AGENDA

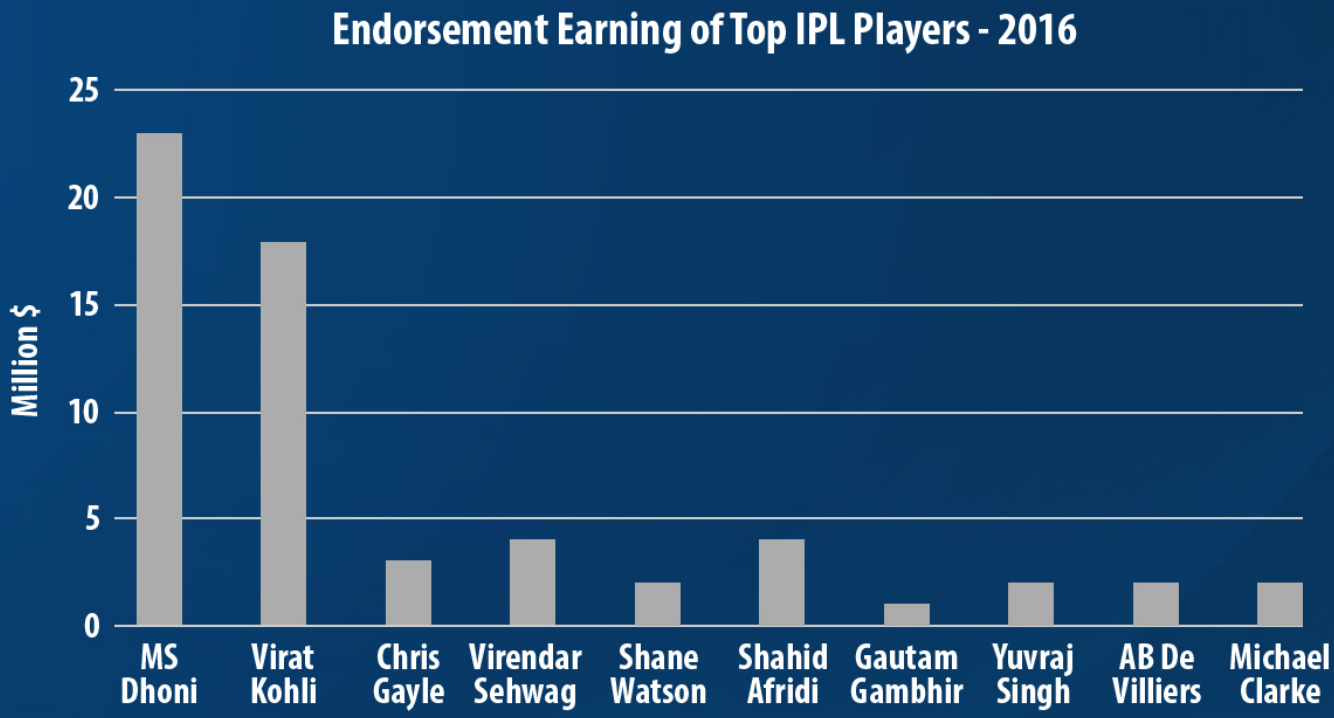
- Measure of Central Tendency and Dispersion
- Probability Distributions & their Physical Relevance
- Important Discrete Distributions
- Important Continuous Distributions

MEASURES OF CENTRAL TENDENCY – THE MEAN

- *What is the typical value of data? Where does the center lie?*
- What does the "Mean" mean? $\bar{x} = \frac{\sum x_i}{n}$
- Gives us an idea of:
  - A typical value of the data
  - A good representation of the sample if no outliers are present
- Where does it falter?
  - If outliers are huge
  - Data is highly variable

MEAN IS MEANINGLESS HERE!

Player	Endorsements 2016 (in Mil USD)
MS Dhoni	23
Virat Kohli	18
Chris Gayle	3
Virender Sehwag	4
Shane Watson	2
Shahid Afridi	4
Gautam Gambhir	1
Yuvraj Singh	2
AB De Villiers	2
Michael Clarke	2
	MEAN = 6.1



Note: The high salary of just two players pull up the mean and hence it does not represent the true picture of the player salaries

Data Source : www.crunchysports.com

MEASURES OF CENTRAL TENDENCY – THE MEDIAN

- *What is the central value of the sample?*
- What does the "Median" mean?
- Consider, the data set of players salaries in Mil USD / Annum

