Probability Distributions and Sampling Techniques

Agenda

- Random Variable
  - Discrete Random Variable
  - Continuous Random Variable
- Probability Distribution
  - Cumulative Distribution Function
  - Discrete Probability Distribution
    - Binomial Distribution
    - Poisson Distribution
  - Continuous Probability Distribution
    - Normal Distribution

Agenda

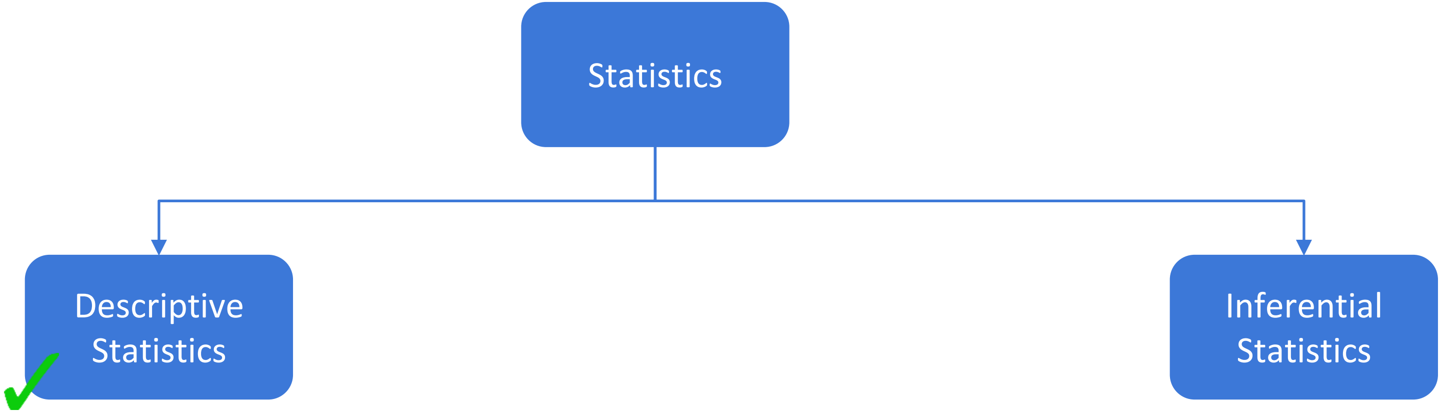
- Population and Sample
- Sampling Techniques
  - Probabilistic Sampling
    - Simple Random Sampling (with and without replacement)
  - Resampling
    - Bootstrap resampling
    - Jackknife resampling

Agenda

- Sampling Distribution
  - Parameter and Statistic
  - Sampling Distribution of Sample Mean
  - Sampling Distribution of Sample Proportion

- Theory of Estimation
  - Point estimation
  - Sampling error
  - Interval estimation

Statistics



Statistics

Descriptive Statistics ✓

Inferential Statistics

It is used to only describe the sample or summarize information about the sample

It is used to make inferences and generalizations about the broader population

Random Variable

Random variable

- The action of rolling a die is called a random experiment

- The set of possible outcomes of an experiment is called the Sample Space and it is denoted by Ω

- Possible outcomes for this experiment are: 1, 2, 3, 4, 5, and 6

- A Random Variable is a function defined from the sample space to real numbers

Random variable

- The usual notation used for a random variable is ‘X’. However, other capital letters like U, V, or Z can be used

- Let us observe the experiment of number appeared on the die. We define our random variable as

  X: the number observed

  X takes values 1, 2, 3, 4, 5, and 6.

  Hence Ω = {1, 2, 3, 4, 5, 6}

Random variable

Question:

Suppose you are the owner of a grocery shop. You wish to improve the shopping experience of your customers, by easing their payment methods and ensuring that whichever products they demand are available in the shop.

Now collect the data for:

- mode of payment
- bill amount

Can you describe the nature of the data collected?

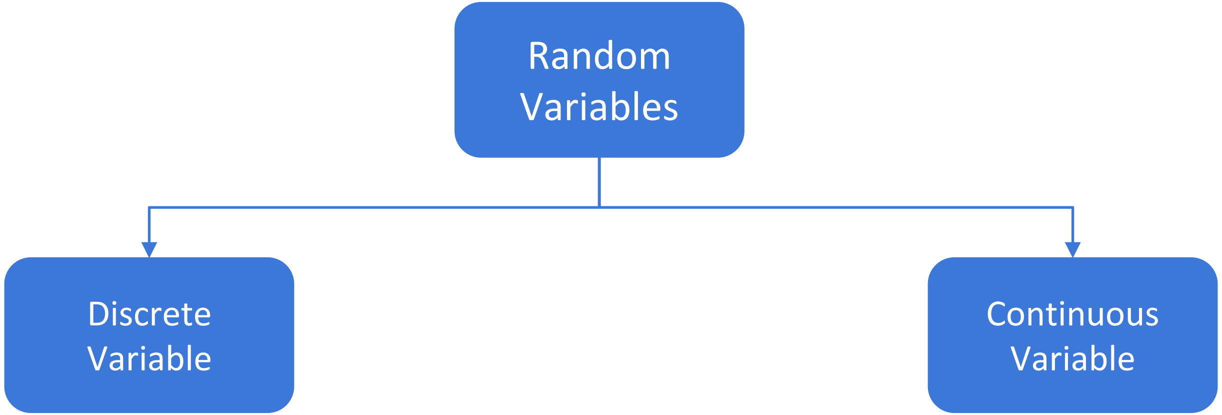
Random variable

Solution:

The nature of the data collected is as follows:

- mode of payment: the data takes two distinct values *Cashless payment* and *Cash payment*
- bill amount: It takes values such as $11.45, $32, $8.

Random variable



Eg: The mode of payment - cash, mobile wallet, card, bank transfer

Eg: The bill amount

Discrete random variable

- A discrete random variable takes discrete values that is value from the set of whole numbers only or a set of categories
- A discrete random variable can either take finite or countably infinite values
- The possible outcomes can be listed effectively

Discrete random variable

Examples

- Let V : The number of pages in a document
  Ω = {1, 2, 3,....}

- Let X: The number of views for a youtube video
  Ω = {0, 1, 2, 3,... }

- Let U: The size of the shirt available
  Ω = {XS, S, M, L, XL}

Note that random variables take values only from the set of whole numbers and are thus discrete random variables

Continuous random variable

- A continuous random variable takes infinitely many values (within a specified range)

- It can either take an uncountably infinite set of values or values in a given interval

Continuous random variable

Examples:

- Let Z: The income of a person
  Ω = (0, maximum income of a person)

- Let V: The level of water in a damn
  Ω = (0, height of the dam)

Note that random variables Z and V take values from a range and are thus continuous random variables

Identify whether the variable is discrete or continuous random variable.

1. Accidents on the New York highway

2. Fuel capacity of a car

3. Amount of milk in 15-ounce bottle

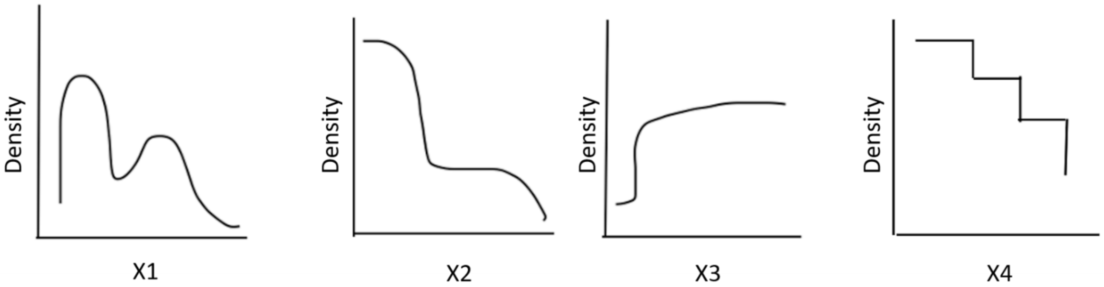
4. Members of the public health society

5. The total duration of a movie

Probability Distributions

Distribution of the data

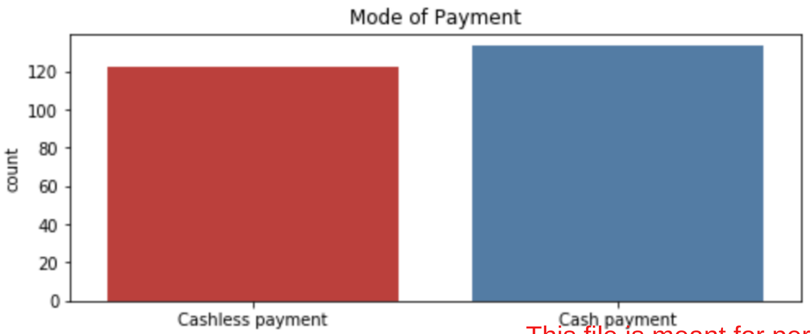
- The distribution is a summary of the frequency of values taken by a random variable
- The distribution of the data gives information on the shape and spread of the data
- On plotting the histogram or a frequency curve for a variable, we actually look at how the data is distributed over its range

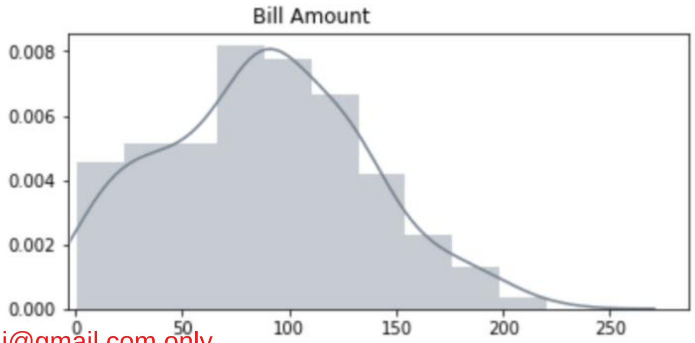


Probability distribution

Recall the example of data collection at the store. The data was collected for the mode of payment and the bill amount.

If we represent the data will have the following figures:





Need for probability distribution

- The statistical tests have assumptions based on the probability distributions that are needed to be verified before using the test

- Once the distribution of the data is known, the descriptive statistics of the data is readily available

- For instance, if the stock trader can determine the probability distribution of a particular stock, then the trader can estimate the potential expected returns that may yield in the future

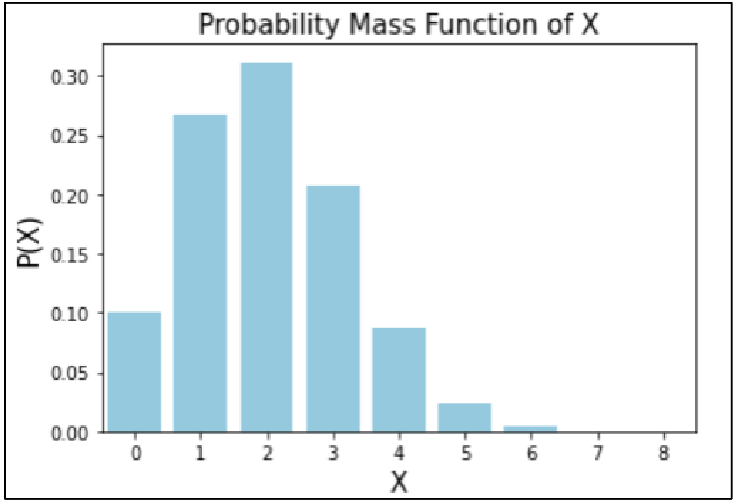
Probability distribution

- A probability distribution is a list of all the possible outcomes of a random variable along with their corresponding probability of occurrence

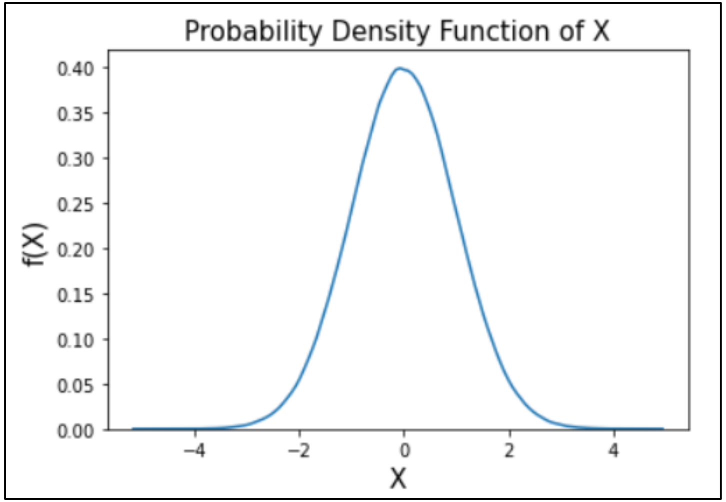
- P(X=x) refers to the probability that the random variable X takes a particular value x

- A probability distribution is defined by its
  - probability mass function (p.m.f) for a discrete random variable
  - probability density function (p.d.f) for a continuous random variable

Probability distribution



Discrete Variable



Continuous Variable

The PMF and PDF difference

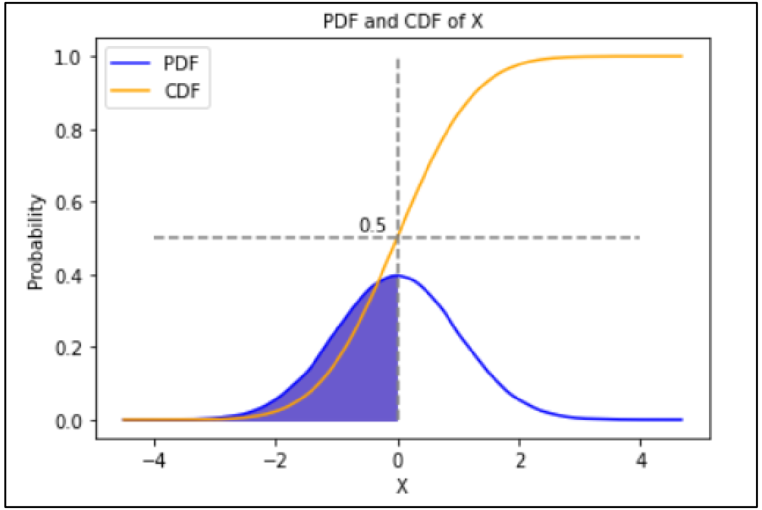
The mathematical difference between a discrete distribution and a continuous distribution is:

- The discrete distribution is defined by the probability mass function (pmf) which is the probability at a particular point

- The continuous distribution is defined by the probability density function (pdf) which is the integration over it’s range

Cumulative distribution function (c.d.f.)

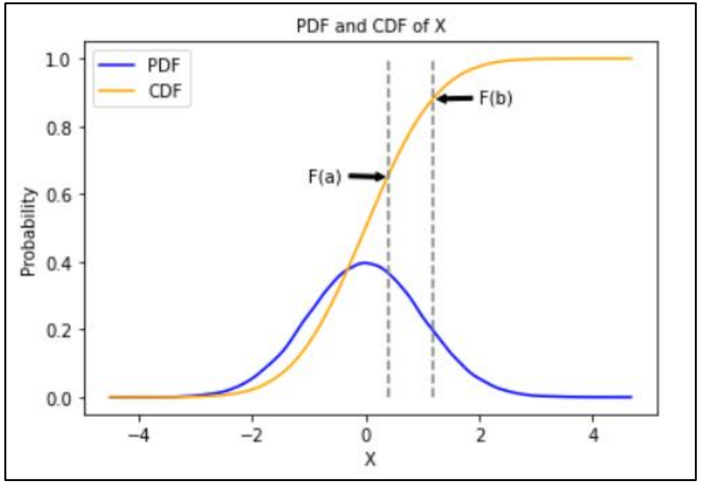
- It gives the cumulative distribution function i.e P(X ≤ x)
- Denoted by F(x)
- That is, F(x) = P(X ≤ x)
- The area of shaded region is 0.5



Properties of c.d.f.

