Data Analytics

Lecture No: 06

What is a distribution?

- Describes the ‘Shape’ of a batch of numbers.
- The characteristics of a distribution can sometimes be defined using a small number of numeric descriptors called “parameters”

Why distribution?

- Can serve as a basis for standardized comparison of empirical distributions.
- Can help us estimate confidence intervals for inferential statistics.
- Form a basis for more advanced statistical methods.
    - “Fit” between observed distributions and certain theoretical distributions is an assumption of many statistical procedures

Random variable

- A variable which contains the outcomes of a chance experiment.
- “ Quantifying the outcomes”
- Example X = (1 for Head, and 0 for Tails)
- A variable that can take on different values in the population according to some “random” mechanism.
- Discrete
  - Distinct values, countable.
  - Year
- Continuous
  - Mass

Probability Distributions

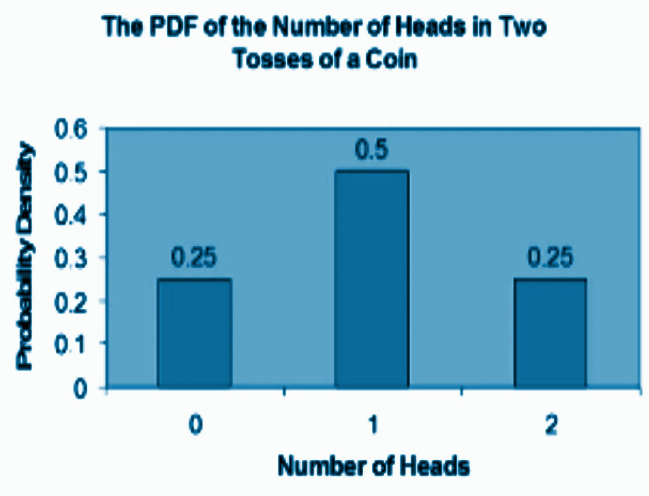
- The probability distribution function or probability density function (PDF) of a random variable X means the values taken by that random variable and their associated probabilities.
- PDF of a discrete random variable (also known as PMF):

Example 1: Let the r.v. X be the number of heads obtained in two tosses of a coin.

Sample Space: {HH,HT,TH,TT}

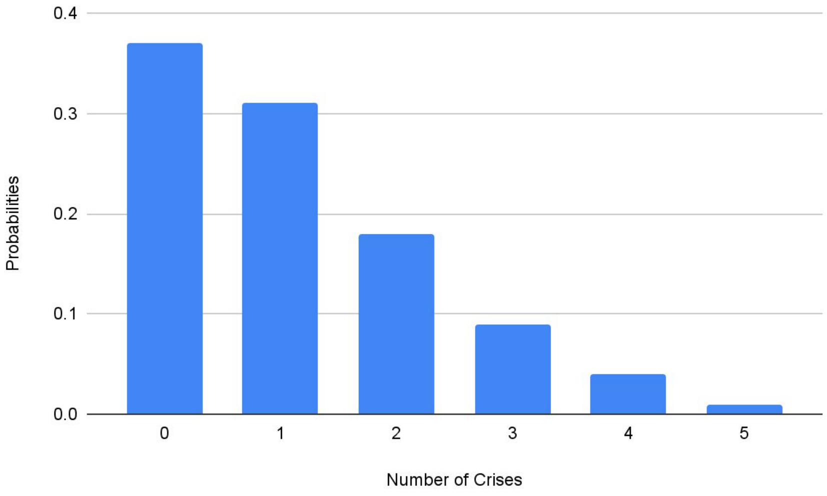
PDF Of Discrete Random Variable

Number of Heads (X):	0	1	2	sum
PDF (P(x)):	¼	½	¼	1



Discrete Distribution – Example

Distribution of Daily Crises	
Number of Crises	Probability
0	0.37
1	0.31
2	0.18
3	0.09
4	0.04
5	0.01



Requirements for a discrete probability function

- Probabilities are between 0 and 1, inclusively
- Total of all probabilities equals 1
  -

$$0 \le P(X) \le 1 \text{ for all } X$$

$$\sum P(x) = 1$$

Cumulative Distribution Function

- The CDF of a random variable X (defined as F(X)) is a graph associating all possible values, or the range of possible values with $P(X \leq x)$.
- CDFs always lie between 0 and 1 i.e., $0 \leq F(X_i) \leq 1$, where $F(X_i)$ is the CDF

Probability Distribution for the random variable X

- A probability distribution for a discrete random variable X:

x	-8	-3	-1	0	1	4	6
P(X=x)	0.13	0.15	0.17	0.20	0.15	0.11	0.09

Find

1. P(X ≤ 0)? 0.65

2. P(-3 ≤ X ≤ 1)? 0.67

The Expected value of X

- Let X be a discrete random variable with set of possible values D and pmf p(x). The expected value or mean value of X, denoted

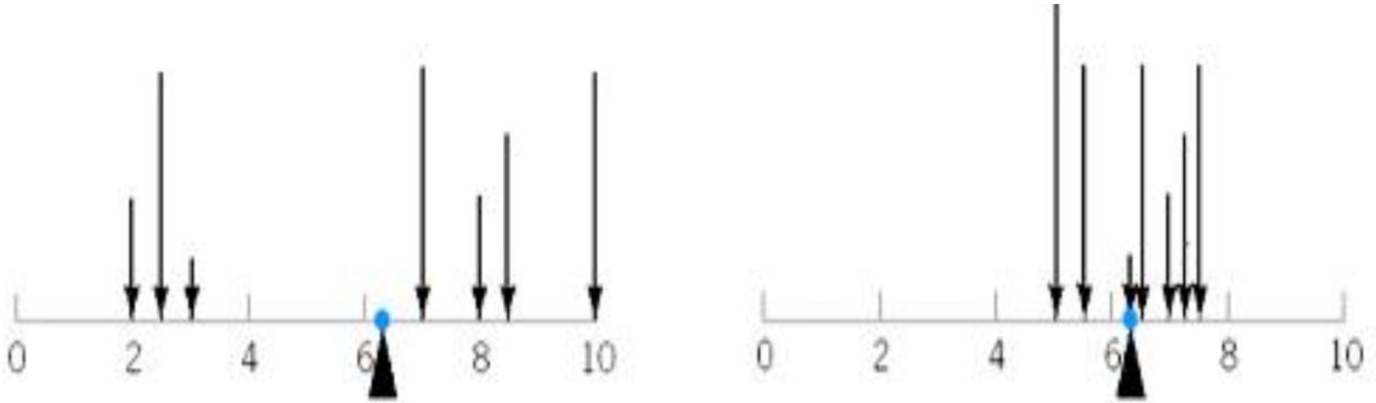
$$E(X) \;\; or \;\; \mu_X \, , \; is$$

$$E(X) = \mu_X = \sum_{x \in D} x \, . \, p(x)$$

where

$$\begin{aligned} E(x) &= \text{long-run average} \\ x &= \text{an outcome} \\ P(x) &= \text{probability of that outcome} \end{aligned}$$

Mean and variance of a discrete random variable

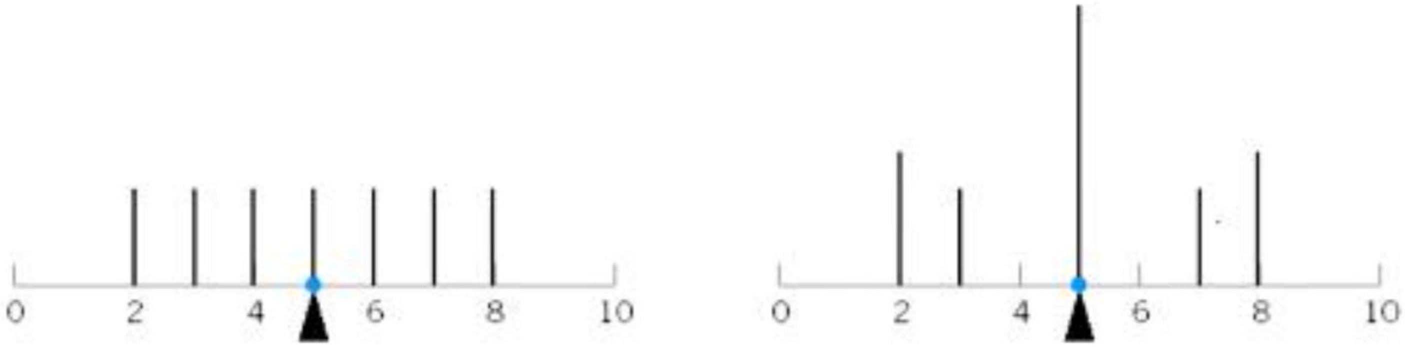


(a)

(b)

A probability distribution can be viewed as a loading with the mean equal to the balance point. Parts (a) and (b) illustrate equal means, but part (a) illustrates a larger variance.

