
3	0.20	0.6	0.36	0.072
4	0.15	1.6	2.56	0.384
5	0.05	2.6	6.76	0.338
Total	$\mu=$ 2.4			$\sigma^2=\sum(x-\mu)^2P(X=x)$ = 1.34

The variance is $\sigma^2=\sum(x-\mu)^2P(X=x)$ = 1.34

Combinations: Notation (n!):

n! is read "n factorial". It defined by:

$$n! = n(n-1)(n-2)\cdots(2)(1) \qquad for \quad n \geq 1$$

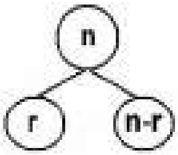
$$0! = 1$$

Example: $5! = (5)(4)(3)(2)(1) = 120$

Combinations:

The number of different ways for selecting r objects from n distinct objects is denoted by ${}_nC_r$ or $\binom{n}{r}$ and is given by:

$${}_nC_r = \frac{n!}{r!\ (n-r)!}; \qquad for \quad r = 0, 1, 2, \ldots, n$$



Notes:

1. ${}_nC_r$ is read as " n " choose " r ".
2. ${}_nC_n = 1$, ${}_nC_0 = 1$,
3. ${}_nC_r = {}_nC_{n-r}$ (for example: ${}_{10}C_3 = {}_{10}C_7$)
4. ${}_nC_r$ = number of unordered subsets of a set of (n) objects such that each subset contains (r) objects.

Example: For n = 4 and r = 2:

$${}_4C_2 = = \frac{4!}{2!\ (4-2)!} = \frac{4!}{2! \times 2!} = \frac{4\times3\times2\times1}{(2\times1)\times(2\times1)} = 6$$

${}_4C_2 = 6$ = The number of different ways for selecting 2 objects from 4 distinct objects.

Example: Suppose that we have the set {a, b, c, d} of (n=4) objects.

We wish to choose a subset of two objects. The possible subsets of this set with 2 elements in each subset are:

{a , b}, {a , c}, {a , d}, {b , d}, {b , c}, {c , d}

The number of these subsets is ${}_4C_2$ = 6.

4.3 Binomial Distribution:

- **Bernoulli Trial**: is an experiment with only two possible outcomes: S = success and F= failure (Boy or girl, Saudi or non-Saudi, sick or well, dead or alive).
- Binomial distribution is a discrete distribution.
- Binomial distribution is used to model an experiment for which:

1. The experiment has a sequence of n Bernoulli trials.
2. The probability of success is $P(S)=p$, and the probability of failure is $P(F)=1-p=q$.
3. The probability of success $P(S)=p$ is constant for each trial.
4. The trials are independent; that is the outcome of one trial has no effect on the outcome of any other trial.

In this type of experiment, we are interested in the discrete r. v. representing the number of successes in the n trials.

X = The number of successes in the n trials

The possible values of X (number of success in n trails) are:

$$x = 0, 1, 2, \dots, n$$

The r.v. X has a binomial distribution with parameters n and p , and we write:

$$X \sim Binomial(n, p)$$

The probability distribution of X is given by:

$$\boxed{P(X=x)=\begin{cases} {}_nC_x\ p^x\ q^{n-x} & for\ x=0,1,2,\dots,n \\ 0 & otherwise \end{cases}}$$

Where: $${}_nC_x = \frac{n!}{x!\,(n-x)!}$$

We can write the probability distribution of X as a table as follows.

x	$P(X=x)$
0	${}_nC_0\ p^0\ q^{n-0}=q^n$
1	${}_nC_1\ p^1\ q^{n-1}$

x	$P(X=x)$
2	${}_nC_2\ p^2\ q^{n-2}$
$\vdots$	$\vdots$
$n-1$	${}_nC_{n-1}\ p^{n-1}\ q^1$
n	${}_nC_n\ p^n\ q^0 = p^n$
Total	1.00

Result:

If X~ Binomial(n, p) , then

The mean: $\mu = np$ (expected value)

The variance: $\sigma_2 = npq$

Example: Suppose that the robability that a Saudi man has high blood pressure is 0.15. Suppose that we randomly select a sample of 6 Saudi men.

(1) Find the probability distribution of the random variable (X)representing the number of men with high blood pressure in the sample.

(2) Find the expected number of men with high blood pressure in the sample (mean of X).

(3) Find the variance X.

(4) What is the probability that there will be exactly 2 men with highblood pressure?

(5) What is the probability that there will be at most 2 men with high blood pressure?

(6)What is the probability that there will be at lease 4 men with highblood pressure?

Solution: We are interested in the following random variable:

X = The number of men with high blood pressure in the sample of 6 men.

Notes:

- Bernoulli trial: diagnosing whether a man has a high blood pressure or not. There are two outcomes for each trial:

S = Success: The man has high blood pressure
F = failure: The man does not have high blood pressure.

- Number of trials = 6 (we need to check 6 men)
- Probability of success: $P(S) = p$ = 0.15
- Probability of failure: $P(F) = q = 1 - p = 0.85$
- Number of trials: n = 6
- The trials are independent because of the fact that the result of each man does not affect the result of any other man since the selection was made ate random.

The random variable X has a binomial distribution with parameters: n=6 and p=0.15, that is:

$$X \sim \text{Binomial (n, p)}$$
$$X \sim \text{Binomial (6, 0.15)}$$

The possible values of X are: x = 0, 1, 2, 3, 4, 5, 6

(1) The probability distribution of X is:

$$P(X = x) = \begin{cases} {}_6C_x \ (0.15)^x (0.85)^{6-x} & ; x = 0,1,2,3,4,5,6 \\ 0 & ; \quad otherwise \end{cases}$$

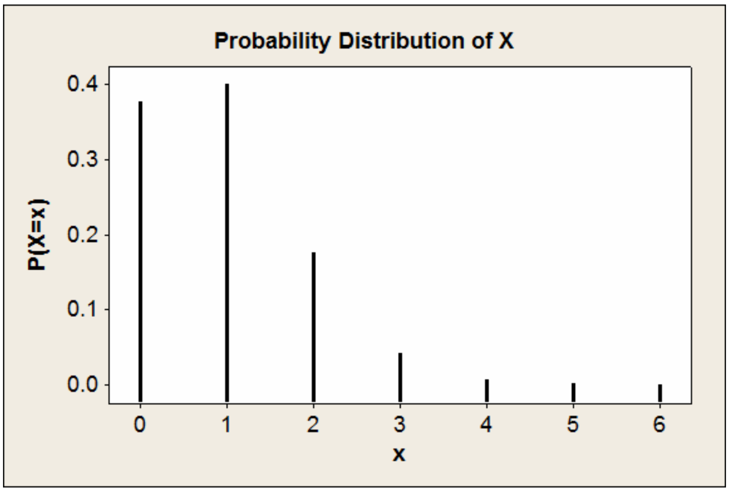
The probabilities of all values of X are:

$$P(X = 0) = {}_6C_0 \ (0.15)^0 (0.85)^6 = (1)(0.15)^0 \ (0.85)^6 = 0.37715$$
$$P(X = 1) = {}_6C_1 \ (0.15)^1 (0.85)^5 = (6)(0.15)(0.85)^5 = 0.39933$$
$$P(X = 2) = {}_6C_2 \ (0.15)^2 (0.85)^4 = (15)(0.15)^2 (0.85)^4 = 0.17618$$
$$P(X = 3) = {}_6C_3 \ (0.15)^3 (0.85)^3 = (20)(0.15)^3 (0.85)^3 = 0.04145$$
$$P(X = 4) = {}_6C_4 \ (0.15)^4 (0.85)^2 = (15)(0.15)^4 (0.85)^2 = 0.00549$$
$$P(X = 5) = {}_6C_5 \ (0.15)^5 (0.85)^1 = (6)(0.15)^5 (0.85)^1 = 0.00039$$
$$P(X = 6) = {}_6C_6 \ (0.15)^6 (0.85)^0 = (1)(0.15)^6 (1) = 0.00001$$

The probability distribution of X can by presented by the following table:

x	$P(X=x)$
0	0.37715
1	0.39933
2	0.17618
3	0.04145
4	0.00549
5	0.00039
6	0.00001
Total	1

The probability distribution of X can by presented by the following graph:



(2) The mean of the distribution (the expected number of men out of 6 with high blood pressure) is:

$$\mu = np = (6)(0.15) = 0.9$$

(3) The variance is:

$$\sigma^2 = npq = (6)(0.15)(0.85) = 0.765$$

(4) The probability that there will be exactly 2 men with high blood pressure is:

$$P(X = 2) = 0.17618$$

(5) The probability that there will be at most 2 men with high blood pressure is:

$$\begin{aligned} P(X \leq 2) &= P(X=0) + P(X=1) + P(X=2) \\ &= 0.37715 + 0.39933 + 0.17618 \\ &= 0.95266 \end{aligned}$$

(6) The probability that there will be at lease 4 men with high blood pressure is:

$$\begin{aligned} P(X \geq 4) &= P(X=4) + P(X=5) + P(X=6) \\ &= 0.00549 + 0.00039 + 0.00001 \\ &= 0.00589 \end{aligned}$$

Example: (Reading Assignment)

Suppose that 25% of the people in a certain population have low hemoglobin levels. The experiment is to choose 5 people at random from this population. Let the discrete random variable X be the number of people out of 5 with low hemoglobin levels.

a) Find the probability distribution of X.
b) Find the probability that at least 2 people have low hemoglobin levels.
c) Find the probability that at most 3 people have low hemoglobin levels.
d) Find the expected number of people with low hemoglobin levels out of the 5 people.
e) Find the variance of the number of people with low hemoglobin levels out of the 5 people

.