- $0 \leq F(x) \leq 1$ since it is a probability
- An increasing function or monotonically non-decreasing function
- $F(-\infty) = 0$ and $F(\infty) = 1$
- For any two numbers a and b such that a < b, the probability is computed as:

$$P(a \leq X \leq b) = F(b) - F(a)$$



Mathematical expectation of a random variable

- It is the mathematical mean of a random variable
- It is the central location
- Also known as expected value
- Denoted by μ
- For a probability distribution, it is $\mu = \sum x.\ f(x)$

  Where f(x) is the probability mass function

Variance of a random variable

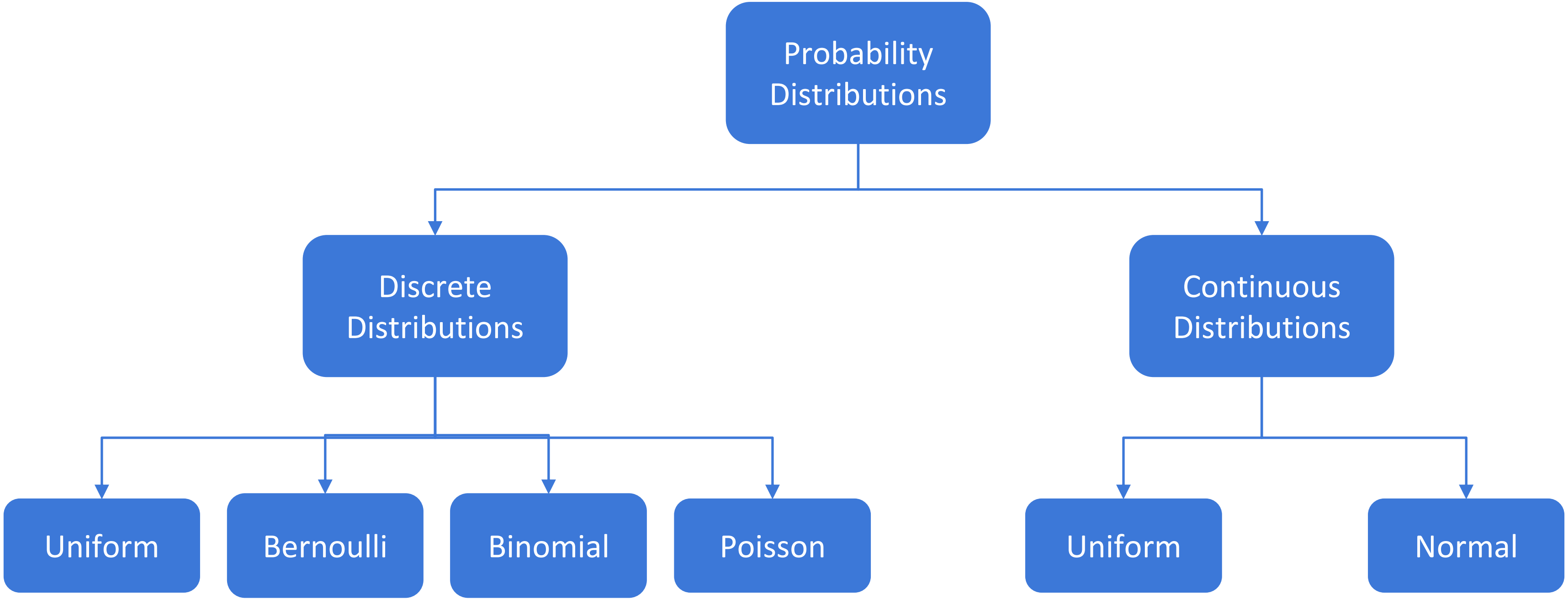
- It is the mathematical mean of squares of deviations taken from the mean

- Denoted by σ^2

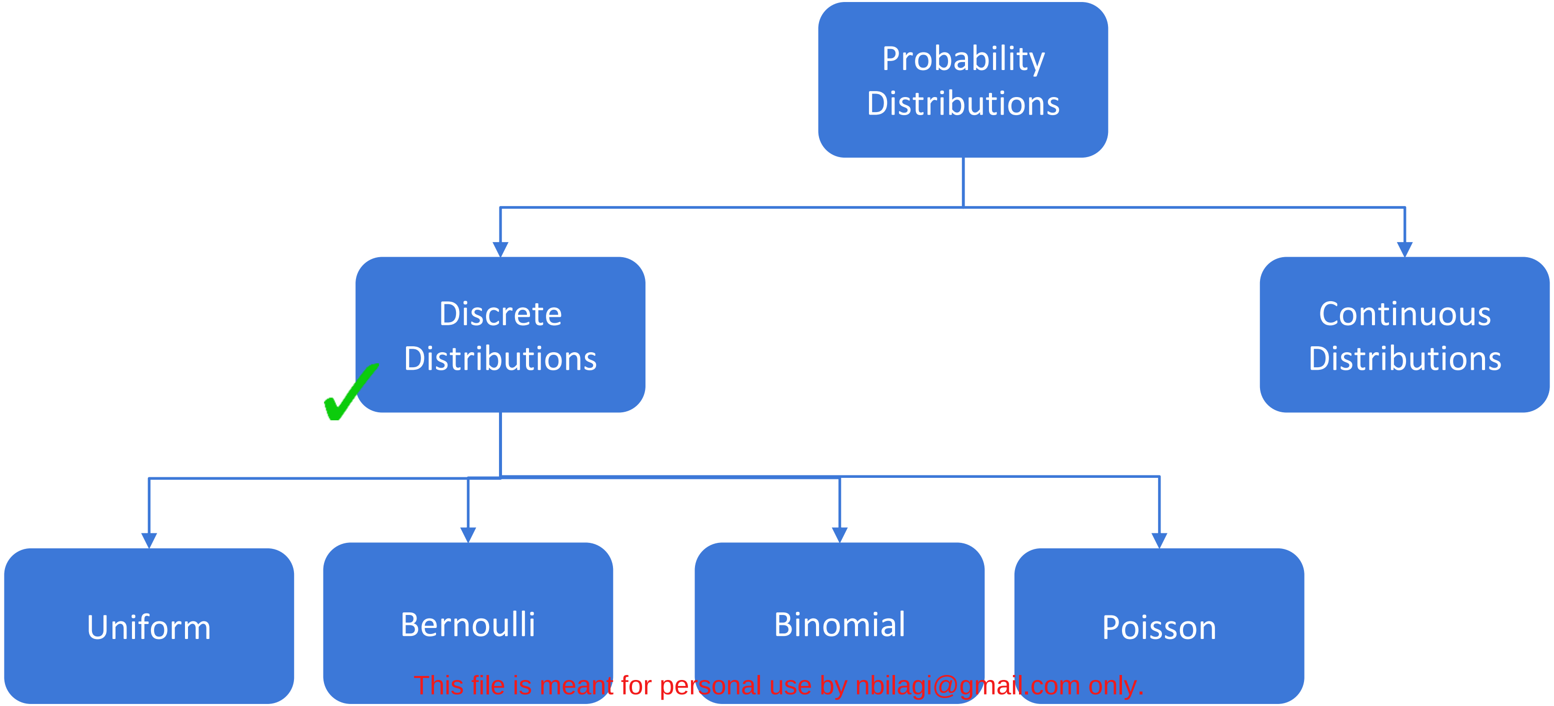
- For a probability distribution, it is $\sigma^2 = \sum (x - \mu)^2 . f(x)$

  where f(x) is the probability mass function

Probability distributions



Probability distributions



Discrete probability distribution

- A discrete distribution comprises of the probabilities of the outcomes of a random variable with finite values and is used to model a discrete random variable

- A discrete probability distribution is defined by its probability mass function (p.m.f)

- The mathematical expectation E(X) and variance V(X) of a discrete probability distribution is given by

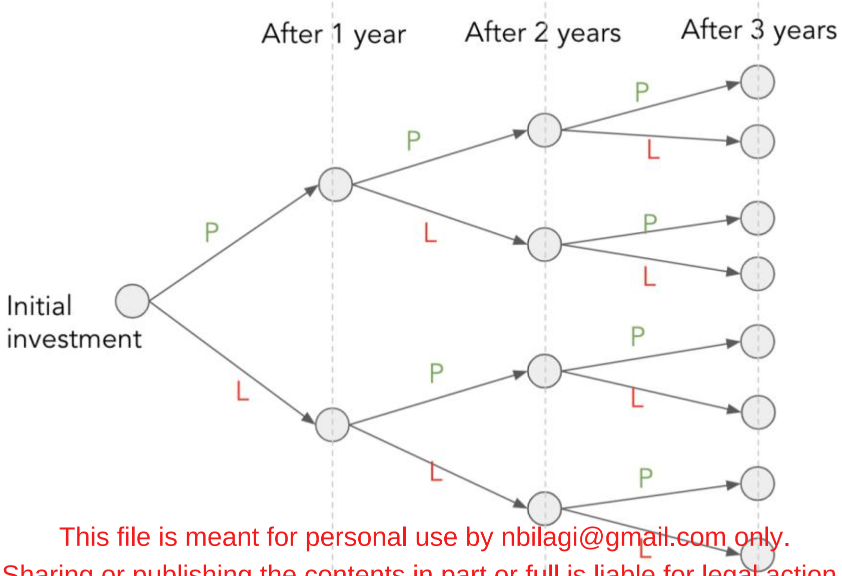
$$E(X) = \sum_{1=1}^{n} x_i P(x_i)$$

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

(Refer appendix A.1)

Binomial distribution

Steve has invested in a scheme. The following tree shows all the possibilities of how his investments may vary year to year over a period of three years



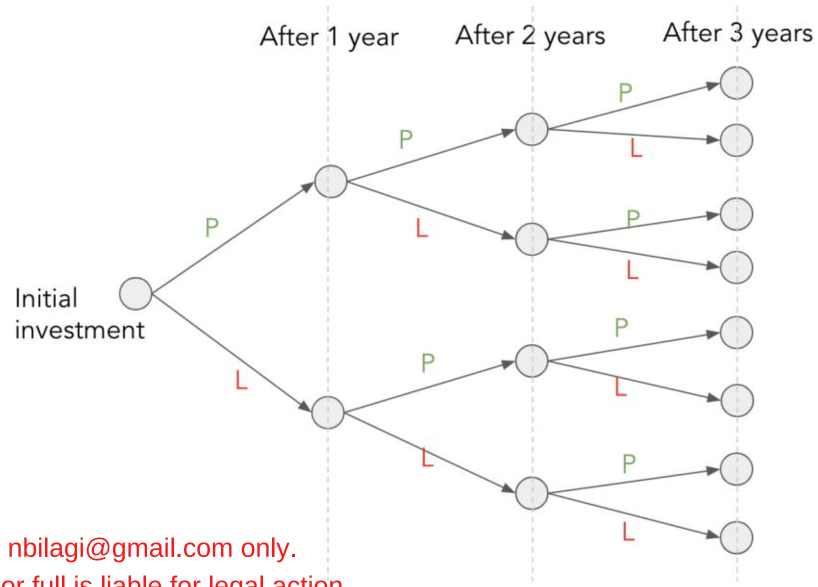
Binomial distribution

- Let X be the event of a change in the initial invested amount

- There is a profit with probability *p*, naturally the loss incurred will have the probability *1-p = q* (say)

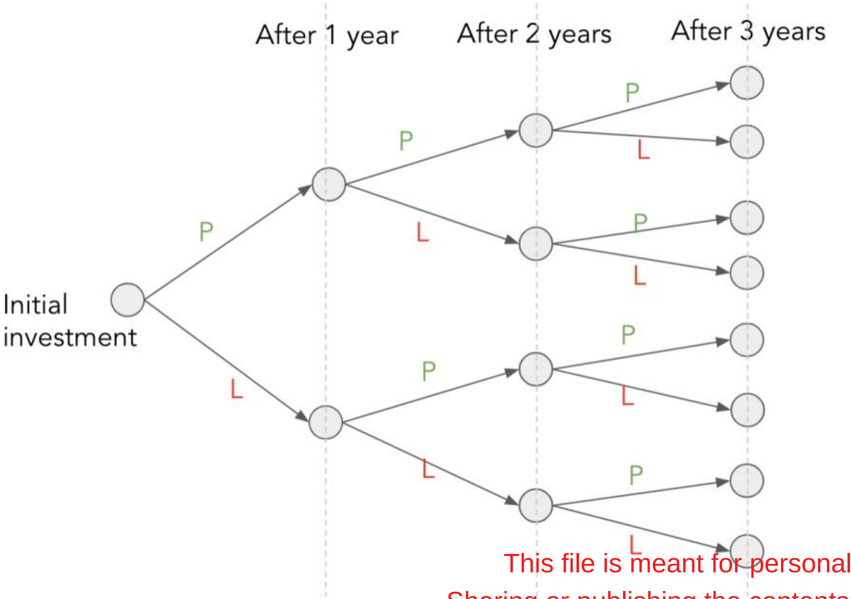
- Let the probability of profit be 0.4 i.e. P(X = P) = p = 0.4

- Let the probability of loss be 0.6 i.e. P(X = L) = 1 - p = 1- 0.4 = 0.6



Binomial distribution

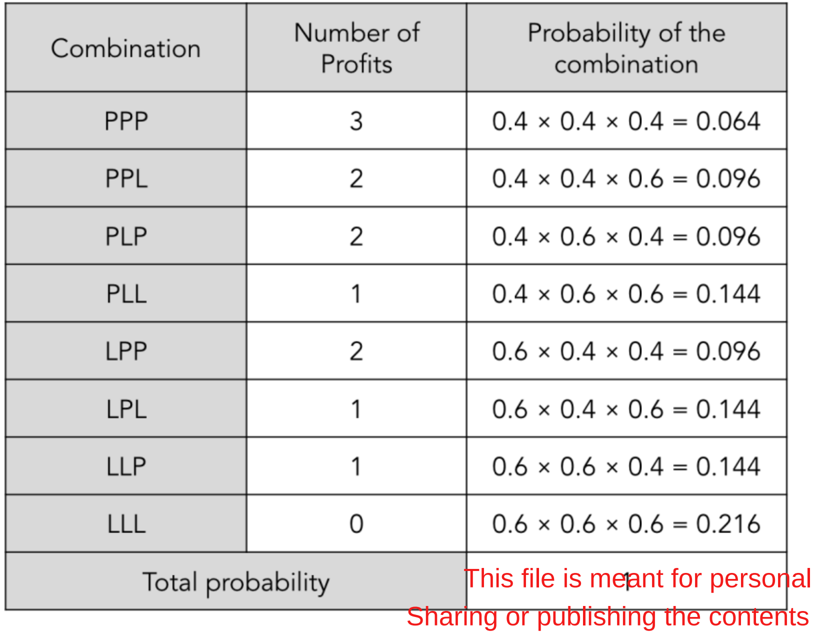
The probability of each of the possible scenarios is calculated as follows:



Combination	Number of Profits	Probability of the combination
PPP	3	0.4 × 0.4 × 0.4 = 0.064
PPL	2	0.4 × 0.4 × 0.6 = 0.096
PLP	2	0.4 × 0.6 × 0.4 = 0.096
PLL	1	0.4 × 0.6 × 0.6 = 0.144
LPP	2	0.6 × 0.4 × 0.4 = 0.096
LPL	1	0.6 × 0.4 × 0.6 = 0.144
LLP	1	0.6 × 0.6 × 0.4 = 0.144
LLL	0	0.6 × 0.6 × 0.6 = 0.216
Total probability		1

Binomial distribution

We can easily tabulate these probabilities.



Combination	Number of Profits	Probability of the combination
PPP	3	0.4 × 0.4 × 0.4 = 0.064
PPL	2	0.4 × 0.4 × 0.6 = 0.096
PLP	2	0.4 × 0.6 × 0.4 = 0.096
PLL	1	0.4 × 0.6 × 0.6 = 0.144
LPP	2	0.6 × 0.4 × 0.4 = 0.096
LPL	1	0.6 × 0.4 × 0.6 = 0.144
LLP	1	0.6 × 0.6 × 0.4 = 0.144
LLL	0	0.6 × 0.6 × 0.6 = 0.216
Total probability		1

Number of Profits	Probability
0	0.216
1	0.144
2	0.096
3	0.064

Binomial distribution

- It is based on the Bernoulli distribution

- A discrete random variable (X) taking values 0,1,2,...,n follows a binomial distribution if the p.m.f. of X is given as:

$$\begin{aligned} P(x) &= \binom{n}{x} p^x q^{n-x} \qquad && x = 0, 1, 2, \ldots, n \\ &= 0 && otherwise \end{aligned}$$

- X follows binomial distribution with parameters *n* and *p*, i.e X ~ Binomial(n,p)

- *p* is the probability of success, *q* is the probability of failure and *n* is the number of times the experiment is conducted

Binomial distribution

Recall Steve's example, multiply each probability with ${}^{n}C_{x}$

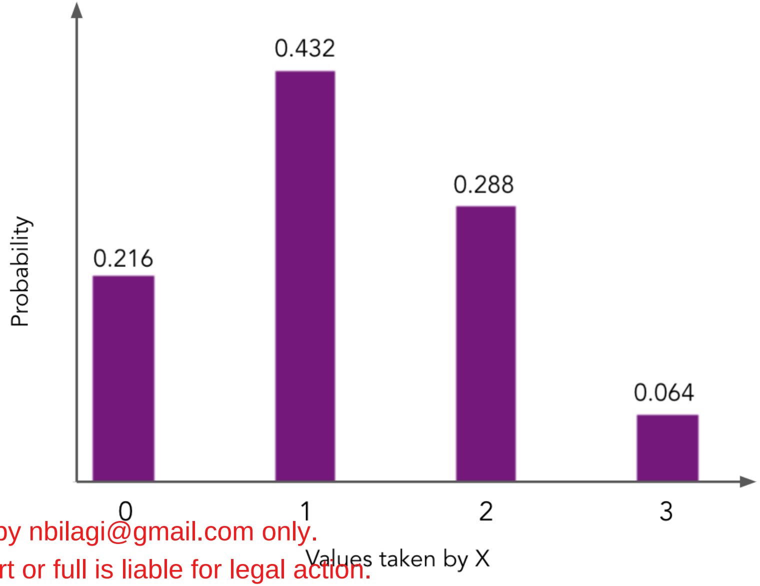
Number of Profits (x)	Probability	${}^{3}C_{x}$	P(x) = ${}^{3}C_{x}$ Probability
0	0.216	${}^{3}C_{0} = 1$	0.216
1	0.144	${}^{3}C_{1} = 3$	0.432
2	0.096	${}^{3}C_{2} = 3$	0.288
3	0.064	${}^{3}C_{3} = 1$	0.064

$$P(x) = \binom{n}{x} p^x q^{n-x} \qquad x = 0, 1, 2, \ldots, n$$
$$= 0 \qquad \textit{otherwise}$$

Binomial distribution

The probability of the Binomial(3, 0.4)

Number of Profits (x)	P(x) = ^3C_x Probability
0	0.216
1	0.432
2	0.288
3	0.064



If *n* independent random variables are Bernoulli distributed, then the sum of these random variables follows a Binomial distribution

Binomial distribution

The mean and the variance of the distribution is given as:

$$Mean = E(X) = \sum_{x=0}^{n} xP(x)$$

$$= \sum_{x=0}^{n} x\binom{n}{x}p^x q^{n-x}$$

$$= np$$

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

$$= \text{npq}$$

For our example, if n = 10, p = ⅔ and q = ⅓ . Then, E(X) = 20/3 and V(X) = 20/9

Binomial distribution - python function

Python function	Description
scipy.stats.binomial.pmf()	returns the pmf of discrete binomial distribution
scipy.stats.binomial.cdf()	returns the cdf of discrete binomial distribution
scipy.stats.binomial.sf()	calculates the value of survival function (1 - cdf)

Binomial distribution

Question:

In a shooting academy, data was collected on the precision shooting of a student. From 15 shots fired 11 were on target. Considering the same student, what is the probability that out of 50 shots fired, exactly 35 will hit the target?