Mean and variance of discrete random variable



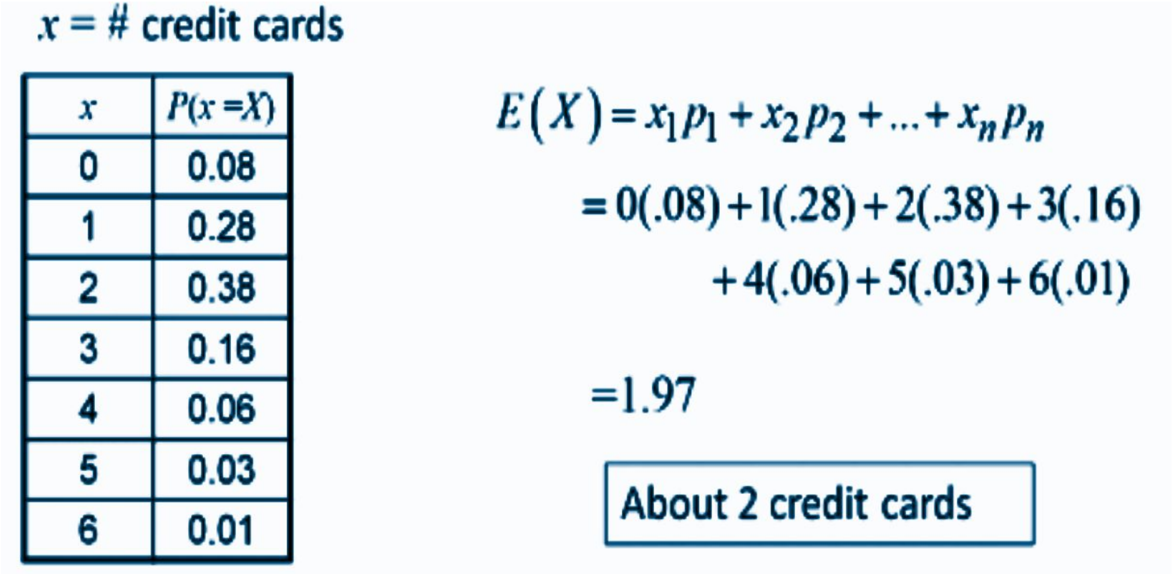
(a)

(b)

The probability distribution illustrated in parts (a) and (b) differ even though they have equal means and variance.

Example – Expected value

- Use the data below to find out the expected number of credit cards that a customer to a retail outlet will possess.



x = # credit cards

x	$P(x=X)$
0	0.08
1	0.28
2	0.38
3	0.16
4	0.06
5	0.03
6	0.01

$$E(X) = x_1p_1 + x_2p_2 + \ldots + x_np_n$$

$$= 0(.08) + 1(.28) + 2(.38) + 3(.16) + 4(.06) + 5(.03) + 6(.01)$$

$$= 1.97$$

About 2 credit cards

The variance and standard deviation

- Let X have pmf p(x), and expected value μ

Then the variance of X, denoted V(X)

(or σ^2_X or σ^2), is

$$V(X) = \sum_{D} (x-\mu)^2 . p(x) = E\left[(X-\mu)^2\right]$$

$$\textit{The standard deviation } (SD) \textit{ of } X \textit{ is}$$

$$\sigma_X = \sqrt{\sigma^2_X}$$

Cont’d

The quiz scores for a particular student are given below:

22,25,20,18,12,20,24,20,20,25,24,25,18.

Find the variance and standard deviation?

Value	12	18	20	22	24	25
Frequency	1	2	4	1	2	3
Probability	0.08	0.15	0.31	0.08	0.15	0.23

$$\mu = 21$$

$$V(X) = p_1(x_1 - \mu)^2 + p_2(x_2 - \mu)^2 + \ldots . + p_n(x_n - \mu)^2$$

$$\sigma = \sqrt{V(X)}$$

Cont’d

$$V(X) = 0.08(12-21)^2 + 0.15(18-21)^2 + 0.31(20-21)^2$$

$$+\ 0.08(22-21)^2 + 0.15(24-21)^2 + 0.23(25-21)^2$$

$$V(X) = 13.25$$

$$\sigma = \sqrt{V(X)} = \sqrt{13.25} \approx 3.64$$

Shortcut formula for variance

$$V(X) = \sigma^2 = \left[\sum_{D} x^2 \,.\, p(x)\right] - \mu^2$$

$$= E\left(X^2\right) - [E(X)]^2$$

Mean of a discrete distribution

$$\mu = E(X) = \sum X.P(X)$$

X	P(X)	X. P(X)
-1	0.1	-0.1
0	0.2	0.0
1	0.4	0.4
2	0.2	0.4
3	0.1	0.3
		1.0

Variance and standard deviation of a discrete distribution

$$\boxed{\sigma^2 = \sum (X - \mu)^2 . P(X) \ = 1.2} \quad \boxed{\sigma = \sqrt{\sigma^2} = \sqrt{1.2} \cong 1.10}$$

X	P(X)	X-µ	$(X-\mu)^2$	$(X-\mu)^2.P(X)$
-1	0.1	-2	4	0.4
0	0.2	-1	1	0.2
1	0.4	0	0	0
2	0.2	1	1	0.2
3	0.1	2	4	0.4
				1.2

Properties of expected value

1. $E(b) = b$, b is a constant
2. $E(X+Y) = E(X) + E(Y)$
3. $E(X/Y) \neq E(X)/E(Y)$
4. $E(XY) \neq E(X)E(Y)$ unless they are independent
5. $E(aX) = aE(X)$, a constant
6. $E(aX + b) = aE(X)+b$, a and b are constants.

Properties of variance

1. var(constant) = 0
2. If X and Y are two independent random variables, then
   a. Var(X+Y) = Var(X) + Var(Y) and
   b. Var(X-Y) = Var(X) + Var(Y) and
3. If b is a constant then Var(b+X) = Var(X)
4. If a is a constant then Var(aX) = a^2Var(X)
5. If a and b are constants then Var(aX+b) = a^2Var(X)
6. If X and Y are two independent random variables and a and b are constants then Var(aX+bY) = a^2Var(X)+ b^2Var(Y)

Covariance

Covariance: For two discrete random variables X and Y with E(X) = μ_x and E(Y) = μ_y, the covariance between X and Y is defined as:

$$Cov(XY) = \sigma_{xy} = E\left(X - \mu_x\right)E\left(Y - \mu_y\right) = E(XY) - \mu_x\mu_y$$

Covariance

- In general, the covariance between two random variables can be positive or negative.
- If two random variables move in the same direction, then the covariance will be positive, if they move in the opposite direction the covariance will be negative.

Properties:

1. If X and Y are independent random variables, their covariance is zero. Since E(XY) = E(X)E(Y)
2. Cov(XX) = Var(X)
3. Cov(YY) = Var(Y)

Correlation coefficient

- The covariance tells the sign but not the magnitude about how strongly the variables are positively or negatively related. The correlation coefficient provides such measure of how strongly the variables are related to each other.
- For two random variables X and Y with E(X) = μ_x and E(Y) =μ_y , the correlation coefficient is defined as

