Introduction to Statistical Methods

Probability Distributions

Revision-1.0

Prof Vineet Garg

BITS Pilani Work Integrated Learning Programmes (WILP)
Bangalore Professional Development Center

Introduction

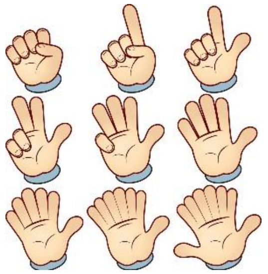
- We have reviewed that an experiment is a process that produces outcomes. A specific outcome is called an event. For example: surveying 20 randomly selected consumers, data collection of a stock exchange etc. are experiments. Negative ratings in the survey, stock exchange crossing 15,000 mark are specific events.
- We have reviewed the basic concepts of probability.
- What is Probability Distribution then?
- A probability distribution is the **mathematical function** that gives the probabilities of occurrence of different possible outcomes for an experiment. Sample space is the input and probability is the output of this function.
- The probability distribution function will be denoted by f(x), g(y), h(z) etc.
- x, y and z in the above probability distribution functions are possible outcomes of an experiment.
- If we say an experiment is conducted and possible outcomes are x, y, z etc. What is the probability that outcome x occurs? It can be denoted as P(X = x). Where P is the probability and x is a specific value of X.
- X is called the Random Variable, which is taking values as x, y, x etc.
- Altogether, f(x) = P(X = x) that it is a Probability Distribution function that assigns probability to each possible outcome.

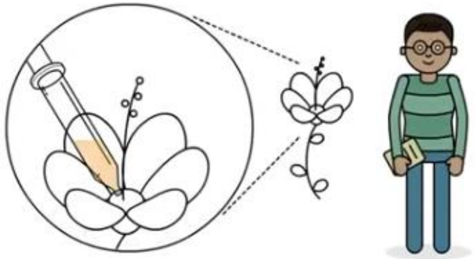
Random Variables

Random Variable is a variable that contains a numerical value to each possible outcome of an experiment.

Example-1: How many cars (X) arrive in a 30 second time interval on a toll booth?

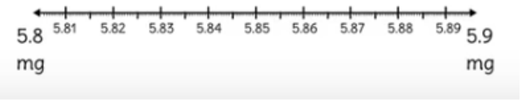
- It could be 0, 1, 2, 3 and so on. Some countable numbers.
- Based on some experiment it is established that the probability that 5 cars arrive on a particular toll booth in 30 second interval is 0.20.
- Then P(X=5) = 0.20
- The **set** of all such possible outcomes of the experiments (0 cars, 1 car, 2 cars etc.) is called the Probability Distribution function.
- Here X, is taking <u>countable</u> values. X is called **Discrete Random Variable**.





Example-2: A botanist measures the amount of nectar available in flowers that makes honey bees attracted towards them.

- Let us say, a flower can have 5.800 mg nectar, another flower can have 5.810 mg, another one 5.811 mg and so on. Essentially **all possible real values**.
- We might be limited by the precision of our measuring scale but the amount of nectar a flower can have could be one from the *possibly infinite values*.
- What is the probability that a randomly selected flower has amount of nectar in the range of 5.80 <= x <= 5.85?
- Here the values of X is <u>measurable</u>. X is called the **Continuous Random Variable**.



Basic Properties

Before any calculations are performed, it is necessary to check if the PDFs are valid using these two properties.

- We reviewed that f(x) = P(X = x).
- There are two important and basic properties associated with f(x).

1. $\mathrm{f(x)} \geq 0$
2. $\sum_{\text{all x}} \mathrm{f(x)} = 1$

Check if the following functions can serve as probability distribution functions for possible outcomes?

Example-1:

$$\mathrm{f(x)} = \frac{x-2}{2} \; for\; x = 1, 2, 3, 4.$$

Answer: No, because for x = 1, f(x) will be negative.

Example-2:

$$\mathrm{f(x)} = \frac{x^2}{25} \; for\; x = 0, 1, 2, 3, 4.$$

Answer: No because for Σf(x) = 30/25 = 6/5 which is > 1.

Example-3:

$$\mathrm{f(x)} = \frac{1}{4} \; for\; x = 10, 11, 12, 13.$$

Answer: Yes, There 4 possible outcomes each has the same probability that sums up to 1.

Different Types of Distribution

There are different distributions depending on the type of Random Variable (Discrete or Continuous):

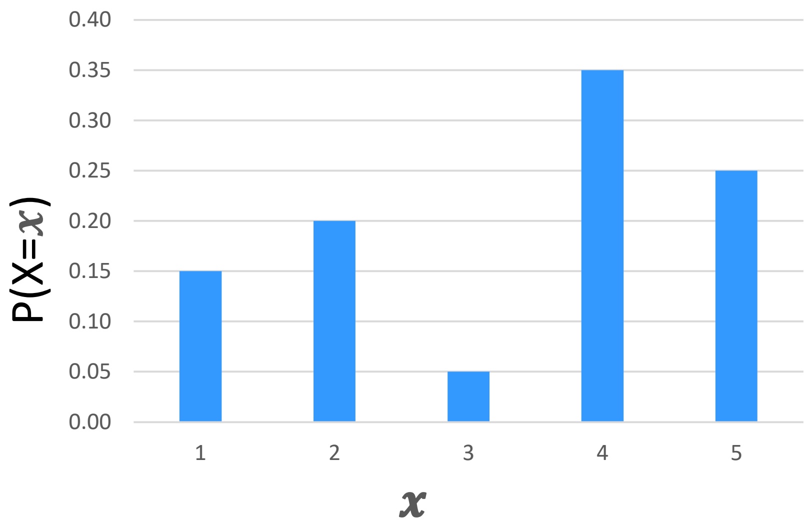
- Discrete Distributions:
    1. Bernoulli Distribution
    2. Binomial Distribution
    3. Poisson Distribution
    4. Hypergeometric Distribution
- Continuous Distributions:
    1. Uniform Distribution
    2. Normal or Gaussian Distribution
    3. General Gamma Distribution and its types like Exponential and Chi-squared Distributions
    4. Beta Distribution
    5. Weibull Distribution
    6. Log Normal Distribution
    7. Student's t Distribution
    8. F Distribution
- We will review few Discrete and few Continuous Distributions in this course.

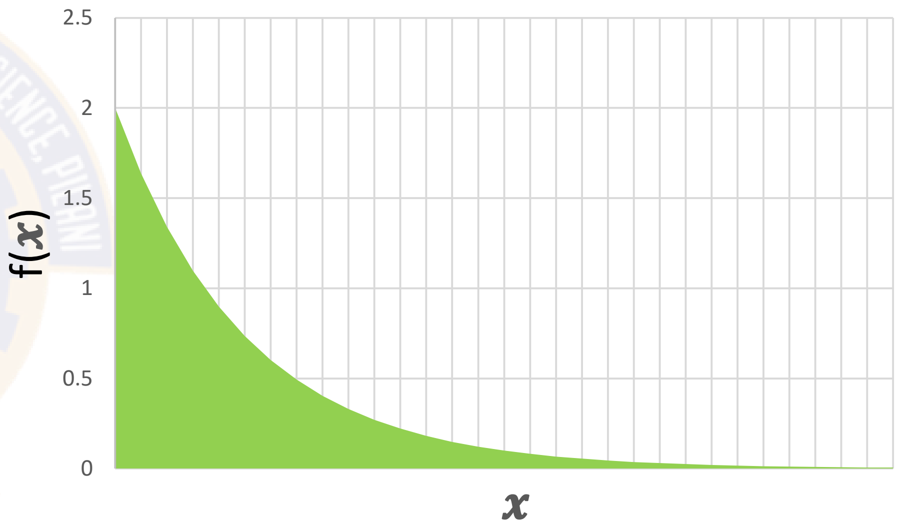
Review Questions

What types of Probability Distributions are these?

Are they valid Probability Distributions?

What are X and x in these two examples?





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

$$f(x) = \begin{cases} 2.e^{-2x} & for\ x \geq 0 \\ 0 & for\ x < 0 \end{cases} \qquad Also, \int_0^{\infty} 2.e^{-2x}dx = 1$$

Why P(X = x) notation is not used above?

Discrete Distributions

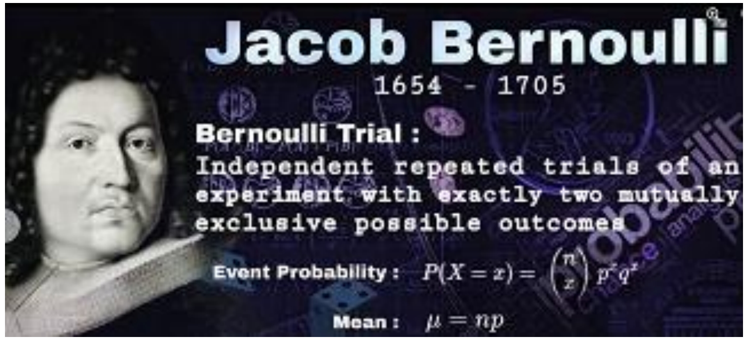
Bernoulli Distribution

- **Bernoulli Distribution** is a basic discrete distribution applicable to those experiments which have *only 2 possible outcomes*. These experiment trials are popularly known as **Bernoulli Trials.**
- It is named after Swiss mathematician **Jacob Bernoulli.**
- The probability of a Bernoulli event is given by the equation:

$$P(X = x) = p^x(1 - p)^{1-x}, \text{where } x \in \{0, 1\} \text{ and } p \text{ is the probability of getting } x = 1$$

Example:

- Let us say for discrete random variable X, there are two possibilities in an experiment:
    - The person will buy an Android mobile phone (X = 0)
    - The person will buy an iOS mobile phone (X = 1)
- Let us say, 20% of all buyers are buying iOS mobile phones (X = x = 1). So,
- The probability that randomly selected person buys:
    - $P(X = 1) = p^x(1 - p)^{1-x} = 0.20^1.(1 - 0.20)^{1-1} = 0.20$
    - $P(X = 0) = p^x(1 - p)^{1-x} = 0.20^0.(1 - 0.20)^{1-0} = 0.80$
    - $P(X \neq 1 \text{ and } X \neq 0) = 0$



Binomial Distribution

Most widely known of all discrete distributions is the **Binomial Distribution**. Several assumptions form the basis of the binomial distribution:

- The experiment involves *n* identical trials. These trials are named as *Bernoulli Trials* with following characteristics:
- Each trial has only two possible and mutually exclusive outcomes. They are normally denoted as ***success*** and as ***failure***. Usually the outcome of interest is labeled as *success*. Depending on the context these two outcomes could be anything: Male/Female, Acceptable/Defective etc.
- Each trial is independent of the previous trials.
- If the term *p* is the probability of getting a success on any one trial and the term *q* = (1 - *p*) is the probability of getting a failure on any one trial. The terms *p* and *q* remain constant throughout the experiment.

