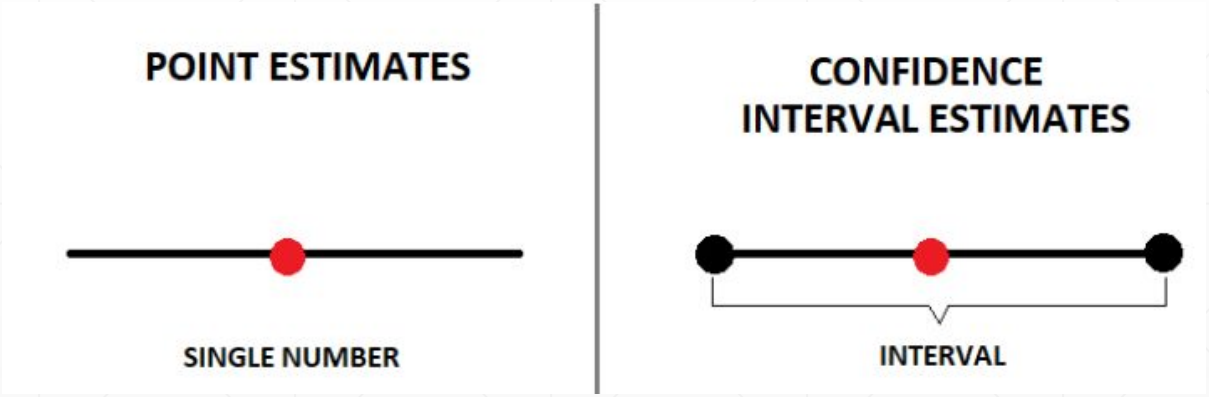
Inferential Statistics

Estimation & Hypothesis Testing

Estimation

- Estimation is the process used to make inferences, from a sample, about an unknown population parameter.
  - Height of students in the class.
  - Pass percentage of past five years
- There are two types of estimates:
  - **Point Estimates**— When the estimate is a single number, the estimate is called a "point estimate“
  - **Confidence Interval Estimates** — when the estimate is a range of scores, the estimate is called an interval estimate.



Point Estimation

- Point estimators are defined as the functions that are used to find an approximate value of a population parameter from random samples of the population.
- The sample mean (**$\bar{x}$**) is a point estimation of the population's mean (**μ**). The same goes for the sample variance, which is an estimate of the population's variance.
- All estimators have two properties, **efficiency** and **bias**:
  - **Bias**— an unbiased estimator has an expected value equal to the population parameter.
  - **Efficiency** — the most efficient estimators are the ones with the least variability of outcomes. The most efficient estimator is the unbiased estimator with the smallest variance.

For example

- 55 is the mean mark obtained by a sample of 5 students randomly drawn from a class of 100 students is considered to be the mean marks of the entire class. This single value 55is a point estimate.
- 50 kg is the average weight of a sample of 10 students randomly drawn from a class of 100 students is considered to be the average weight of the entire class. This single value 50 is a point estimate.

Confidence Interval Estimates

- **Point estimators are not very reliable!** A confidence interval is a much more accurate representation of reality.
- Imagine you decide to randomly measure 40 men in your city, and you get a sample average height of $\bar{x}$ =175 cm. You might get close to the population's real height (**μ)**, but the chances are that the true value is somewhere between 170 cm and 180 cm. It is most accurate to say that the average height for men in your city is somewhere between a specific interval [170 cm, 180 cm].

I.I Sampling distribution of the sample mean

Traditional Approach

1. Draw s (a large number) independent samples $\{\mathbf{x}^1, \ldots, \mathbf{x}^s\}$ from the population where each element $\mathbf{x}^j$ is composed of $\{x_i^j\}_{i=1,\ldots,n}$.
2. Evaluate the sample mean $\hat{\mu}^j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$ of each sample.
3. Estimate the sampling distribution of $\hat{\mu}$ by the empirical distribution of the sample replications.