$$\rho_{xy} = \frac{Cov(XY)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

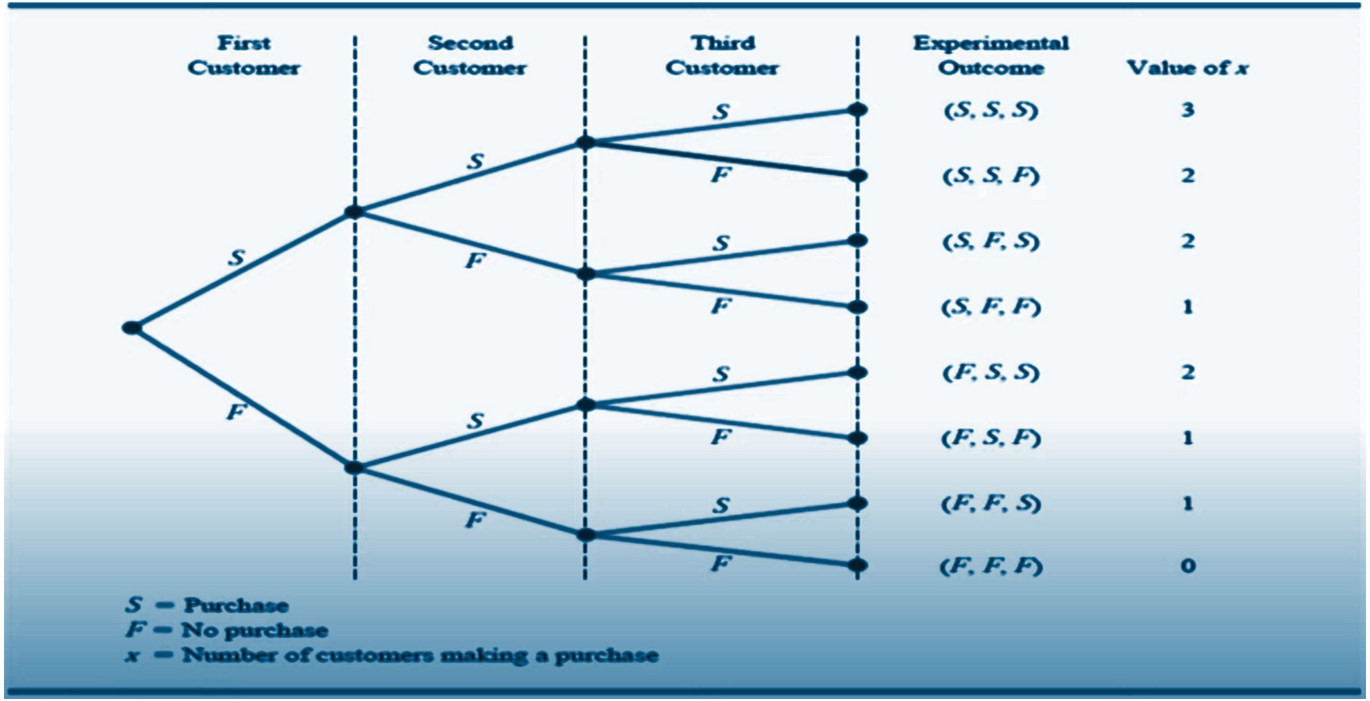
Some special distributions

- Discrete
  - Binomial
  - Poisson
  - Hyper geometric
- Continuous
  - Uniform
  - Exponential
  - Normal

Binomial distribution

- Let us consider the purchase decisions of the next three customers who enter a store.
- On the basis of past experience, the store manager estimates the probability that any one customer will make a purchase is 0.30.
- What is the probability that two of the next three customers will make a purchase?

Tree diagram for the Mr. X clothing store problem



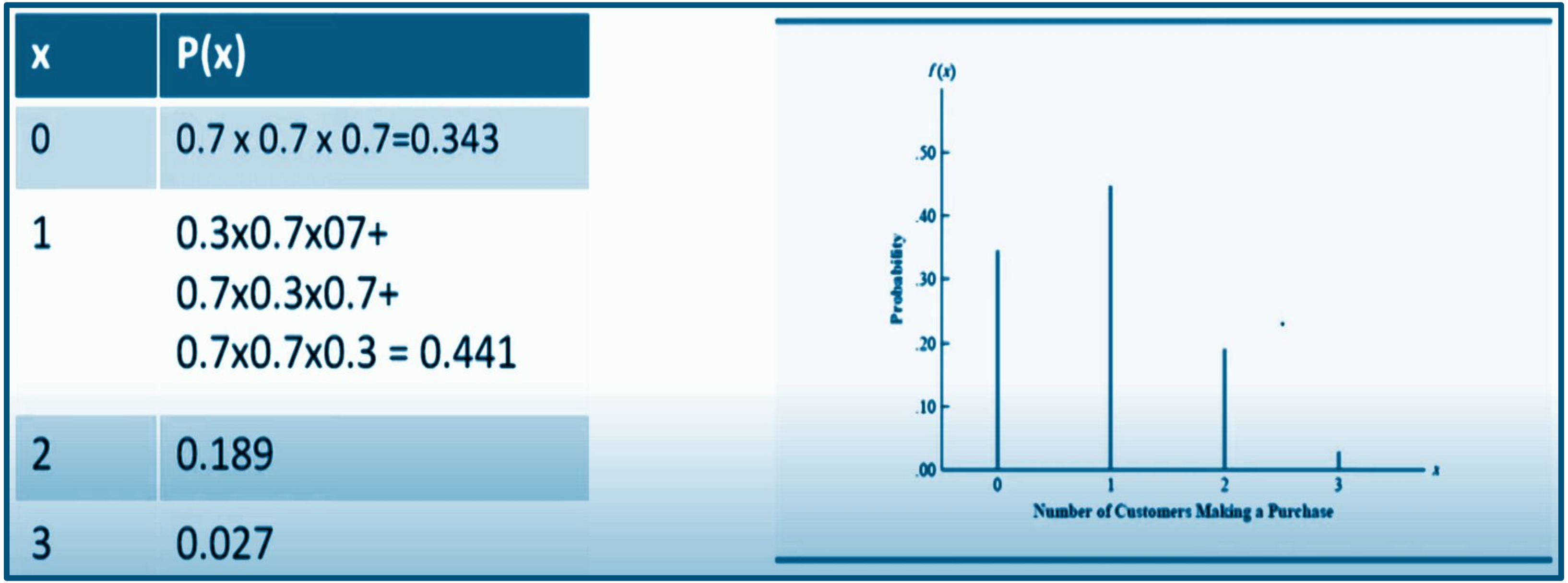
Experimental Outcome	Value of *x*
(*S*, *S*, *S*)	3
(*S*, *S*, *F*)	2
(*S*, *F*, *S*)	2
(*S*, *F*, *F*)	1
(*F*, *S*, *S*)	2
(*F*, *S*, *F*)	1
(*F*, *F*, *S*)	1
(*F*, *F*, *F*)	0

S = Purchase
F = No purchase
x = Number of customers making a purchase

Trial Outcomes

Trial outcomes				
1st Customer	**2nd Customer**	**3rd Customer**	**Experimental Outcome**	**Probability of Experimental Outcome**
Purchased	Purchased	No Purchased	(S,S,F)	$pp(1-p)=p^2(1-p)$ $=(0.30)^2(0.70) = 0.063$
Purchased	No Purchased	Purchased	(S,F,S)	$p(1-p)p = p^2(1-p)$ $=(0.30)^2(0.70)=0.063$
No purchased	Purchased	Purchased	(F,S,S)	$(1-p)pp = p^2(1-p)$ $=0.30)^2(0.70)=0.063$

Graphical representation of the probability distribution for the number of customers making a purchase



x	P(x)
0	0.7 x 0.7 x 0.7=0.343
1	0.3x0.7x07+ 0.7x0.3x0.7+ 0.7x0.7x0.3 = 0.441
2	0.189
3	0.027

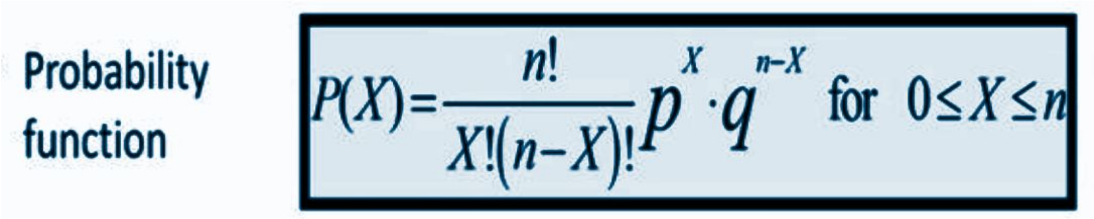
Binomial distribution – Assumptions

- Experiment involves n identical trials.
- Each trial has exactly two possible outcomes: success and failure.
- Each trial is independent of the previous trials.
- p is the probability of a success on any one trial.

  q = (1-p) is the probability of a failure on any one trial.

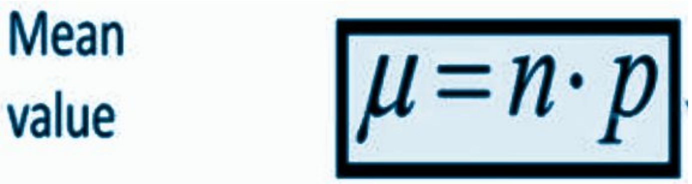
- p and q are constant throughout the experiment
- X is the number of successes in the n trials.