Example: A Human Resources Consultancy firm conducted a survey and found several reason why employee reject job offers when they need to relocate from their present locations. These reasons are family considerations, financial reasons, relocation support, emotional reasons, children schooling etc. 4% percent of the respondents said they rejected relocation offers because they received very little relocation support from the employer. Suppose 5 candidates who just rejected relocation offers are randomly selected and surveyed, what is the probability that exactly one candidate interviewed rejected the offer because of very little relocation support from the employer?

Given that, the probability of rejecting an offer because of very little relocation support = 0.04, so the probability for rejecting the offer for other reasons = 0.96.Let us say T represents the relocation support and R other reasons. The possible sequences of getting exactly one candidate selecting T and other four as R with their corresponding probabilities are following:

T1, R2, R3, R4, R5, P = **(0.04)**.(0.96).(0.96).(0.96).(0.96) = 0.034

R1, **T2**, R3, R4, R5, P = (0.96).**(0.04)**.(0.96).(0.96).(0.96) = 0.034

R1, R2, **T3**, R4, R5, P = (0.96).(0.96).**(0.04)**.(0.96).(0.96) = 0.034

R1, R2, R3, **T4**, R5, P = (0.96).(0.96).(0.96).**(0.04)**.(0.96) = 0.034

R1, R2, R3, R4, **T5**, P = (0.96).(0.96).(0.96).(0.96).**(0.04)** = 0.034

= 5*0.034 = 0.17 **= $(5).(0.04)^1.(0.96)^4$**

Binomial Distribution: Mathematical Formulation

- *n identical and independent trials.*
- *Each trial has two possible but mutually exclusive outcomes.*

Let for the previous example:

n = the number of trials (or the sample size) = 5

x = the number of successes desired = 1

p = the probability of getting a success in one trial = 0.04

q = (1 – p) = the probability of getting a failure in one trial = 0.96

There are ${}^{n}C_{x}$ ways to select x out of n candidates who rejected the offer because of very little relocation support.

In the last example, the probability was = $\mathbf{(5).(0.04)^1.(0.96)^4}$

So generalizing the example, the formula for probability for **Binomial Distribution** can be written as:

$$P(x) = {}^{n}C_{x}.\, p^{x}.q^{(n-x)}$$

Example-1: A survey found that 65% of all consumers were very satisfied with their primary saving bank. Suppose 25 consumers are sampled; what is the probability that exactly 19 are very satisfied with their primary saving bank?

Answer:

n = 25, x = 19, p = 0.65, q = (1 - 0.65) = 0.35

$P(19) = {}^{25}C_{19}.(0.65)^{19}.(0.35)^{(25-19)} = \mathbf{0.091}$

Example-2: According to the Tribal Commission, approximately 6% of adult population in Kalahandi find it difficult to earn their livelihood. What is the probability of getting two or fewer adults in a sample of 20 who find it difficult?

Answer:

n = 20, x = 0, 1 or 2, p = 0.06, q = (1 - 0.06) = 0.94

$P(0, 1, \text{or } 2) = {}^{20}C_{0}.(0.06)^{0}.(0.94)^{20} + {}^{20}C_{1}.(0.06)^{1}.(0.94)^{19} + {}^{20}C_{2}.(0.06)^{2}.(0.94)^{18} = 0.290 + 0.370 + 0.225 = \mathbf{0.89}$

Exercise

What is the big impact that hit the Indian urban society because of COVID-19 pandemic? According to a survey, 30% said that it was cutting on recreational activities. However, 27% said that it was completing the household chores themselves. If 20 household are randomly sampled and asked what is the first big change that hit:

1. What is the probability that exactly 8 say that it was completing the household chores themselves?

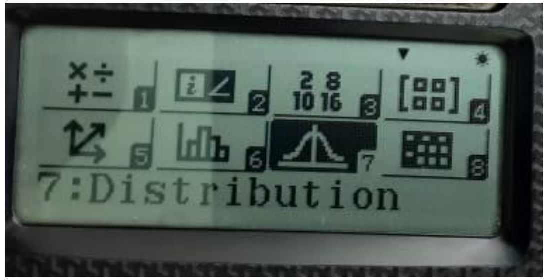
   Answer = 0.081

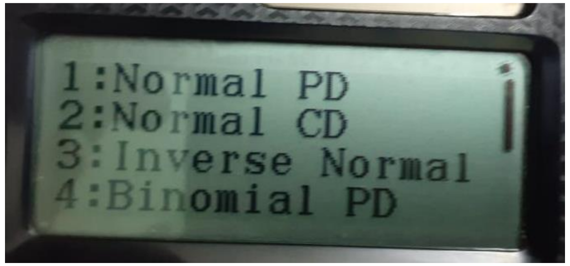
2. What is the probability that none of them say that it was cutting on recreational activities?

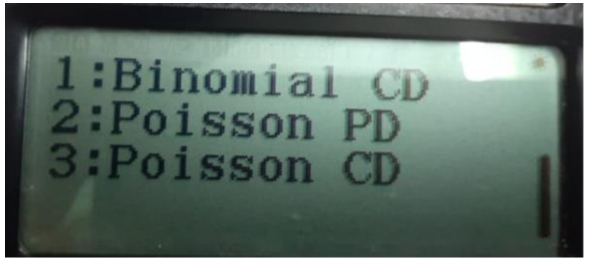
   Answer = 0.00080

3. What is the probability that more than 7 say that it was cutting on recreational activities?

   Answer = 0.23







Poisson Distribution

- The **Poisson Distribution** is another discrete distribution. It is named after Simeon-Denis Poisson (1781–1840), a French mathematician, who published its essentials in a paper in 1837.
- Poisson distribution and the binomial distribution have some similarities but also several differences. The binomial distribution describes a distribution of two possible outcomes designated as successes and failures from a given number of trials. The Poisson distribution focuses only on the number of discrete occurrences over some **interval**.
- In a Poisson experiment, count of trials or samples (n) are not needed.

Example: On a toll booth:

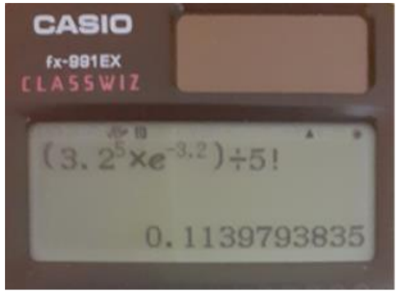
- In a random sample of 20 cars, how many are Hyundai cars? ✓ Binomial Distribution
- How many Hyundai cars arrived in a 10 minute interval? ✓ Poisson Distribution
- The Poisson Distribution has it utility in identifying the probability of ***improbable*** or ***rare*** events. For example: serious accidents at a chemical plant are rare, and the count of such accidents per year might be described by the Poisson Distribution.
- The Poisson Distribution has the following characteristics:
  - It is a discrete distribution that describes the count of events over a time interval.
  - Each occurrence is independent of the other occurrences.
  - The occurrences in each interval can range from zero to infinity.
  - The expected (average) number of occurrences must be constant throughout the experiment.

Poisson Distribution: Mathematical Formulation

- If a Poisson Distributed phenomenon is studied over a long period of time, a long-run average can be determined. This average is denoted lambda **(λ)** or mu (μ). Each Poisson problem contains a long-run average value from which the probabilities of particular occurrences are determined. Here $\lambda > 0$.
- A Poisson distribution can be described by **λ** alone. The Poisson formula is used to compute the probability of occurrences over an interval is following: (Proof in the appendix).

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

If $x \longrightarrow \infty$ then $P(x) \longrightarrow 0$



Where:
$x = 0, 1, 2, 3\ldots$
λ= long run average
e = Base of natural log / Euler's Constant

Example-1: Bank customers arrive randomly on weekday afternoons at an average of 3.2 customers every 4 minutes. What is the probability that exactly 5 customers arriving in a 4-minute interval on a weekday afternoon?

Answer:
Given that, $\lambda = 3.2$ and $x = 5$
So $P(5) = (3.2)^5 \times e^{-3.2} / 5! =$ **0.1140**

Example-2: If a real estate office sells 1.6 houses on an average weekday and sales of houses on weekdays are Poisson distributed. What is the probability of selling more than 5 houses in a day?

Answer: 0.0060

λ = 1.6	
x	**Probability**
0	0.2019
1	0.3230
2	0.2584
3	0.1378
4	0.0551
5	0.0176
6	0.0047
7	0.0011
8	0.0002
9	0.0000

0.0060

Exercise

1. Evaluate the following expression for Poisson Distribution: $P(4 < x < 8$, given that $\lambda = 4.4)$.

Answer: 0.3702

2. 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?

Answer: 0.0169

3. The average number of annual trips per family to amusement parks in the India is Poisson distributed, with a mean of 0.6 trips per year. What is the probability of randomly selecting an Indian family and finding the that the family did not make a trip to an amusement park last year?

Answer: 0.5488

Hypergeometric Distribution

- There is a lot of ***N*** units in which there are ***a*** defective units. ***n*** units are sampled (trials), and we are interested in the number of defective units. (being defective is just an example here; **a** is basically the count of successes desired depending on the context)
- The experiment is conducted **without replacement**. That means sampled item is not mixed back to the lot.
- The probability that the first draw will yield a defective unit = **a / N.** What is the probability of getting a defective in the second draw?
- It is **(a-1)/(N-1)** if the first unit was defective or **(a/N-1)** if the first unit was not defective. **It means the trials are not independent.**
- In such cases, **Hypergeometric Distribution** is used to calculate the probability. Its mathematical formulation is as follows:
- Total ways to select n units out of N = ${}^{N}C_{n}$ ----- (i)
- Ways to select x defectives out of a = ${}^{a}C_{x}$ ----- (ii)
- Ways to select (n-x) non-defectives out of (N-a) = ${}^{(N-a)}C_{(n-x)}$ ----- (iii)
- So the probability of getting x defective in n trials:

$$P(x) = \frac{{}^{a}C_{x} \cdot {}^{(N-a)}C_{(n-x)}}{{}^{N}C_{n}}$$

Where:

N = size of the population
n = sample/trial size
a = count of successes in the population
x = count of desired successes in the sample/trials

Examples

Example-1: 24 people, of whom 8 are women, apply for a job. If 5 of the applicants are sampled randomly one-after-another, what is the probability that exactly 3 of those sampled are women?