  23 18 3 4 2 4 1 2 2 2

- Let us rearrange it in Ascending Order,

  1 2 2 2 2 3 4 5 18 23

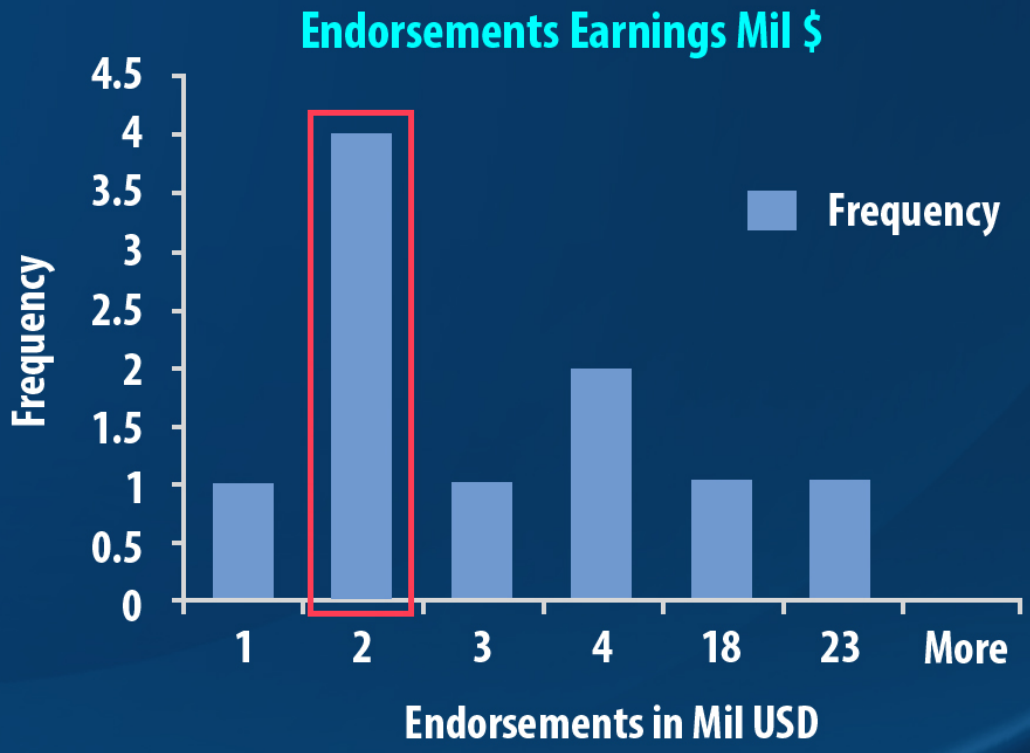
- Now, pick the middle number / numbers.
  Thus median value is Avg(2,3) = 2.5.
- This gives us a more real picture of as to how the salaries are for the group, in conjunction with the mean.

MEASURES OF CENTRAL TENDENCY – THE MODE

Mode: It refers to the most frequently occurring number found in a set of data

What does the mode tell us?

Endorsement	Frequency
1	1
2	4
3	1
4	2
18	1
23	1



Mean = 6.5 Median = 2.5 Mode = 2

AMITY UNIVERSITY ONLINE
CAREERS OF TOMORROW

PERCENTILE: RELATIVE STANDING

- Consider the following data set of N = 10 values

Endorsement in Mil USD
1
2
2
2
2
3
4
4
18
23

Let us find out the 80th Percentile

Find the ordinal rank

$$n = \left\lceil \frac{P \times N}{100} \right\rceil$$

Thus, we have,

$$n = \left\lceil \frac{80 \times 10}{100} \right\rceil = 8$$

Thus, 80th Percentile Value = 4

What does this mean ?

Thus 80% of the population have earnings less than or equal to 4 Mil USD

QUARTILES: WHAT ARE THEY ?

- Quartiles as three points which divide the data set into four equal groups

Consider the following example of an ordered data set:

6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

6, 7, 15, 36, 39, 40 Lower Half

40, 41, 42, 43, 47, 49 Upper Half

1) Find the Median

2) Split the data into two sets from the median

(Include median in both if odd numbered set)

AMITY UNIVERSITY ONLINE
CAREERS OF TOMORROW

QUARTILES: WHAT ARE THEY ?

Ordered data set:

6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

6, 7, 15, 36, 39, 40 Lower Half

40, 41, 42, 43, 47, 49 Upper Half

Now the Quartiles are obtained as medians of the two new data sets!