Binomial distribution

Solution:

Let X denote the shot fired hits the target

The probability of success, i.e. is hitting the target is 11/15
Thus, p = 0.733

To find: the probability that out of 50 shots fired, 35 will hit the target

Here, n = 50

We have X~Binomial(50, 0.733)

Binomial distribution

Solution:

The required probability is P(X = 35)

$P(X = x) = {}^{n}C_{x}\ p^{x}\ (1\text{-}p)^{n\text{-}x}$

$P(X = 35) = {}^{50}C_{35}\ (0.733)^{35}\ (1\text{-}0.733)^{50\text{-}35}$

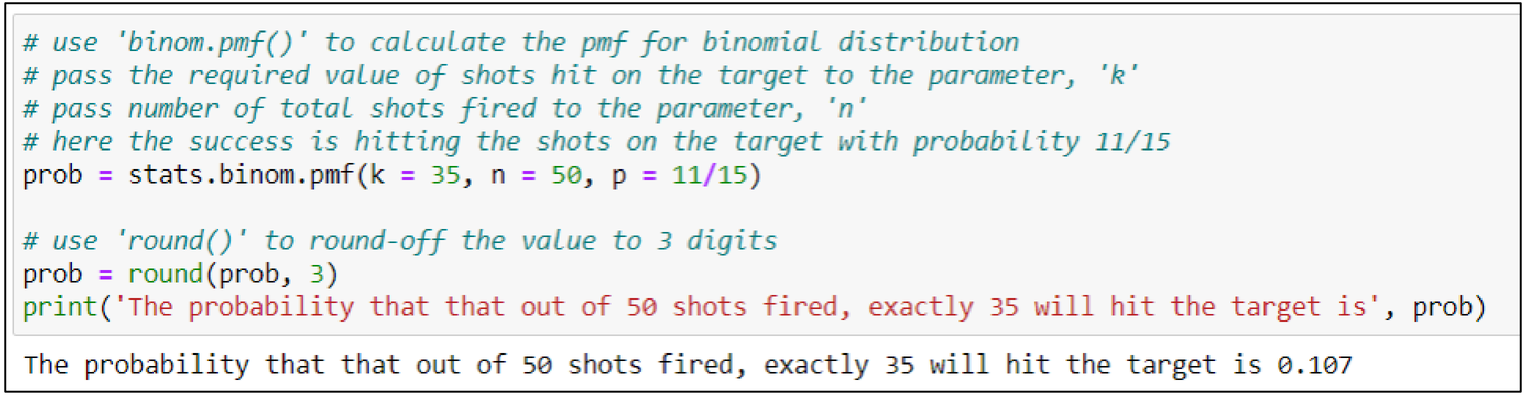
$= {}^{50}C_{35}\ (0.733)^{35}\ (0.267)^{15}$

$= 0.107$

Thus, the probability that out of 50 shots fired, exactly 35 will hit the target is 0.107

Binomial distribution

Python solution:



```
# use 'binom.pmf()' to calculate the pmf for binomial distribution
# pass the required value of shots hit on the target to the parameter, 'k'
# pass number of total shots fired to the parameter, 'n'
# here the success is hitting the shots on the target with probability 11/15
prob = stats.binom.pmf(k = 35, n = 50, p = 11/15)

# use 'round()' to round-off the value to 3 digits
prob = round(prob, 3)
print('The probability that that out of 50 shots fired, exactly 35 will hit the target is', prob)
```

```
The probability that that out of 50 shots fired, exactly 35 will hit the target is 0.107
```

Poisson distribution

- It is used to determine the probability of events that occur in a fixed interval of time/space
- It is used to describe the distribution of rare events in a large population
- A discrete random variable X follows a Poisson distribution with parameter λ, if p.m.f. of X is given as:

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

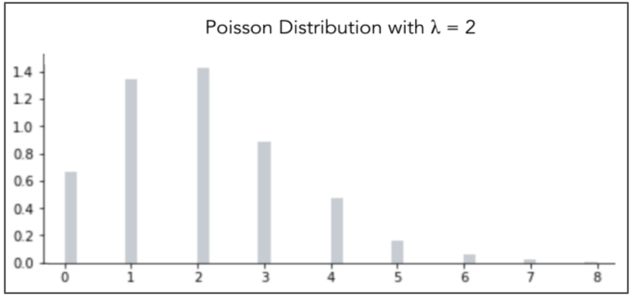
Where,

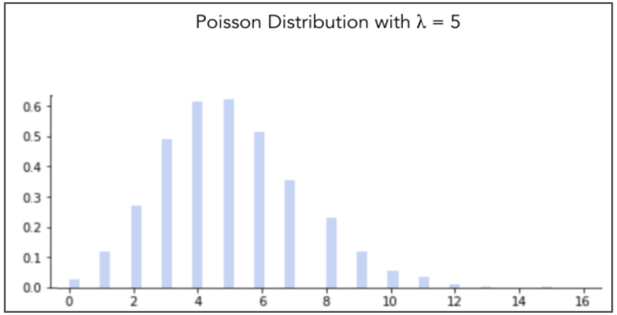
e = Euler's number (= 2.718)

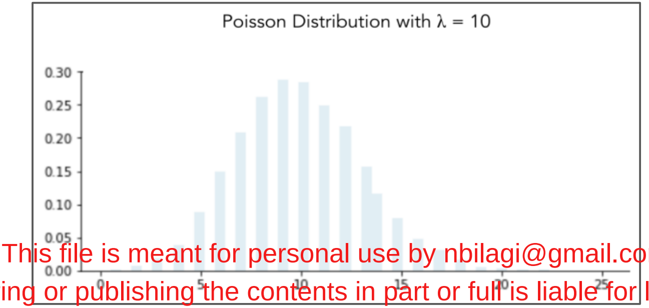
x! = Factorial of x

Poisson distribution

The Poisson distribution is positively skewed.







Poisson distribution

The mean and the variance of the distribution is given as:

$$Mean = E(X) = \sum_{x=0}^{\infty} xP(x)$$
$$= \lambda$$

$$\text{Variance} = \text{E}(\text{X}^2) - [\text{E}(\text{X})]^2$$
$$= \lambda$$

Mean and variance of the poisson distribution are equal and it is the parameter of distribution.

did you know?

Poisson as a limiting case of binomial distribution

Let X~Binomial(n, p) such that for large n ($n \to \infty$) and small p, $np = m$ is a constant then X approaches Poisson(m)

Poisson distribution - python function

Python function	Description
scipy.stats.poisson.pmf()	returns the pmf of discrete poisson distribution
scipy.stats.poisson.cdf()	returns the cdf of discrete poisson distribution
scipy.stats.poisson.sf()	calculates the value of survival function (1 - cdf)

Poisson distribution

Question:

The number of calls received at a telephone exchange in a day follows a Poisson distribution. The probability that the telephone exchange receives 5 calls is three times the probability that the telephone exchange receives 10 calls. Obtain the average calls that the telephone exchange receives in a day.

Poisson distribution

Solution:

Let X: number of calls received at a telephone exchange

We know X~Poisson(λ)

It is known that probability the exchange receives 5 calls is three times that of the exchange receiving 10 calls.
i.e. $P(X=5) = 3\ P(X=10)$

To find: the average calls that the telephone exchange receives in a day

Poisson distribution

Solution:

The distribution of X is given by

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

We have,

$$\frac{e^{-\lambda}\lambda^{5}}{5!} = 3\,\frac{e^{-\lambda}\lambda^{10}}{10!}$$

Poisson distribution

Solution:

$$\frac{e^{-\lambda}\lambda^5}{5!} = 3\,\frac{e^{-\lambda}\lambda^{10}}{10!}$$

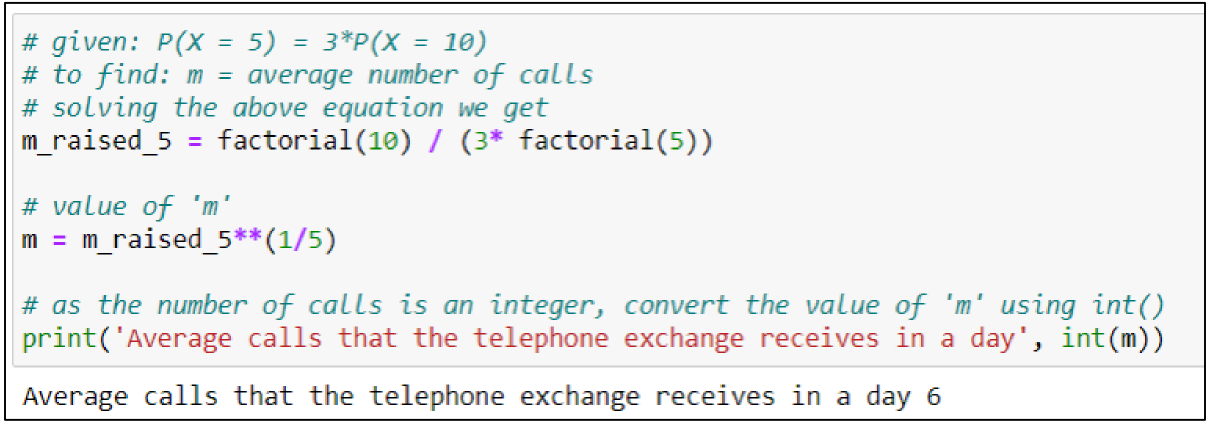
$$\Rightarrow \lambda^5 = \frac{10!}{3\times 5!}$$

$$\Rightarrow \lambda = \sqrt[5]{10080} = 6.32$$

Thus, the average calls received at the telephone exchange in a day are 6 (≈ 6.32)

Poisson distribution

Python solution:



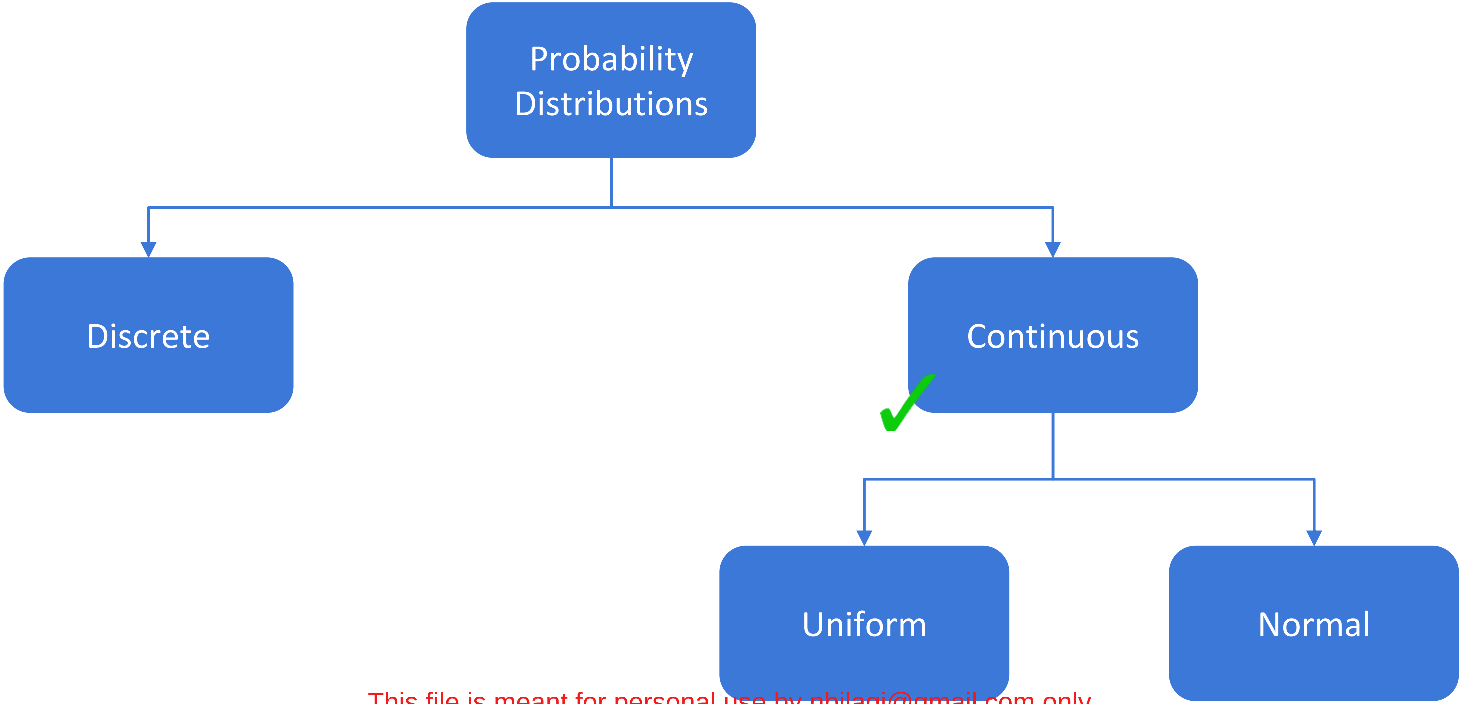
```
# given: P(X = 5) = 3*P(X = 10)
# to find: m = average number of calls
# solving the above equation we get
m_raised_5 = factorial(10) / (3* factorial(5))

# value of 'm'
m = m_raised_5**(1/5)

# as the number of calls is an integer, convert the value of 'm' using int()
print('Average calls that the telephone exchange receives in a day', int(m))
```

```
Average calls that the telephone exchange receives in a day 6
```

Summary

- Some of the discrete distributions are: Uniform, Bernoulli, Binomial and Poisson
- Discrete distribution is determined using probability mass function (p.m.f.)
- In uniform distribution, all the values of the variable are equally likely
- The sum of independent and identically distributed Bernoulli variables follows binomial distribution
- Poisson distribution can be considered as a limiting case of a binomial distribution

Probability distributions



Continuous probability distribution

- It is a probability distribution of the continuous random variable
- The continuous distributions are defined by their probability density function (p.d.f.) denoted as f(x)
- The probability is given by the area under the pdf on a specified range. Thus, the value of the probability density function is 0 at a particular point
- The area under the curve is always equal to 1. Therefore, is considered as a probability distribution

Normal distribution

The probability density function for the normal distribution is,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{where} \quad -\infty < x, \mu < \infty, \sigma > 0$$

x and μ take value between -∞ to +∞

σ is strictly positive

μ is the mean and σ^2 is the variance

It is said that X ~ N($\boldsymbol{\mu}$, σ^2), i.e the variable X follows the normal distribution with parameter $\boldsymbol{\mu}$ and σ^2