Answer:

Given that: N = 24, a = 8, n = 5, x = 3

So, $P(3) = {}^{8}C_{3} \cdot {}^{(24-8)}C_{(5-3)} / {}^{24}C_{5} = \mathbf{0.1581}$

Example-2: An online retailer sells mobile phone accessories. Out of 20 car chargers, 5 are defective. If retailer selects 10 of these chargers one-after-another what is the probability that 2 will be defective?

Answer:

Given that: N = 20, a = 5, n = 10, x = 2

So, $P(3) = {}^{5}C_{2} \cdot {}^{(20-5)}C_{(10-2)} / {}^{20}C_{10} = \mathbf{0.3483}$

$$P(x) = \frac{{}^{a}C_{x} \cdot {}^{(N-a)}C_{(n-x)}}{{}^{N}C_{n}}$$

Where:
N = size of the population
n = sample size
a = count of successes in the population
x = count of successes desired in the sample

Exercise

Out of 18 major IT companies operate in the US and 12 are located in Silicon Valley. If 3 IT are selected randomly from the entire list one-after-another, what is the probability that <u>one or more</u> of the selected companies out of 3 are located in the Silicon Valley?

<u>Solution:</u>

This can be solved in a different way:

Let us find out the probability that no company is located in Silicon Valley (x=0).

So, N = 18, a = 12, n = 3, x = 0

$P(X) = {}^{12}C_0 \cdot {}^{(18-12)}C_{(3-0)} \,/\, {}^{18}C_3 = 0.0245$

So the probability that one or more out of 3 companies are located in Silicon Valley = 1 – 0.0245 = **0.9755**

$$P(x) = \frac{{}^{a}C_{x} \cdot {}^{(N-a)}C_{(n-x)}}{{}^{N}C_{n}}$$

Where:

N = size of the population

n = sample size

a = count of successes in the population

x = count of successes desired in the sample

Discrete Distributions: Comparison

Binomial Distribution

Experiment: *n* independent trials with two mutually exclusive outcomes (success/failure).

The probability of getting *x* success from these *n* trials:

$$P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{(n-x)}$$

Where:
n = the number of trials (samples)
x = the number of successes desired
p = the probability of getting a success in one trial
q = (1 – p) = the probability of getting a failure in one trial

Population size not needed. Trial count (n) and probability of success/failure in each trial are known!

Poisson Distribution

Experiment: Count of trials are not needed but long run average ($\lambda > 0$) of some event occurrence is available for some time interval.

The probability of the occurrence of an event x in that time interval:

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

Where:
x = 0, 1, 2, 3…
λ= long run average
e = Base of natural log / Euler's Constant

Concept of long run average (λ) involved. Population size and trial count not needed!

Hypergeometric Distribution

Experiment: *n* dependent samples (trials) from population N with two mutually exclusive outcomes (success/failure).

Experiment without replacement.

The probability of getting *x* success from these *n* trials:

$$P(x) = \frac{{}^{a}C_{x} \cdot {}^{(N-a)}C_{(n-x)}}{{}^{N}C_{n}}$$

Where:
N = size of the population
n = sample size
a = count of successes in the population
x = count of successes desired in the sample

Population size (N) and success size (a) in this population are known!

Describing a Discrete Distribution

- We have reviewed few discrete distributions. Now let us review how we can describe these distributions.
- *Describing something in statistics? What is that sub-area called? What do we do in that?*

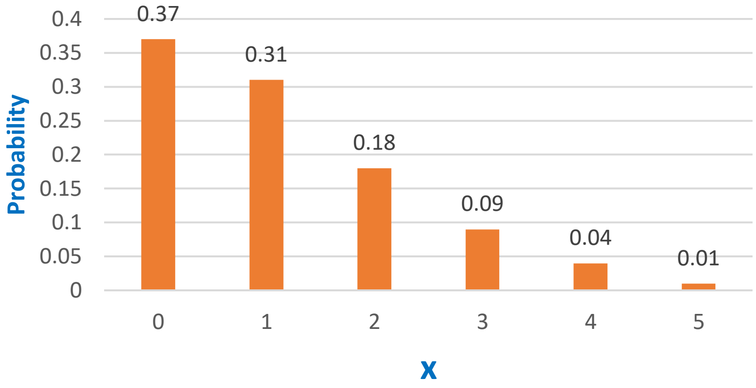
Example: The lady in the house that is full of kids needs to boil the milk several times in a day. Few times the milk boils over and spills. The probability how many times it can happen in a single day is captured in the table.

- Is it a discrete distribution?
- What is the random variable and what values it assumes?
- Is it a valid distribution function?
- *How can we describe this distribution?*

Spill Incidents (x)	Probability P(x)	x.P(x)	$(x-\mu)^2$	$(x-\mu)^2.P(x)$
0	0.37	0.00	1.32	0.49
1	0.31	0.31	0.02	0.01
2	0.18	0.36	0.72	0.13
3	0.09	0.27	3.42	0.31
4	0.04	0.16	8.12	0.32
5	0.01	0.05	14.82	0.15
Σ	**1.00**	**μ = 1.15**		**σ² = 1.41**

- Bar chart is the most common way to describe the distribution.
- We get a quick idea how many spills are most frequent. Will be helpful in large dataset.
- The mean or **expected value** of a discrete distribution is the long-run average of occurrences $\mu = \sum x.P(x) = 1.15$
- The variance $\sigma^2 = \sum (x-\mu)^2.P(x)$ = 1.41
- The standard deviation $\sigma = \sqrt{1.41}$ = 1.19

How these formula are established?



Binomial Distribution: Average and Std. Deviation

- We have reviewed that in Binomial Distribution, the probability of x:

$$P(x) = {}^{n}C_{x} \, . \, p^{x}.q^{(n-x)}$$

Where:
n = the number of trials / the number being sampled
x = the number of successes desired
p = the probability of getting a success in one trial
q = (1 – p) = the probability of getting a failure in a trial

- A binomial distribution has an **expected value** or a **long-run average**, which is denoted by **μ**. The value of μ is determined:

$$\mu = n.p$$

Example: According to one study, 64% of all bank customers believe that banks are more competitive today than they were a decade ago. If 23 financial customers are selected randomly, what is the expected number who believe it?

Given that n = 23, p = 0.64 so q = (1-p) = 0.36
So, long run average or expected value = n.p = 23x0.64 = **14.72**

In the long run, if 23 financial customers are selected randomly over and over and if 64% of all financial customers believe banks are more competitive today, then the experiment should average 14.72 financial customers out of 23 who believe banks are more competitive today. This is evident from the table because the probability is highest around x = 14 and 15.

- The standard deviation of Binomial Distribution is denoted by **σ** and defined as below:

$$\sigma = \sqrt{n.p.q}$$

For the above example, σ = √ (23x0.64x0.36) = **2.3**

- What does the standard deviation signify?

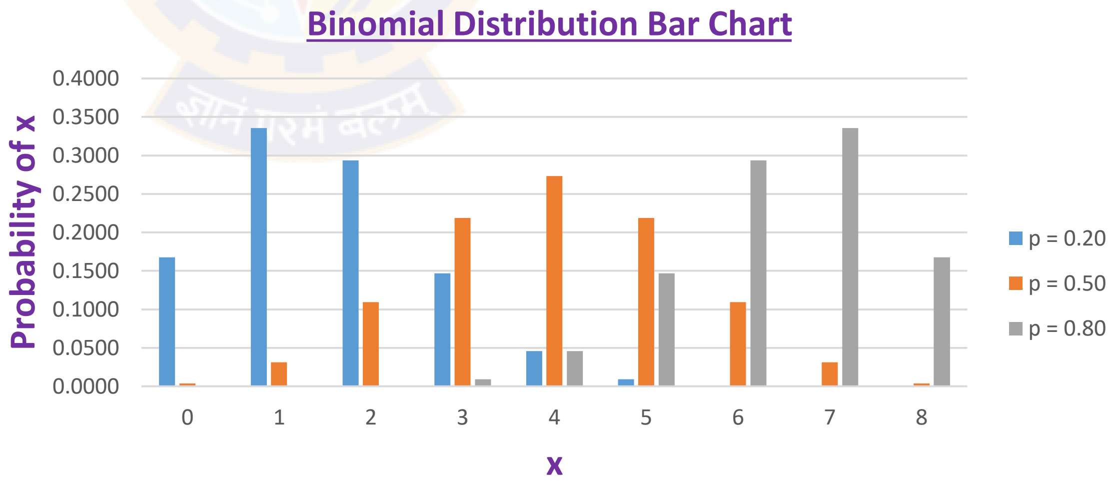
n = 23, p = 0.64, q = 0.36	
x	P(x)
0	0.0000000
1	0.0000000
2	0.0000000
3	0.0000006
4	0.0000055
5	0.0000373
6	0.0001988
7	0.0008581
8	0.0030511
9	0.0090402
10	0.0225000
11	0.0472728
12	0.0840406
13	0.1264200
14	0.1605333
15	0.1712355
16	0.1522094
17	0.1114212
18	0.0660274
19	0.0308900
20	0.0109831
21	0.0027894
22	0.0004508
23	0.0000348

Graphing Binomial Distribution

$P(x) = {}^{n}C_{x} \cdot p^{x} \cdot q^{(n-x)}$

- A table and bar chart for the Binomial Distribution for n = 8 and three different values of p (0.20, 0.50 and 0.80) for the different values of x from 0 to 8 are shown below:
- For p = 0.50, the distribution is symmetrical. For p = 0.20 the distribution is skewed right (positive) and for p = 0.80 the distribution is skewed left (negative).
- The pattern is intuitive because:
  - The mean of the binomial distribution n=8 and p = 0.50 is 4, which is in the middle of the distribution.
  - The mean of the binomial distribution n=8 and p = 0.20 is 1.6. Therefore the probabilities near x = 1 and x = 2 are high.
  - The mean of the binomial distribution n=8 and p = 0.80 is 6.4. Therefore the probabilities near x = 6 and x = 7 are high.

n = 8			
x	p = 0.20	p = 0.50	p = 0.80
0	0.1678	0.00391	0.0000
1	**0.3355**	0.03125	0.0001
2	**0.2936**	0.10938	0.0011
3	0.1468	0.21875	0.0092
4	0.0459	**0.27344**	0.0459
5	0.0092	0.21875	0.1468
6	0.0011	0.10938	**0.2936**
7	0.0001	0.03125	**0.3355**
8	0.0000	0.00391	0.1678



Example

Observe these tables from the different books and keep them handy!