Binomial distribution



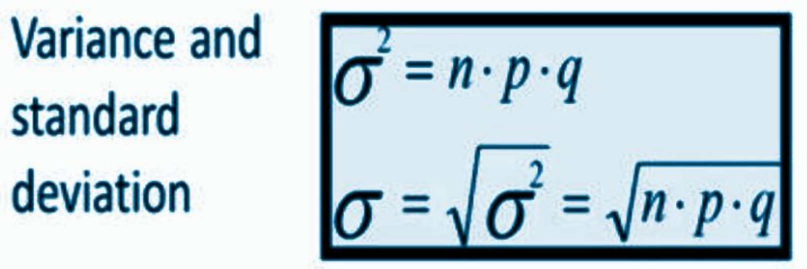
Probability function

$$P(X)=\frac{n!}{X!(n-X)!}p^{X}\cdot q^{n-X} \quad \text{for} \quad 0\leq X\leq n$$



Mean value

$$\mu=n\cdot p$$



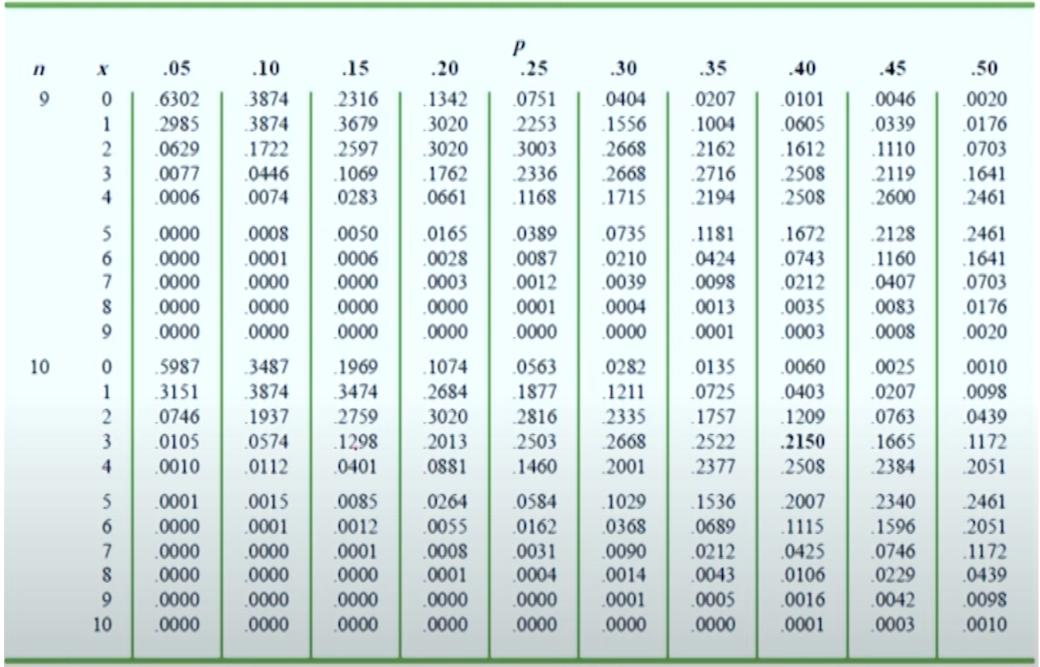
Variance and standard deviation

$$\sigma^{2}=n\cdot p\cdot q$$

$$\sigma=\sqrt{\sigma^{2}}=\sqrt{n\cdot p\cdot q}$$

Binomial Table

- Selected values from the binomial probability table.
- Example: n = 10, x = 3, p = 0.40; f(3) = 0.2150



						p					
n	*x*	**.05**	**.10**	**.15**	**.20**	**.25**	**.30**	**.35**	**.40**	**.45**	**.50**
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	**.2150**	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010

Mean and variance

- Suppose that for the next month the clothing store forecasts 1000 customers will enter the store.
- What is the expected number of customers who will make a purchase?
- The answer is $\mu = np = (1000)(0.3) = 300$
- For the next 1000 customers entering the store, the variance and standard deviation for the number of customers who will make a purchase are
  - $\sigma^2 = np(1-p) = 1000(0.3)(0.7) = 210$
  - $\sigma = \sqrt{210} = 14.49$

Poisson distribution

- Describes discrete occurrences over a continuum or interval.
- A discrete distribution
- Describes rare events
- Each occurrence is independent any other occurrences
- The number of occurrences in each interval can vary from zero to infinity.
- The expected number of occurrences must hold constant throughout the experiment.

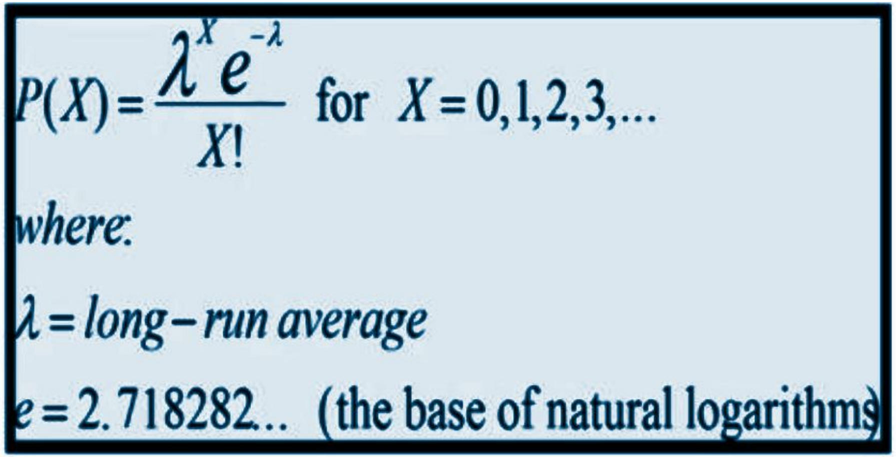
Poisson distribution: Applications

- Arrivals at queueing systems
  - airports – people, airplanes, automobiles, baggage.
  - Banks – people, automobiles, loan applications
  - Computer file servers – read and write operations.

- Defects in manufacturing goods.
  - Number of defects per 1000 feet of copper wire
  - Number of blemishes per square foot of painted surface
  - Number of errors per typed page.

Poisson distribution

- Probability function



$$P(X) = \frac{\lambda^{X} e^{-\lambda}}{X!} \quad \text{for} \quad X = 0,1,2,3,\ldots$$

where:

$\lambda = long-run\ average$

$e = 2.718282\ldots$ (the base of natural logarithms)

Example

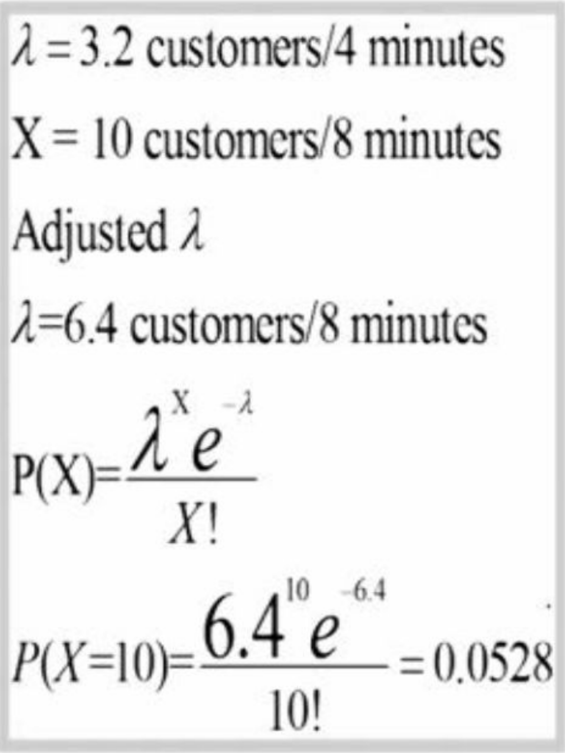
- Suppose bank customers arrive randomly on weekday afternoons at an average of 3.2 customers every 4 minutes. What is the probability of exactly 5 customers arriving in a 4-minute interval on a weekday afternoon?

- Bank customers arrive randomly on weekday afternoons at an average of 3.2 customers every 4 minutes. What is the probability of having more than 7 customers in a 4-minute interval on a weekday afternoon?

Example

- A bank has an average random arrival rate of 3.2 customers every 4 minutes. What is the probability of getting exactly 10 customers during an 8-minute interval?