Normal distribution

The mean and the variance of the distribution is given as:

$$Mean = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

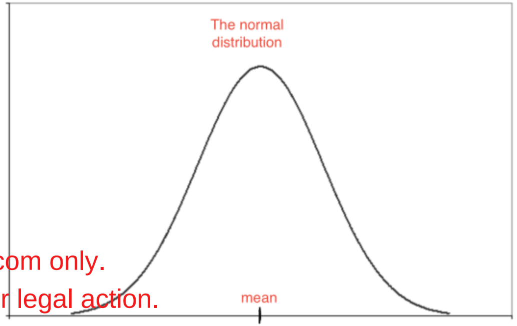
$$= \mu$$

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

$$= \sigma^2$$

Normal distribution

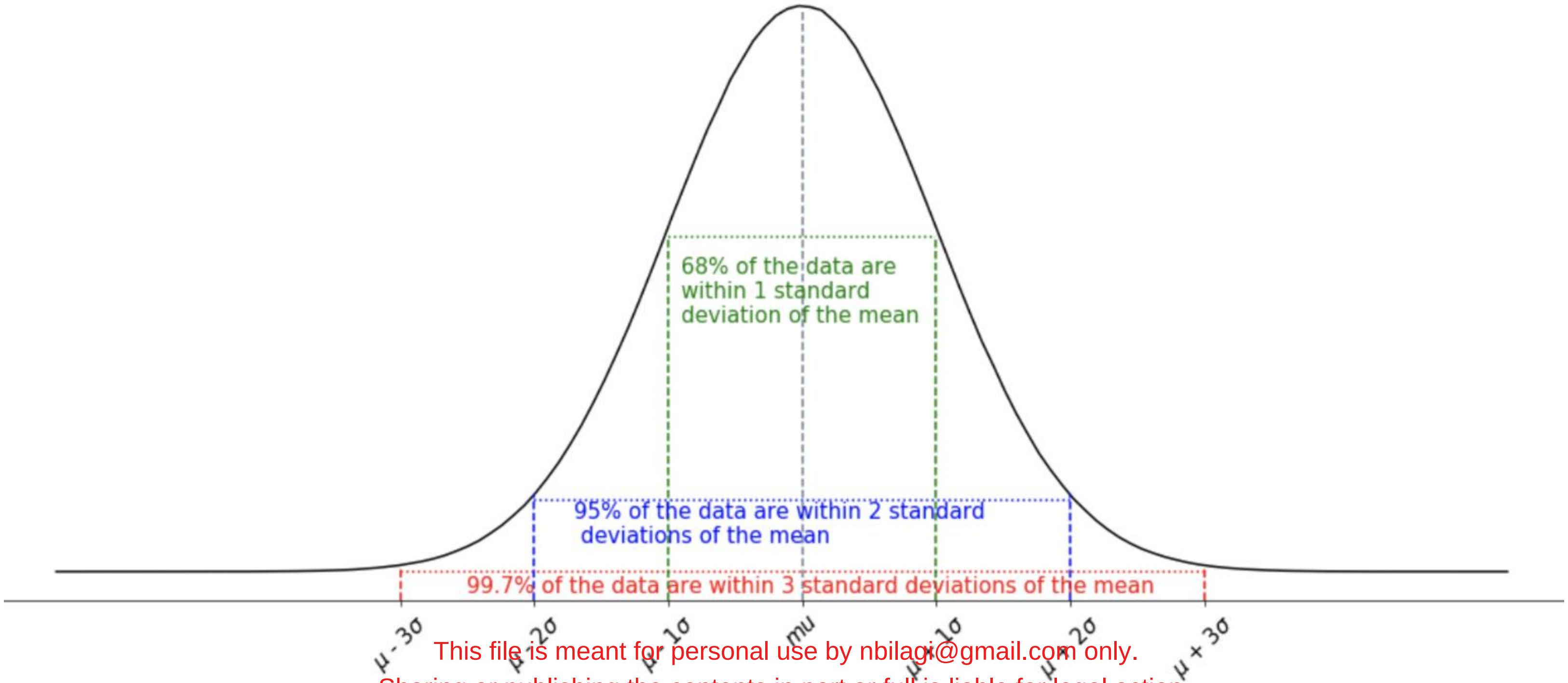
- Also called as the Gaussian distribution
- The frequent observations are found in the middle of the distribution, and further decrease away towards the tails
- A normal distribution has a bell-shaped density curve described by its mean and standard deviation
- For normally distributed data,
  - Mean = Mode = Median
  - Symmetric about the center



Characteristics of normal distribution

- The distribution is symmetric about the mean
- For X ~ $N(\mu, \sigma^2)$, $P(X > a) = P (X < -a)$ due to symmetricity
- The standard normal distribution has mean 0 and variance 1
- For X ~ $N(\mu, \sigma^2)$, if $Z = (x - \mu)/\sigma$ then Z ~ $N(0, 1)$ that is the standard normal distribution

The normal distribution spread

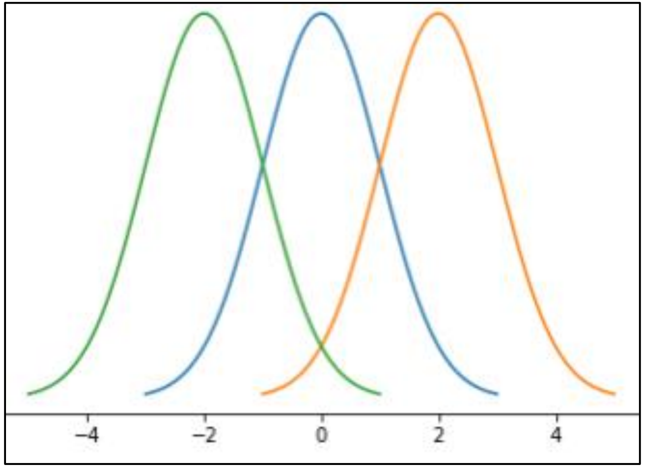


Normal distribution

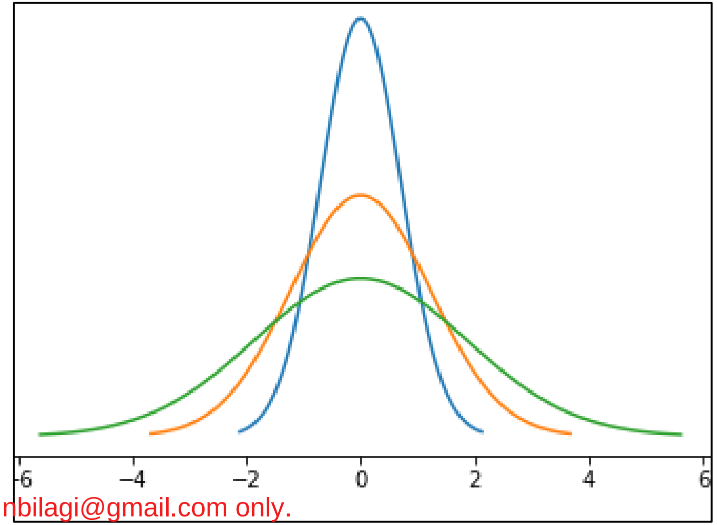
- IQ of the human population is the example of normal distribution
- Most of the people in the specific population have average IQ
- The number of people with higher IQ and lower IQ than the average IQ is almost equal, and a very less number of people have extremely less IQ or extremely high IQ

Different shapes of normal distribution

Normal distributions with same standard deviation but different means:



Normal distributions with same mean but different standard deviations:



did you know?

Normal approximation of binomial distribution

Let X~Binomial(n, p) such that for large $np \geq 5$ and $n(1 - p) \geq 5$ then X approaches ________

Normal distribution with mean $\mu = np$ and variance $\sigma^2 = \sqrt{np(1 - p)}$

Normal distribution - python function

Python function	Description
scipy.stats.normal.pmf()	returns the pmf of discrete normal distribution
scipy.stats.normal.cdf()	returns the cdf of discrete normal distribution
scipy.stats.normal.sf()	calculates the value of survival function (1 - cdf)

Normal distribution

Question:

A monthly balance in the bank account of credit card holders is assumed to be normally distributed with mean \$500 and variance \$100. What is the probability that the balance can be more than \$513.5?

Normal distribution

Solution:

Let X: monthly balance in the bank account of a credit card holders

$X \sim N(\mu = 500, \sigma^2 = 100)$

To find: the probability that the balance can be more than $513.5

The required probability is $P(X > 513.5)$

Normal distribution

Solution:

$$P(X > 513.5) = \int_{513.5}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}} dx$$

$$= \int_{513.5}^{\infty} \frac{1}{\sqrt{2\pi \times 100}} e^{-\frac{(x-500)^2}{100}} dx$$

$$= \int_{513.5}^{\infty} \frac{1}{25.07} e^{-\frac{(x-500)^2}{100}} dx$$

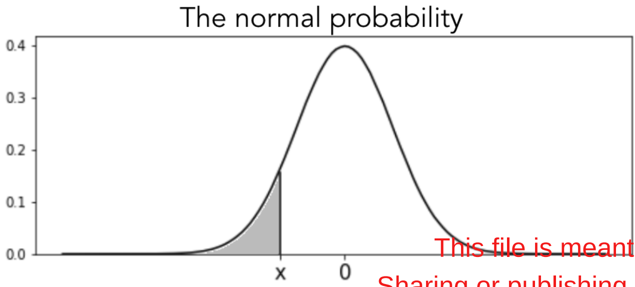
… Here is a problem. We see it is difficult to solve the integral, so we make use of the normal tables.

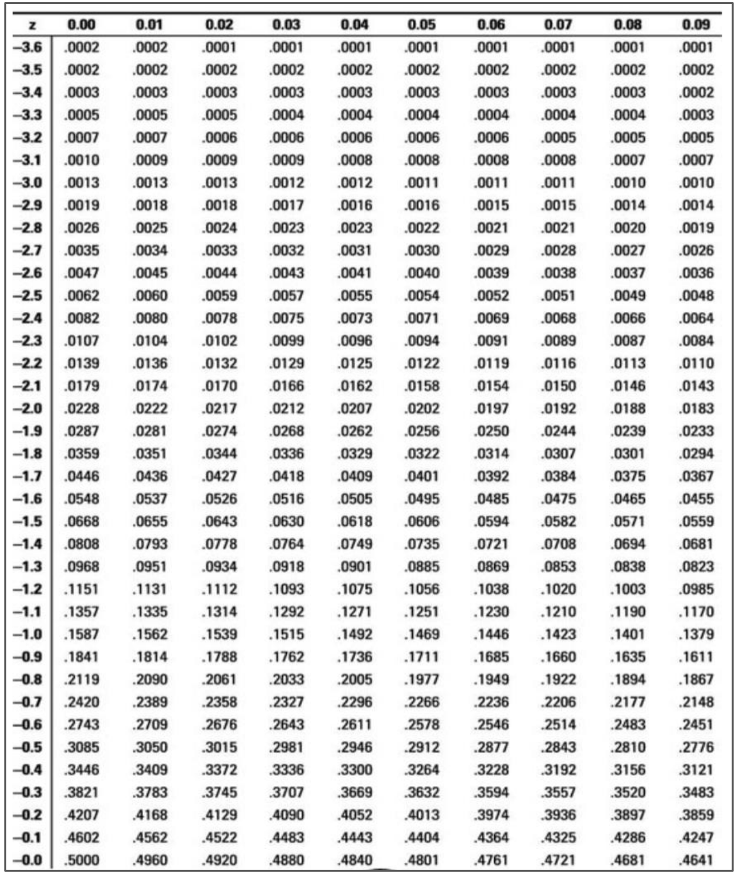
The normal table

- Let the r.v. X strictly follow the standard normal distribution

- Let r.v. X take a negative value x

- This table gives the lower tail probabilities; P(X ≤ x)





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
–3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
–3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
–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	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

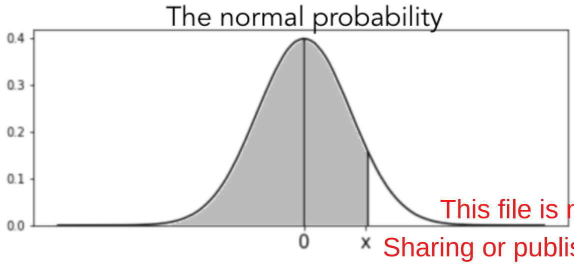
Table Ref: Link

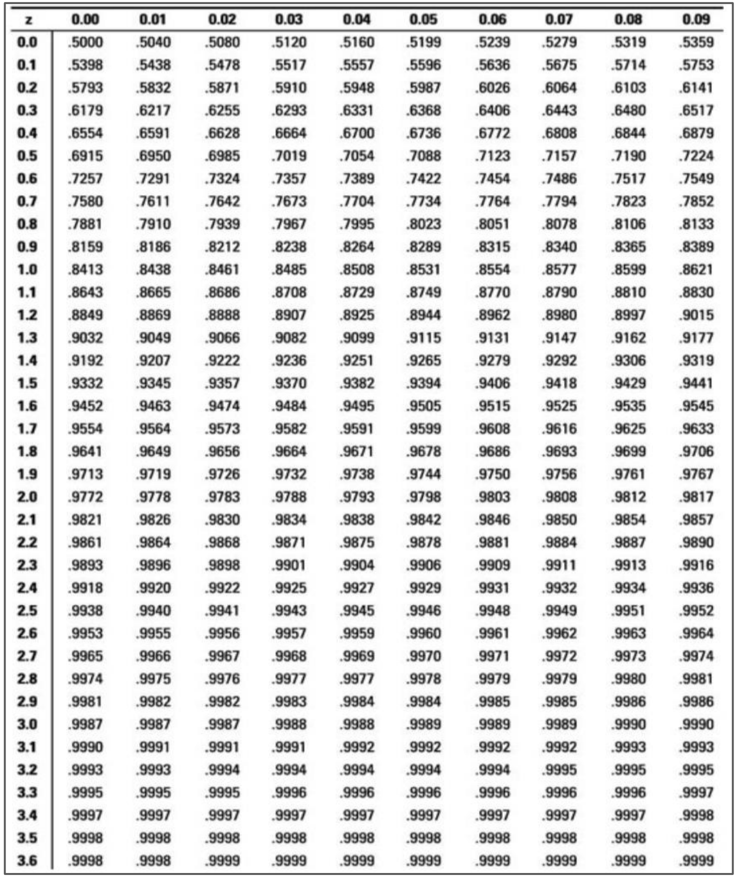
The normal table

- Let the r.v. X strictly follow the standard normal distribution

- Let r.v. X take a positive value x

- This table gives the lower tail probabilities; P(X ≤ x)





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Table Ref: Link

Normal distribution

Solution:

We know that for $X \sim N(\mu, \sigma^2)$, if $Z = (x - \mu)/\sigma$ then $Z \sim N(0, 1)$ that is the standard normal distribution

$X \sim N(\mu = 500, \sigma^2 = 100)$ thus $Z = (x - 500)/\sqrt{100} \sim N(0, 1)$

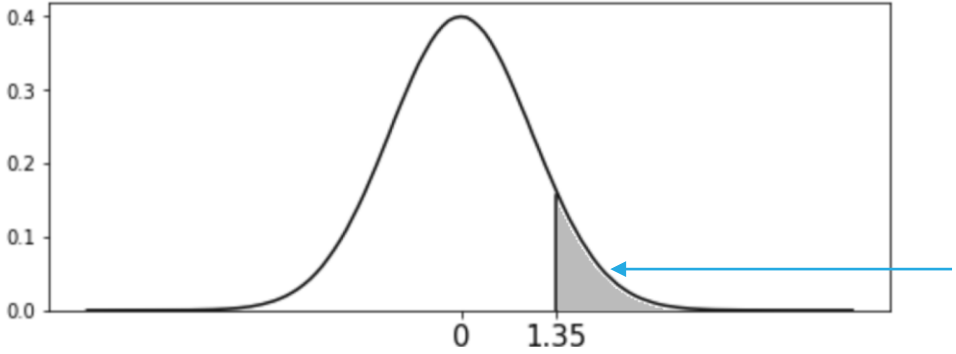
The required probability is $P(X > 513.5)$, now becomes

$$P(X > 513.5) = P(\frac{X-\mu}{\sigma} > \frac{513.5-500}{\sqrt{100}}) = P(Z > \frac{513.5-500}{\sqrt{100}}) = P(Z > 1.35)$$

Normal distribution

Solution:

Thus the required probability now is P(Z > 1.35)