Many statistics books compile the pre-calculated probability distribution tables in their appendix. This exercise shows the usage of the Binomial Distribution Table.

A manufacturing company produces 10,000 baby rides per month. This company supplies rides to another wholesaler. The wholesaler randomly samples 10 rides per lot sent from the manufacturer. If 2 or fewer of the randomly sampled rides are defective in the lot, the wholesaler accepts the lot. In case, manufacturer supplies 4 lots which have 10%, 20%, 30% and 40% defective rides respectively, what are the probabilities that the wholesaler will accept the lots?

Answer:

The question indirectly asks to calculate the cumulative probability (P) for x = 0, 1 and 2 for probabilities(p) of 10%, 20%, 30% and 40% defective rides when n = 10.

$n = 10$

	Probability								
x	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	.349	.107	.028	.006	.001	.000	.000	.000	.000
1	.387	.268	.121	.040	.010	.002	.000	.000	.000
2	.194	.302	.233	.121	.044	.011	.001	.000	.000
3	.057	.201	.267	.215	.117	.042	.009	.001	.000
4	.011	.088	.200	.251	.205	.111	.037	.006	.000
5	.001	.026	.103	.201	.246	.201	.103	.026	.001
6	.000	.006	.037	.111	.205	.251	.200	.088	.011
7	.000	.001	.009	.042	.117	.215	.267	.201	.057
8	.000	.000	.001	.011	.044	.121	.233	.302	.194
9	.000	.000	.000	.002	.010	.040	.121	.268	.387
10	.000	.000	.000	.000	.001	.006	.028	.107	.349

p	P(x = 0)	P(x = 1)	P(x = 2)	ΣP
0.10	0.349	0.387	0.194	**0.930**
0.20	0.107	0.268	0.302	**0.677**
0.30	0.028	0.121	0.233	**0.382**
0.40	0.006	0.040	0.121	**0.167**

*When there are 10% defective rides, there are highest chances that the lot will be accepted (**93%**), when 40% rides are defective, the chances are the least (just **16.7%**).*

Poisson Distribution: Average and Std. Deviation

- We have reviewed that in Poisson Distribution, the probability of x:

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

Where:
x = 0, 1, 2, 3…
λ= long run average and λ > 0
e = Base of natural log / Euler's Constant

- The λ is defined as the mean and $\sqrt{\lambda}$ as the standard deviation in the Poisson Distribution.

This is the reason, why few books represent λ as μ in the Poisson Distribution formula.

Example: In an experiment given that λ = 6.5. With this long run average, what will be the range of 75% of the values of the random variable? **Hint:** Chebyshev's Formula for range.

Given that λ = 6.5, so standard deviation (σ) = √6.5 = 2.55

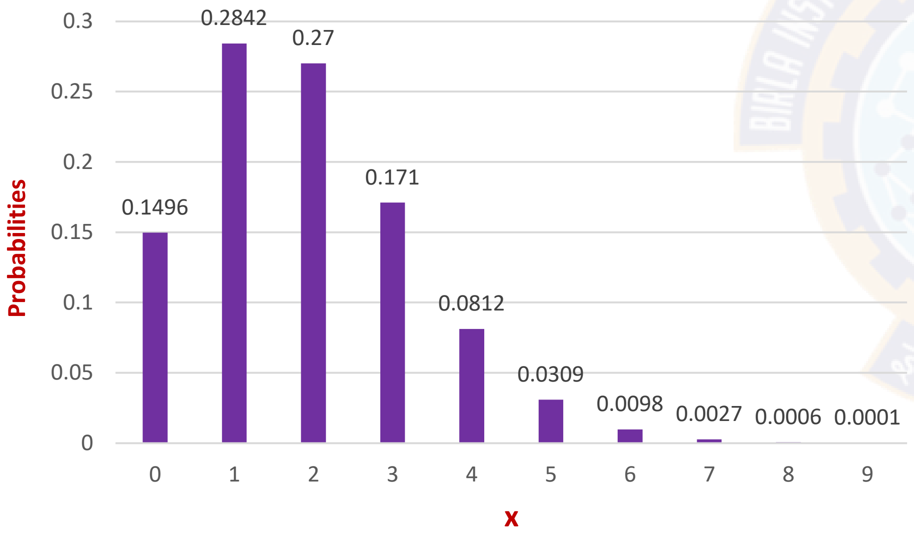
From the Chebyshev's Theorem, we know that at least $1-(1/k)^2$ values fall within the range of (μ ± k.σ).

For $1-(1/k)^2 = 0.75$, so k = 2

So, the range will be (μ ± k.σ) or 6.5 ± 2x2.55 = 6.5 ± 5.1 or from **1.4 to 11.6**

Graphing Poisson Distribution

National Center for Health Statistics (NCHS) claims that on average an American has 1.9 acute illnesses or injuries per year. If these cases are Poisson distributed, λ is 1.9 per year. Chart the distribution using the Poisson Distribution Table.



Positively Skewed Distribution. Intuitively most values are near 1 and 2.

	λ									
x	.005	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.9950	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
1	.0050	.0099	.0196	.0291	.0384	.0476	.0565	.0653	.0738	.0823
2	.0000	.0000	.0002	.0004	.0008	.0012	.0017	.0023	.0030	.0037
3	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001
x	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

Poisson Approximation to Binomial Distribution

- As an acceptable rule of thumb when **sample space (n) ≥ 20** and **probability of success (p) ≤ 0.05** the Poisson Approximation can be used for Binomial Distribution.
- If **sample space (n) ≥ 100**, the approximation renders excellent results as long as **np ≤ 10**.

Example: A printing press is known to have 5% of the books printed incorrectly. What is the probability that 2 out of 100 books will be printed incorrectly?

Binomial Distribution

$$n = 100, p = 0.05, x = 2$$

$$P(x) = {}^{n}C_{x}\,.\,p^{x}.q^{(n-x)}$$

$$P(2) = 100C_{2}\,.(0.05)^{2}.(0.95)^{(100-2)} = \mathbf{0.081}$$

Poisson Approximation to Binomial Distribution

$$n = 100, p = 0.05, x = 2$$

$$\lambda = \mu = n.p = (100).(0.05) = 5$$

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

$$P(2) = \frac{5^{2}.e^{-5}}{2!} = \mathbf{0.084}$$

PDF (PMF) and CDF: Discrete Distributions

IMPORTANT

- In the beginning of this module we reviewed what a Probability Distribution Function is. Now let us review few more details in the context of Discrete Distributions.
- **Probability Distribution or Probability Mass Function (PDF or PMF):** It specifies the probability of every possible value of a discrete random variable.
- **Example:** In a lab 6 high-end computers are reserved for Data Science experiments. The probabilities are shown below that how many out of 6 are occupied between 9-11 am on any weekday. The discrete random variable is X. It takes a value from x = 0 to 6.

x	0	1	2	3	4	5	6
P (X=x)	0.05	0.10	0.15	0.25	0.20	0.15	0.10

Note that the sum of probabilities is 1.0 and no probability value is negative, so it is a valid PDF (or PMF).

- **Cumulative Distribution Function (CDF):** F(x) of a discrete random variable X with PDF P(X=x) is defined for every number x such that:

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} P(y)$$

- **Example:** For the above computers example, CDF is shown below:

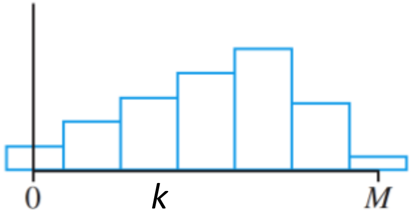
x	0	1	2	3	4	5	6
F(x)	0.05	0.15	0.30	0.55	0.75	0.90	1.00

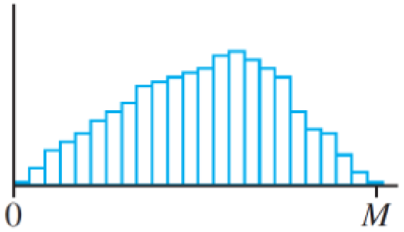
CDF answers questions like: *what is the probability that at most 3 computers are occupied in the aforementioned duration?*

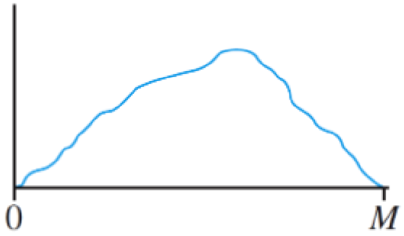
Continuous Distributions

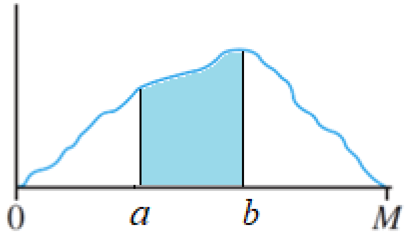
Introduction

- Suppose the Random Variable X of interest is the depth of a lake at a randomly chosen point. Let M = the maximum depth (in meters), so that any number in the interval [0, M] is a possible value of X.
- If we *discretize* X by measuring depth to the nearest meter, then possible values are non-negative integers less than or equal to M.
- A histogram can be drawn in such a way so that the area of the rectangular bar above any integer k is the proportion of the lake whose depth is k meters. Example:
  - 1/10th (= 10%) of the lake is 5 meters deep so histogram bar area at 5 will be 0.1
  - 1/20th (= 5%) of the lake is 10 meters deep, so histogram bar area at 10 will be 0.05
  - and so on….for all bars in the histograms.
  - In general, area of each bar at x is the probability of the lake for being x meters deep.
  - Sum of all bar areas in the histogram will be 1 (= 100%)
- If the depth is measured much more accurately (e.g. in centimeters), the histogram bars are much narrower, though the sum of all bars areas will still be 1.
- If we continue in this way to measure depth more and more accurately, the resulting sequence of histograms approaches to a *smooth curve*. The total area under the smooth curve will still be 1.
- It can also be said that the probability that depth of the lake will be between a and b, will be the area under the curve between points a and b (shaded region).