Solution: X = the number of people out of 5 with low hemoglobin levels

The Bernoulli trail is the process of diagnosing the person

Success = the person has low hemoglobin
Failure = the person does not have low hemoglobin

$n = 5$ (no. of trials)
$p = 0.25$ (probability of success)
$q = 1 - p = 0.75$ (probability of failure)

a) X has a binomial distribution with parameter $n = 5$ and $p = 0.25$

$$X \sim Binomial(n, p)$$
$$X \sim Binomial(5, 0.25)$$

The possible values of X are: x=0, 1, 2, 3, 4, 5

The probability distribution is:

$$P(X = x) = \begin{cases} {}_nC_x \; p^x \; q^{n-x} \; ; & for\, x = 0, 1, 2, \ldots, n \\ 0 \; ; & otherwise \end{cases}$$

$$P(X = x) = \begin{cases} {}_5C_x \; (0.25)^x \; (0.75)^{5-x} \; ; & for \;\; x = 0, 1, 2, 3, 4, 5 \\ 0 \; ; & otherwise \end{cases}$$

x	P(X = x)
0	${}_5C_0 \times 0.25^0 \times 0.75^{5-0} \quad = 0.23730$
1	${}_5C_1 \times 0.25^1 \times 0.75^{5-1} \quad = 0.39551$
2	${}_5C_2 \times 0.25^2 \times 0.75^{5-2} \quad = 0.26367$
3	${}_5C_3 \times 0.25^3 \times 0.75^{5-3} \quad = 0.08789$
4	${}_5C_4 \times 0.25^4 \times 0.75^{5-4} \quad = 0.01465$
5	${}_5C_5 \times 0.25^5 \times 0.75^{5-5} \quad = 0.00098$
Total	$\sum P(X=x)=1$

b) The probability that at least 2 people have low hemoglobin levels:

P(X≥2) = P(X=2)+P(X=3)+P(X=4)+P(X=5)
= 0.26367+ 0.08789+ 0.01465+ 0.00098
= 0. 0.36719

c) The probability that at most 3 people have low hemoglobin levels:

P(X≤3) = P(X=0)+P(X=1)+P(X=2)+P(X=3)
= 0.23730+ 0.39551+ 0.26367+ 0.08789
= 0.98437

d) The expected number of people with low hemoglobin levels out of the 5 people (the mean of X):

$$\mu = n\,p = 5\times 0.25 = 1.25$$

e) The variance of the number of people with low hemoglobin levels out of the 5 people (the variance of X) is:

$$\sigma^2 = n\,pq = 5\times 0.25\times 0.75 = 0.9375$$

4.4 The Poisson Distribution:

- It is a discrete distribution.
- The Poisson distribution is used to model a discrete r. v. representing the number of occurrences of some random event in an interval of time or space (or some volume of matter).
- The possible values of X are: x= 0, 1, 2, 3, …
- The discrete r. v. X is said to have a Poisson distribution with parameter (average or mean) λ if the probability distribution of X is given by

$$P(X=x)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{x}}{x!} & ; \quad for \quad x=0,1,2,3,\ldots \\ 0 & ; \quad otherwise\end{cases}$$

where e = 2.71828 . We write :

X ~ Poisson (λ)

Result: (Mean and Variance of Poisson distribution)

If X ~ Poisson (λ), then:

- The mean (average) of X is : $\mu=\lambda$ (Expected value)
- The variance of X is: $\sigma^2=\lambda$

Stander deviation = sqrt(lambda)

Example:

Some random quantities that can be modeled by Poisson distribution:

- No. of patients in a waiting room in an hours.
- No. of surgeries performed in a month.
- No. of rats in each house in a particular city.

Note:

- λ is the average (mean) of the distribution.
- If X = The number of patients seen in the emergency unit in a day, and if X ~Poisson (λ), then:
  1. The average (mean) of patients seen every day in the emergency unit = λ.
  2. The average (mean) of patients seen every month in the emergency unit =30λ.
  3. The average (mean) of patients seen every year in the emergency unit = 365λ.
  4. The average (mean) of patients seen every hour in the emergency unit = λ/24.

Also, notice that:

(i) If Y = The number of patients seen every month, then:

$Y \sim$ Poisson (λ^*), where $\lambda^* = 30\lambda$

(ii) W = The number of patients seen every year, then:

$W \sim$ Poisson (λ^*), where $\lambda^* = 365\lambda$

(iii) V = The number of patients seen every hour, then:

$V \sim$ Poisson (λ^*), where $\lambda^* = \frac{\lambda}{24}$

Example: Suppose that the number of snake bites cases seen at KKUH in year has a Poisson distribution with average 6 bite cases.

(1) What is the probability that in a year:

(i) The no. of snake bite cases will be 7?

(ii) The no. of snake bite cases will be less than 2?

(2) What is the probability that there will be 10 snake bite cases in 2 years?

(3) What is the probability that there will be no snake bite cases in a month?

There are additional questions !!

Solution:

(1) X = no. of snake bite cases in a year.

$X \sim$ Poisson (6) $\quad (\lambda=6)$

$$P(X=x) = \frac{e^{-6}6^x}{x!}\ ; \quad x = 0, 1, 2, \ldots$$

(i) $$P(X=7) = \frac{e^{-6}6^7}{7!} = 0.13768$$

(ii) $$P(X<2) = P(X=0) + P(X=1)$$

$$= \frac{e^{-6}6^0}{0!} + \frac{e^{-6}6^1}{1!} = 0.00248 + 0.01487 = 0.01735$$

(2) Y = no of snake bite cases in 2 years

$Y \sim$ Poisson(12) $\quad (\lambda^* = 2\lambda = (2)(6) = 12)$

$$P(Y=y) = \frac{e^{-12}12^y}{y!} : \quad y = 0\ ,1\ ,2 \ldots$$

$$\therefore P(Y=10) = \frac{e^{-12}12^{10}}{10!} = 0.1048$$

(3) W = no. of snake bite cases in a month.

$W \sim$ Poisson (0.5) $\quad (\lambda^{**} = \frac{\lambda}{12} = \frac{6}{12} = 0.5)$

$$P(W = w) = \frac{e^{-0.5}(0.5)^{w}}{w!}: \quad w = 01,2, \ldots .$$

$$P(W = 0) = \frac{e^{-0.5}(0.5)^{0}}{0!} = 0.6065$$

Extra questions (**Page 71)**:

(4) Find the probability that there will be more than or equal one snake bite cases in a month_ $\lambda^{*} = \frac{\lambda}{12} = \frac{6}{12} = 0.5$

$$P(X \geq 1) = 1 - P(x < 1)$$

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

$$= 1 - \frac{e^{-0.5}(0.5)^{0}}{0!} = 1 - 0.6065 = 0.3935$$

(5) The mean of snake bite cases in a year

$$\mu = \lambda = 6$$

(6) The variance of snake bite cases in a month

$$\sigma^{2} = \lambda^{*} = \frac{\lambda}{12} = \frac{6}{12} = 0.5$$

(7) The standard deviation of snake bite cases in 2 years

$$\sigma = \sqrt{\lambda^{*}} = \sqrt{2\lambda} = \sqrt{2(6)} = \sqrt{12} = 3.4641$$

(8) Find the probability that there will be more than

3 snake bite cases in 2 years $\lambda^* = 2\lambda = 2(6) = 12$

$$P(x > 3) = 1 - P(X \leq 3)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$

$$= 1 - 0.0023$$

$$= 0.9977$$

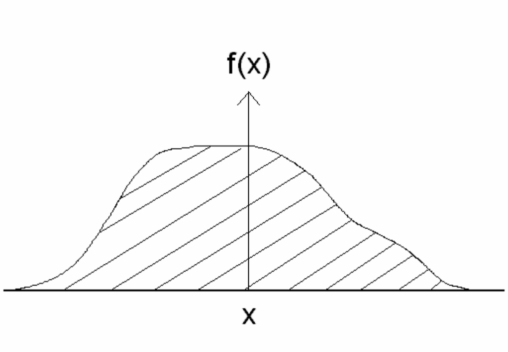
By calculator:

$$\sum_{x=0}^{3} \left(\frac{e^{-12} \times 12^{x}}{x!} \right) = 2.29 \times 10^{-3} = 0.00229 \approx 0.0023$$

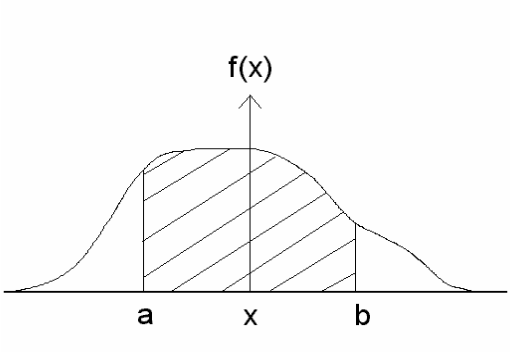
4.5 Continuous Probability Distributions:

For any continuous r. v. X, there exists a function $f(x)$, called the probability density function (pdf) of X , for which:

(1) The total area under the curve of $f(x)$ equals to 1.



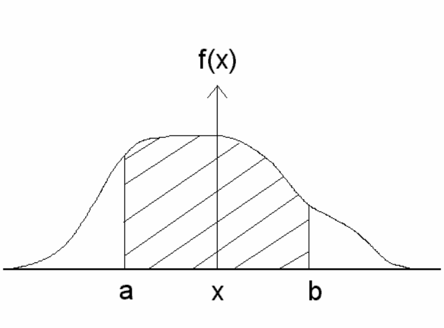
$Total\ area = \int_{-\infty}^{\infty} f(x)\, dx = 1$



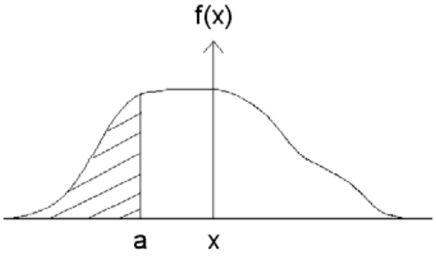
$P(a \le X \le b) = \int_{a}^{b} f(x)dx = area$

(2) The probability hat X is between the points (a) and (b) equals to the area under the curve of f(x) which is bounded by the point a and b.