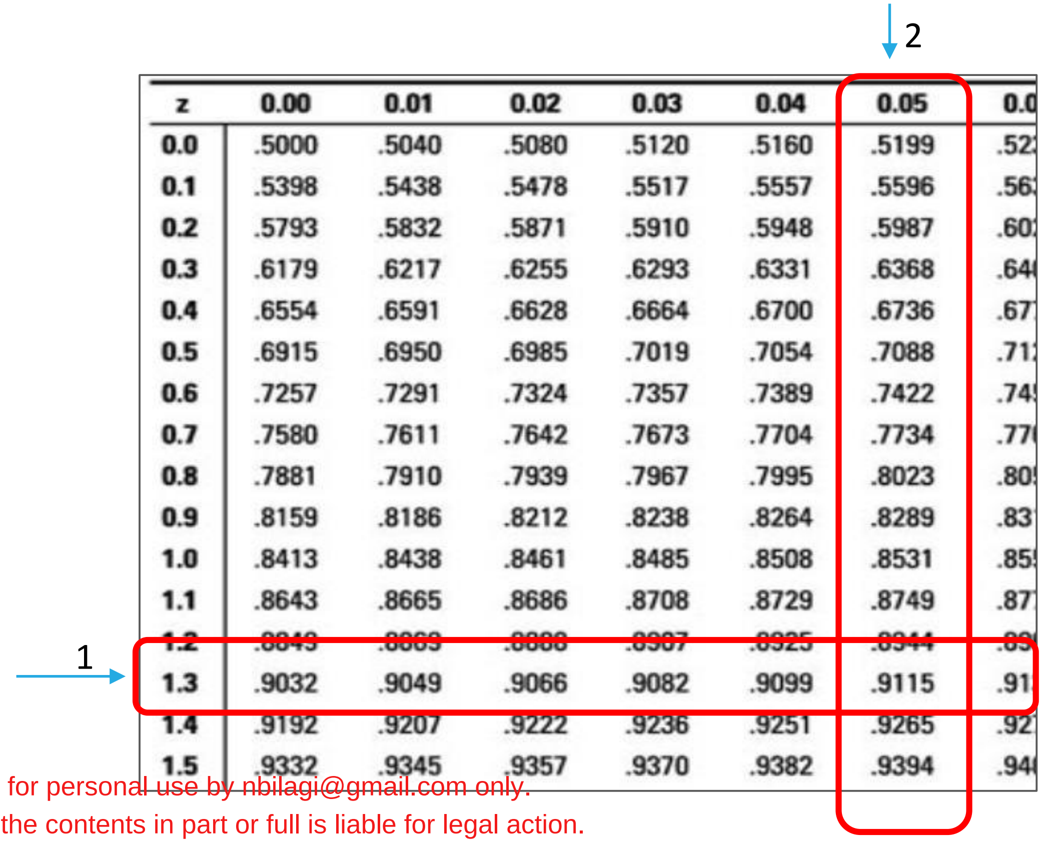
The shaded region represents the required probability

Now we refer to the normal table to the probability

Normal distribution

Solution:

The table gives P(Z ≤ 1.35).



z	0.00	0.01	0.02	0.03	0.04	0.05	0.0
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.52
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.56
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.60
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.64
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.67
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.71
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.74
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.77
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.80
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.83
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.85
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.87
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.89
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.91
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.92
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.94

1. Look horizontally in the row of 1.3
2. Consider the value corresponding column to 0.05

Normal distribution

Solution:

Since we get the lower tail probability as 0.9115

The required probability $P(Z > 1.35) = 1 - P(Z \leq 1.35)$

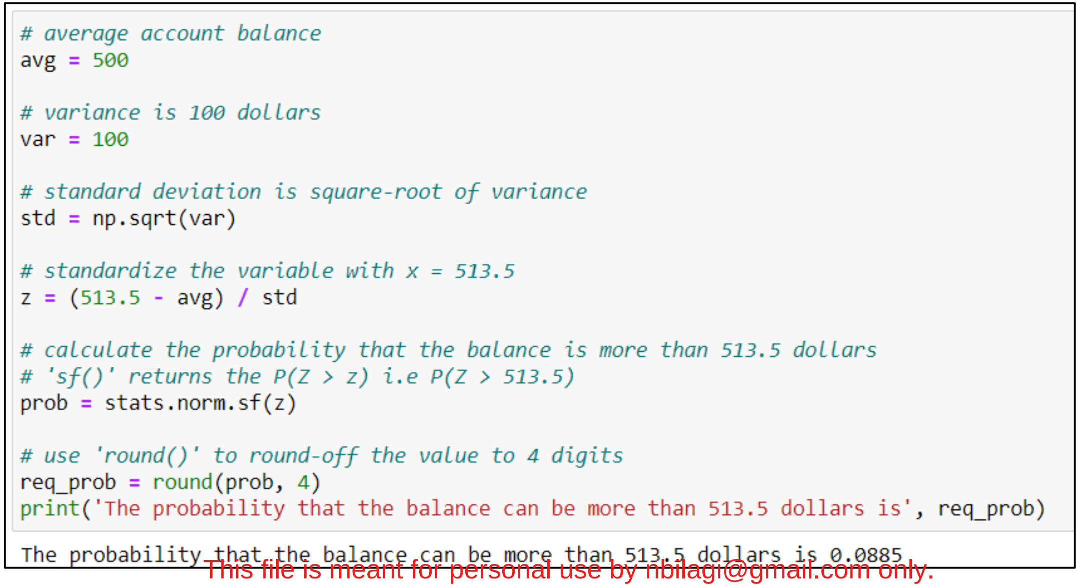
$= 1 - 0.9115$

$= 0.0885$

The probability that the balance can be more than \$513.5 is 0.0885

Normal distribution

Python solution:



```python
# average account balance
avg = 500

# variance is 100 dollars
var = 100

# standard deviation is square-root of variance
std = np.sqrt(var)

# standardize the variable with x = 513.5
z = (513.5 - avg) / std

# calculate the probability that the balance is more than 513.5 dollars
# 'sf()' returns the P(Z > z) i.e P(Z > 513.5)
prob = stats.norm.sf(z)

# use 'round()' to round-off the value to 4 digits
req_prob = round(prob, 4)
print('The probability that the balance can be more than 513.5 dollars is', req_prob)
```

```
The probability that the balance can be more than 513.5 dollars is 0.0885
```

The scipy.stats.norm.cdf() in python gives the lower tail probabilities

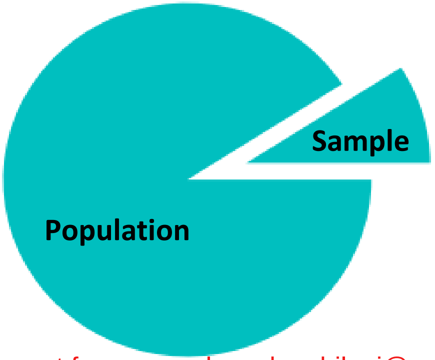
Summary

- Continuous distributions: Uniform and normal
- Continuous distribution is determined using probability density function (p.d.f.)
- The value of pdf at a particular point is always zero
- For a normal distribution, Mean = Median = Mode
- Standard normal distribution has mean zero and variance one

Population and Sample

Population & Sample

- Population is a collection of all the individuals or objects
- Sample is a subset of the population which is a representative of the population



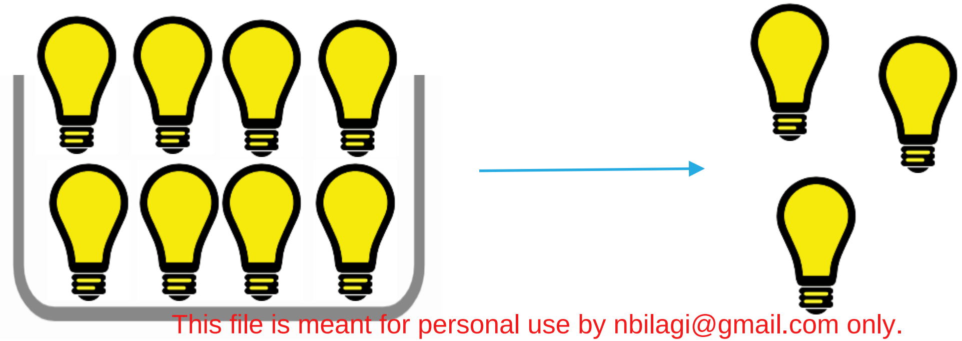
Need to draw a sample

- In some of the real case scenarios we may not be able to gather all the observations in the population
- In such cases we can consider a sample which can be used to draw inferences about the population
- Sampling can reduce the computational time and cost
- In machine learning models, a sample obtained using an appropriate sampling strategy can provide accurate results with the less number of observations

Need to draw a sample

Suppose a company producing electric bulbs wants to know the average life of a bulb. If the details of all the bulbs are available, then this information is regarded as the population.

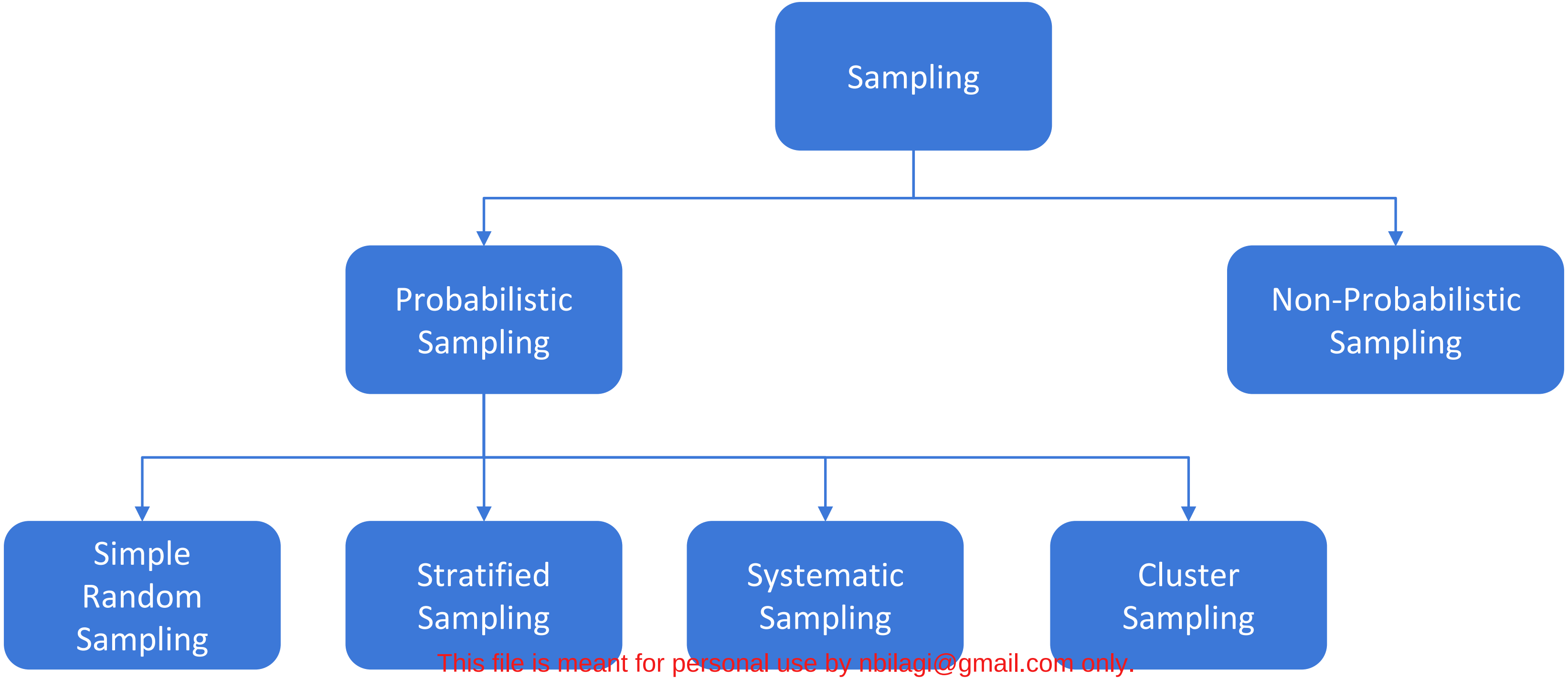
It would be challenging for the company to test each and every bulb produced. In such a scenario, the company would draw a sample from the produced bulbs to test.



Terminologies

- Sampling unit: An individual and indivisible unit of the population eligible to be in the sample
- Sampling frame: An exhaustive list of all members or elements of a population
- Sampling: To consider a subset of the population
- Sampling fraction (f): Ratio of sample size (n) to population size (N), i.e, f = n/N

Sampling techniques



Simple random sampling (SRS)

- Simplest and most common method
- Sample is drawn unit by unit
- Each unit has equal probability of being selected
- Two ways of selecting units:
  - Without replacement: if the selected unit is not returned to the population
  - With replacement: if the selected unit is returned to the population

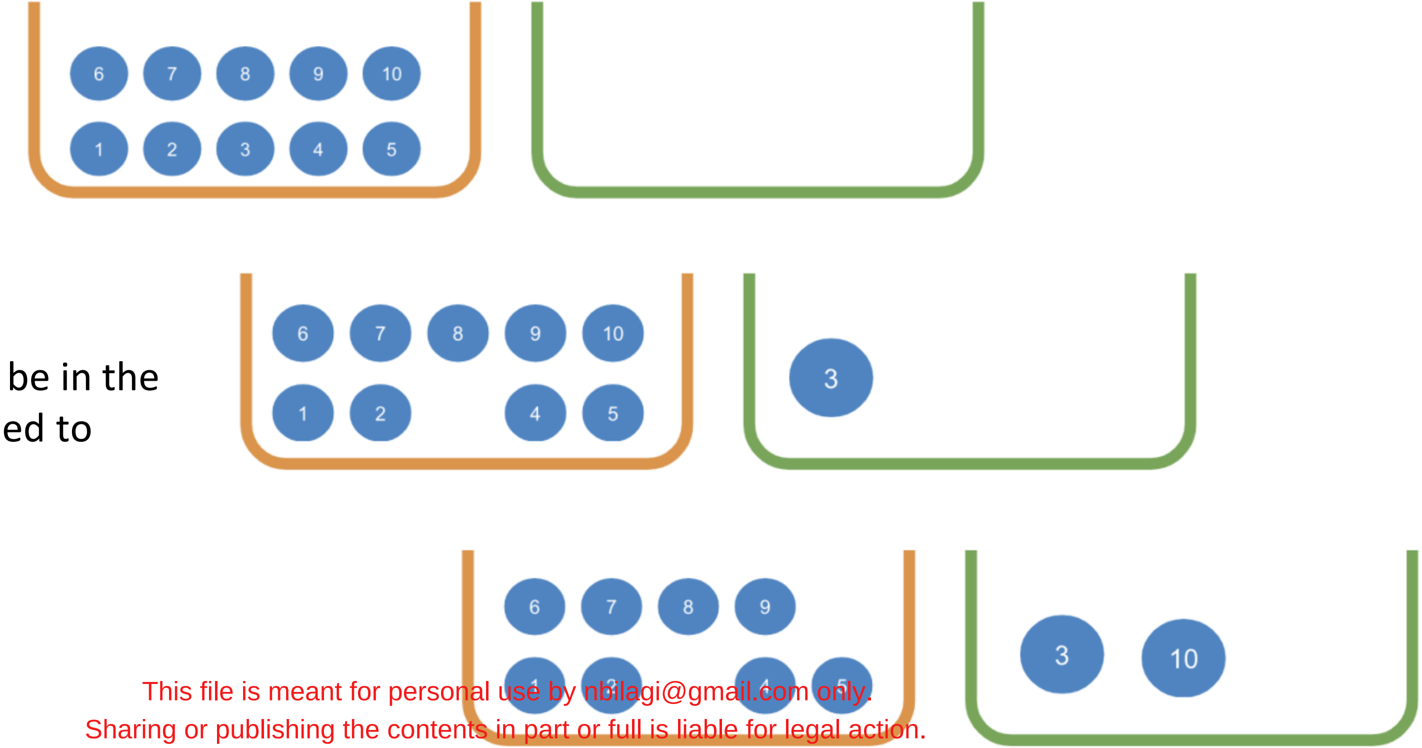
SRS without replacement

- At any specific stage in the sampling, the probability of any of the remaining units is 1/N, where N is the total number of sample units

- From N sample units, if a sample of size n is drawn, there are ${}^{N}C_{n}$ possible samples

- The probability of selecting one of them is $1/{}^{N}C_{n}$

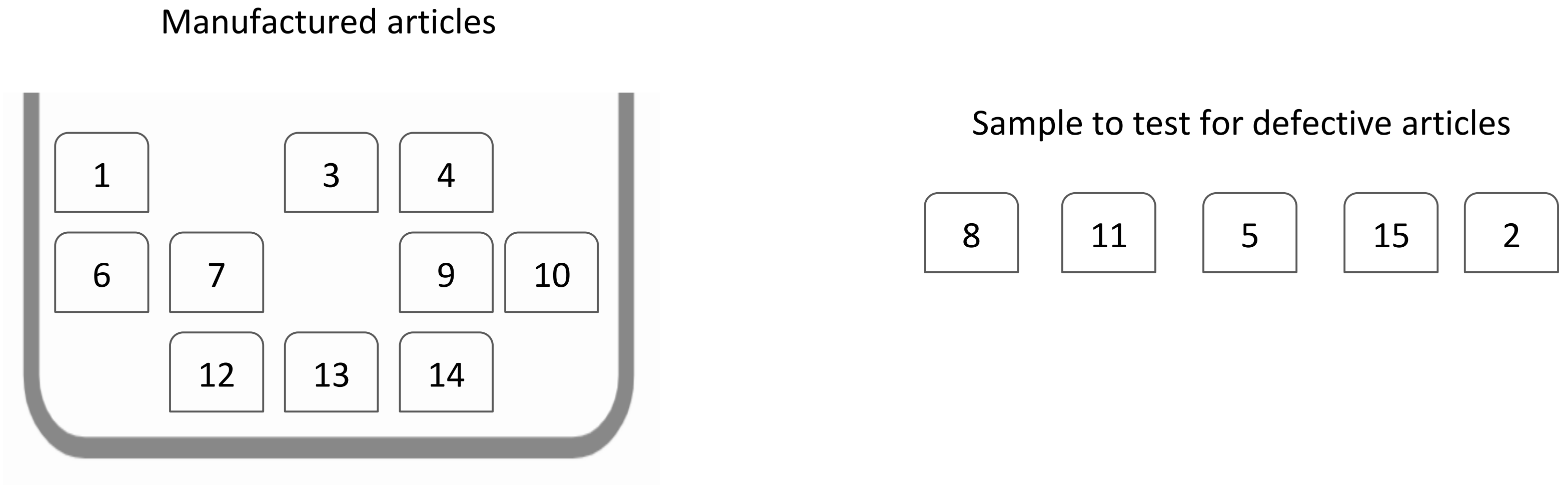
SRS without replacement



The unit selected to be in the sample is not returned to the population.