PDF and CDF: Continuous Distributions

IMPORTANT

- **Probability Distribution Function:** is the function f(x) is such that for any two numbers a and b where a ≤ b, the probability of continuous random variable X:

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

Area under the curve between $x = a$ to $x = b$.

- Taking the inspirations from Physics, f(x) is also called the Density Curve and P(X) is also called the **Probability Density Function (PDF).** The acronym PDF stands for both the expansions. Probability Mass Function (PMF) term is **not** used for continuous distributions.

Note: unlike discrete random variable the probability for continuous random variables does not exist for a single value. E.g. the probability that the depth of the lake will be exactly 4.9812 meters, because there are infinite possibilities of depth depending on the precision of the measurement. So effectively such probability will be 0.

- As we have already reviewed, any valid PDF must meet two conditions:
  - $f(x) >= 0$
  - $\int_{-\infty}^{\infty} f(x)dx = \text{the area under the complete curve} = 1$

- **Cumulative Distribution Function:** is the function F(x) that gives the probability P(X ≤ x) and it is obtained integrating the PDF between the limits -∞ to x (or the valid limits for the given continuous function)

$$F(X \le x) = \int_{-\infty}^{x} f(y)\,dy$$

Example

The distribution function for a random variable is defined as follows (assume PDF is valid):

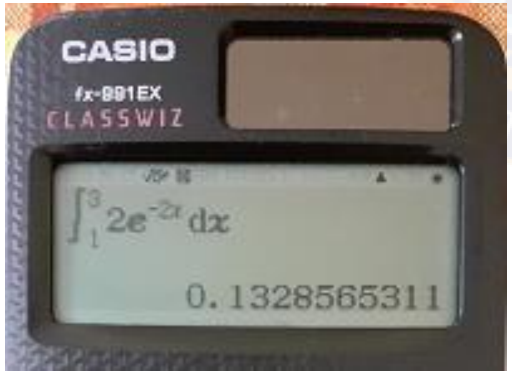
$$f(x) = \begin{cases} 2.e^{-2x} & for\ x > 0 \\ 0 & for\ x \leq 0 \end{cases}$$

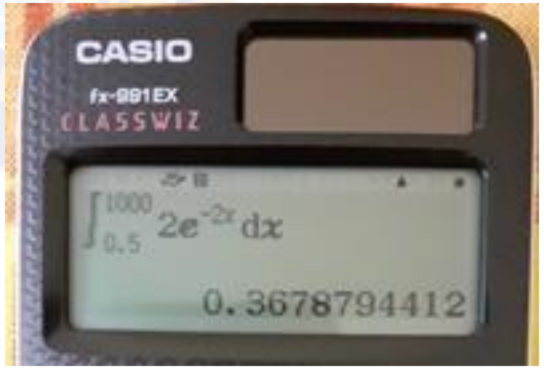
(1) Find the probability that it will take on a value between 1 and 3

$P(1 \leq X \leq 3) = \int_1^3 2.e^{-2x}dx =$ **0.133**

(2) Find the probability that it will take on a value greater than 0.5

$P(X > 0.5) = \int_{0.5}^{\infty} 2.e^{-2x}dx =$ **0.368**





Appendix: Few Useful Integration Formulae

The Uniform Distribution

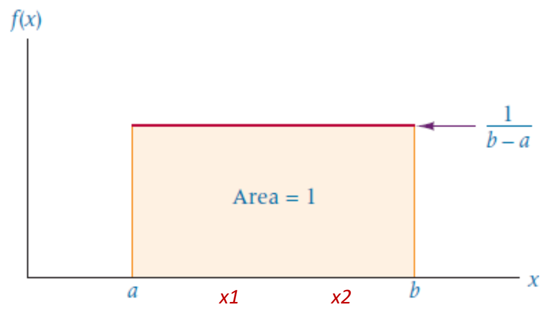
- The **Uniform Distribution**, sometimes referred to as the *Rectangular Distribution*, is a relatively simple continuous distribution in which the same height of f (x), is obtained over a range of values.
- The following distribution function defines a Uniform Distribution:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & for\ a \le x \le b \\ 0 & all\ other\ values \end{cases}$$

- We know that the area under the curve has to be 1.
- Length of the shaded rectangle = (b-a), so:
- Area = Height x Length = 1 or Height = $\dfrac{1}{(b-a)}$
- Mean (μ) and Standard Deviation (σ) is given by:

$$\text{Mean } (\mu) = \frac{a+b}{2}$$

$$\text{Std. Deviation } (\sigma) = \frac{b-a}{\sqrt{12}}$$



- The probability for an in-between pair of values (x_1 and x_2):

$$P(x_2, x_1) = \frac{x_2 - x_1}{b-a} \qquad where\ a \le x1 \le x2 \le b$$

Example

A machine manufactures a part and the part weight ranges from 41 to 47 grams in the uniform distribution.

1. What are the mean and standard deviation?

   Mean (μ) = (41+47)/2 = **44**

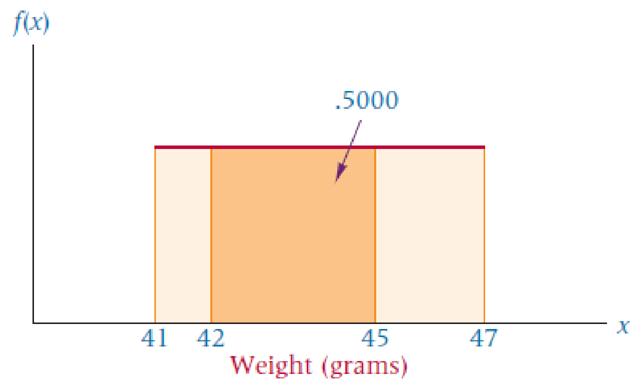
   Std. Deviation (σ) = (47-41) / √12 = 6/3.46 = **1.73**

2. What is the probability that the part would weigh between 42 and 45 grams?

   P(45, 42) = (45-42) / (47-41) = **0.50**

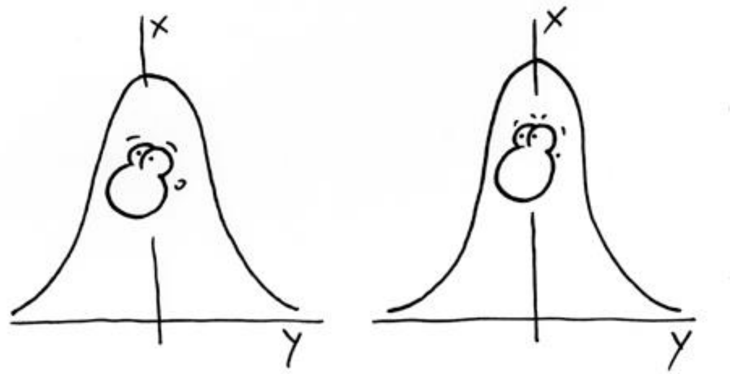
3. What is the probability that the part would weigh less than 42 grams?

   P(42, 41) = (42-41) / (47-41) = 1/6 = **0.17**



Normal Distribution

- The most widely known and used of all continuous distributions is the **Normal Distribution**.
- Many daily life observations and human characteristics such as height, weight, length, speed, IQ, years of work experience and salaries exhibit Normal Distribution.
- Many living things in nature such as trees, animals, insects and others have many characteristics that are normally distributed.
- Many variables in business and industry also are normally distributed: annual cost of insurance, the real estate cost per square foot, produce of machines etc.
- Normal Distribution is credited to mathematician and astronomer Carl Gauss (1777-1855) and therefore it is also called **Gaussian Distribution**.
- The normal distribution exhibits the following characteristics:
  - It is a continuous distribution
  - It is a symmetrical distribution about its mean
  - It is asymptotic to the horizontal axis i.e. it never touches the x-axis
  - Area under the curve is 1



"I always feel so normal, so bored, you know. Sometimes I would like to do something... you know... something... mmm... Poissonian."

PDF for Normal Distribution

*Standard deviation is shown as **σ** but for sample it has to be **s**.*

- The normal distribution is described by two parameters: the **mean (μ)** and the **standard deviation (σ)**. Where $-\infty < x < \infty$ and $\sigma > 0$.
- The Normal distribution function is defined as ($-\infty < x < \infty$ and $\sigma > 0$):

$$\mathbf{f(x) = \frac{1}{\sigma.\sqrt{2\pi}} e^{-1/2\left[\frac{(x-\mu)}{\sigma}\right]^2}}$$

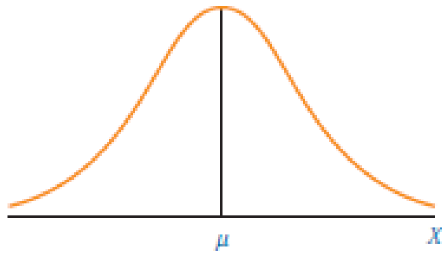
Where:

μ = mean of x

σ = standard deviation of x

π = 3.14159..

e = Euler's Constant



Symmetrical about mean

- To find the probability that under normal distribution the value of x lies between a and b, the above distribution function needs to be integrated with limits from a to b (to determine area under the curve). None of the standard integration techniques can be used to accomplish this.
- So, for μ = 0 and σ = 1 the values are tabulated. This table can also be used to compute probabilities for any other values of μ and σ under consideration.
- The normal distribution with parameter values μ = 0 and σ = 1 and is called the **Standard Normal Distribution.** A random variable having a standard normal distribution is called a standard normal random variable and will be denoted by z. The distribution function of z is:

$$\mathbf{f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}}$$

Solving Normal Distribution Problems

- Every unique pair of μ and σ values defines a different normal distribution. So, standard normal distribution table is used to solve the problems transforming the x values with respect to μ and σ. This transformation is called the **z-Transformation.** It uses the following formula:

$$z = \frac{x - \mu}{\sigma}$$

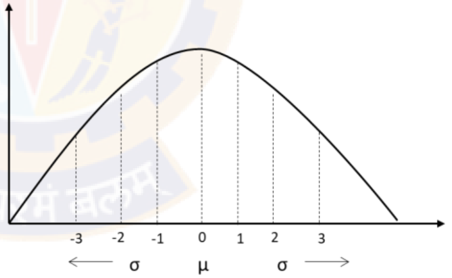
- This transformation basically tells: *how far (above or below) a value is from the mean (μ) in the units of standard deviation (σ).*

Example: In a recent year, the mean score in a test was 494 and the standard deviation was about 100. What is the probability that a randomly selected score from this test is between 600 and the mean?