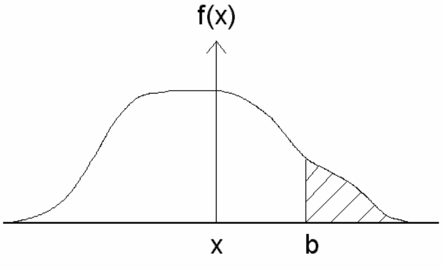
(3) In general, the probability of an interval event is given by the area under the curve of $f(x)$ and above that interval.



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



$P(X \le a) = \int_{-\infty}^{a} f(x)dx = area$

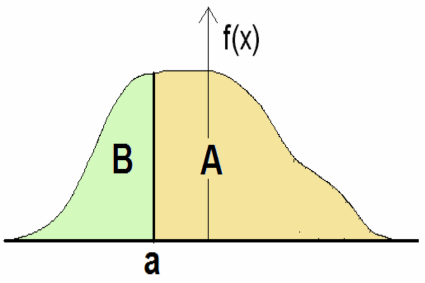


$P(X \ge b) = \int_{b}^{\infty} f(x)\, dx = area$

a

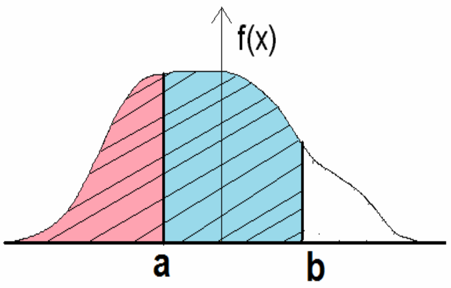
Note: If X is continuous r.v. then:

1. $P(X = a) = 0$ for any a.
2. $P(X \le a) = P(X < a)$
3. $P(X \ge b) = P(X > b)$
4. $P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = P(a < X < b)$
5. $P(X \le x) =$ cumulative probability
6. $P(X \ge a) = 1 - P(X < a) = 1 - P(X \le a)$
7. $P(a \le X \le b) = P(X \le b) - P(X \le a)$



$P(X \ge a) = 1 - P(X \le a)$

$A = 1 - B$, Total area = 1



$P(a \le X \le b) = P(X \le b) - P(X \le a)$

$$\int_a^b f(x)dx = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx$$

4.6 The Normal Distribution:

- One of the most important continuous distributions.
- Many measurable characteristics are normally or approximately normally distributed.

(Examples: height, weight, …)

- The probability density function of the normal distribution is given by:

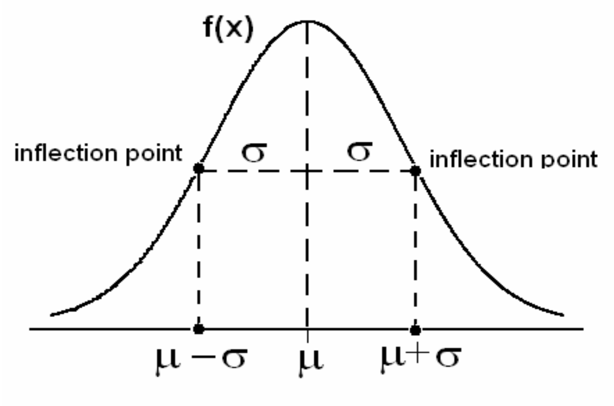
$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad ; -\infty < x < \infty}$$

where (e=2.71828) and (π=3.14159).

The parameters of the distribution are the mean (μ) and the standard deviation (σ).

- The continuous r.v. X which has a normal distribution has several important characteristics:

1. $-\infty < X < \infty$,
2. The density function of X , $f(x)$, has a bell-Shaped curve:



mean = μ

standard deviation = σ

variance = σ^2

3. The highest point of the curve of f(x) at the mean μ . (Mode = μ)
4. The curve of f(x) is symmetric about the mean μ .

   μ = mean = mode = median
5. The normal distribution depends on two parameters:

   mean = μ (determines the location)

   standard deviation = σ (determines the shape)
6. If the r.v. X is normally distributed with mean μ and standard deviation σ (variance σ^2), we write:

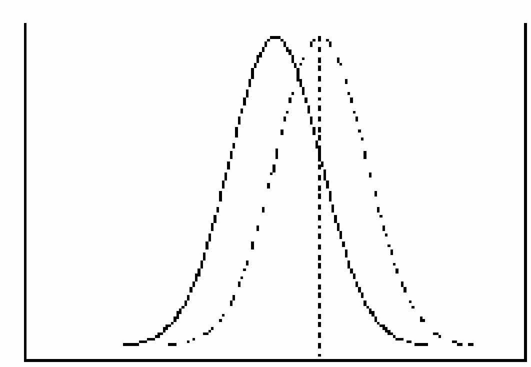
   $X \sim \text{Normal}\left(\mu, \sigma^2\right)$ or $X \sim \text{N}\left(\mu, \sigma^2\right)$
7. The location of the normal distribution depends on μ . The shape of the normal distribution depends on σ.

Note: The location of the normal distribution depends on μ and its shape depends on σ.

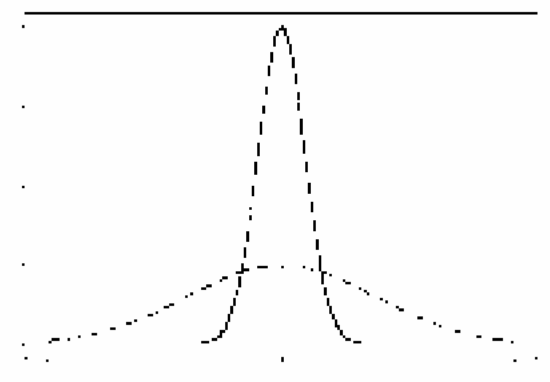
Suppose we have two normal distributions:

________ $N(\mu_1, \sigma_1)$

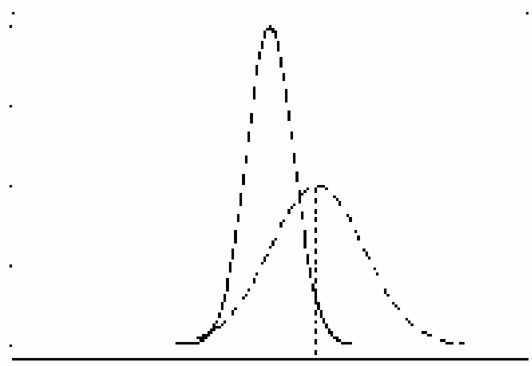
----------- $N(\mu_2, \sigma_2)$



$\mu_1 < \mu_2, \sigma_1=\sigma_2$



$\mu_1 = \mu_2, \sigma_1 < \sigma_2$



$\mu_1 < \mu_2, \sigma_1 < \sigma_2$

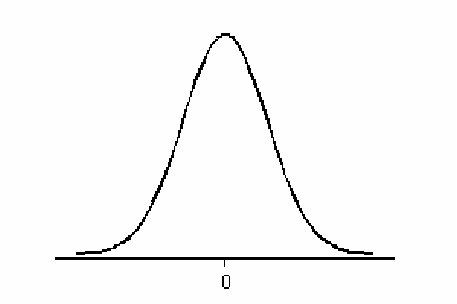
The Standard Normal Distribution:

The normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the standard normal distribution and is denoted by Normal (0,1) or N(0,1). The standard normal random variable is denoted by (Z), and we write:

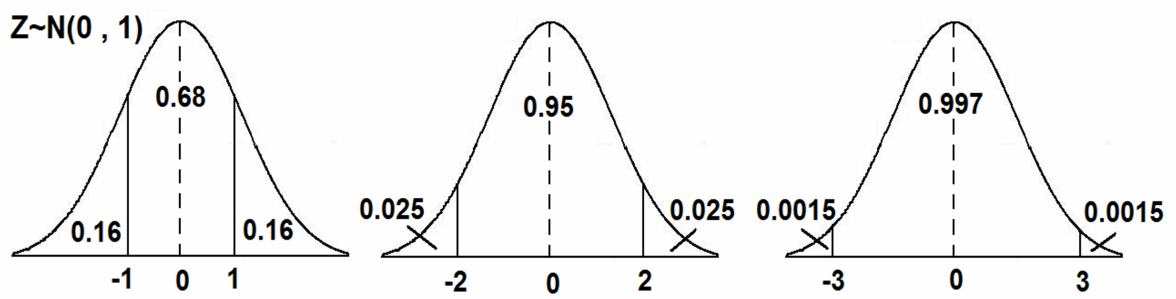
$$Z \sim N(0, 1)$$

The probability density function (pdf) of Z~N(0,1) is given by:

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



The standard normal distribution, Normal (0,1), is very important because probabilities of any normal distribution can be calculated from the probabilities of the standard normal distribution.



68% of the area is between -1 and 1 (approximately)

95% of the area is between -2 and 2 (approximately)

99.7% of the area is between -3 and 3 (approximately)

Result:

If $X \sim$ Normal μ, σ^2 , then $Z = \frac{X -}{} \sim$ Normal (0,1).

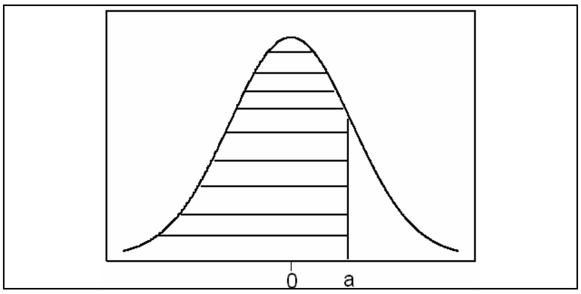
Calculating Probabilities of Normal (0,1):

Suppose Z ~ Normal (0,1).