Q1, Lower Quartile, 25th Percentile

6, 7, 15, 36, 39, 40

Q1 = 25.5

Q2, Second Quartile, Median, 25th Percentile

Q2 = 40

Q3, Upper Quartile, 75th Percentile

40, 41, 42, 43, 47, 49

Q3 = 42.5

And Finally !

The Interquartile Range:

IQ = Q3 – Q1

= 42.5 – 25.5

= 17

The interquartile range (IQR) is a measure of statistical dispersion or the spread of the data.

MEASURES OF DISPERSION - STANDARD DEVIATION

Why is Standard Deviation so important?

- The mean does not tell us about the spread of the data

Endorsement 2016 (in Mil USD)
23
18
3
4
2
4
1
2
2
2

MEAN = 6.1

- The standard deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

where,
x = data point
$\bar{x}$ = Mean
n = number of observations

Standard Deviation of the same data set is, $S = 7.34$

Thus we see how hugely varying the data is!

MEASURES OF DISPERSION - COEFFICIENT OF VARIATION

What is it?

- Value obtained by normalising the Standard Deviation with the Mean of the sample.
- The coefficient of variation (C_v), also known as relative standard deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage.

$$c_v = \frac{\sigma}{\mu}$$

Where

σ = Standard Deviation

μ = Mean

PRO'S Dimensionless quantity can be used to compare between two sets of data, irrespective of units.

CON'S Shoots up to infinity if mean is close to zero.

MEASURES OF DISPERSION - STANDARDIZED VARIABLE (Z - SCORE)

- A standard variable (Z) is one that has been re - scaled to have a mean of zero and a Standard Deviation of one.

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

Where:

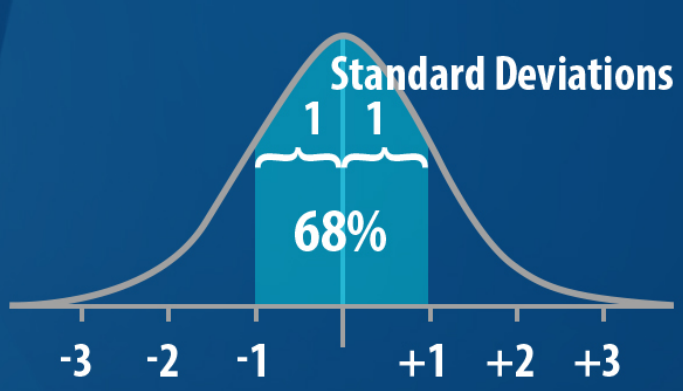
X = Data point

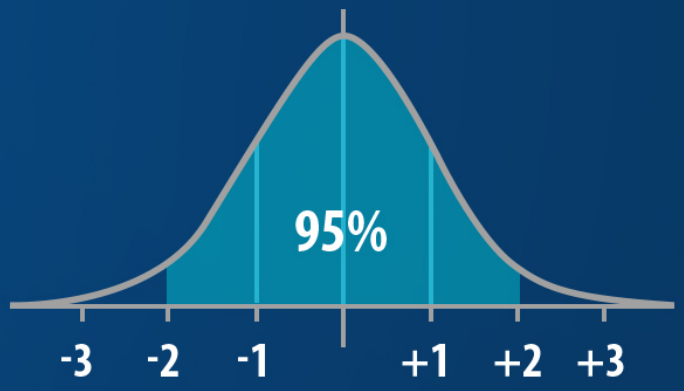
σ = Standard Deviation

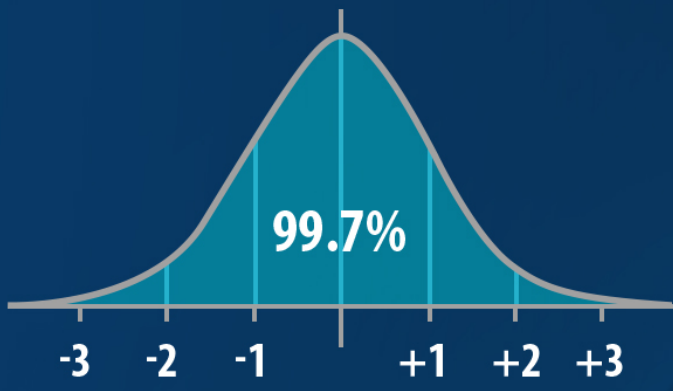
μ = Mean

- Applications:
  - Standard normal table
  - Normal distributions.