Poisson distribution– Example



$\lambda = 3.2$ customers/4 minutes

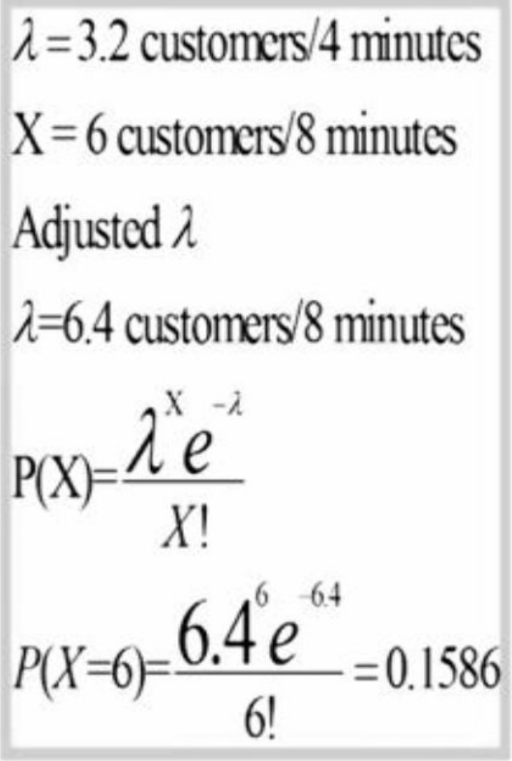
X = 10 customers/8 minutes

Adjusted λ

λ=6.4 customers/8 minutes

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

$$P(X=10)=\frac{6.4^{10} e^{-6.4}}{10!}=0.0528$$



$\lambda = 3.2$ customers/4 minutes

X = 6 customers/8 minutes

Adjusted λ

λ=6.4 customers/8 minutes

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

$$P(X=6)=\frac{6.4^{6} e^{-6.4}}{6!}=0.1586$$

Hypergeometric Distribution

- The **binomial distribution** is applicable when selecting from a finite population with replacement or from an infinite population without replacement.
- The hypergeometric distribution is applicable when selecting from a **finite population without replacement**

Hypergeometric Distribution

- Sampling without replacement from a finite population.
- The population, N, is finite and known.
- Each trial has exactly two possible outcomes, success and failure
- Trials are not independent
- X is the number of successes in the n trials and known.
-

Hypergeometric Distribution

- Probability function
    - N is the population size
    - n is the sample size
    - A is the number of successes in the population
    - x is the number of successes in the sample.

$$\boxed{P(x) = \frac{\left({}_{A}C_{x}\right)\left({}_{N-A}C_{n-x}\right)}{{}_{N}C_{n}}}$$

- Mean value

$$\boxed{\mu = \frac{A \cdot n}{N}}$$

- Variance and standard deviation

$$\boxed{\begin{aligned} \sigma^2 &= \frac{A(N-A)n(N-n)}{N^2(N-1)} \\ \sigma &= \sqrt{\sigma^2} \end{aligned}}$$

Hypergeometric distribution – Example

- Different computers are checked in the department. 4 out of 10 computers have illegal software loaded.
- What is the probability that 2 of the 3 selected computers have illegal software loaded?
- So, N = 10, n = 3, A = 4, X = 2

$$P(X = 2) = \frac{\binom{A}{X}\binom{N-A}{n-X}}{\binom{N}{n}} = \frac{\binom{4}{2}\binom{6}{1}}{\binom{10}{3}} = \frac{(6)(6)}{120} = 0.3$$

- The probability that 2 of the 3 computers have illegal software loaded is 0.30 or 30%

Example

- Suppose 18 major computer companies operate in the United States and that 12 are located in California’s Silicon Valley. If three computer companies are selected randomly from the entire list, what is the probability that one or more of the selected companies are located in the Silicon Valley?

Continuous probability distribution

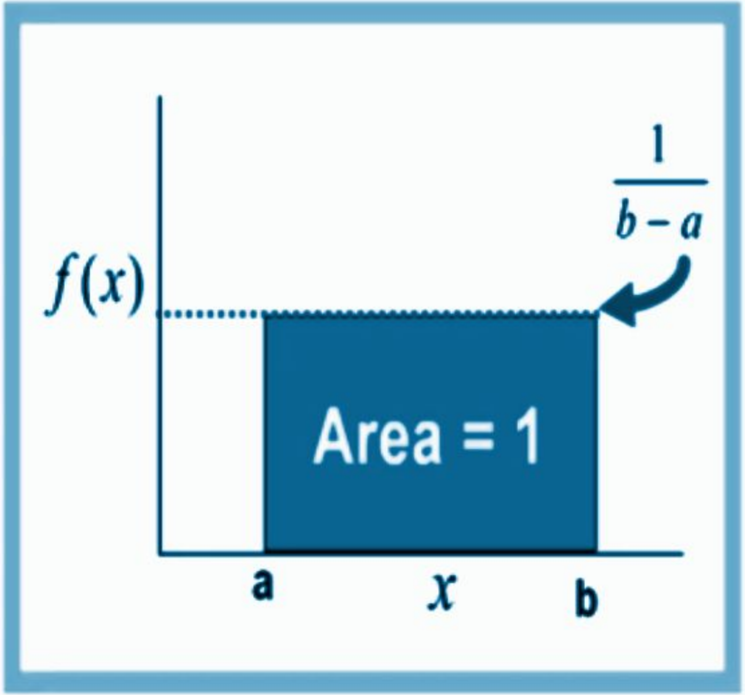
- A continuous random variable is a variable that can assume any value on a continuum (can assure an uncountable number of values)
    - Thickness of an item
    - Time required to complete a task
    - Temperature of a solution
    - Height
- These are the potentially take on any value, depending only on the ability to measure precisely and accurately.

The Uniform distribution

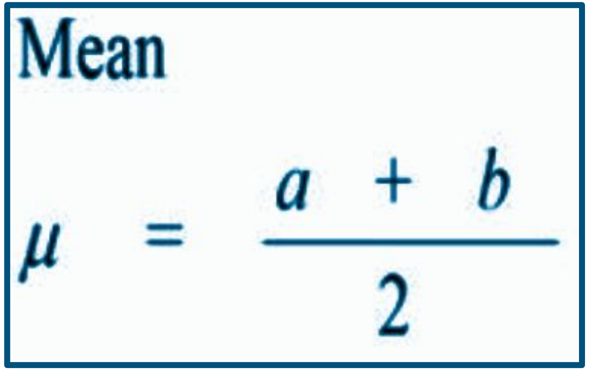
- The uniform distribution is the probability distribution that has equal probabilities for all the outcomes of the random variable.
- Because of its shape it is also called as rectangle distribution

Uniform distribution

$$f(x)=\begin{cases}\dfrac{1}{b-a} & \text{for} \quad a\le x\le b\\[2ex] 0 & \text{for} \quad \text{all other valu es}\end{cases}$$