SRS without replacement - example

Drawing a sample from manufactured articles to know the proportion of defectives.



Manufactured articles

Sample to test for defective articles

SRS with replacement

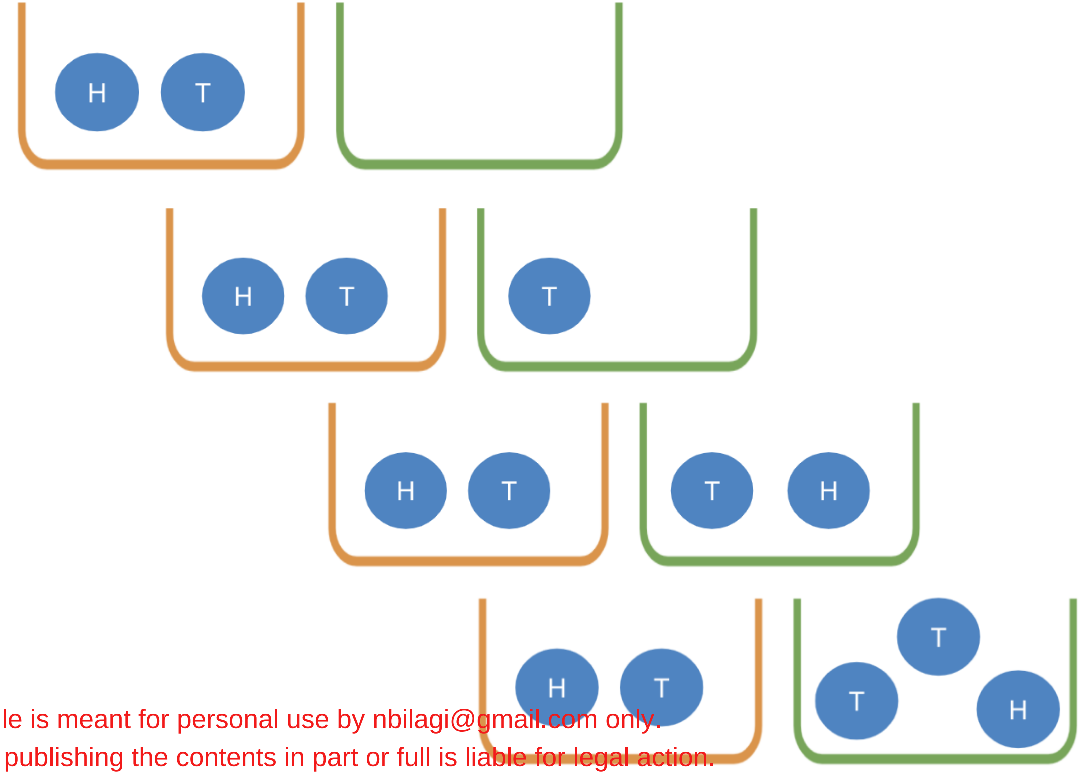
- The probability of selecting a specific unit at any given draw is always 1/N, where N is total number of sample units

- From N sample units, if sample of size n is drawn, there are N^n possible samples

- The probability of selecting one of them is $1/N^n$

- Example: You have 5 fruits and want to take a sample of 2. You can choose a fruit more than once.

SRS with replacement



The unit selected to be in the sample is returned to the population.

SRS with replacement - example

- Generally used in simulation studies

- A car insurance company wants to set the premium amount. The probabilities for major and minor accident are known to be 0.02% and 2% respectively
(assume that an insuree can meet with an accident only once)

- A SRS with replacement is conducted to simulate 1 million such scenarios

Simple random sample - python function

Python function	Description
random.choices()	draw sample with replacement
random.sample()	draw sample without replacement

Sampling variability

- Let us draw four different samples from the data
- The sample mean varies from sample to sample
- This intuitive idea is called sampling variability
- The sample mean is close to the population mean

Population Mean
15.125

Sample	Sample Mean
14.7, 15.1, 15.0, 15.6, 14.9, 15.4, 15.8, 15.1	15.20
15.8, 15.5, 15.3, 15.4, 15.8, 14.8, 14.6, 14.8	15.250
15.7, 15.6, 14.8, 14.8, 14.7,14.9, 15.6, 15.1	15.150
14.8, 15.1, 15.7, 14.6, 15.6, 15.4, 15.6, 15.0	15.225

A machine is used for packing grains in a sack. The weight of every bag is recorded as follows:

15.6	15.7	14.6	14.7	15.8	15.0	15.1	14.9
14.8	15.5	15.4	15.3	14.8	15.1	14.8	14.9

Draw at least the two samples and obtain their means. Compare this with the population mean.

Summary

Sampling Method	Summary
Simple Random	Mostly used method. The sample can be selected with or without replacement.
Stratified	Sample point is selected from each subgroup in the population.
Systematic	Easy to implement for large population. The first sample point is selected randomly and remaining points are selected at fixed intervals.
Cluster	An entire subgroup in the population is selected as a sample point.

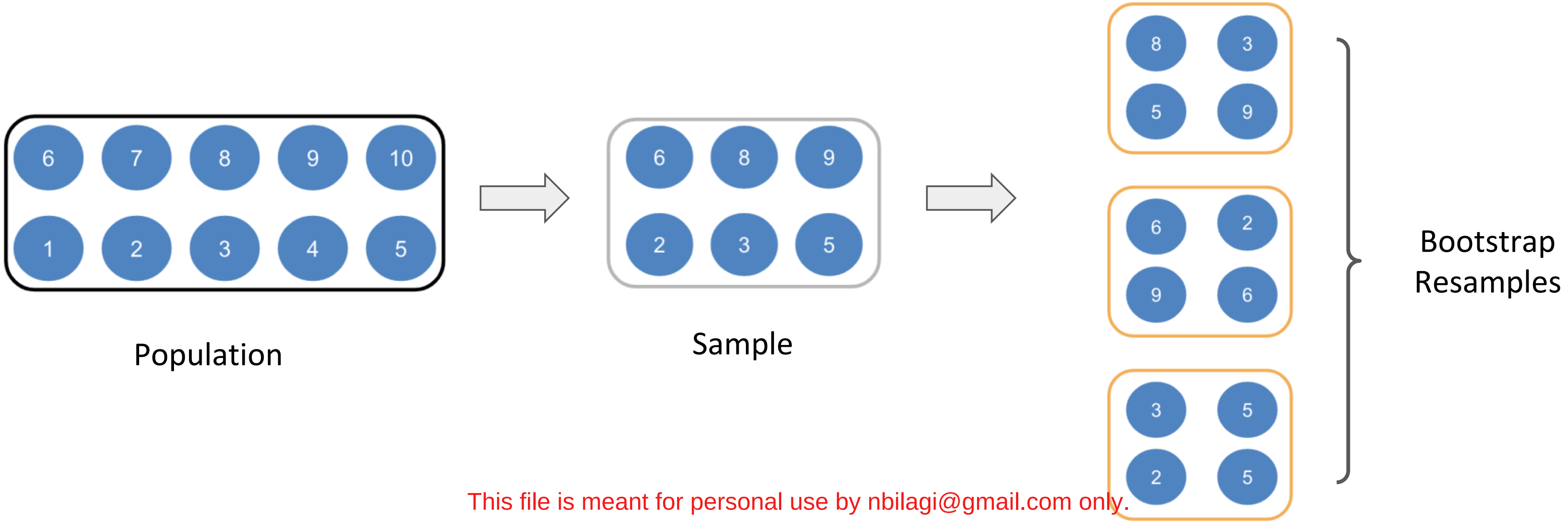
Resampling

Resampling

- Resampling is a technique in which we draw a sample from the available sample
- The resampling techniques are:
  - Bootstrap resampling
  - Jackknife resampling

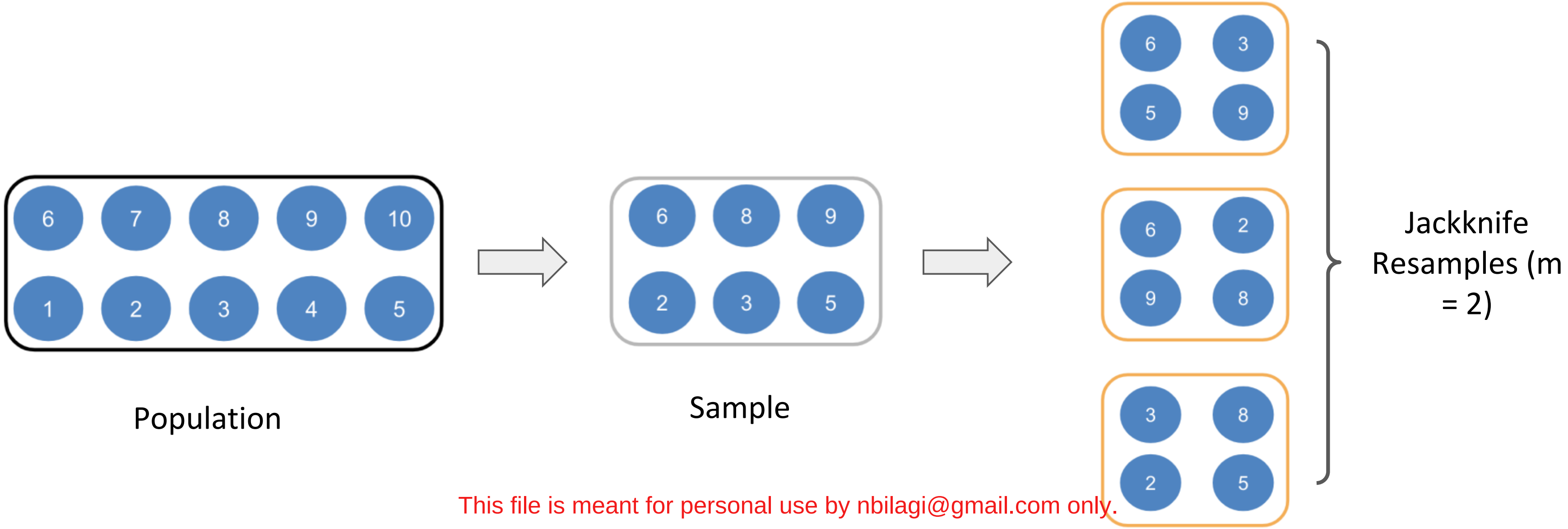
Bootstrap resample

Bootstrap resample is a simple random sample with replacement drawn from a sample



Jackknife resample

Jackknife resample is a method in which one deletes m observations from the original sample



Population

Sample

Jackknife Resamples (m = 2)

Sampling Distributions

Parameter and Statistic

- Parameter:
  - The population quantities are termed as a parameter
  - In inferential statistics, sample is used to estimate the population parameters
  - Example: Population mean

- Statistic:
  - Statistical measure computed from sample observations
  - Also known as an estimator
  - Example: Sample mean

Parameter and Statistic

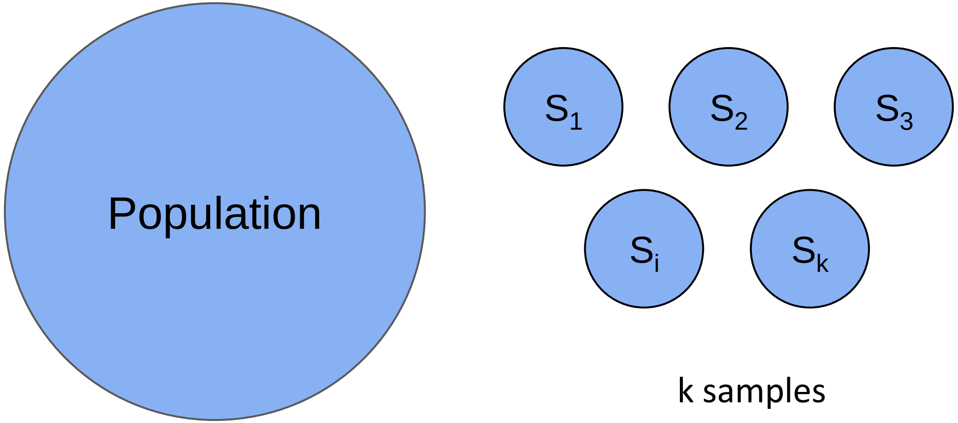
The statistical notations for population parameters and sample statistics

	Population Parameter	Sample Statistic
Mean	μ	x̄
Variance	σ^2	s^2
Standard Deviation	σ	s
Proportion	p	$\bar{p}$

Sampling distribution

- A total of ${}^{N}C_{n}$ (= k say) samples can be drawn from a population of size N and sample size n

- Obtain the mean and variance of each of these samples



k samples

Sampling distribution of sample mean

- The set of sample means $\bar{x}_1$, $\bar{x}_2$,..., $\bar{x}_k$ are the observations for sampling distribution of sample mean

- These sample means follow the normal distribution

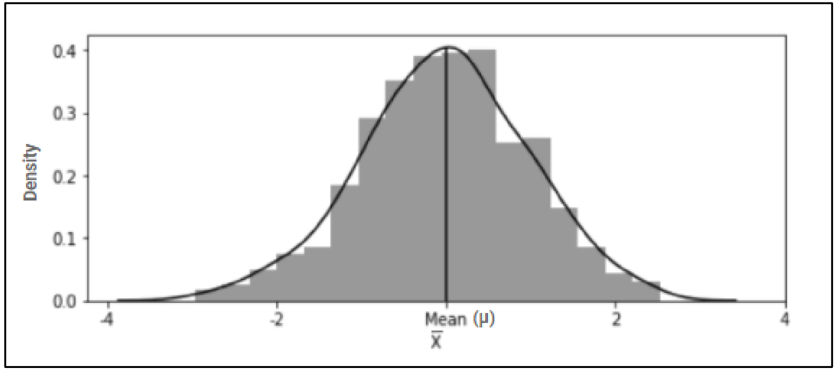
- Also, the mean of sample means is the population mean $\mu_{\bar{x}} = \mu$

Sample No	Sample Mean
1	$\bar{x}_1$
2	$\bar{x}_2$
3	$\bar{x}_3$
⋮	⋮
k	$\bar{x}_k$

Sampling distribution of sample mean

- The mean of sampling distribution of sample mean is $\mu_{\bar{x}}$

- The standard deviation of sampling distribution is σ/√n



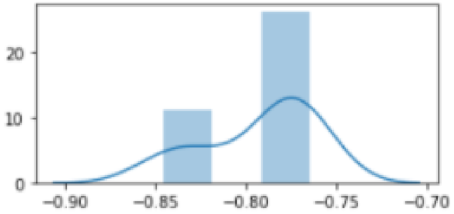
Central limit theorem

The central limit states that for random samples of large size n (usually n > 30) taken from a single population with mean μ and standard deviation σ, the sampling distribution of mean follows a normal distribution with mean μ and standard deviation σ/√n.