For the standard normal distribution Z ~ N(0,1), there is a special table used to calculate probabilities of the form:

$$P(Z \leq a)$$

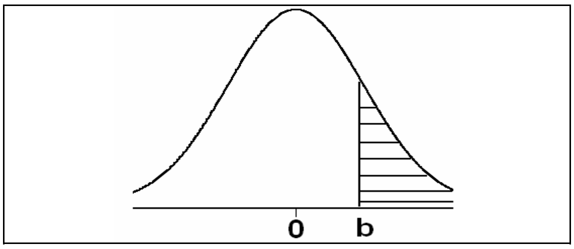
(i) $P(Z \leq a) =$ From the table



(ii) $P(Z \geq b) = 1 - P(Z \leq b)$

Where:

$P(Z \leq b \quad) =$ From the table

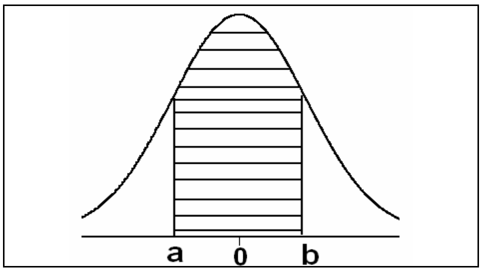


(iii) $P(a \leq Z \leq b) = P(Z \leq b) - P(z \leq a)$

Where:

$P(Z \leq b)$ = from the table

$P(z \leq a)$ = from the table

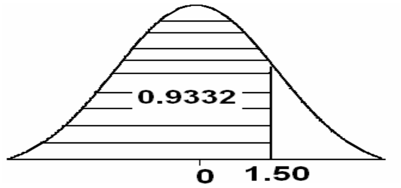


(iv) $P(Z=a)=0$ for every a .

Example:

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

(1) $P(Z \le 1.50)=0.9332$



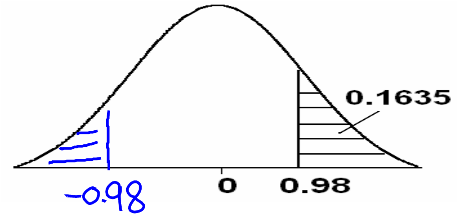
Z	0.00	0.01	…
:	⇓		
1.50 ⇒	0.9332		
:			

(2)

$P(Z \ge 0.98)=1-P(Z \le 0.98)$

$=1-0.8365$

$=0.1635$



Z	0.00	…	0.08
:	:	:	⇓
:	…	…	⇓
0.90⇒	⇒	⇒	0.8365

or

P(Z>0.98)=P(Z< -0.98)= 0.1635

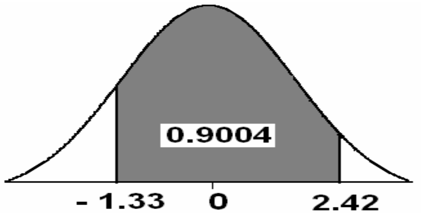
(3)

$P(-1.33 \le Z \le 2.42)=$

$P(Z \le 2.42)-P(Z \le -1.33)$

$=0.9922-0.0918$

$=0.9004$

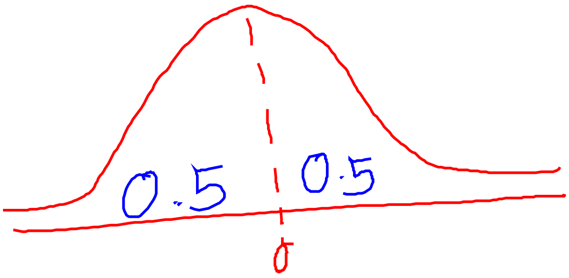


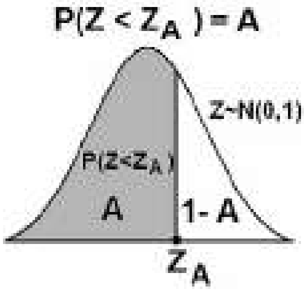
Z	…		-0.03
:	:		⇓
-1.30	⇒		0.0918
:			

(4) $P(Z \le 0)=P(Z \ge 0)=0.5$

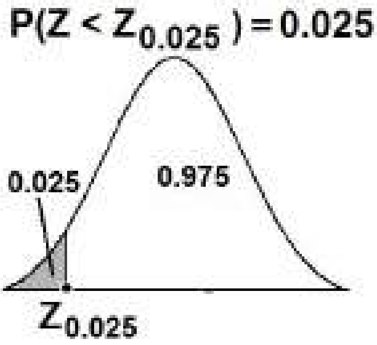
Notation:

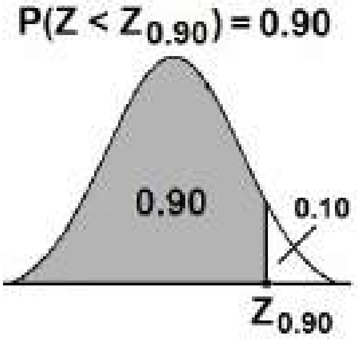
$P(Z \le Z_A)=A$





For example:





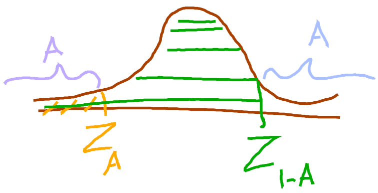
Result:

Since the pdf of Z~N(0,1) is symmetric about 0, we have:

$$Z_A = -\,Z_{1-A}$$

For example: $Z_{0.35} = -\,Z_{1-0.35} = -\,Z_{0.65}$

$Z_{0.86} = -\,Z_{1-0.86} = -\,Z_{0.14}$



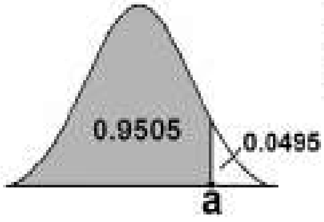
Example:

Suppose that Z ~ N(0,1).

If $P(Z \le a) = 0.95053$

Then $a = 1.65$

Z	…	0.05	…
:		⇑	
1.60	⇐	0.9505	
:			



$P(Z < a) = 0.9505$

$P(Z < Z_{0.9505}) = 0.9505$

$a = Z_{0.9505}$

$a = Z_{0.9505} = 1.65$

Example:

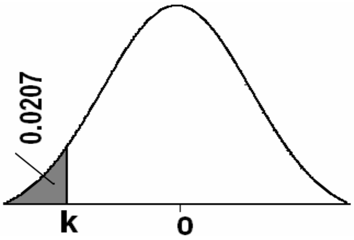
Suppose that Z~N(0,1). Find the value of *k* such that P(Z≤*k*)= 0.0207.

Solution:

.*k* = –2.04

Notice that *k*= $Z_{0.0207}$ = –2.04

Z	…	-0.04	
:	:	⇑ ⇑	
–2.0	⇐⇐	0.0207	
:			



Example:

If Z ~ N(0,1), then:

$Z_{0.90}$ = 1.285

$Z_{0.95}$ = 1.645

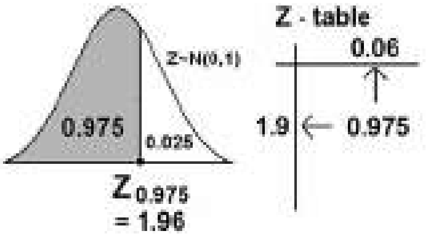
$Z_{0.975}$ = 1.96

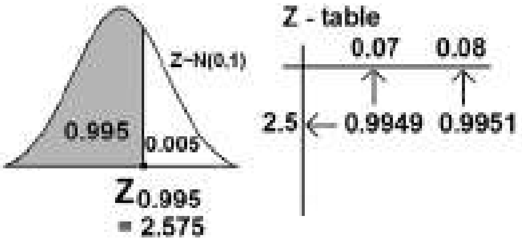
$Z_{0.99}$ = 2.325

$$Z_{0.90} = (Z_{0.89973} + Z_{0.90147})/2 = (1.28 + 1.29)/2 = 1.285$$

$$Z_{.95} = (Z_{0.94950} + Z_{0.95053})/2 = (1.64 + 1.65)/2 = 1.645$$

$$Z_{0.99} = (Z_{0.98983} + Z_{0.99010})/2 = (2.32 + 2.33)/2 = 2.325$$





Using the result: $Z_A = -\,Z_{1-A}$

$Z_{0.10}$ = – $Z_{0.90}$ = – 1.285

$Z_{0.05}$ = – $Z_{0.95}$ = – 1.645

$Z_{0.025}$ = – $Z_{0.975}$ = – 1.96

$Z_{0.01}$ = – $Z_{0.99}$ = – 2.325

Calculating Probabilities of Normal $\left(\mu,\sigma^2\right)$:

- Recall the result:

$$X \sim \text{Normal}\ \left(\mu,\sigma^2\right) \quad \Leftrightarrow \quad Z = \frac{X-\mu}{\sigma} \sim \text{Normal (0,1)}$$

- $X \le a \Leftrightarrow \frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma} \Leftrightarrow Z \le \frac{a-\mu}{\sigma}$

1. $P(X \le a) = P\left(Z \le \frac{a-\mu}{\sigma}\right)$ = From the table.
2. $P(X \ge a) = 1 - P(X \le a) = 1 - P\left(Z \le \frac{a-\mu}{\sigma}\right)$
3. $P(a \le X \le b) = P(X \le b) - P(X \le a)$
   $= P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right)$
4. $P(X = a) = 0$, for every a.

4.7 Normal Distribution Application:

Example

Suppose that the hemoglobin levels of healthy adult males are approximately normally distributed with a mean of 16 and a variance of 0.81.

(a) Find that probability that a randomly chosen healthy adult male has a hemoglobin level less than 14.

(b) What is the percentage of healthy adult males who have hemoglobin level less than 14?

(c) In a population of 10,000 healthy adult males, how many would you expect to have hemoglobin level less than 14?