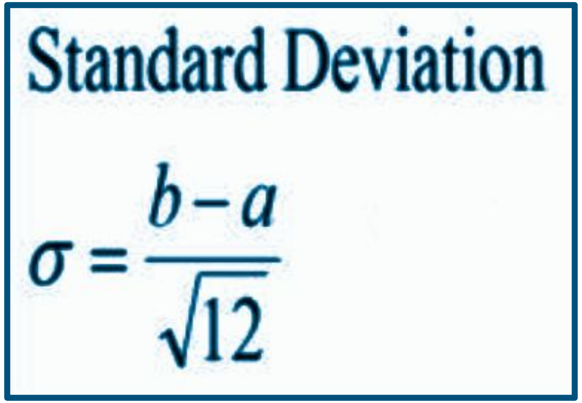


Uniform distribution – Mean and standard deviation



Mean

$$\mu = \frac{a + b}{2}$$

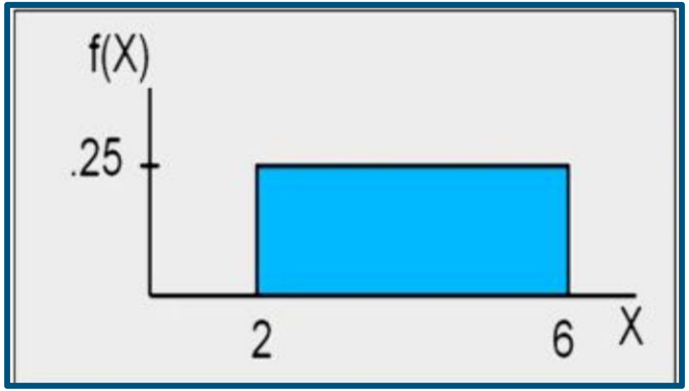


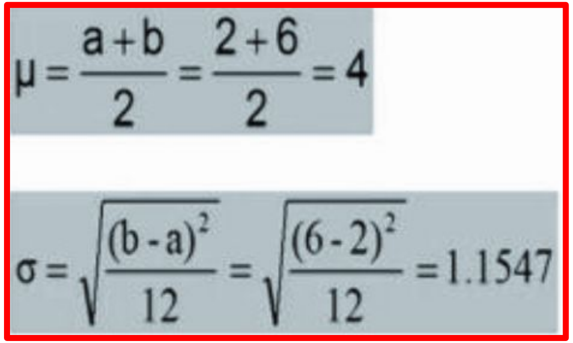
Standard Deviation

$$\sigma = \frac{b - a}{\sqrt{12}}$$

Uniform distribution – Example

- Uniform probability distribution over the range 2 ≤ X ≤ 6



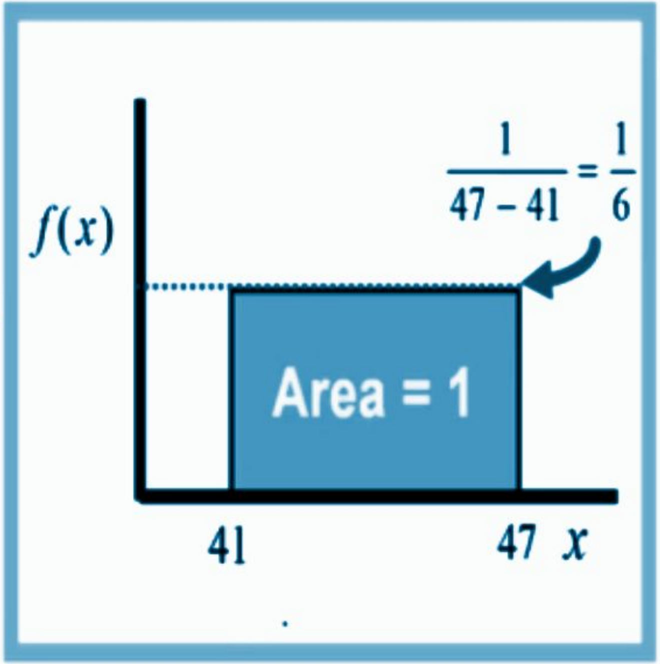


$$\mu = \frac{a+b}{2} = \frac{2+6}{2} = 4$$

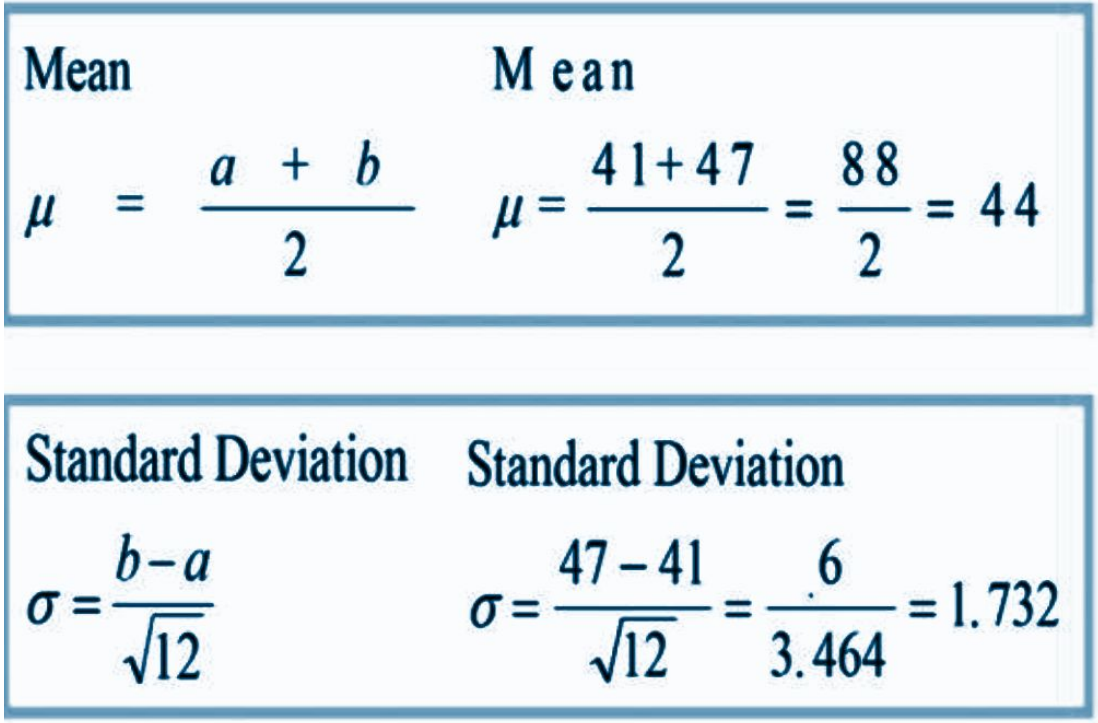
$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(6-2)^2}{12}} = 1.1547$$

Uniform distribution – Example

$$f(x)=\begin{cases}\dfrac{1}{47-41} & \text{for} \quad 41 \le x \le 47 \\ 0 & \text{for} \quad \text{all other values}\end{cases}$$