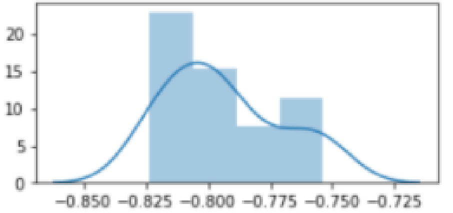
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N\,(0, 1)$$

Central limit theorem

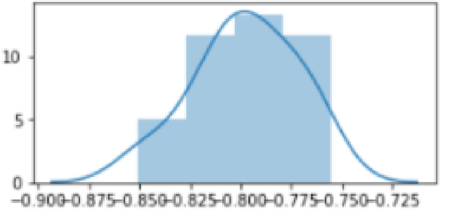
nSamples: 10



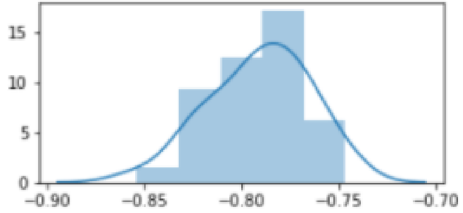
nSamples: 15



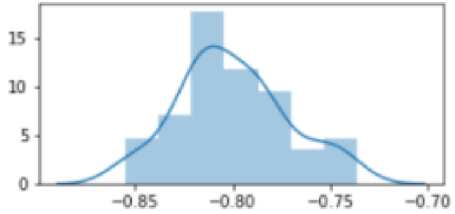
nSamples: 25



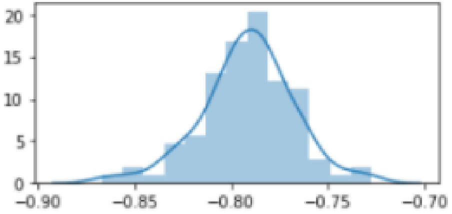
nSamples: 30



nSamples: 50



nSamples: 100



Proportion

- Computed for a categorical variable
- Proportion (p) is the ratio of the number of observations of the desired category to the total number of observations
- It is given by

$$p = \frac{x}{n}$$

where x is the number of observations of the desired category and n is the total number of observations

Sample proportion

- The sample proportion is an estimate of the population proportion

- Like the sampling distribution of mean is obtained from samples, the sampling distribution for proportions can also be obtained

- The set of sample proportions are the observations for the sampling distribution of the sample proportion

Sample No	Sample Proportion
1	$\bar{p}_1$
2	$\bar{p}_2$
3	$\bar{p}_3$
⋮	⋮
k	$\bar{p}_k$

Sampling distribution of sample proportion

- For a SRS from a large population, X follows Binomial distribution representing the number of observations of desired category in the sample

- Let the sample size be n, which is a constant

- For Y = X/n, P(Y = y) = P(X/n = x/n) is same as P(X=x). i.e. P(x/n) = P(x)

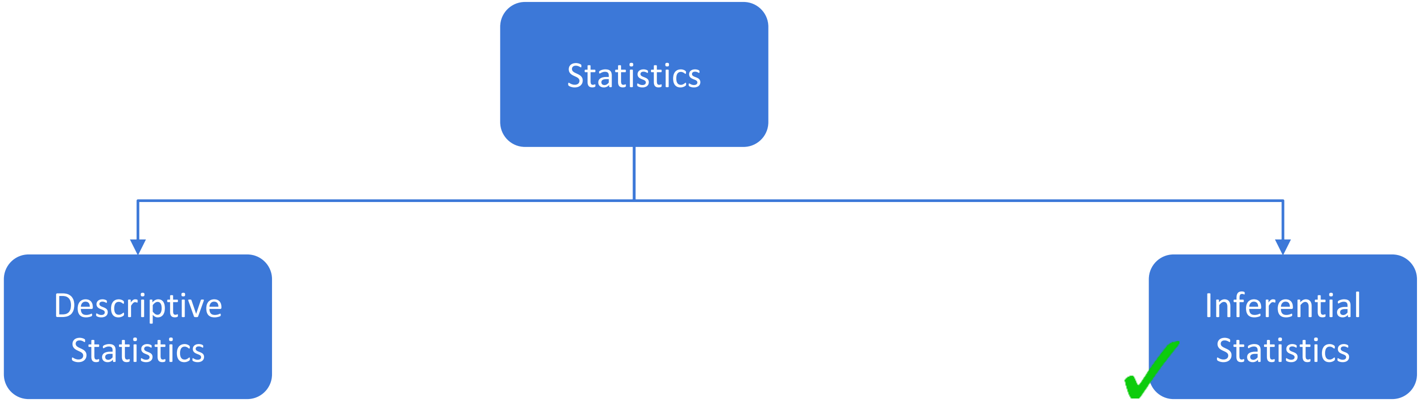
- We have E($\overline{p}$)= p
  Where $\overline{p}$ is the sample proportion and p is the population proportion

- For np ≥ 5 and n(1 - p) ≥ 5, follows normal distribution

Summary

- The parameter describes the population and the statistic describes the sample
- There are different notations for population parameter and sample statistic
- Sampling distribution is the distribution associated with sample statistic
- Sampling distribution of the sample mean follows normal distribution with mean $\mu_{\bar{x}}$ and standard deviation $\sigma/\sqrt{n}$
- Proportion is calculated for the categorical variable

Statistics



Statistics

Descriptive Statistics

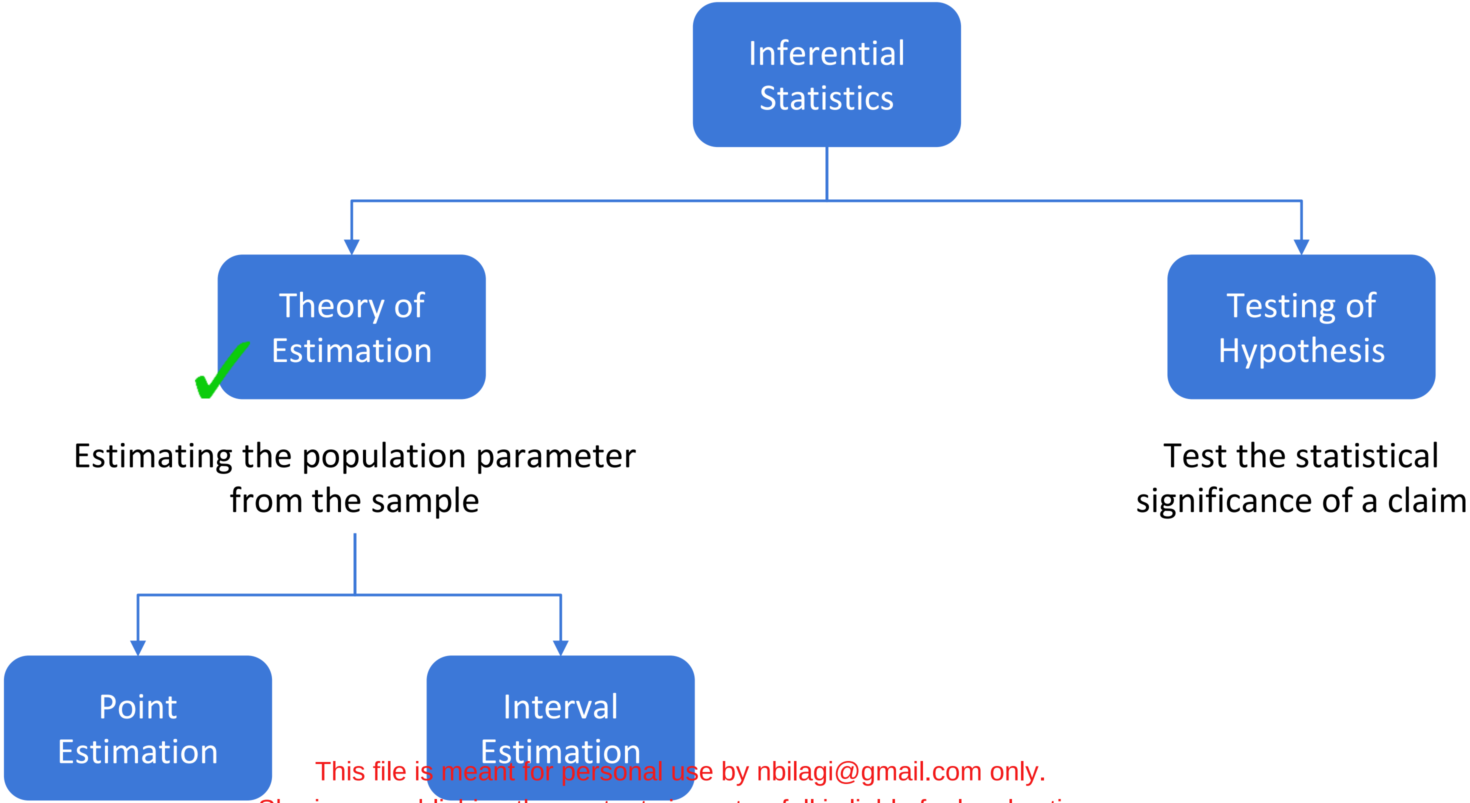
✓ Inferential Statistics

It is used to only describe the sample or summarize information about the sample

It is used to make inferences and generalizations about the broader population from the sample

Inferential Statistics

Inferential statistics



Inferential Statistics

Theory of Estimation ✓

Estimating the population parameter from the sample

Testing of Hypothesis

Test the statistical significance of a claim

Point Estimation

Interval Estimation

Point Estimation

Estimate

Question:

What would be the average income of a 27-year-old in London? How would you calculate it?

Estimate

Solution:

The trivial way is to collect the income data of every 27-year-old, further compute its mean.

This mean would be the population mean.

Point estimation

Is this scenario feasible? Is it possible to collect data of every 27-year-old person?

- To do so, draw a sample of size n and compute the mean.
- This sample mean is the point estimate of the population mean

Point estimation

- The sample estimate is considered to be the point estimate of the population parameter

- If population parameters are not known they are estimated from the sample

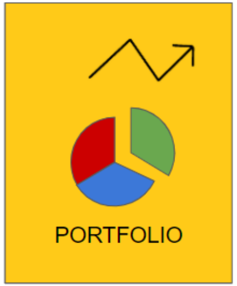
- The point estimate of
  - population mean ( ) is sample mean (μ)

    $\bar{X}$
  - population variance (σ) is sample variance s^2
  - Population proportion (P) is sample proportion (p)

Point Estimate

Question:

A financial firm has created 50 portfolios. From them a sample of 13 portfolios was selected, out of which 8 were found to be underperforming. Estimate the number of underforming portfolios?



Point Estimate

Solution:

A financial firm has created 50 portfolios. Thus N = 50

From them a sample of 13 portfolios was selected. Thus n = 13

The desired category is of underperforming portfolios. 8 of 13 were found to be underperforming. Thus x = 8

To find: Estimate the number of underperforming portfolios

Point Estimate

Solution:

The estimate of proportion of underperforming portfolios is given by

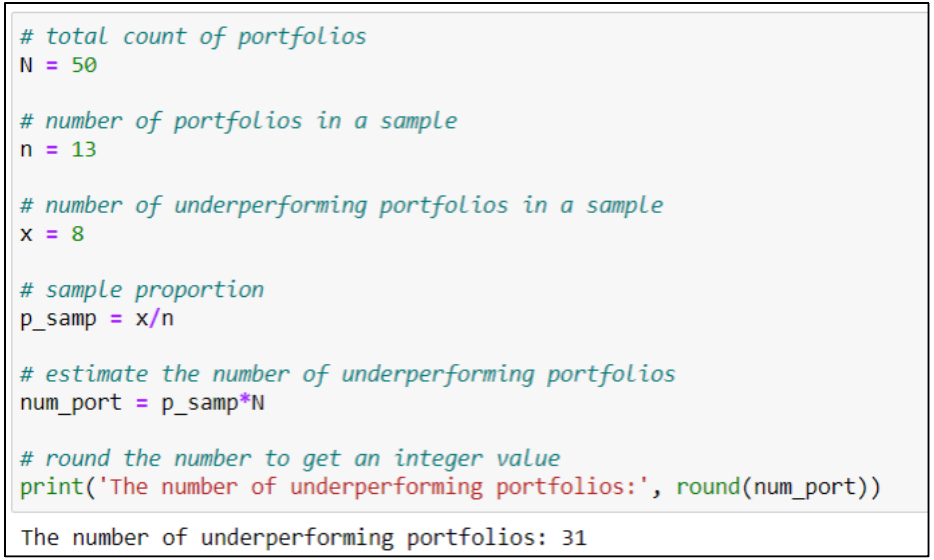
$$\hat{p} = \frac{x}{n} = \frac{8}{13}$$

To estimate the number of underperforming portfolios multiply by the estimate by N.

$$\text{The number of estimated underperforming samples} = \hat{p}\,.\,N = \frac{8}{13}\,.\,50 = 30.76 \approx 31$$

Point Estimate

Python solution:



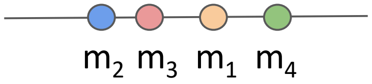
```
# total count of portfolios
N = 50

# number of portfolios in a sample
n = 13

# number of underperforming portfolios in a sample
x = 8

# sample proportion
p_samp = x/n

# estimate the number of underperforming portfolios
num_port = p_samp*N

# round the number to get an integer value
print('The number of underperforming portfolios:', round(num_port))
```

```
The number of underperforming portfolios: 31
```

Demerits of point estimate

- In most of the scenarios, the point estimate is not accurate
- It is the single value representation of the population that differs with each sample
- The deviation of the point estimate from the population parameter gives rise to sampling error

Sample means of different samples



Sampling error

- Sampling error is the absolute difference between the population parameter and its sample statistic (point estimate)

- Since the entire population is not considered as the sample, the values of mean, median, quantiles, and so on calculated on sample differ from the actual population values

- In order to reduce sampling error increase the sample size

Sample size

Question:

The average temperature on a summer morning in Alaska is 55°F. The research student of climate studies records the temperature for 15 summer mornings in °F.

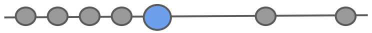
Data = [49.8, 52.3, 51, 56.9, 54.8, 63, 58.2, 54.1, 50.4, 49.2, 47, 51.3, 43.5, 56, 55]

Find the sampling error for the average temperature.

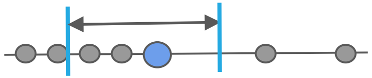
Interval Estimation

Need of interval estimation

- The point estimate is taken from a particular sample
- For a different sample, we get a different estimate (sampling variability)
- This problem is resolved if we obtained the interval estimate



Point estimate



Interval estimate

Interval Estimation

- Interval estimate is a range or an interval within which the parameter lies for a stated confidence level

- Based on the central limit theorem

- It is known as the confidence interval

Confidence level

- Confidence level (1-α) is the percentage of all possible point estimates that can be expected to include the actual population parameter

- α is known as the level of significance

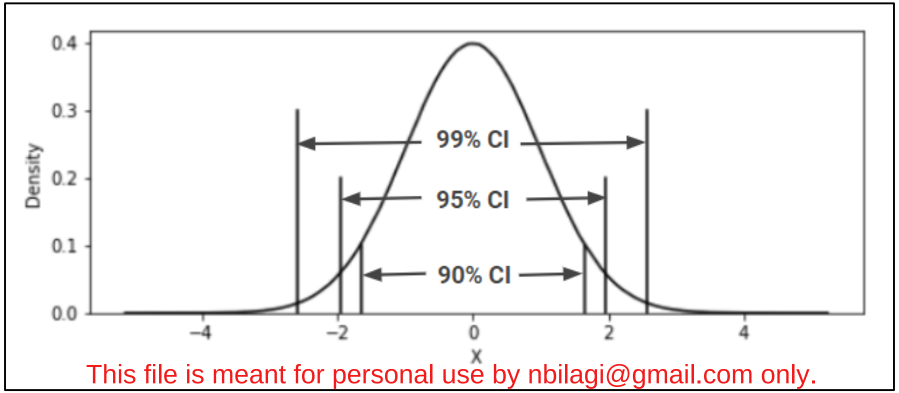
- The higher confidence level illustrates the wider confidence interval

- A 95% confidence level implies that 95% of the confidence intervals would include the actual population parameter

Confidence level

Example:

The average income estimate for a 27-year-old in London is £45,000. After taking repeated samples, the 99% confidence intervals of the average income include the actual population parameter for 99% of the times.