Solution:

X = hemoglobin level for healthy adults males

Mean: $\mu = 16$

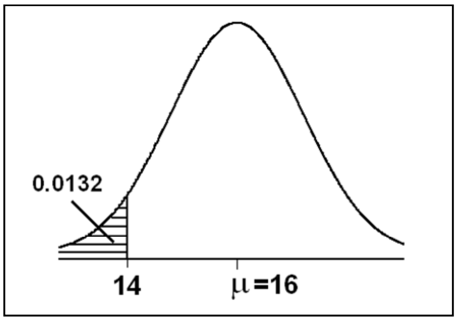
Variance: $\sigma^2 = 0.81$

Standard deviation: $\sigma = 0.9$

$X \sim$ Normal (16, 0.81)

(a) The probability that a randomly chosen healthy adult male has hemoglobin level less than 14 is $P(X \le 14)$.

$$
\begin{aligned}
P(X \le 14) &= P\left(Z \le \frac{14-\mu}{\sigma}\right) \\
&= P\left(Z \le \frac{14-16}{0.9}\right) \\
&= P(Z \le -2.22) \\
&= 0.0132
\end{aligned}
$$



(b) The percentage of healthy adult males who have hemoglobin level less than 14 is:

$$P(X \le 14) \times 100\,\% = 0.0132 \times 100\,\% = 1.32\,\%$$

(c) In a population of 10000 healthy adult males, we would expect that the number of males with hemoglobin level less than 14 to be:

$$P(X \le 14) \times 10000 = 0.0132 \times 10000 = 132 \ \textit{males}$$

Example:

Suppose that the birth weight of Saudi babies has a normal distribution with mean μ=3.4 and standard deviation σ=0.35.

(a) Find the probability that a randomly chosen Saudi baby has a birth weight between 3.0 and 4.0 kg.

(b) What is the percentage of Saudi babies who have a birth weight between 3.0 and 4.0 kg?

(c) In a population of 100000 Saudi babies, how many would you expect to have birth weight between 3.0 and 4.0 kg?

Solution:

X = birth weight of Saudi babies

Mean: μ = 3.4

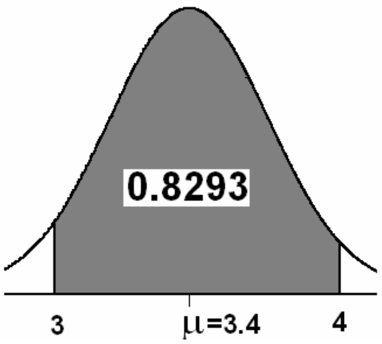
Standard deviation: σ = 0.35

Variance: $\sigma^2 = (0.35)^2 = 0.1225$

X ~ Normal (3.4, 0.1225)

(a) The probability that a randomly chosen Saudi baby has a birth weight between 3.0 and 4.0 kg is $P(3.0 < X < 4.0)$

$$
\begin{aligned}
P(3.0 < X < 4.0) &= P(X \le 4.0) - P(X \le 3.0) \\
&= P\left(Z \le \frac{4.0-\mu}{\sigma}\right) - P\left(Z \le \frac{3.0-\mu}{\sigma}\right) \\
&= P\left(Z \le \frac{4.0-3.4}{0.35}\right) - P\left(Z \le \frac{3.0-3.4}{0.35}\right) \\
&= P(Z \le 1.71) - P(Z \le -1.14) \\
&= 0.9564 - 0.1271 = 0.8293
\end{aligned}
$$



(b) The percentage of Saudi babies who have a birth weight between 3.0 and 4.0 kg is

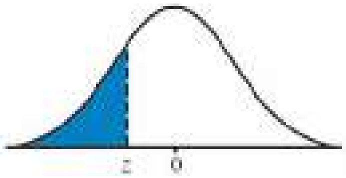
P(3.0<X<4.0) × 100%= 0.8293× 100%= 82.93%

(c) In a population of 100,000 Saudi babies, we would expect that the number of babies with birth weight between 3.0 and 4.0 kg to be:

P(3.0<X<4.0) × 100000= 0.8293× 100000= 82930 babies

Standard Normal Table

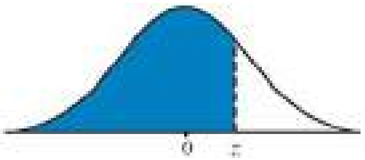
Areas Under the Standard Normal Curve



z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00	z
-3.50	0.00017	0.00017	0.00018	0.00019	0.00019	0.00020	0.00021	0.00022	0.00022	0.00023	**-3.50**
-3.40	0.00024	0.00025	0.00026	0.00027	0.00028	0.00029	0.00030	0.00031	0.00032	0.00034	**-3.40**
-3.30	0.00035	0.00036	0.00038	0.00039	0.00040	0.00042	0.00043	0.00045	0.00047	0.00048	**-3.30**
-3.20	0.00050	0.00052	0.00054	0.00056	0.00058	0.00060	0.00062	0.00064	0.00066	0.00069	**-3.20**
-3.10	0.00071	0.00074	0.00076	0.00079	0.00082	0.00084	0.00087	0.00090	0.00094	0.00097	**-3.10**
-3.00	0.00100	0.00104	0.00107	0.00111	0.00114	0.00118	0.00122	0.00126	0.00131	0.00135	**-3.00**
-2.90	0.00139	0.00144	0.00149	0.00154	0.00159	0.00164	0.00169	0.00175	0.00181	0.00187	**-2.90**
-2.80	0.00193	0.00199	0.00205	0.00212	0.00219	0.00226	0.00233	0.00240	0.00248	0.00256	**-2.80**
-2.70	0.00264	0.00272	0.00280	0.00289	0.00298	0.00307	0.00317	0.00326	0.00336	0.00347	**-2.70**
-2.60	0.00357	0.00368	0.00379	0.00391	0.00402	0.00415	0.00427	0.00440	0.00453	0.00466	**-2.60**
-2.50	0.00480	0.00494	0.00508	0.00523	0.00539	0.00554	0.00570	0.00587	0.00604	0.00621	**-2.50**
-2.40	0.00639	0.00657	0.00676	0.00695	0.00714	0.00734	0.00755	0.00776	0.00798	0.00820	**-2.40**
-2.30	0.00842	0.00866	0.00889	0.00914	0.00939	0.00964	0.00990	0.01017	0.01044	0.01072	**-2.30**
-2.20	0.01101	0.01130	0.01160	0.01191	0.01222	0.01255	0.01287	0.01321	0.01355	0.01390	**-2.20**
-2.10	0.01426	0.01463	0.01500	0.01539	0.01578	0.01618	0.01659	0.01700	0.01743	0.01786	**-2.10**
-2.00	0.01831	0.01876	0.01923	0.01970	0.02018	0.02068	0.02118	0.02169	0.02222	0.02275	**-2.00**
-1.90	0.02330	0.02385	0.02442	0.02500	0.02559	0.02619	0.02680	0.02743	0.02807	0.02872	**-1.90**
-1.80	0.02938	0.03005	0.03074	0.03144	0.03216	0.03288	0.03362	0.03438	0.03515	0.03593	**-1.80**
-1.70	0.03673	0.03754	0.03836	0.03920	0.04006	0.04093	0.04182	0.04272	0.04363	0.04457	**-1.70**
-1.60	0.04551	0.04648	0.04746	0.04846	0.04947	0.05050	0.05155	0.05262	0.05370	0.05480	**-1.60**
-1.50	0.05592	0.05705	0.05821	0.05938	0.06057	0.06178	0.06301	0.06426	0.06552	0.06681	**-1.50**
-1.40	0.06811	0.06944	0.07078	0.07215	0.07353	0.07493	0.07636	0.07780	0.07927	0.08076	**-1.40**
-1.30	0.08226	0.08379	0.08534	0.08691	0.08851	0.09012	0.09176	0.09342	0.09510	0.09680	**-1.30**
-1.20	0.09853	0.10027	0.10204	0.10383	0.10565	0.10749	0.10935	0.11123	0.11314	0.11507	**-1.20**
-1.10	0.11702	0.11900	0.12100	0.12302	0.12507	0.12714	0.12924	0.13136	0.13350	0.13567	**-1.10**
-1.00	0.13786	0.14007	0.14231	0.14457	0.14686	0.14917	0.15151	0.15386	0.15625	0.15866	**-1.00**
-0.90	0.16109	0.16354	0.16602	0.16853	0.17106	0.17361	0.17619	0.17879	0.18141	0.18406	**-0.90**
-0.80	0.18673	0.18943	0.19215	0.19489	0.19766	0.20045	0.20327	0.20611	0.20897	0.21186	**-0.80**
-0.70	0.21476	0.21770	0.22065	0.22363	0.22663	0.22965	0.23270	0.23576	0.23885	0.24196	**-0.70**
-0.60	0.24510	0.24825	0.25143	0.25463	0.25785	0.26109	0.26435	0.26763	0.27093	0.27425	**-0.60**
-0.50	0.27760	0.28096	0.28434	0.28774	0.29116	0.29460	0.29806	0.30153	0.30503	0.30854	**-0.50**
-0.40	0.31207	0.31561	0.31918	0.32276	0.32636	0.32997	0.33360	0.33724	0.3409	0.34458	**-0.40**
-0.30	0.34827	0.35197	0.35569	0.35942	0.36317	0.36693	0.37070	0.37448	0.37828	0.38209	**-0.30**
-0.20	0.38591	0.38974	0.39358	0.39743	0.40129	0.40517	0.40905	0.41294	0.41683	0.42074	**-0.20**
-0.10	0.42465	0.42858	0.43251	0.43644	0.44038	0.44433	0.44828	0.45224	0.45620	0.46017	**-0.10**
-0.00	0.46414	0.46812	0.47210	0.47608	0.48006	0.48405	0.48803	0.49202	0.49601	0.50000	**-0.00**

Standard Normal Table (continued)

Areas Under the Standard Normal Curve



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586	**0.00**
0.10	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535	**0.10**
0.20	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409	**0.20**
0.30	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173	**0.30**
0.40	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793	**0.40**
0.50	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240	**0.50**
0.60	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490	**0.60**
0.70	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524	**0.70**
0.80	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327	**0.80**
0.90	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891	**0.90**
1.00	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214	**1.00**
1.10	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298	**1.10**
1.20	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147	**1.20**
1.30	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774	**1.30**
1.40	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189	**1.40**
1.50	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408	**1.50**
1.60	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449	**1.60**
1.70	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327	**1.70**
1.80	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062	**1.80**
1.90	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670	**1.90**
2.00	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169	**2.00**
2.10	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574	**2.10**
2.20	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899	**2.20**
2.30	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158	**2.30**
2.40	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361	**2.40**
2.50	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520	**2.50**
2.60	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643	**2.60**
2.70	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736	**2.70**
2.80	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807	**2.80**
2.90	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861	**2.90**
3.00	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.9990	**3.00**
3.10	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929	**3.10**
3.20	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950	**3.20**
3.30	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965	**3.30**
3.40	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976	**3.40**
3.50	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983	**3.50**

CHAPTER 5: Probabilistic Features of the Distributions of Certain Sample Statistics

5.1 Introduction:

In this Chapter we will discuss the probability distributions of some statistics.