Uniform distribution – mean and S.D



Mean

$$\mu = \frac{a+b}{2}$$

Mean

$$\mu = \frac{41+47}{2} = \frac{88}{2} = 44$$

Standard Deviation

$$\sigma = \frac{b-a}{\sqrt{12}}$$

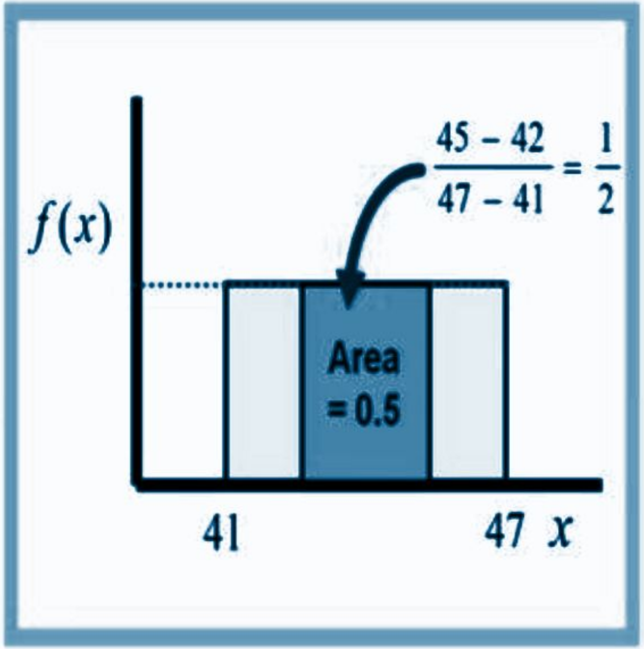
Standard Deviation

$$\sigma = \frac{47-41}{\sqrt{12}} = \frac{6}{3.464} = 1.732$$

Uniform distribution probability

$$P(x_1 \le X \le x_2) = \frac{x_2 - x_1}{b - a}$$

$$P(42 \le X \le 45) = \frac{45-42}{47-41} = \frac{1}{2}$$



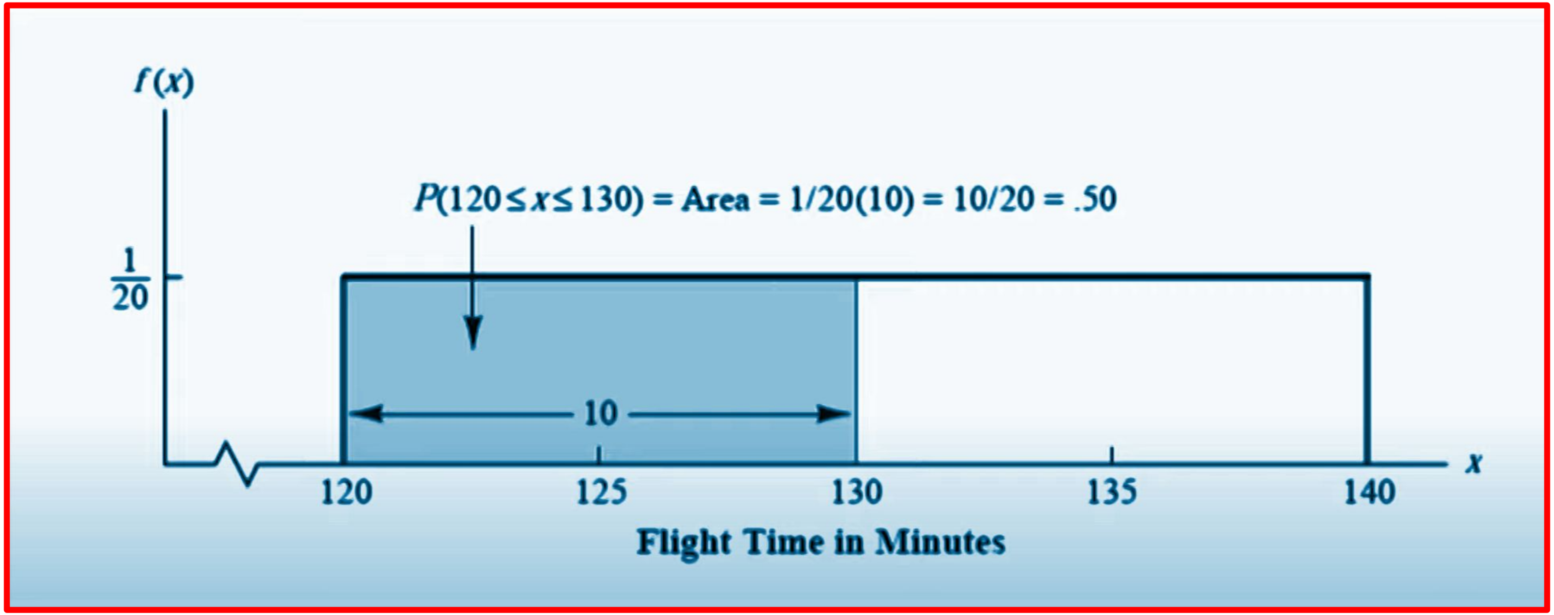
Example – Uniform distribution

- Consider the random variable x representing the flight time of an airplane traveling from Peshawar to Karachi.
- Suppose the flight time can be any value in the interval from 120 mins to 140 mins.
- Because the random variable x can assume any value in that interval, x is a continuous rather than a discrete random variable

Example – Uniform distribution (Cont'd)

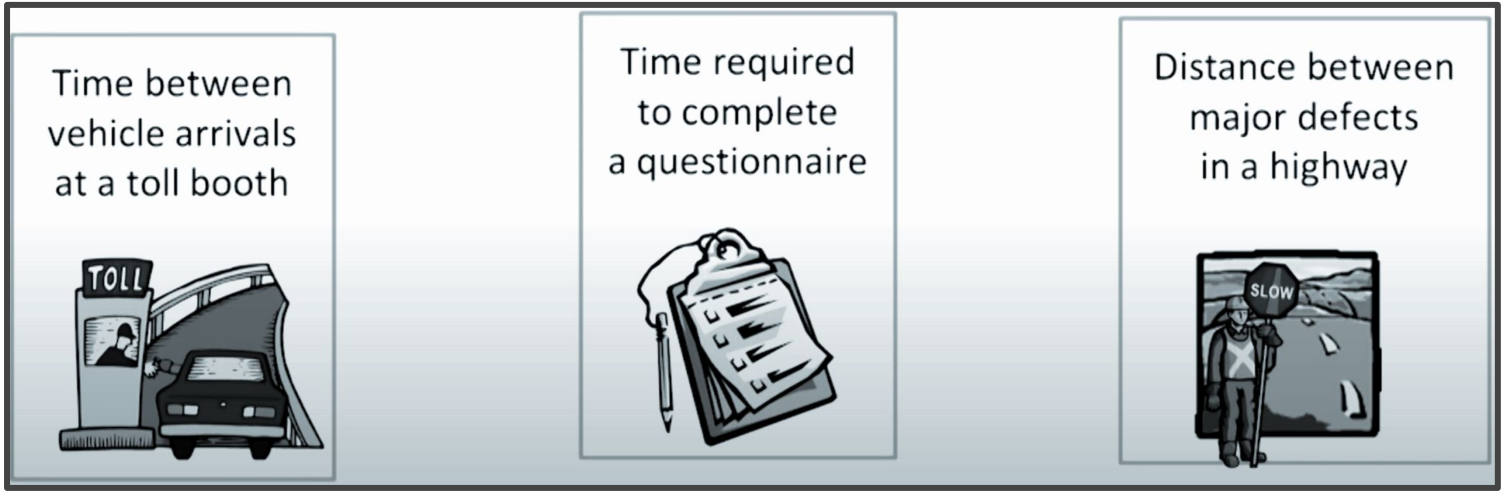
- Let us assume that sufficient actual flight data are available to conclude that the probability of a flight time within any 1- minute interval is the same as the probability of a flight time within any other 1-min interval contained in the larger interval from 120 to 140 minutes.
- With every 1- minute interval being equally likely, the random variable x is said to have a uniform probability distribution

Probability of a flight time between 120 and 140 minutes



Exponential probability distribution

- The exponential probability distribution is useful in describing the time it takes to complete a task.
- The exponential random variables can be used to describe:



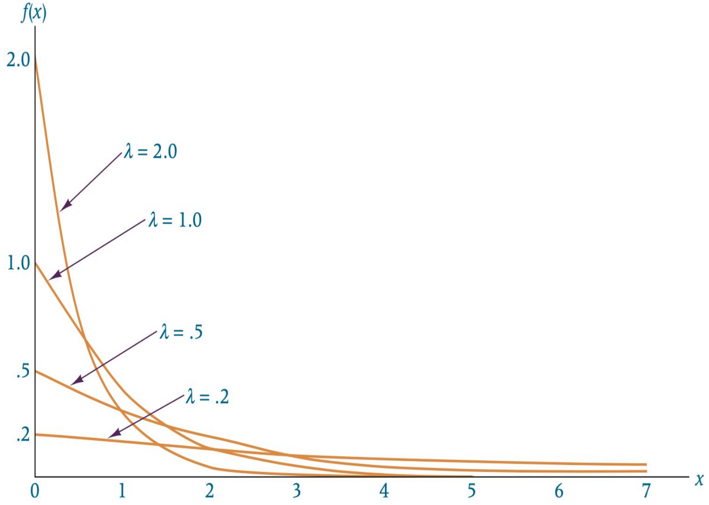
Time between vehicle arrivals at a toll booth

Time required to complete a questionnaire

Distance between major defects in a highway

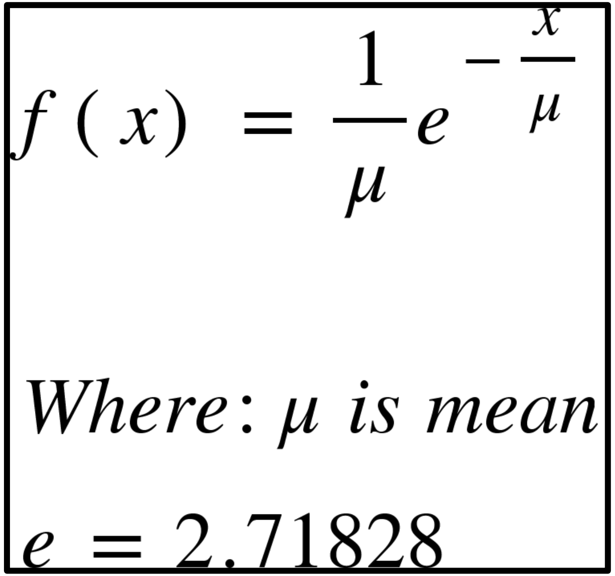
Exponential distribution

- It is a continuous distribution.
- It is a family of distributions.
- It is skewed to the right.
- The x values range from zero to infinity.
- The curve steadily decreases as x gets larger.