MEASURES OF DISPERSION - THE BELL CURVE







68% of values are within 1 standard deviation of the mean

95% of values are within 2 standard deviations of the mean

99.7% of values are within 3 standard deviations of the mean

AMITY UNIVERSITY ONLINE
CAREERS OF TOMORROW

PROBABILITY CONCEPTS

Definition:

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

For a Single Coin

P(Head) = 0.5

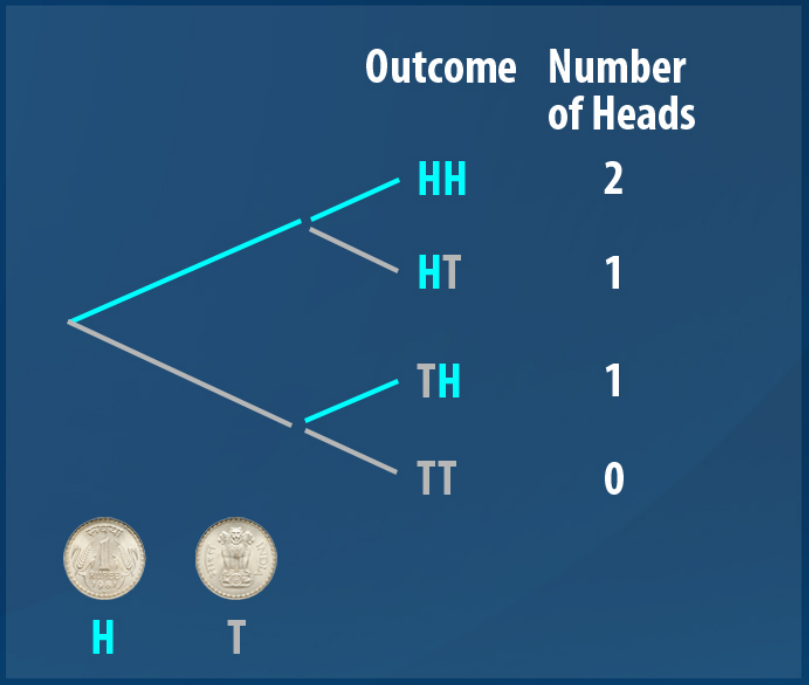
P(Tail) = 0.5

Fundamentals for Independent Events A and B

P(Event A AND Event B) = P(Event A) * P(Event B)

P(Event A OR Event B) = P(Event A) + P(Event B)

Outcome of 2 coin Toss

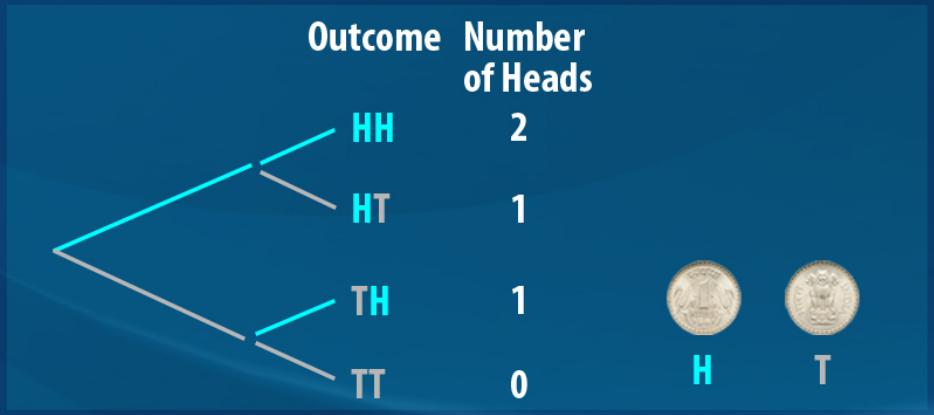


THE COIN TOSS CASE

The Two coin Toss is a typical Example of a Discrete Probability Distribution.

Number of Heads	Logical Combination	Probablity Combinations	Probability Computations	Probability of Event	Cumulative Probablity	Cumulative Probabilty Event Description
0	T+T	P(Tail)* P(Tail)	0.5*0.5	0.25	0.25	0 heads
1	H+T OR T+H	P(Head)*P(Tail) + P(Tail)*P(Head)	0.5*0.5+0.5*0.5	0.5	0.75	1 or less than 1 head
2	H+H	P(Head)*P(Head)	0.5*0.5	0.25	1	2 or less than 2 heads

Outcome of 2 coin Toss



AMITY UNIVERSITY ONLINE
CAREERS OF TOMORROW

DISCRETE vs CONTINUOUS

If a variable can take on any value between two specified values, it is called a continuous variable; otherwise, it is called a discrete variable.