Given that μ = 494 and σ = 100

We need to find P(494 <= x <= 600)

z(494) = (494 - 494) / 100 = **0** and z(600) = (600 - 494) / 100 = **1.06**

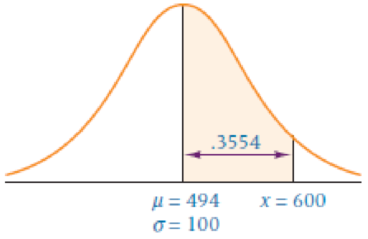


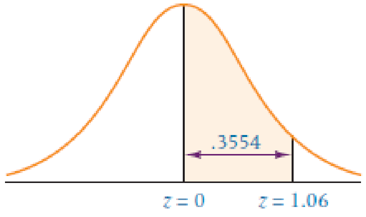
From the table:

The probability value for 0 = 0.0

The probability value for 1.06 = 0.3554

So, P(494 <= Z <= 600) = 0.3554 – 0.0 = **0.3554**





SECOND DECIMAL PLACE IN *z*

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									
6.0	.499999999									

z-Distribution Table

Examples

Same as in the last slide, in a recent year, the mean score in a test was 494 and the standard deviation was about 100.

Example-1: What is the probability that a randomly selected score from this test is between 300 and 600?

Given that μ = 494 and σ = 100

We need to find P(300 <= x <= 600)

z(300) = (300 - 494) / 100 = **-1.94** and z(600) = (600 - 494) / 100 = **1.06**

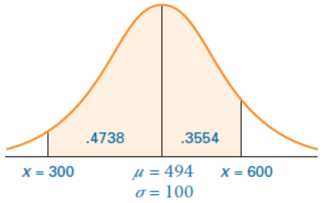
Note: *negative values are not present in the z-table. But distribution is symmetrical around the mean so in place of -1.94, the value of +1.94 would be taken from the table and the negative sign will be put before it.*

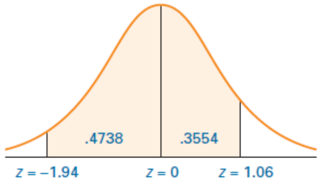
From the table:

The probability value for -1.94 = -0.4738

The probability value for 1.06 = 0.3554

So, P(300 <= Z <= 600) = 0.3554 – (-0.4738) = **0.8292**





Example-2: What is the probability that a randomly selected score from this test is between 350 and 450?

We need to find P(350 <= x <= 450)

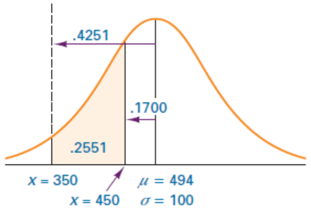
z(350) = (350 - 494) / 100 = **- 1.44** and z(450) = (450 - 494) / 100 = - **0.44**

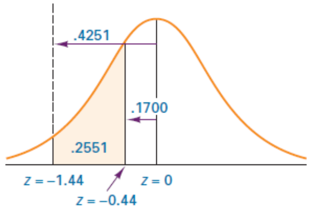
From the table:

The probability value for -1.44 = - 0.4251

The probability value for -0.44 = - 0.1700

So, P(350 <= Z <= 450) = - 0.1700 – (- 0.4251) = **0.2551**





Example-3: What is the probability that a randomly selected score from this test is 550 or less?

We need to find P (0 <= x <= 550)

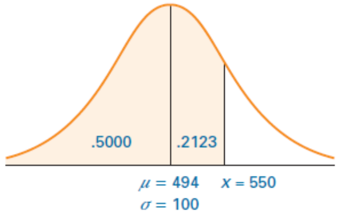
z(0) = (0 - 494) / 100 = **- 4.94** and z(550) = (550 - 494) / 100 = **0.56**

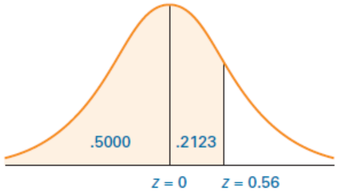
From the table:

The probability value for - 4.94 = - 0.4999

The probability value for 0.56 = 0.2123

So, P(0 <= Z <= 550) = 0.2123 – (-0.4999) = **0.7122**





SECOND DECIMAL PLACE IN z										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									
6.0	.499999999									

z-Distribution Table

Exercise

According to the Income Tax Department the average amount of tax refund is ₹1,332 in a year, with a standard deviation of ₹725. Assume the values are normally distributed. Using the z-Distribution table find out:

a. What proportion of the tax returns show a refund between ₹100 and ₹700?

   Answer = 14.70%

b. What proportion of tax returns show a refund greater than ₹2,000?

   Answer = 17.88%

c. What proportion of the tax returns show that the taxpayer owes money to the government?

   Answer = 3.29%

Gamma Function and Distributions

- Let Γ(n) is a function defined as Γ(n) = (n-1)! for the integers n ∈ {1, 2, 3, …}

Γ: capital gamma, γ: small gamma

- Leonhard Euler (1707-1783) introduced the **Gamma Function**, that can extend the values of n to all real numbers (α > 0) and it is defined as:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} . e^{-x} dx$$

- It has the following important properties:
  - $\Gamma(1) = 1$
  - $\Gamma(\alpha) = (\alpha - 1)!$ ------------------------- For positive integers (α > 0):
  - $\Gamma(\alpha) = (\alpha - 1).\Gamma(\alpha-1)$ ---------------- For any (α > 1):
  - $\Gamma(1/2) = \sqrt{\pi}$

Examples:

$\Gamma(5) = 4! = 24$

$\Gamma(7/2) = (5/2).\Gamma(5/2)$

$= (5/2).(3/2).\Gamma(3/2)$

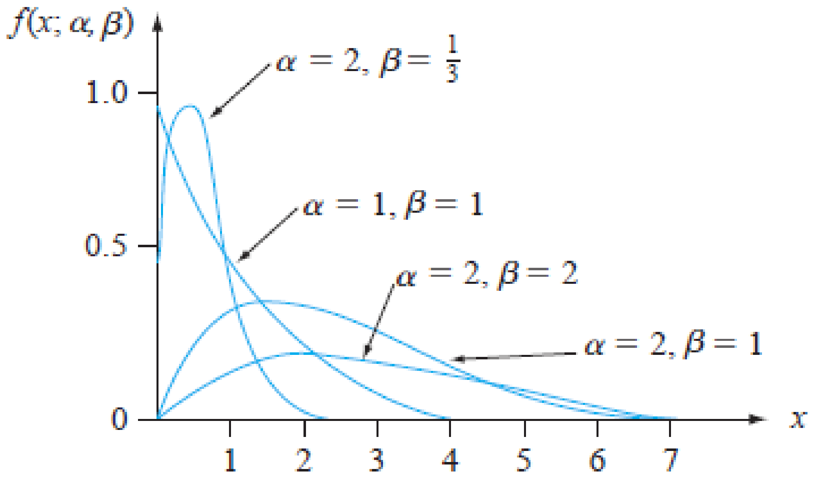
$= (5/2).(3/2)(1/2).\Gamma(1/2)$

$= 15/8.\sqrt{\pi}$

- Gamma Function is an important function in the field of mathematics because it provides many ***continuous skewed distributions*** (unlike ***symmetrical*** normal distribution) which have applications in the real life, e.g. load on web servers, nuclear radiation, accumulated rainfall in reservoir, performance of a new electronic device over time etc. Gamma distribution function is given by:

$$f(x;\ \alpha, \beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}.\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} & \text{for } x \geq 0, \alpha, \beta > 0 \\ 0 & \text{otherwise} \end{cases}$$

Gamma Distribution Function



- α is the **shape** and β is the **scale** parameter. β has stretching/compressing impact.
- Mean and variance of Gamma Distribution: **μ = α. β** and **σ² = α. β²**
- In the above Gamma Distribution Function, if β = 1, it is called **Standard Gamma Distribution Function**.
- In the above Gamma Distribution Function, if α = 1, it provides **Exponential Distribution Function**.
- In the above Gamma Distribution Function, if α = v/2, where v is a positive integer and β =2 it provides **Chi-Squared Distribution Function**, where v is the Greek alphabet called nu.

Exponential Distribution

- Gamma Distribution is written as:

$$f(x;\ \alpha,\beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}.\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} & for\ x \geq 0, \alpha, \beta > 0 \\ 0 & otherwise \end{cases}$$

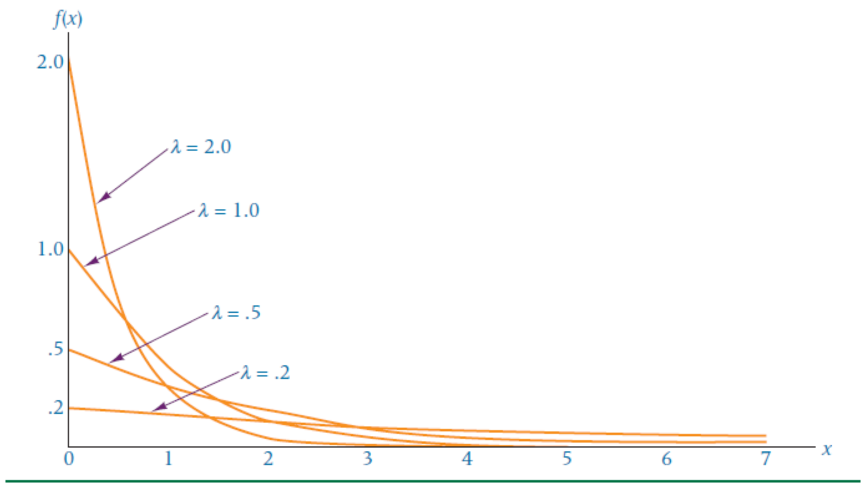
- In the above Gamma Distribution Function, if α = 1, it provides **Exponential Distribution Function** and it can be rewritten as:

$$f(x;\ \beta) = \begin{cases} \dfrac{1}{\beta} e^{\frac{-x}{\beta}} & for\ x \geq 0, \beta > 0 \\ 0 & otherwise \end{cases}$$

- **Exponential Distribution** is closely related with **Poisson Distribution** where long run average λ = 1/β. The Exponential Distribution Function can be re-written as:

$$f(x;\lambda) = \begin{cases} \lambda.e^{-\lambda x} & for\ x\ \geq 0, \lambda > 0 \\ 0 & otherwise \end{cases}$$

Where the Poisson Distribution is discrete and describes random occurrences over some interval, the Exponential Distribution is continuous and describes a probability distribution of the time intervals between random occurrences.