Exponential probability distribution

- Density function



$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$

$$Where: \mu \ is \ mean$$

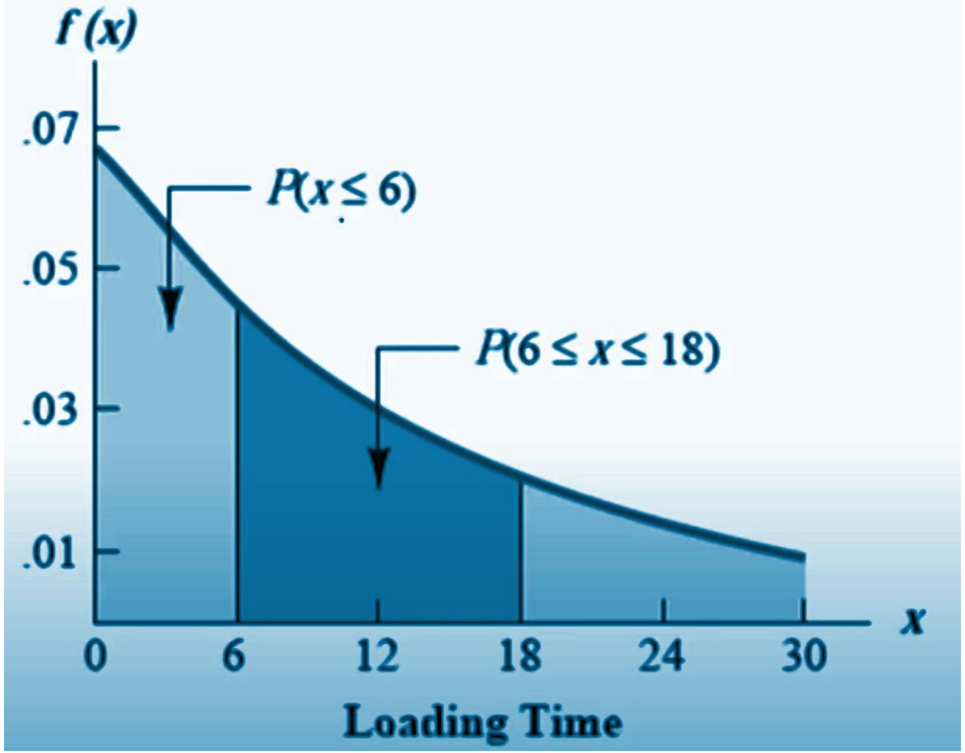
$$e = 2.71828$$

Exponential probability distribution

- Suppose that x represents the loading time for a truck at loading dock and follows such a distribution.
- If the mean, or average, loading time is 15 minutes (μ = 15), the appropriate probability density function for x is

$$f(x) = \frac{1}{15} e^{-\frac{x}{15}}$$

Exponential distribution – loading dock example



Exponential probability distribution
Cumulative probabilities

$$P\left(x \leq x_0\right) = 1 - e^{\frac{-x_0}{\mu}}$$

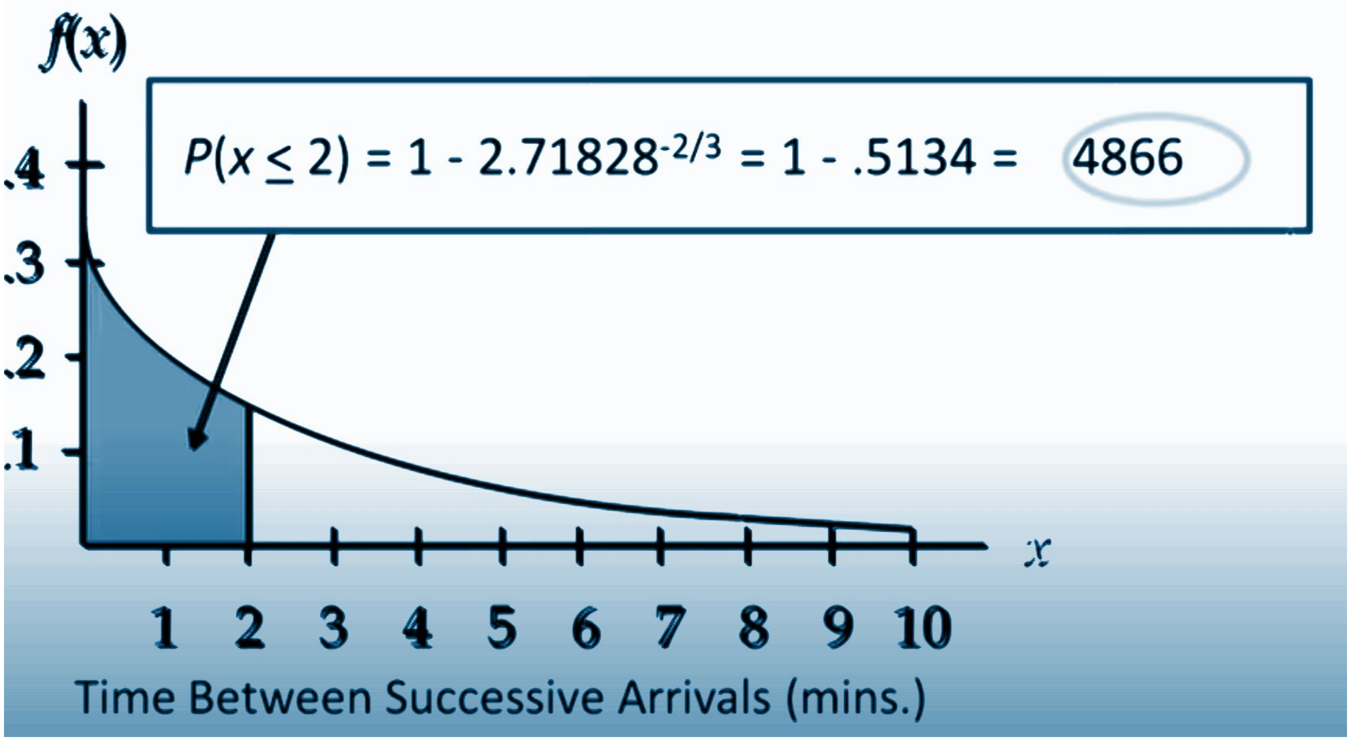
Where:

$$x_0 = \textit{some specific value of } x$$

Example – exponential distribution

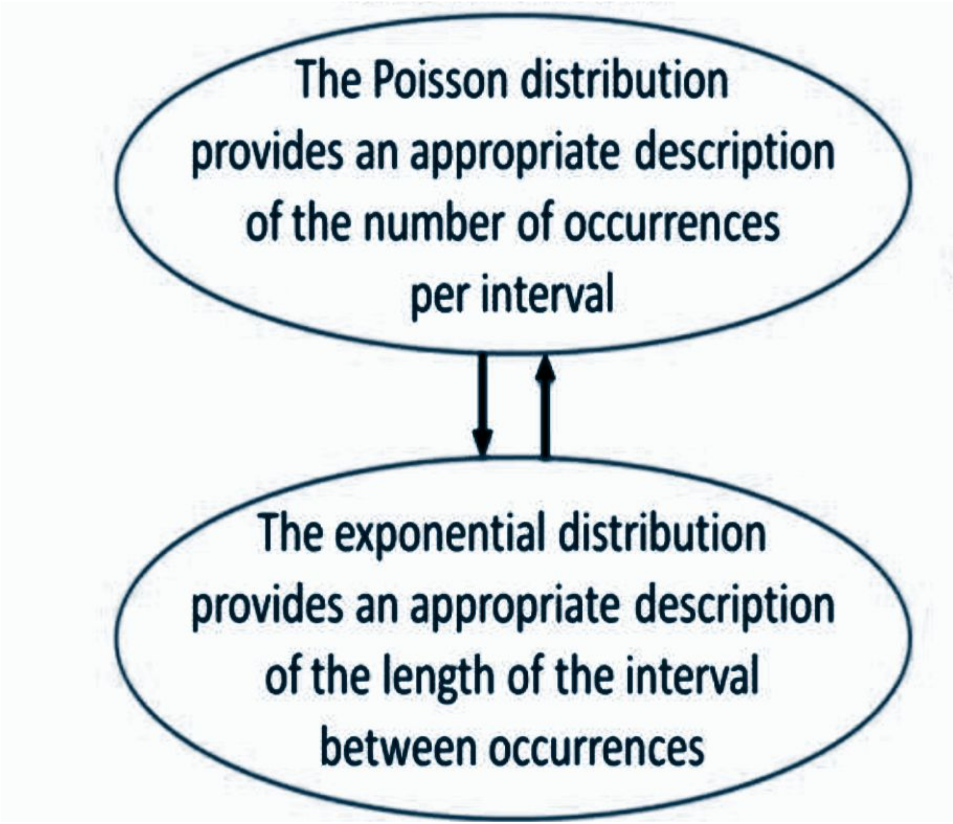
- The time between arrivals of the cars at a petrol pump follows an exponential probability distribution with a mean time between arrivals of 3 minutes.
- The petrol pump owner would like to know the probability that the time between two successive arrivals will be 2 minutes or less.

Example – petrol pump problem



$P(x \le 2) = 1 - 2.71828^{-2/3} = 1 - .5134 = 4866$

Relationship between poisson and exponential distributions



The Poisson distribution provides an appropriate description of the number of occurrences per interval

The exponential distribution provides an appropriate description of the length of the interval between occurrences