Discrete Variable

Flip a coin and count the number of heads : Head Count would be a discrete variable.

Continuous

Weight of soldiers enlisted in army between minimum and maximum weight.

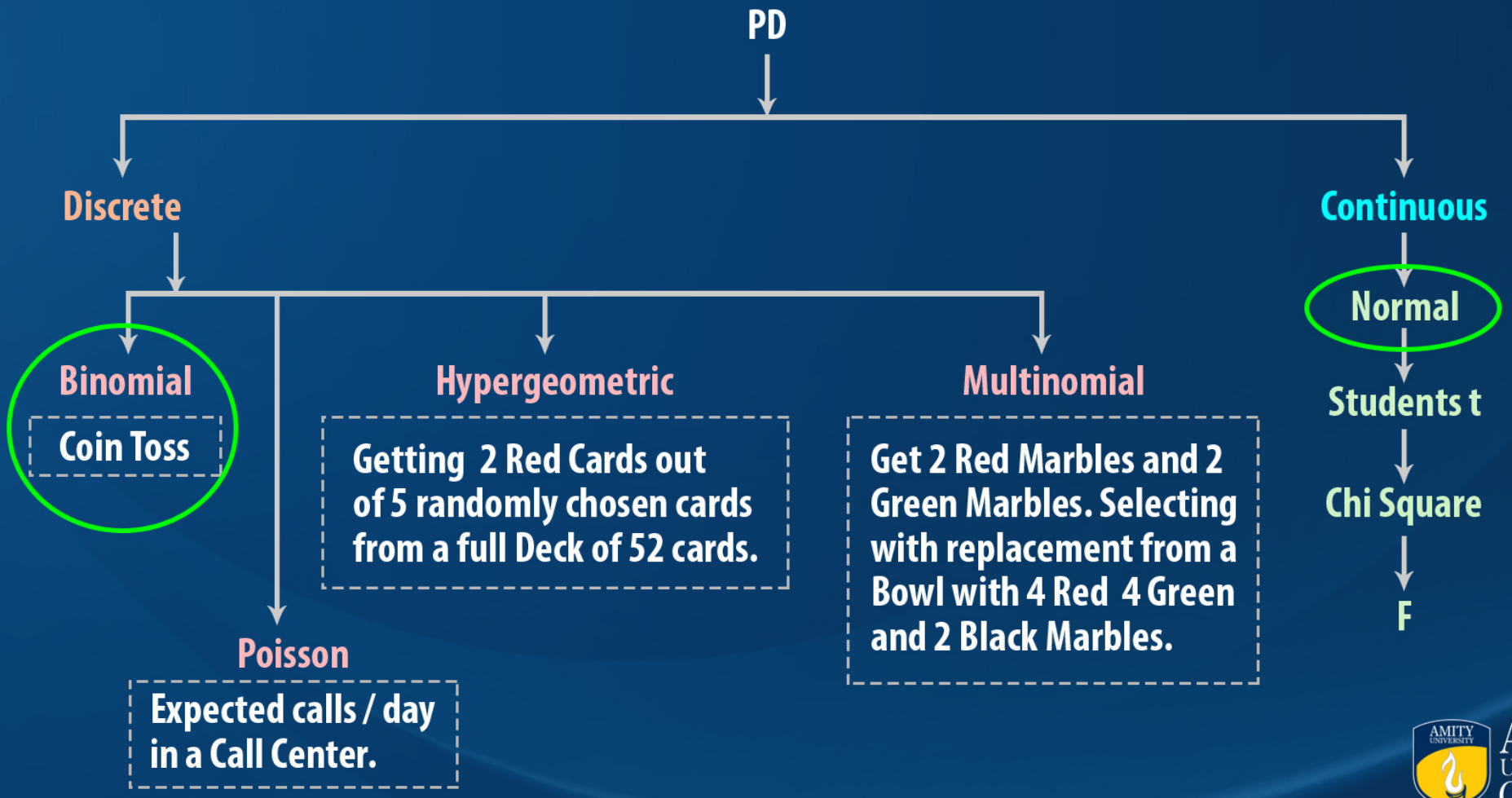
Discrete Probability Distribution

If a random variable is a discrete variable, its probability distribution is called a discrete probability distribution.