Examples

Students are expected to learn how to perform definite integration using a calculator. Direct answers without showing calculations will not be accepted in the exams.

Example-1: The survival time in weeks of a randomly selected mouse exposed to 240 rads (unit of absorbed radiation) in a radioactive environment has a Gamma Distribution with α = 8 and β = 15. What is the probability that a mouse survives between 60 and 120 weeks?

Gamma Distribution Function is given as: $\mathbf{f}(\mathbf{x};\ \boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} \dfrac{\mathbf{1}}{\boldsymbol{\beta}^{\boldsymbol{\alpha}}.\boldsymbol{\Gamma}(\boldsymbol{\alpha})}\mathbf{x}^{\boldsymbol{\alpha}-\mathbf{1}}\mathbf{e}^{\frac{-\mathbf{x}}{\boldsymbol{\beta}}} & \text{for } x \geq 0, \alpha, \beta > 0 \\ \mathbf{0} & \text{otherwise} \end{cases}$

Given that: α = 8 and β = 15, so:

$\beta^{\alpha} = 15^8$

$\Gamma(\alpha) = \Gamma(8) = 7!$

$$\text{So}, P(60 \leq X \leq\ 120) = \frac{1}{\beta^{\alpha}.\Gamma(\alpha)}\int_{60}^{120} x^{\alpha-1}\, e^{-\frac{x}{\beta}}\, dx = \frac{1}{15^8.7!}\int_{60}^{120} x^{8-1} e^{\frac{-x}{15}}\, dx \quad = \mathbf{0.4959}$$

Example-2: On the average 3 trucks arrive per hour to unload a warehouse. What is the probability that the time between the arrival of successive trucks will be at least 45 minutes.

Exponential Distribution Function is given as: $\mathrm{f}(x;\lambda) = \begin{cases} \lambda.e^{-\lambda x} & for\ x, \geq 0, \lambda > 0 \\ 0 & otherwise \end{cases}$

Given that:

λ = 3

$$\text{So}, P(0.75 \leq X \leq \infty) = \int_{0.75}^{\infty} \lambda\, e^{-\lambda x}\, dx = 3.\int_{0.75}^{\infty} e^{-3x}\, dx \quad = \mathbf{0.1054}$$

Exercise

1. **A busy restaurant determined that between 6:30 P.M. and 9:00 P.M. on Friday nights, the arrivals of customers are Poisson distributed with an average arrival rate of 2.44 per minute.**

a. What is the probability that at least 10 minutes will elapse between arrivals?

Answer: 2.31×10^{-14}. It is very small; can be taken a 0.

b. What is the probability that at least 5 minutes will elapse between arrivals?

Answer: 5.03×10^{-6}. It is very small; can be taken a 0.

c. What is the probability that at least 1 minute will elapse between arrivals?

Answer: 0.08716

d. What is the expected amount of time between arrivals?

Answer: $\mu = \alpha.\, \beta = 1.(1/\lambda) = 1/2.44 = 0.4098$

$$f(x;\, \alpha, \beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}.\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} & for\ x \geq 0, \alpha, \beta > 0 \\ 0 & otherwise \end{cases}$$

2. **Suppose that when a transistor of a certain type is subjected to an accelerated life test, the lifetime X (in weeks) has a gamma distribution with mean 24 weeks and standard deviation 12 weeks.**

a. What is the probability that a transistor will last between 12 and 24 weeks?

Answer: 0.4237

b. What is the probability that a transistor will last at most 24 weeks?

Answer: 0.5665

Remember: Gamma Distribution is defined for x >=0, so lower limit starts from 0.

Gamma Distribution Table

- Gamma Distribution is written as:

$$f(x;\ \alpha,\beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}.\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} & for\ x \geq 0, \alpha, \beta > 0 \\ 0 & otherwise \end{cases}$$

- When β=1, it is called the Standard Gamma Distribution.
- The Standard Gamma Distribution concept can be used to look into the pre-calculated tables to simplify the integration related calculations. Here the value of x is taken in the units of β (as x/ β).

Example: Suppose when a transistor of a certain type is subjected to an accelerated life test, the lifetime X (in weeks) has a gamma distribution with α = 4 and β = 6. What is the probability that the transistor will last between 12 to 24 weeks.

P(12, α=4, β=6)

= P(12/6, α=4)

= P(2, α=4)

= **0.143**

P(24, α=4, β=6)

= P(24/6, α=4)

= P(4, α=4)

= **0.567**

So P(12<= X <= 24) = 0.567 – 0.143 = **0.424**

x \ α	1	2	3	4	5	6	7	8	9	10
1	.632	.264	.080	.019	.004	.001	.000	.000	.000	.000
2	.865	.594	.323	$\boxed{.143}$	.053	.017	.005	.001	.000	.000
3	.950	.801	.577	.353	.185	.084	.034	.012	.004	.001
4	.982	.908	.762	$\boxed{.567}$	.371	.215	.111	.051	.021	.008
5	.993	.960	.875	.735	.560	.384	.238	.133	.068	.032
6	.998	.983	.938	.849	.715	.554	.394	.256	.153	.084
7	.999	.993	.970	.918	.827	.699	.550	.401	.271	.170
8	1.000	.997	.986	.958	.900	.809	.687	.547	.407	.283
9		.999	.994	.979	.945	.884	.793	.676	.544	.413
10		1.000	.997	.990	.971	.933	.870	.780	.667	.542
11			.999	.995	.985	.962	.921	.857	.768	.659
12			1.000	.998	.992	.980	.954	.911	.845	.758
13				.999	.996	.989	.974	.946	.900	.834
14				1.000	.998	.994	.986	.968	.938	.891
15					.999	.997	.992	.982	.963	.930

Gamma Distribution Table

PDF & CDF Example

The PDF is given by the following continuous function: $f(x) = \begin{cases} \frac{6}{5}(x^2 + x) & for\ 0 \le x \le 1 \\ 0 & otherwise \end{cases}$

$P(-\infty \le X \le \infty) = \int_{-\infty}^{\infty} f(x)$ f(x) is +ve for all values of x.

$= \int_{-\infty}^{0} f(x)dx + \int_{0}^{1} f(x)dx + \int_{1}^{\infty} f(x)dx$

$= 0 + \frac{6}{5}\left[\frac{x^3}{3} + \frac{x^2}{2}\right]_0^1 + 0$ $\because \int x^n\, dx = ((x^{n+1})/(n+1)) + C$

$= \frac{6}{5} \cdot \frac{5}{6}$

$= 1$ ∴ It is a valid PDF

Cumulative Distribution Function (CDF):

$$F(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 1 \\ & and \\ \frac{6}{5}\left(\frac{x^3}{3} + \frac{x^2}{2}\right), & 0 \le x \le 1 \end{cases}$$

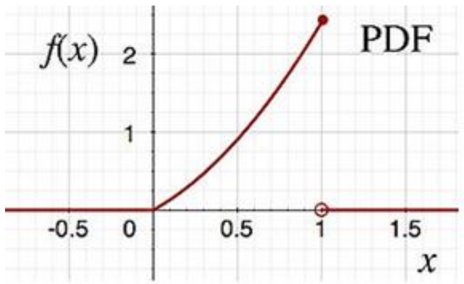
Find the probability that $X \le 0.5$:

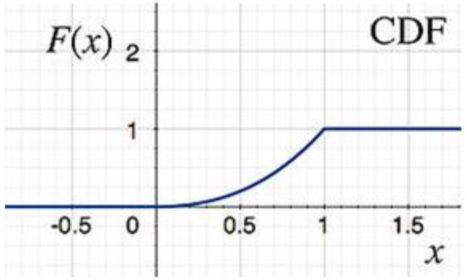
Method-1 (Definite Integration of PDF):

$P(0 \le X \le 0.5) = \int_0^{0.5} \frac{6}{5}(x^2 + x)dx = 0.2$

Method-2 (Directly From CDF):

$P(0 \le X \le 0.5) = F(0.5) - F(0) = \frac{6}{5}\left(\frac{0.125}{3} + \frac{0.25}{2}\right) - 0 = 0.2$





Exercise

The PDF is given by the following function: $f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & for\ 0 \le x \le 2 \\ 0 & otherwise \end{cases}$

a) Find out the CDF.

Answer = $F(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 2 \\ & and \\ \left(\frac{x}{8} + \frac{3x^2}{16}\right), & 0 \le x \le 2 \end{cases}$

b) Find the Probability for 1<= X <=1.5.

Answer = 0.297

c) Find the Probability for X > 1.

Answer = 0.688

Normal Distribution: Classification & Bayes' Theorem

- We have reviewed in the Probability module how to use the Bayes' Theorem for classification when attributes are non-numeric.
- This is a classical example borrowed from Wikipedia.
- This example shows the usage of Bayes' Theorem and Normal Distribution for classification when attributes are numeric with the assumption that attributes are independent given the class.
- Considering dataset as sample, the mean and variance are also calculated for each attribute.
- The objective is to classify the given record (R). In other words, the probabilities for P(C = male | R) and P(C = female | R). The higher probability will classify the record to either male or female.
- From the data set: P (male) = 0.50 and P (female) = 0.50
- From the Bayes' Theorem:

$$P(C = male \mid R) = \frac{P(male) \text{ x } P(height = 6|male) \text{ x } P(weight = 130|male) \text{ x } P(foot\ size = 8|male)}{P(R)}$$

$$P(height = 6 \mid male) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} = \frac{1}{\sqrt{2x(22/7)x0.035}} e^{\frac{-(6-5.86)^2}{2(0.035)}} = 1.58$$

- P(R) will remain the same for both male and female, so not required to be calculated.
- Solve the remaining part as an exercise and check the solution from Wikipedia.