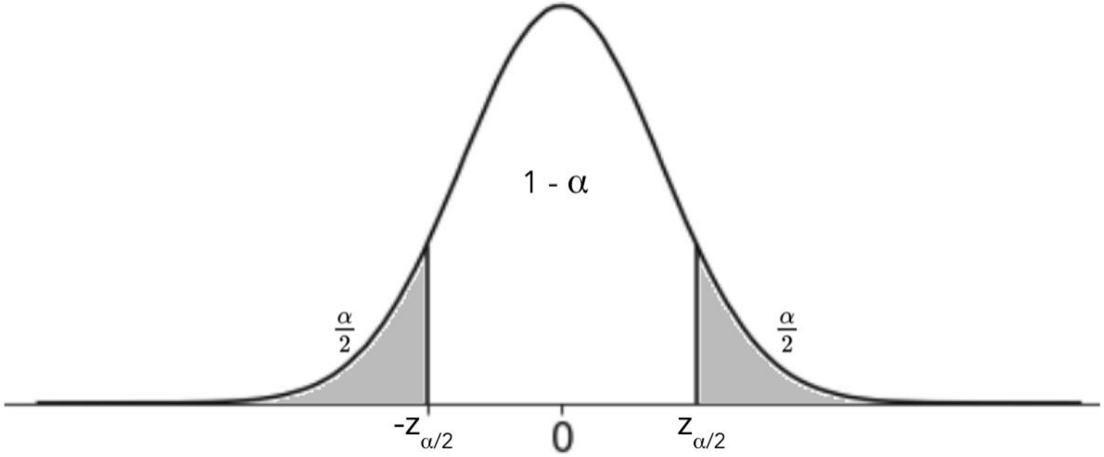
Recall - CLT

For random samples of large size n taken from a single population with mean μ and standard deviation σ, the sampling distribution of mean follows a normal distribution with mean μ and standard deviation σ/√n

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

Interval estimate

From the CLT, $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$

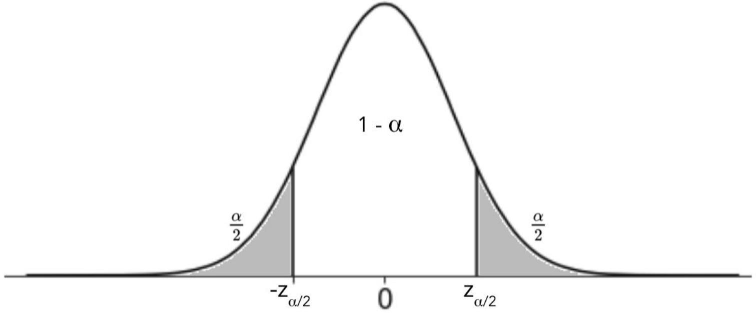


Where α is is the level of significance

Interval estimate

From the CLT, $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$

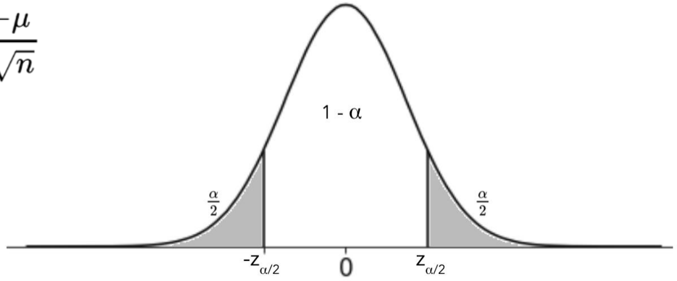
Value of α	$Z_{\alpha/2}$	Value of $Z_{\alpha/2}$
0.10	$z_{0.05}$	1.645
0.05	$z_{0.025}$	1.96
0.01	$z_{0.005}$	2.575



Note: To obtain these values in python use scipy.stats.norm.isf()

Interval estimate

From the CLT, P(-$z_{\alpha/2}$ ≤ Z ≤ $z_{\alpha/2}$) = 1 - α and value of Z is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$



Thus, P(-$z_{\alpha/2}$ ≤ $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ ≤ $z_{\alpha/2}$) = 1 - α

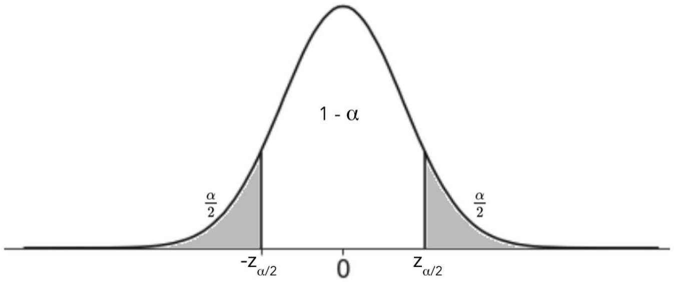
Rearranging the terms, P($\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$) = 1 - α

This defines a 100(1 - α)% confidence interval as $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Interval estimate

The 100(1 - α)% confidence interval is

$$\underbrace{\bar{X}}_{\text{Point Estimate}} \pm \underbrace{z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}}_{\text{Margin of error}}$$



Margin of error

Question:

100 bags of coal were tested and had an average of 35% of ash with a standard deviation of 15%. Calculate the margin of error for a 90% confidence level.

Margin of error

Solution:

100 bags of coal were tested and had an average of 35% of ash with a standard deviation of 15%.

Here $n = 100$

$\sigma = 0.15$

To find: The margin of error for a 90% confidence level.

Here $\alpha = 0.10$. Thus $z_{\alpha/2} = 1.64$

Margin of error

Solution:

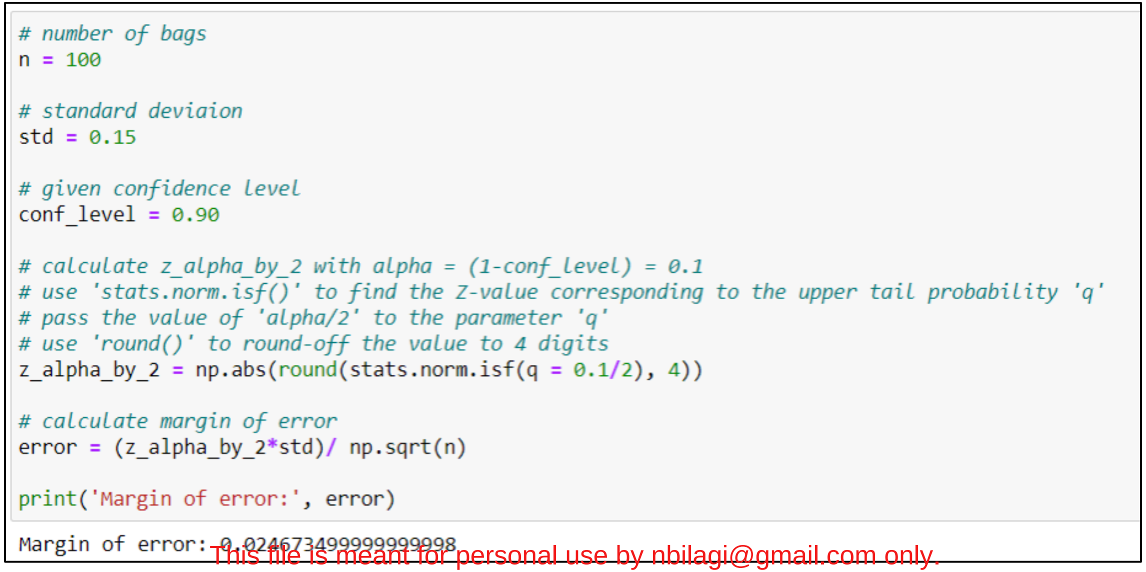
The margin of error for a 90% confidence level is

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.64\frac{0.15}{\sqrt{100}} = 0.0246$$

A 90% CI with 2.46% margin of error implies that the sample mean will be within 2.46% points of the real population value 90% of the time

Margin of error

Python solution:



```
# number of bags
n = 100

# standard deviaion
std = 0.15

# given confidence level
conf_level = 0.90

# calculate z_alpha_by_2 with alpha = (1-conf_level) = 0.1
# use 'stats.norm.isf()' to find the Z-value corresponding to the upper tail probability 'q'
# pass the value of 'alpha/2' to the parameter 'q'
# use 'round()' to round-off the value to 4 digits
z_alpha_by_2 = np.abs(round(stats.norm.isf(q = 0.1/2), 4))

# calculate margin of error
error = (z_alpha_by_2*std)/ np.sqrt(n)

print('Margin of error:', error)
```

```
Margin of error: 0.024673499999999998
```

- As sample size increases the margin of error decreases

- If margin of error is fixed then the sample size can be obtained

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow n = z^2_{\alpha/2}\left(\frac{\sigma^2}{E^2}\right)$$

Sample size

Question:

For the previous question, for a margin of error 2.46. Verify whether you get sample size of 100.

Previous Question: 100 bags of coal were tested and had an average of 35% of ash with a standard deviation of 15%. Calculate the margin of error for a 90% confidence level.

Interval estimate

Question:

From a sample of 250 observations it is found that average income of a 27-year-old Londoner is £45,000 with a sample standard deviation of £4000. Obtain the 95% confidence interval to estimate the average income.

Interval estimate

Solution:

A sample of 250 observations of a 27-year-old Londoner has an average income of £45,000 with a sample standard deviation of £4000

Thus, $\bar{X}$= 45,000, σ = 4000 and n = 250

To find: The 95% confidence interval.

Here α = 0.05. So $z_{0.025}$ = 1.96

Interval estimate

Solution:

The 95% confidence interval is

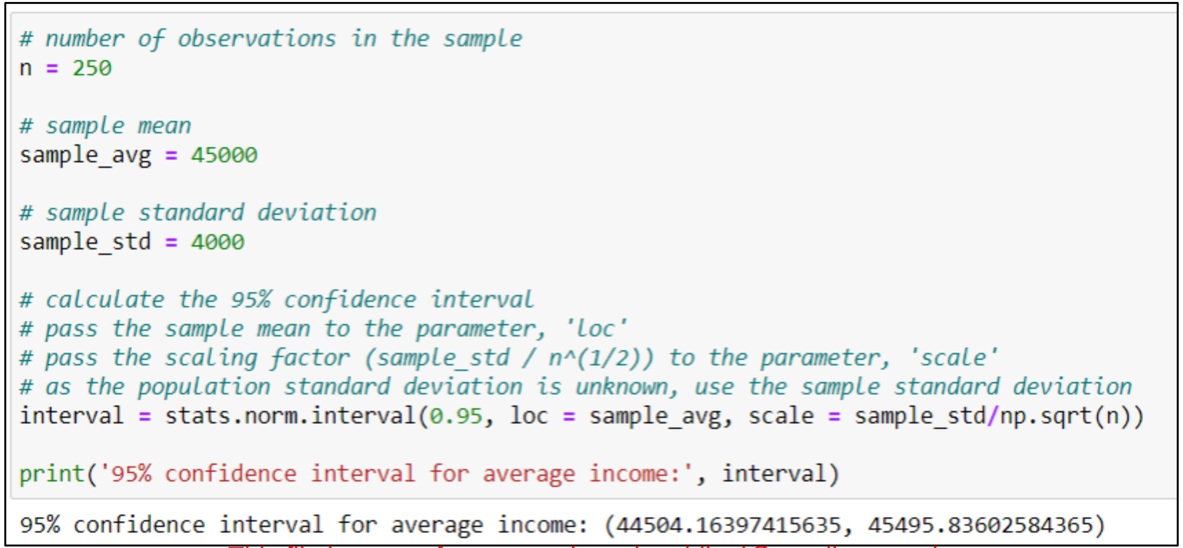
$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\Rightarrow 45000 \pm 1.96 \frac{4000}{\sqrt{250}}$$

$$\Rightarrow (44504.155, 45495.845)$$

Interval estimate

Python solution:



```
# number of observations in the sample
n = 250

# sample mean
sample_avg = 45000

# sample standard deviation
sample_std = 4000

# calculate the 95% confidence interval
# pass the sample mean to the parameter, 'loc'
# pass the scaling factor (sample_std / n^(1/2)) to the parameter, 'scale'
# as the population standard deviation is unknown, use the sample standard deviation
interval = stats.norm.interval(0.95, loc = sample_avg, scale = sample_std/np.sqrt(n))

print('95% confidence interval for average income:', interval)
```

```
95% confidence interval for average income: (44504.16397415635, 45495.83602584365)
```

- It cannot be said that P(44504.155≤ μ ≤ 45495.845) = 0.95

- The population parameter μ is a constant

- It can be said that from many samples drawn of size 250, construct a 95% CI for each sample, roughly 95% of the intervals would actually contain the population parameter μ do it YOURSELF

Confidence Interval

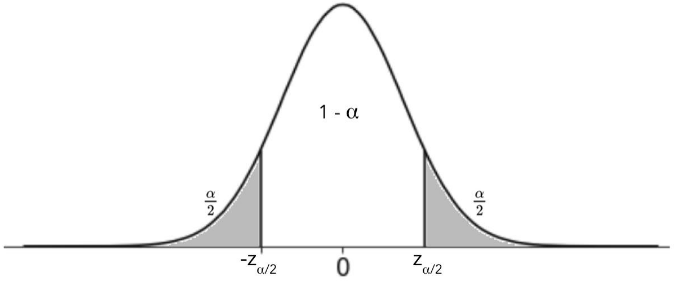
Question:

For the previous question, obtain the 90% and 99% confidence intervals. Further, logically explain which one of them is the widest.

Interval estimate

The 100(1 - α)% confidence interval for proportion is

$$\underbrace{\hat{p}}_{\text{Point Estimate}} \pm \underbrace{z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{\text{Margin of error}}$$



Interval estimate

Question:

A financial firm has created 50 portfolios. From them a sample of 13 portfolios was selected, out of which 8 were found to be underperforming. Construct a 99% confidence interval to estimate the population proportion.

Interval estimate

Solution:

From them a sample of 13 portfolios was selected. Thus n = 13

The desired category is of underperforming portfolios. 8 of 13 were found to be underperforming. Thus x = 8

Thus the estimate of sample proportion is x/n = 8/13 =0.615

To find: Construct a 99% confidence interval to estimate the population proportion.

Interval estimate

Solution:

Here α = 0.01. So $z_{0.005}$ = 2.575

The 95% confidence interval is

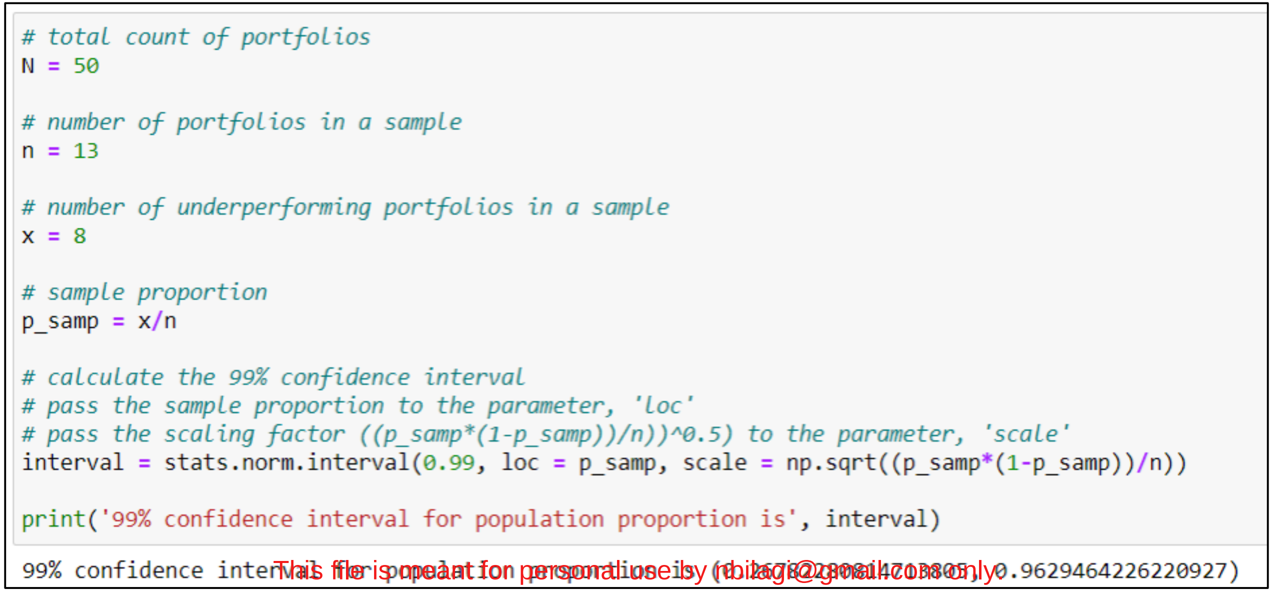
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\Rightarrow 0.615 \pm 2.575\sqrt{\frac{0.615(1-0.615)}{13}}$$

$$\Rightarrow (0.267, 0.963)$$

Interval estimate

Python solution:



```
# total count of portfolios
N = 50

# number of portfolios in a sample
n = 13

# number of underperforming portfolios in a sample
x = 8

# sample proportion
p_samp = x/n

# calculate the 99% confidence interval
# pass the sample proportion to the parameter, 'loc'
# pass the scaling factor ((p_samp*(1-p_samp))/n))^0.5) to the parameter, 'scale'
interval = stats.norm.interval(0.99, loc = p_samp, scale = np.sqrt((p_samp*(1-p_samp))/n))

print('99% confidence interval for population proportion is', interval)
```

```
99% confidence interval for population proportion is (0.26782280814713805, 0.9629464226220927)
```

Summary

- Sample estimate of a population parameter is considered as a point estimate
- Sample mean, median, standard deviation are the examples of the point estimate
- Point estimate varies significantly for different samples
- Confidence interval returns the interval estimate for the population parameter
- The confidence level describes the uncertainty of the estimate
- Confidence interval is calculated using the point estimate and the margin of error
- The optimal sample size can be obtained by deciding the required margin of error

Thank You

Appendix

Relation of expectation and variance (A.1)

$$V(X) = \sum (x - \mu)^2 . f(x)$$

$$= \sum (x^2 - 2\mu.x + \mu^2). f(x)$$

$$= \sum (x^2 - 2\mu.x + \mu^2). f(x)$$

$$= \sum x^2 . f(x) - \sum 2\mu. x . f(x) + \sum \mu^2. f(x)$$

$$= \sum x^2 . f(x) - 2\mu \sum x . f(x) + \sum \mu^2. f(x)$$

$$= E(X^2) - 2 E(X) . E(X) + [E(X)]^2$$

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