Coin Toss

Number of Heads x	Probability of Event P (X=x)
0	0.25
1	0.5
2	0.25

TYPES OF PROBABILITY DISTRIBUTIONS (PD)



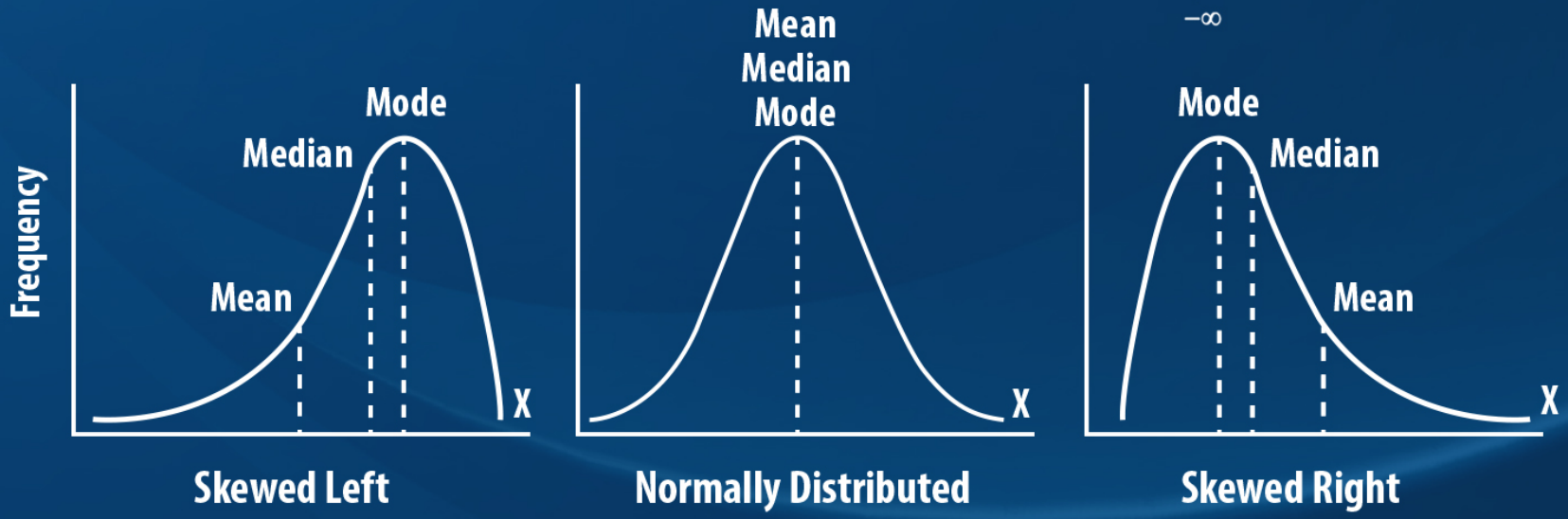
NORMAL DISTRIBUTION

The integration values are available through statistical tables.

Normal Distribution is defined by the probability distribution function. A random variable X with probability density function is given as below,

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

The probability of X being less than x is given by following integration, $A = \int_{-\infty}^{x} f(x)dx$



NORMAL DISTRIBUTION (contd)

Key Features

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable X equals any particular value is 0.
- The probability that X is greater than a equals the area under the normal curve bounded by a and plus infinity (as indicated by the non-shaded area in the figure below).
- The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (as indicated by the shaded area in the figure below).

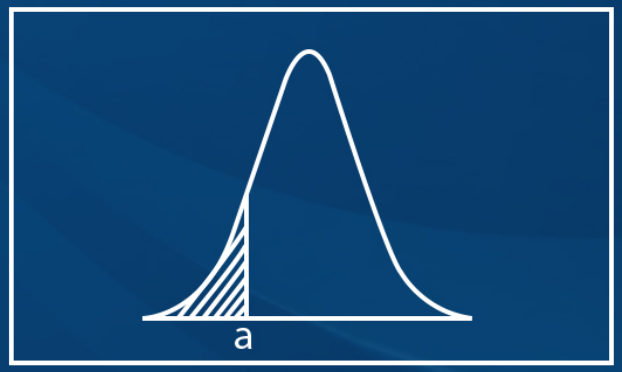


Figure (a)