As we mention earlier, a statistic is measure computed form the random sample. As the sample values vary from sample to sample, the value of the statistic varies accordingly.

A statistic is a random variable; it has a probability distribution, a mean and a variance.

5.2 Sampling Distribution:

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

The sampling distribution of the statistic is used to make statistical inference about the unknown parameter.

5.3 Distribution of the Sample Mean:
(Sampling Distribution of the Sample Mean $\overline{X}$):

Suppose that we have a population with mean μ and variance σ^2. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size (*n*) selected randomly from this population. We know that the sample mean is:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Suppose that we select several random samples of size *n*=5.

	1st sample	2nd sample	3rd sample	…	Last sample
Sample values	28 30 34 34 17	31 20 31 40 28	14 31 25 27 32	. . . .	17 32 29 31 30
Sample mean $\overline{x}$	28.4	29.9	25.8	…	27.8

- The value of the sample mean $\overline{X}$ varies from random sample to another.
- The value of $\overline{X}$ is random and it depends on the random sample.
- The sample mean $\overline{X}$ is a random variable.
- The probability distribution of $\overline{X}$ is called the sampling distribution of the sample mean $\overline{X}$.
- Questions:
  - What is the sampling distribution of the sample mean $\overline{X}$?
  - What is the mean of the sample mean $\overline{X}$?
  - What is the variance of the sample mean $\overline{X}$?

Some Results about Sampling Distribution of $\overline{X}$:

Result (1): (mean & variance of $\overline{X}$)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from any distribution with mean μ and variance σ^2; then:

1. The mean of $\overline{X}$ is: $\mu_{\overline{X}} = \mu$.
2. The variance of $\overline{X}$ is: $\sigma^2_{\overline{X}} = \dfrac{\sigma^2}{n}$.
3. The Standard deviation of $\overline{X}$ is call the standard error and is defined by: $\sigma_{\overline{X}} = \sqrt{\sigma^2_{\overline{X}}} = \dfrac{\sigma}{\sqrt{n}}$.

Result (2): (Sampling from normal population)

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal population with mean μ and variance σ^2; that is Normal(μ, σ^2), then the sample mean has a normal distribution with mean μ and variance σ^2/n, that is:

1. $\overline{X} \sim$ Normal $\left(\mu, \dfrac{\sigma^2}{n}\right)$.
2. $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim$ Normal (0,1).

We use this result when sampling from normal distribution with known variance σ^2.

Result (3): (Central Limit Theorem: Sampling from Non-normal population)

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from non-normal population with mean μ and variance σ^2. If the sample size n is large $(n \geq 30)$, then the sample mean has approximately a normal distribution with mean μ and variance σ^2/n, that is

1. $\bar{X} \approx$ Normal $\left(\mu, \dfrac{\sigma^2}{n}\right)$ (approximately)
2. $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx$ Normal $(0,1)$ (approximately)

Note: "≈" means "approximately distributed".
We use this result when sampling from non-normal distribution with known variance σ^2 and with large sample size.

Result (4): (used when σ² is unknown + normal distribution) **, n < 30**

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and unknown variance σ^2; that is Normal(μ, σ^2), then the statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t- distribution with $(n-1)$ degrees of freedom, where S is the sample standard deviation given by:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

We write:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Notation: degrees of freedom = df = ν

The t-Distribution: (Section 6.3. pp 172-174)

- Student's t distribution.
- t-distribution is a distribution of a continuous random variable.

Result 2:

- Recall that, if $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu,\sigma^2)$, then

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

We can apply this result only when σ^2 is known!

- If σ^2 is unknown, we replace the population variance σ^2 with the sample variance $S^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$ to have the following statistic

$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}$$

Recall:

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2 is unknown, i.e. $N(\mu,\sigma^2)$, then the statistic:

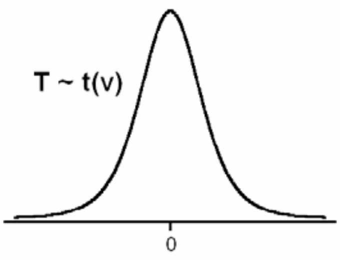
$$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}$$

has a t-distribution with $(n-1)$ degrees of freedom ($df = \; = n-1$), and we write $T\sim t(\nu)$ or $T\sim t(n-1)$.

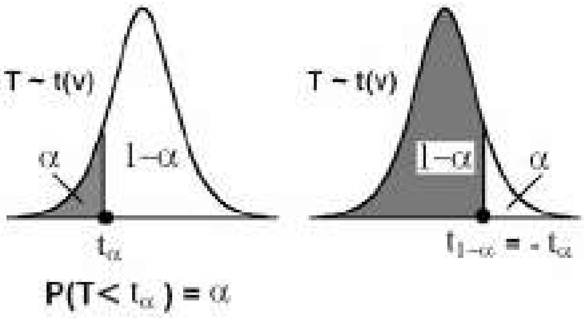
Note:

- t-distribution is a continuous distribution.
- The value of t random variable range from $-\infty$ to $+\infty$ (that is, $-\infty<t<\infty$).
- The mean of t distribution is 0.
- It is symmetric about the mean 0.
- The shape of t-distribution is similar to the shape of the standard normal distribution.
- t-distribution → Standard normal distribution as n → ∞.

i.e. If (n) go to infinity , the t distribution approximately normal distribution



Notation: (t_α)



- t_α = The t-value under which we find an area equal to α
  = The t-value that leaves an area of α to the left.
- The value t_α satisfies: $P(T< t_\alpha) = \alpha$.
- Since the curve of the pdf of T~ t(ν) is symmetric about 0, we have

$$t_{1-\alpha} = -\,t_\alpha$$

For example:

$$t_{0.1} = -t_{1-0.1} = -\,t_{0.9}$$

$$t_{0.975} = -t_{1-0.975} = -t_{0.025}$$

- Values of t_α are tabulated in a special table for several values of α and several values of degrees of freedom. (Table E, appendix p. A-40 in the textbook).

Example:

Find the t-value with ν=14 (df) that leaves an area of:

(a) 0.95 to the left.

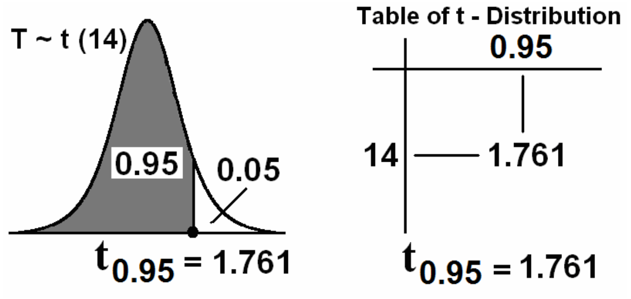
(b) 0.95 to the right.

Solution:

ν = 14 (df); T~ t(14)

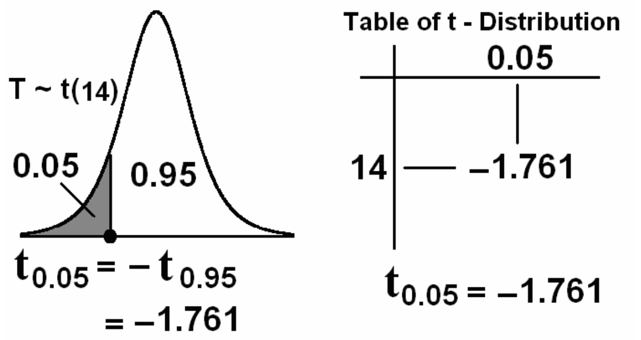
(a) The t-value that leaves an area of 0.95 to the left is

$t_{0.95} = 1.761$.



(b) The t-value that leaves an area of 0.95 to the right is

$t_{0.05} = -\, t_{1-0.05} = -\, t_{0.95} = -\, 1.761$



Note: Some t-tables contain values of α that are greater than or equal to 0.90. When we search for small values of α in these tables, we may use the fact that:

$$t_{1-\alpha} = -\, t_{\alpha}$$

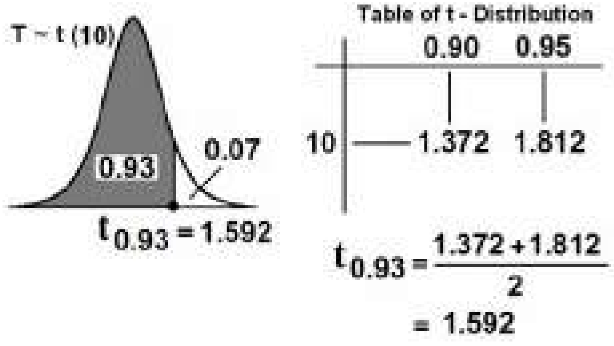
Example:

For ν = 10 degrees of freedom (df), find $t_{0.93}$ and $t_{0.07}$.