How come $P(height = 6 \mid male)$ value > 1?
Should it not be 0?

height (feet)	weight (lbs)	foot size (inches)	Gender (C)
6.00	180	12	male
5.92	190	11	male
5.58	170	12	male
5.92	165	10	male
5.00	100	6	female
5.50	150	8	female
5.42	130	7	female
5.75	150	9	female

Person	male	female
mean (height)	5.86	5.42
variance (height)	0.035	0.097
mean (weight)	176.25	132.50
variance (weight)	122.92	558.33
mean (foot size)	11.25	7.50
variance (foot size)	0.92	1.67

Test Record (R)		
height (feet)	weight (lbs)	foot size (inches)
6	130	8

Method of Moments

Before method of moments is applied it is necessary to check if the PDFs are valid.

- Method of moments are used extensively in statistics to calculate the mean and variance of distributions.
- The n^{th} order moment $E(X^n)$ for discrete random variable X and P(x) PDF (PMF) is given by the following:

$$\mathbf{E(X^n) = \sum_i x_i^n . P(x_i)}$$

- The n^{th} order moment $E(X^n)$ for continuous random variable X and f(x) PDF is given by the following:

$$\mathbf{E(X^n) = \int_{-\infty}^{+\infty} x^n . f(x)\ dx}$$

- The n^{th} order moment $E(X^n)$ for discrete random variable X **about the mean (μ)** is given by the following:

$$\mathbf{E(X-\mu)^n = \sum_i (xi-\mu)^n . P(x_i)}$$

- The n^{th} order moment $E(X^n)$ for continuous random variable X **about the mean (μ)** is given by the following:

$$\mathbf{E(X-\mu)^n = \int_{-\infty}^{+\infty} (x-\mu)^n . f(x)\ dx}$$

- The zero order moment is the **total probability**:

$$\mathbf{E(X^0) = \int_{-\infty}^{+\infty} x^0 . f(x)\ dx = 1}$$

- The first order moment is the **mean**:

$$\mathbf{E(X^1) = \int_{-\infty}^{+\infty} x^1 . f(x)\ dx = \mu}$$

- The second order moment about the mean is the **variance**:

$$\mathbf{E(X-\mu)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 . f(x)\ dx = \sigma^2}$$

Similar expressions can be written for discrete distributions also.

Example

- First order of moment is called the **mean**. It is also called the **expected value** of random variable X.

$$E(X) = \int_{-\infty}^{+\infty} x.f(x)\ dx = \mu$$

- The second order moment about the mean is the **variance**:

$$E(X-\mu)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2.f(x)\ dx = \sigma^2$$

The following relationship can be shown:

$\mathbf{E(X-\mu)^2}$

$= \mathrm{E}(\mathrm{X}^2 + \mu^2 - 2\mathrm{X}\mu)$

$= \mathrm{E}(\mathrm{X}^2) + \mu^2 - 2\mu.\mathrm{E}(\mathrm{X})$

$= \mathrm{E}(\mathrm{X}^2) + \mu^2 - 2\mu.\mu$

$= \mathbf{E(X^2) - \mu^2}$

This is very useful in finding out the expressions for variance.

Example: Uniform distribution is given by the following probability density function. Calculate its mean and variance.

$$f(x) = \begin{cases} \dfrac{1}{b-a} & for\ a \le x \le b \\ 0 & all\ other\ values \end{cases}$$

Expected Value (mean) E(X):

$= \int_a^b x.\frac{1}{b-a}\ dx$

$= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b$

$= \frac{1}{b-a}\left[\frac{b^2}{2} - \frac{a^2}{2}\right]$

$= \frac{(b+a).(b-a)}{2(b-a)}$

$= \frac{(b+a)}{2}$

Variance:

$= \int_a^b (x-\mu)^2.\frac{1}{b-a}\ dx$

$= \int_a^b \left(x - \frac{(b+a)}{2}\right)^2.\frac{1}{b-a}\ dx$

$= \frac{1}{3(b-a)}\left[\left\{x - \frac{(b+a)}{2}\right\}^3\right]_a^b$

$= \frac{(b-a)^2}{12}$

Standard Deviation:

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

Example

Example: Establish the expressions for the mean and variance for Binomial Distribution.

Mean (μ):

$$E(X) = \sum_{x=0}^{n} x.nCx.p^x.q^{n-x}$$

$$= \sum_{x=0}^{n} x.\frac{n!}{(n-x)!\,x!}.p^x.q^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n(n-1)!}{(n-x)!\,(x-1)!}.p.p^{x-1}.q^{n-x}$$

$$= np\sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!\,(x-1)!}.p^{x-1}.q^{n-x}$$

$$= np\sum_{x=1}^{n} (n-1)C(x-1).p^{x-1}.q^{n-x}$$

$$= \mathbf{np}$$

Second Order Moment:

$$E(X^2) = \sum_{x=0}^{n} x^2.nCx.p^x.q^{n-x}$$

$$= \sum_{x=0}^{n} (x.(x-1)+x).nCx.p^x.q^{n-x}$$

$$= \sum_{x=0}^{n} x.(x-1).nCx.p^x.q^{n-x} + \sum_{x=0}^{n} x.nCx.p^x.q^{n-x}$$

$$= \sum_{x=2}^{n} \frac{x.(x-1).n!}{(n-x)!\,x.(x-1)(x-2)!}.p^x.q^{n-x} + np$$

$$= n.(n-1).p^2\sum_{x=2}^{n} \frac{(n-2)!}{(n-x)!\,(x-2)!}.p^{x-2}.q^{n-x} + np$$

$$= n.(n-1).p^2\sum_{x=2}^{n} (n-2)C(x-2).p^{x-2}.q^{n-x} + np$$

$$= n.(n-1).p^2 + np$$

Variance:

$$= E(X^2) - \mu^2$$

$$= n.(n-1).p^2 + np - n^2p^2$$

$$= np.(1-p)$$

$$= \mathbf{npq}$$

Standard Deviation (σ):

$$= \sqrt{\mathbf{npq}}$$

Example

- $\Gamma(1) = 1$
- $\Gamma(\alpha) = (\alpha - 1)!$ --------------For positive integers ($\alpha > 0$):
- $\Gamma(\alpha) = (\alpha - 1).\Gamma(\alpha-1)$ ------For any ($\alpha > 1$):
- $\Gamma(1/2) = \sqrt{\pi}$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}.e^{-x}dx$$

Example: Calculate the mean and variance for the Gamma Distribution that is given as:

$$f(x;\ \alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha.\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} & \text{for } x \geq 0, \alpha, \beta > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean (μ):

$$E(X) = \int_0^\infty x.\frac{1}{\beta^\alpha.\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} dx$$

$$= \frac{1}{\beta^\alpha.\Gamma(\alpha)} \int_0^\infty x^\alpha.e^{\frac{-x}{\beta}} dx$$

Let, $y = \frac{x}{\beta}$, so:

$$E(X) = \frac{1}{\beta^\alpha.\Gamma(\alpha)} \int_0^\infty (y.\beta)^\alpha.e^{-y}(\beta.dy)$$

$$= \frac{\beta^{\alpha+1}}{\beta^\alpha.\Gamma(\alpha)} \int_0^\infty y^\alpha.e^{-y}dy$$

$$= \frac{\beta}{\Gamma(\alpha)}.\Gamma(\alpha+1)$$

$$= \frac{\beta}{\Gamma(\alpha)}.\alpha.\Gamma(\alpha)$$

$$= \boldsymbol{\alpha.\beta}$$

Second Order Moment:

$$E(X^2) = \int_0^\infty x^2.\frac{1}{\beta^\alpha.\Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} dx$$

$$= \frac{\alpha.(\alpha+1).\beta^2}{\beta^{(\alpha+2)}.\Gamma(\alpha+2)} \int_0^\infty x^{(\alpha+2)-1}.e^{\frac{-x}{\beta}} dx$$

Let, $y = \frac{x}{\beta}$, so:

$$E(X^2) = \frac{\alpha.(\alpha+1).\beta^2}{\beta^{(\alpha+2)}.\Gamma(\alpha+2)} \int_0^\infty (y.\beta)^{(\alpha+2)-1}.e^{-y}(\beta.dy)$$

$$= \frac{\alpha.(\alpha+1).\beta^2}{\beta^{(\alpha+2)}.\Gamma(\alpha+2)}.\beta^{(\alpha+2)}.\Gamma(\alpha+2)$$

$$= \alpha.(\alpha+1).\beta^2$$

$$\because \Gamma(\alpha+2) = \int_0^\infty y^{\alpha+1}.e^{-y}dy$$

Variance (σ²):

$$E(X^2) - \mu^2$$

$$= \alpha.(\alpha+1).\beta^2 - \alpha^2.\beta^2$$

$$= \alpha^2.\beta^2 + \alpha.\beta^2 - \alpha^2.\beta^2$$

$$= \boldsymbol{\alpha.\beta^2}$$

Standard Deviation (σ):

$$= \boldsymbol{\sqrt{\alpha}.\beta}$$

Thank You

Appendix

Few Useful Integration Formulae

- $\int 1\,dx = x + C$
- $\int a\,dx = ax + C$
- $\int x^n\,dx = ((x^{n+1})/(n+1)) + C$
- $\int \sin x\,dx = -\cos x + C$
- $\int \cos x\,dx = \sin x + C$
- $\int \sec^2 dx = \tan x + C$
- $\int \csc^2 dx = -\cot x + C$
- $\int \sec x\,(\tan x)\,dx = \sec x + C$
- $\int \csc x\,(\cot x)\,dx = -\csc x + C$
- $\int (1/x)\,dx = \ln|x| + C$
- $\int e^x\,dx = e^x + C$
- $\int a^x\,dx = (a^x/\ln a) + C$; $a > 0,\ a \neq 1$

Poisson Distribution

- λ is the known long term average in the Poisson Distribution, where $\lambda > 0$.
- For all the possible outcomes (x) for the known λ, the sum of probabilities will be 1.
- As per the Maclaurin Series:

$$e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$$

- How the above expression can have the sum of 1?
- Multiplying with $e^{-\lambda}$ on both the extreme sides:

$$1 = \sum_{x=0}^{\infty} \frac{e^{-\lambda}.\lambda^x}{x!}$$

- Where the x^{th} term or the probability is given by:

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