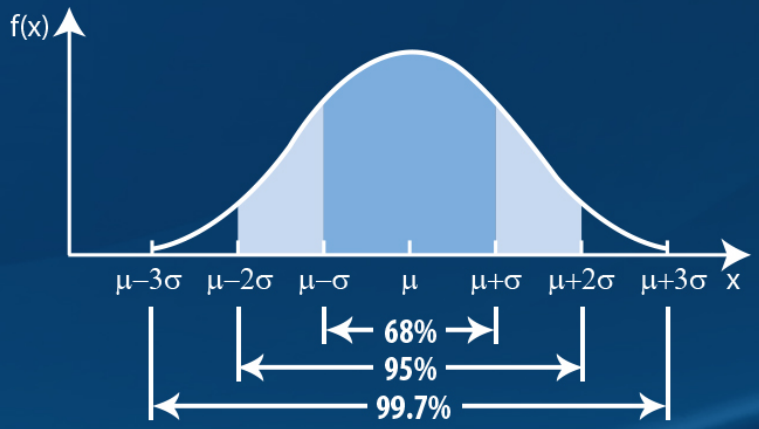
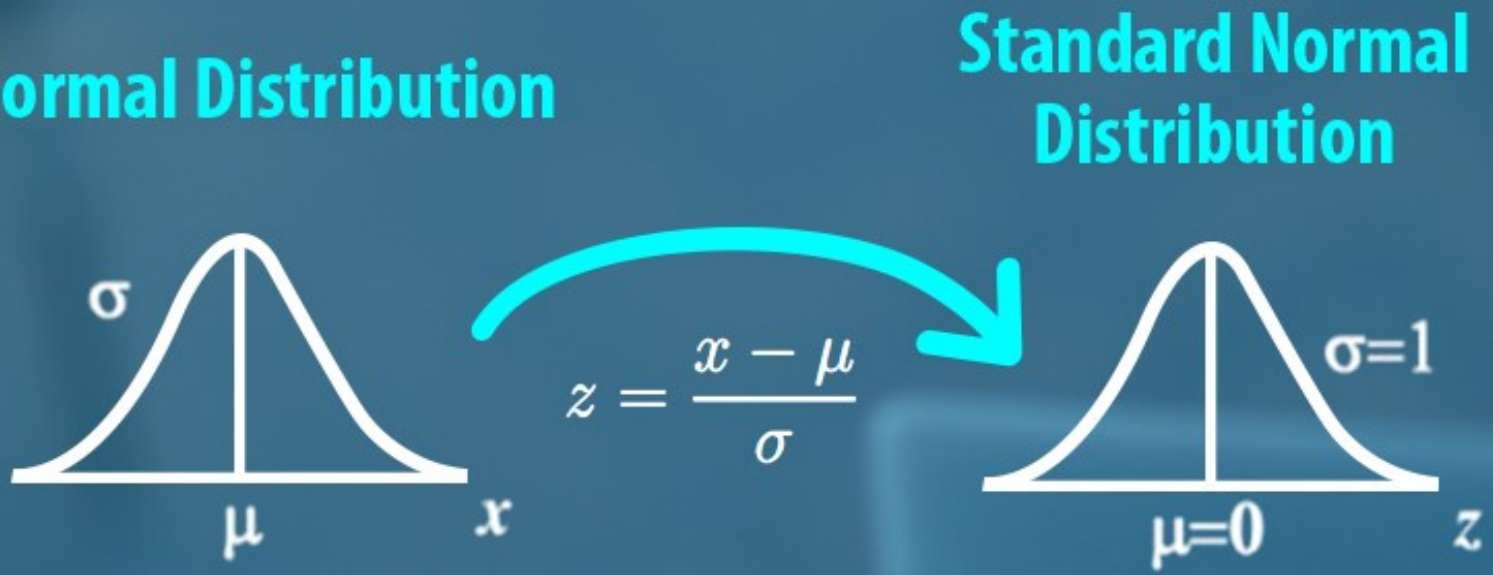


Figure (b)

NORMAL DISTRIBUTION – GETTING THE Z SCORE

Transforming normal distribution to standard normal distribution i.e., transform any *x*-value into *z*-score or critical values.



A z-score always reflects the number of standard deviations above or below the mean, a particular score is.

NORMAL DISTRIBUTION – EXAMPLE

A person scored 70 in a test where, mean score of the population is 50 and standard deviation is 10.

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

$$z = \frac{70 - 50}{10} = 2$$

Person scored 2 standard deviation above the mean.

To find out number of readings/score within the bounded between the two z-scores, and the curve i.e., area, you can use z-table.

NORMAL DISTRIBUTION – GETTING THE Z SCORE - PROBLEM

Acme Light Bulb Company has found that an average light bulb lasts 1000 hours with a standard deviation of 100 hours. Assume that bulb life is normally distributed. What is the probability that a randomly selected light bulb will burn out in 1200 hours or less?