Solution:

$t_{0.93} = (1.372+1.812)/2 = 1.592$ (from the table)

$t_{0.07} = -\, t_{1-0.07} = -\, t_{0.93} = -\, 1.592$ (using the rule: $t_{1-\alpha} = -\, t_{\alpha}$)



The t-Distribution

Find :

The t-value that leaves an area of 0.975 to the left (use $v = 12$) is

$t_{0.975}$ =2.179

The t-value that leaves an area of 0.90 to the right (use $v = 16$) is

$t_{0.10} =$ - $t_{1-0.10}$=- $t_{0.90}$= -1.337

The t-value that leaves an area of 0.025 to the right $(use\ v = 8)$ is

$t_{0.975} = 2.306$

The t-value that leaves an area of 0.025 to the left $(use\ v = 8)$ is

$t_{0.025} = -t_{1-0.025} = -t_{0.975} = -2.306$

The t-value that leaves an area of 0.93 to the left $(use\ v = 10)$ is

$t_{0.93} = \frac{t_{0.90}+t_{0.95}}{2} = \frac{1.372+1.812}{2} = 1.592$

The t-value that leaves an area of 0.07 to the left $(use\ v = 10)$ is

$t_{0.07} = -t_{0.93} = -\left(\frac{t_{0.90}+t_{0.95}}{2}\right) = -1.592$

$P(T < K) = 0.90 \quad , df = 10$

K=1.372

$P(T \geq K) = 0.95 \quad , df = 15$

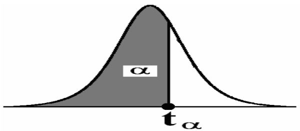
K= -1.753

$P(T < 2.110) =? \quad (df = 17)$

$P(T < 2.110) = 0.975$

$P(T \leq 2.718) =? \quad (df = 11) \quad P(T \leq 2.718) = 0.99$

Critical Values of the t-distribution (t_α)



ν=df	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.3062	1.6896	2.0301	2.4377	2.7238
40	1.3030	1.6840	2.0210	2.4230	2.7040
45	1.3006	1.6794	2.0141	2.4121	2.6896
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.2901	1.6602	1.9840	2.3642	2.6259
120	1.2886	1.6577	1.9799	2.3578	2.6174
140	1.2876	1.6558	1.9771	2.3533	2.6114
160	1.2869	1.6544	1.9749	2.3499	2.6069
180	1.2863	1.6534	1.9732	2.3472	2.6034
200	1.2858	1.6525	1.9719	2.3451	2.6006
∞	1.282	1.645	1.960	2.326	2.576

Application:

Example: (Sampling distribution of the sample mean)
Suppose that the time duration of a minor surgery is approximately normally distributed with mean equal to 800 seconds and a standard deviation of 40 seconds. Find the probability that a random sample of 16 surgeries will have average time duration of less than 775 seconds.

Solution:

X= the duration of the surgery

μ=800 , σ=40 , $\sigma^2 = 1600$

X~N(800, 1600)

Sample size: n=16

Calculating mean, variance, and standard error (standard deviation) of the sample mean $\overline{X}$:

Mean of $\overline{X}$: $\mu_{\overline{X}} = \mu$=800

Variance of $\overline{X}$: $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{1600}{16} = 100$

Standard error (standard deviation) of $\overline{X}$: $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$

Using result 2
~~Using the central limit theorem,~~ $\overline{X}$ has a normal distribution with mean $\mu_{\overline{X}} = 800$ and variance $\sigma^2_{\overline{X}} = 100$, that is:

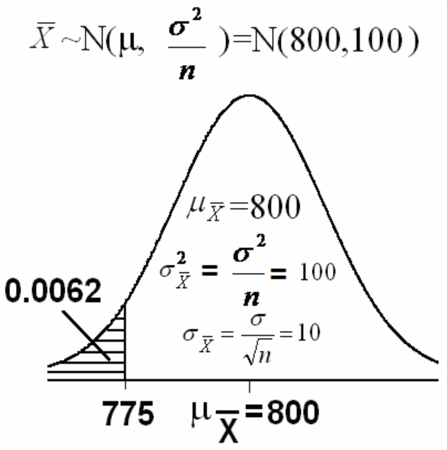
$$\overline{X} \sim \text{N}(\mu, \frac{\sigma^2}{n})=\text{N}(800,100)$$

$$\Leftrightarrow \quad Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\overline{X} - 800}{10} \sim \text{N}(0,1)$$

The probability that a random sample of 16 surgeries will have an average time duration that is less than 775 seconds equals to:

$$P(\overline{X} < 775) = P\left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < \frac{775 - \mu}{\sigma / \sqrt{n}}\right) = P\left(\frac{\overline{X} - 800}{10} < \frac{775 - 800}{10}\right)$$

$$= P\left(Z < \frac{775 - 800}{10}\right) = P(Z < -2.50) = 0.0062$$



Example:

If the mean and standard deviation of serum iron values for healthy men are 120 and 15 microgram/100ml, respectively, what is the probability that a random sample of size 50 normal men will yield a mean between 115 and 125 microgram/100ml?

Solution:

X= the serum iron value

$\mu=120$, $\sigma=15$, $\sigma^2 = 225$, n is large

$X \approx N(120, 225)$

Sample size: $n=50$

Calculating mean, variance, and standard error (standard deviation) of the sample mean $\bar{X}$:

Mean of $\bar{X}$: $\mu_{\bar{X}} = \mu = 120$

Variance of $\bar{X}$: $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{225}{50} = 4.5$

Standard error (standard deviation) of $\bar{X}$: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{50}} = 2.12$

Using the central limit theorem, $\bar{X}$ has a normal distribution with mean $\mu_{\bar{X}} = 120$ and variance $\sigma^2_{\bar{X}} = 4.5$, that is:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(120, 4.5)$$

$$\Leftrightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 120}{2.12} \sim N(0,1)$$

The probability that a random sample of 50 men will yield a mean between 115 and 125 microgram/100ml equals to:

$$P(115 < \bar{X} < 125) = P\left(\frac{115-\mu}{\sigma/\sqrt{n}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{125-\mu}{\sigma/\sqrt{n}}\right)$$

$$= P\left(\frac{115-120}{2.12} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{125-120}{2.12}\right) = P(-2.36 < Z < 2.36)$$

$= \ P(Z < 2.36)$ - $P(Z < -2.36)$

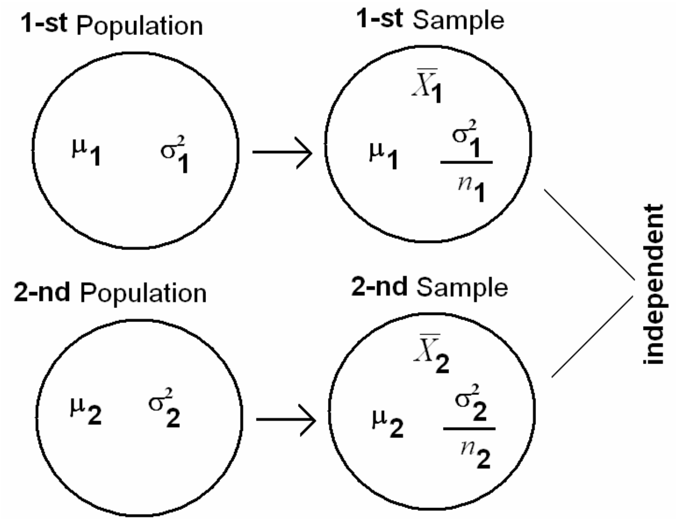
= 0.9909 – 0.0091

= 0.9818

5.4 Distribution of the Difference Between Two Sample Means ($\overline{X}_1 - \overline{X}_2$):

Suppose that we have two populations:

- 1-st population with mean μ_1 and variance σ_1^2
- 2-nd population with mean μ_2 and variance σ_2^2
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means ($\mu_1 - \mu_2$).
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let $\overline{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1-st sample.
- Let $\overline{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\overline{X}_1 - \overline{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.



Note: Square roots distribute over multiplication or division, but not addition or subtraction. $\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$

In general: Z= (value - Mean)/ Standard deviation

The sampling distribution of $\bar{X}_1 - \bar{X}_2$:

Result:

The mean, the variance and the standard deviation of $\bar{X}_1 - \bar{X}_2$ are:

Mean of $\bar{X}_1 - \bar{X}_2$ is: $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

Variance of $\bar{X}_1 - \bar{X}_2$ is: $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Standard error (standard) deviation of $\bar{X}_1 - \bar{X}_2$ is:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1 - \bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Result:

If the two random samples were selected from normal distributions (or non-normal distributions with large sample sizes) with known variances σ_1^2 and σ_2^2, then the difference between the sample means $(\bar{X}_1 - \bar{X}_2)$ has a normal distribution with mean $(\mu_1 - \mu_2)$ and variance $((\sigma_1^2 / n_1) + (\sigma_2^2 / n_2))$, that is:

- $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2\,,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
- $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$

Application:

Example:

Suppose it has been established that for a certain type of client (type A) the average length of a home visit by a public health nurse is 45 minutes with standard deviation of 15 minutes, and that for second type (type B) of client the average home visit is 30 minutes long with standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first type and 40 clients from the second type, what is the probability that the average length of home visit of first type will be greater than the average length of home visit of second type by 20 or more minutes?

Solution: $\bar{X}_1 > \bar{X}_2 + 20$

For the first type:

$\mu_1 = 45$

$\sigma_1 = 15$

$\sigma_1^2 = 225$

$n_1 = 35$ is large

For the second type:

$\mu_2 = 30$

$\sigma_2 = 20$

$\sigma_2^2 = 400$

$n_2 = 40$ is large

The mean, the variance and the standard deviation of $\bar{X}_1 - \bar{X}_2$ are:

Mean of $\bar{X}_1 - \bar{X}_2$ is:

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 = 45 - 30 = 15$$

Variance of $\bar{X}_1 - \bar{X}_2$ is:

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{225}{35} + \frac{400}{40} = 16.4286$$

Standard error (standard) deviation of $\bar{X}_1 - \bar{X}_2$ is:

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1-\bar{X}_2}} = \sqrt{16.4286} = 4.0532$$

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is:

$$\bar{X}_1 - \bar{X}_2 \sim N(15\,,16.4286)$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 15}{\sqrt{16.4286}} \sim \mathrm{N}(0,1)$$

The probability that the average length of home visit of first type will be greater than the average length of home visit of second type by 20 or more minutes is:

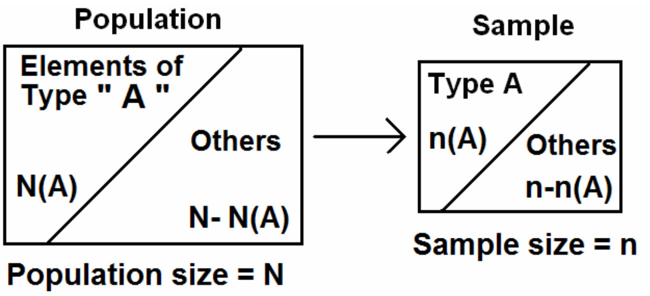
$$P(\overline{X}_1 - \overline{X}_2 > 20) = P\left(\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{20 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

$$= P\left(Z > \frac{20 - 15}{4.0532}\right) = \text{P(Z>1.23)} = 1 - \text{P(Z<1.23)}$$

$$= 1 - 0.8907$$

$$= 0.1093$$

5.5 Distribution of the Sample Proportion ($\hat{p}$):



- For the population:

  $N(A)$ = number of elements in the population with a specified characteristic "A"

  N = total number of elements in the population (population size)

The population proportion is

$$p = \frac{N(A)}{N} \qquad \text{(p is a parameter)}$$

- For the sample:

  $n(A)$ = number of elements in the sample with the same characteristic "A"

  n = sample size

The sample proportion is

$$\hat{p} = \frac{n(A)}{n} \qquad (\hat{p} \text{ is a statistic})$$

- The sampling distribution of $\hat{p}$ is used to make inferences about p.

Result:

The mean of the sample proportion ($\hat{p}$) is the population proportion (p); that is:

$$\mu_{\hat{p}} = p$$

The variance of the sample proportion ($\hat{p}$) is:

$$\sigma^2_{\hat{p}} = \frac{p(1-p)}{n} = \frac{pq}{n}. \qquad \text{(where q=1 −p)}$$

The standard error (standard deviation) of the sample proportion ($\hat{p}$) is:

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

Result:

For large sample size ($n \geq 30, np > 5, nq > 5$), the sample proportion ($\hat{p}$) has approximately a normal distribution with mean $\mu_{\hat{p}} = p$ and a variance $\sigma^2_{\hat{p}} = pq/n$, that is:

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right) \qquad \text{(approximately)}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \qquad \text{(approximately)}$$

Example:

Suppose that 45% of the patients visiting a certain clinic are females. If a sample of 35 patients was selected at random, find the probability that:

1. the proportion of females in the sample will be greater than 0.4.
2. the proportion of females in the sample will be between 0.4 and 0.5.

Solution:

- .n = 35 (large)
- p = The population proportion of females = $\frac{45}{100}$ = 0.45

- $\hat{p}$ = The sample proportion

  (proportion of females in the sample)

- The mean of the sample proportion ($\hat{p}$) is p = 0.45
- The variance of the sample proportion ($\hat{p}$) is:

$$\frac{p(1-p)}{n} = \frac{pq}{n} = \frac{0.45(1-0.45)}{35} = 0.0071.$$

- The standard error (standard deviation) of the sample proportion ($\hat{p}$) is:

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{0.0071} = 0.084$$

- $n \geq 30, \;\; np = 35 \times 0.45 = 15.75 > 5, nq = 35 \times 0.55 = 19.25 > 5$

1. The probability that the sample proportion of females ($\hat{p}$) will be greater than 0.4 is:

$$P(\hat{p} > 0.4) = 1 - P(\hat{p} < 0.4) = 1 - P\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0.4-p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= 1 - P\left(Z < \frac{0.4-0.45}{\sqrt{\frac{0.45(1-0.45)}{35}}}\right) = 1 - P(Z < -0.59)$$

$$= 1 - 0.2776 = 0.7224$$

2. The probability that the sample proportion of females ($\hat{p}$) will be between 0.4 and 0.5 is:

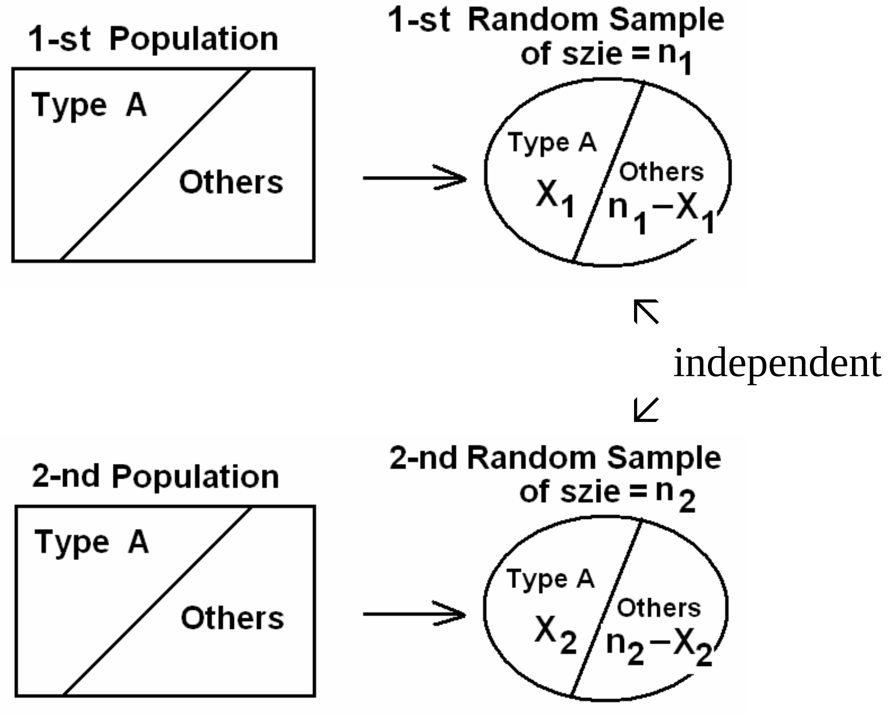
$$P(0.4 < \hat{p} < 0.5) = P(\hat{p} < 0.5) - P(\hat{p} < 0.4)$$

$$= P\left(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0.5-p}{\sqrt{\frac{p(1-p)}{n}}}\right) - 0.2776$$

$$= P\left(Z < \frac{0.5-0.45}{\sqrt{\frac{0.45(1-0.45)}{35}}}\right) - 0.2776$$

$$= P(Z < 0.59) - 0.2776$$
$$= 0.7224 - 0.2776$$
$$= 0.4448$$

5.6 Distribution of the Difference Between Two Sample Proportions ($\hat{p}_1 - \hat{p}_2$):



Suppose that we have two populations:

- p_1 = proportion of elements of type (A) in the 1-st population.
- p_2 = proportion of elements of type (A) in the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about $p_1 - p_2$.
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let X_1 = no. of elements of type (A) in the 1-st sample.
- Let X_2 = no. of elements of type (A) in the 2-nd sample.
- $\hat{p}_1 = \frac{X_1}{n_1}$ = sample proportion of the 1-st sample

- $\hat{p}_2 = \dfrac{X_2}{n_2}$ = sample proportion of the 2-nd sample
- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about $p_1 - p_2$.

The sampling distribution of $\hat{p}_1 - \hat{p}_2$:

Result:

The mean, the variance and the standard error (standard deviation) of $\hat{p}_1 - \hat{p}_2$ are:

- Mean of $\hat{p}_1 - \hat{p}_2$ is:

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

- Variance of $\hat{p}_1 - \hat{p}_2$ is:

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}$$

- Standard error (standard deviation) of $\hat{p}_1 - \hat{p}_2$ is:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}}$$

- $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$

Result:

For large samples sizes $(n_1 \geq 30, n_2 \geq 30, n_1 p_1 > 5, n_1 q_1 > 5, n_2 p_2 > 5, n_2 q_2 > 5)$, we have that $\hat{p}_1 - \hat{p}_2$ has approximately normal distribution with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}$, that is:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}\right) \quad \text{(Approximately)}$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1\,q_1}{n_1} + \frac{p_2\,q_2}{n_2}}} \sim \text{N}(0,1) \quad \text{(Approximately)}$$

Example:

Suppose that 40% of Non-Saudi residents have medical insurance and 30% of Saudi residents have medical insurance in a certain city. We have randomly and independently selected a sample of 130 Non-Saudi residents and another sample of 120 Saudi residents. What is the probability that the difference between the sample proportions, $\hat{p}_1 - \hat{p}_2$, will be between 0.05 and 0.2?

Solution:

p_1 = population proportion of non-Saudi with medical insurance.
p_2 = population proportion of Saudi with medical insurance.
$\hat{p}_1$ = sample proportion of non-Saudis with medical insurance.
$\hat{p}_2$ = sample proportion of Saudis with medical insurance.

q1=0.6 $p_1 = 0.4$ n_1=130 > 30
q2=0.7 $p_2 = 0.3$ n_2=120 >30

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 = 0.4 - 0.3 = 0.1$$

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} = \frac{(0.4)(0.6)}{130} + \frac{(0.3)(0.7)}{120} = 0.0036$$

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = \sqrt{0.0036} = 0.06$$

The probability that the difference between the sample proportions, $\hat{p}_1 - \hat{p}_2$, will be between 0.05 and 0.2 is:

$$P(0.05 < \hat{p}_1 - \hat{p}_2 < 0.2) = P(\hat{p}_1 - \hat{p}_2 < 0.2) \; - \; P(\hat{p}_1 - \hat{p}_2 < 0.05)$$

$$= P\left(\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1\, q_1}{n_1} + \frac{p_2\, q_2}{n_2}}} < \frac{0.2 - (p_1 - p_2)}{\sqrt{\frac{p_1\, q_1}{n_1} + \frac{p_2\, q_2}{n_2}}} \right)$$

$$- \mathrm{P}\left(\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1\ q_1}{n_1} + \frac{p_2\ q_2}{n_2}}} < \frac{0.05 - (p_1 - p_2)}{\sqrt{\frac{p_1\ q_1}{n_1} + \frac{p_2\ q_2}{n_2}}} \right)$$

$$= \mathrm{P}\left(Z < \frac{0.2 - 0.1}{0.06} \right) - \mathrm{P}\left(Z < \frac{0.05 - 0.1}{0.06} \right)$$

$$= \mathrm{P}(Z < 1.67) - \mathrm{P}(Z < -0.83)$$

= 0.95254 - 0.20327

= 0.74927