Solution:

Given Data:

Mean = 1000 hours
Standard deviation = 100 hours

Need to Find:

P_Life ≤ 1200 hours

Compute Z score

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

$$z = \frac{1200 - 1000}{100} = 2$$

From Normal Tables

$$P = 0.97$$

Thus a 97 % probability exists that a randomly selected bulb will burn out within 1200 hours

Doing it in R !

```
> Pnorm(1200, Mean=1000, sd=100, lower.tail=TRUE)
[1] 0.9772499
>
```

AMITY UNIVERSITY ONLINE
CAREERS OF TOMORROW

POSITIVE Z TABLE

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

NEGATIVE Z TABLE

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3	.483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3	.859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4	.247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.468	.4641

DISCRETE PROBABILITY DISTRIBUTION: BINOMIAL DISTRIBUTION

Key Features

The binomial distribution is a special discrete distribution where there are two distinct complementary outcomes, a success and a failure.

Probability Mass Function

In general, if the random variable X follows the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0,1]$, we write $X \sim B(n, p)$.

The probability of getting exactly k successes in n trials is given by the probability mass function:

$$f(k, n, p) = \Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where $\binom{n}{k}$ is the binomial coefficient, hence the name of the distribution.

k successes occur with probability p^k and $n - k$ failures occur with probability $(1 - p)n - k$.

However, the k successes can occur anywhere among the n trials, and there are $\binom{n}{k}$ different ways of distributing k successes in a sequence of n trials.

DISCRETE PROBABILITY DISTRIBUTION: BINOMIAL DISTRIBUTION (contd)

Cumulative Distribution Functions

$$F(k;n,p) = \Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

Probability Mass Function

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Notation Notes!

Mean: $\mu = n\pi$

Standard Dev: $\sigma = \sqrt{n\pi(1-\pi)}$

CONDITIONS FOR BINOMIAL EXPERIMENT TO BE SATISFIED

- The experiment consists of n identical trials
- Each trial results in one of the two outcomes, called success and failure
- The probability of success, denoted π, remains the same from trial to trial
- The n trials are independent. That is, the outcome of any trial does not affect the outcome of the others

BINOMIAL DISTRIBUTION: PROBLEM

Problem Statement

An FBI survey shows that about 80% of all property crimes go unsolved. Suppose that in your town 3 such crimes are committed and they are each deemed independent of each other.

a) 1 of 3 of these crimes will be solved.

b) At least one of the crimes will be solved.

Conditions Satisfied check !

- Does it satisfy fixed number of trials?

  YES. The number of trials is fixed at 3 (n = 3.)
- Does it have only 2 outcomes?

  YES.Solved and unsolved
- Do all the trials have the same probability of success?

  YES. p = 0.2
- Are all crimes independent?

  YES. Stated in the description.

*Source: https: //newonlinecourses.science.psu.edu/stat500/node/22/

BINOMIAL DISTRIBUTION: SOLUTION

Solution: Part a: At least one of the crimes will be solved.

We have the binomial formula as follows :

$$\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Let's apply this formula in our example. In our example $n = 3$ and $X = 1$.
If we fill in the formula above using the data from our example it would be:

$$\frac{3!}{1!(3-1)!}0.2^1(1-0.2)^{3-1} = 3(0.2)(0.8)^2 = 0.384$$

BINOMIAL DISTRIBUTION: SOLUTION

Solution: Part b: At least one of the crimes will be solved.

Here we are looking to solve $P(X \geq 1)$.

The long way to solve for $P(X \geq 1)$.
This would be to solve: as follows:

$P(x = 1) + P(x = 2) + P(x = 3)$

$$P(x = 1) = 3!/1!2! \times 0.2^1 \times 0.8^2$$

$$P(x = 2) = 3!/2!1! \times 0.2^2 \times 0.8^1$$

$$P(x = 3) = 3!/3!0! \times 0.2^3 \times 0.8^0$$

We add up all of the above probabilities and get 0.488.

Alternate Approach:

Here the complement to $P(X \geq 1)$
is equal to 1 - $P(X < 1)$
which is equal to 1 - $P(X = 0)$.

$$= 1 - P(x < 1) = 1 - P(x = 0)$$

$$= 1 - \frac{3!}{0!(3-0)!} 0.2^0 (1-0.2)^3$$

$$= 1 - 1(1)(0.8)^3 = 1 - 0.512 = 0.488$$

AMITY UNIVERSITY ONLINE
CAREERS OF TOMORROW