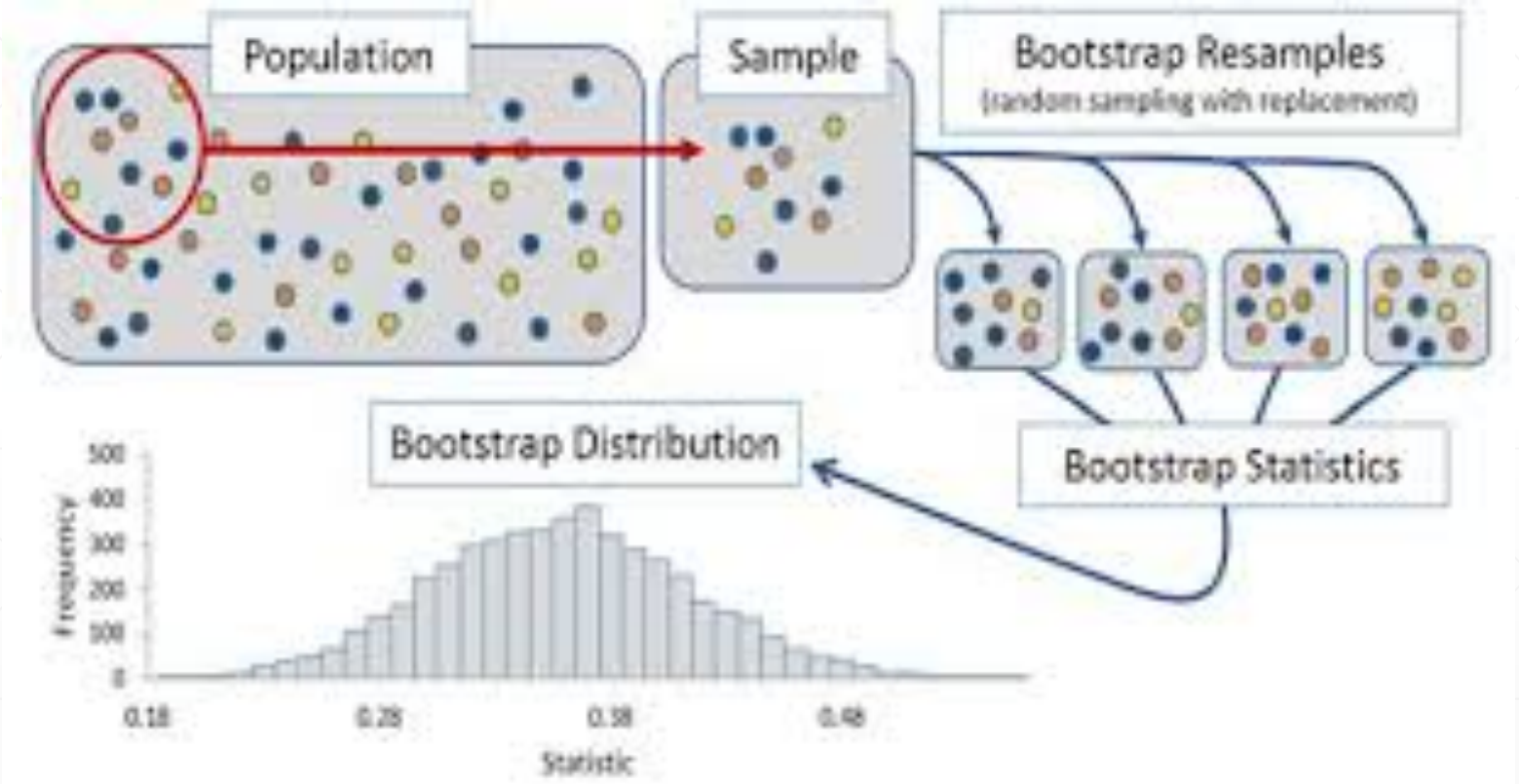
The standard deviation of the sample mean $\sigma_{\bar{x}}$, or *standard error*, can be approximated by this formula:

$$SE = \frac{\sigma_x}{\sqrt{n}}$$

S.E mean = sample standard deviation (s) / √sample size

I.2 Bootstrapping method

Computationally Intensive Approach



Central Limit Theorem

- *Given a population with a finite mean μ and a finite non-zero variance σ2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ2/n as n, the sample size, increases.*

$$CI = [\Theta - 1.96 \times SE, \Theta + 1.96 \times SE]$$

https://www.menti.com/uu23nvdt8x

Problem 1

Let X be the height of a randomly chosen individual from a population. In order to estimate the mean and variance of X, we observe a random sample $X_1, X_2, \cdots, X_7$. Thus, X_i's are i.i.d. and have the same distribution as X. We obtain the following values (in centimeters):

$$166.8, 171.4, 169.1, 178.5, 168.0, 157.9, 170.1$$

Find the values of the sample mean, the sample variance, and the sample standard deviation for the observed sample.

Solution

$$\begin{aligned}\overline{X} &= \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7}{7} \\ &= \frac{166.8 + 171.4 + 169.1 + 178.5 + 168.0 + 157.9 + 170.1}{7} \\ &= 168.8\end{aligned}$$

The sample variance is given by

$$S^2 = \frac{1}{7-1}\sum_{k=1}^{7}(X_k - 168.8)^2 = 37.7$$

Finally, the sample standard deviation is given by

$$= \sqrt{S^2} = 6.1$$

In a certain property investment company with an international presence, workers have a mean hourly wage of $12 with a population standard deviation of $3. Given a sample size of 30, estimate and interpret the SE of the sample mean:

$$\begin{aligned}\sigma_x &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{3}{\sqrt{30}} \\ &= \$0.55\end{aligned}$$

A sample of 30 latest returns on XYZ stock reveals a mean return of $4 with a sample standard deviation of $0.13. Estimate the SE of the sample mean.

$$\begin{aligned}S_x &= \frac{S}{\sqrt{n}} \\ &= \frac{0.13}{\sqrt{30}} \\ &= \$0.02\end{aligned}$$

A confidence interval (CI) is an interval of numbers believed to contain the parameter value.

Formula

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

CI = confidence interval

$\bar{x}$ = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

Critical Values Z_α	Level of significance (α)			
	1%	2%	5%	10%
Two-tailed test	$\|Z_\alpha\|=2.58$	$\|Z_\alpha\|=2.33$	$\|Z_\alpha\|=1.96$	$\|Z_\alpha\|=1.645$
Right tailed test	$Z_\alpha=2.33$	$Z_\alpha=2.055$	$Z_\alpha=1.645$	$Z_\alpha=1.28$
Left tailed test	$Z_\alpha=-2.33$	$Z_\alpha=-2.055$	$Z_\alpha=-1.645$	$Z_\alpha=-1.28$

In a given sample of 97 females having the SD of 64.9 and mean cholesterol of 261.75 .Calculate the confidence interval of the mean cholesterol level of the female population with 95% confidence level provided

Critical Values Z_α	Level of significance (α)			
	1%	2%	5%	10%
Two-tailed test	$\|Z_\alpha\|=2.58$	$\|Z_\alpha\|=2.33$	$\|Z_\alpha\|=1.96$	$\|Z_\alpha\|=1.645$
Right tailed test	$Z_\alpha=2.33$	$Z_\alpha=2.055$	$Z_\alpha=1.645$	$Z_\alpha=1.28$
Left tailed test	$Z_\alpha=-2.33$	$Z_\alpha=-2.055$	$Z_\alpha=-1.645$	$Z_\alpha=-1.28$

The CI came out to be 248.83 and 274.67.

That means the true mean of the cholesterol of the female population will fall between 248.83 and 274.67

A machine produces a component of a product with a standard deviation of 1.6 cm in length. A random sample of 64 componentsvwas selected from the output and this sample has a mean length of 90 cm. The customer will reject the part if it is either less than 88 cm or more than 92 cm. Does the 95% confidence interval for the true mean length of all the components produced ensure acceptance by the customer?

A machine produces a component of a product with a standard deviation of 1.6 cm in length. A random sample of 64 componentsvwas selected from the output and this sample has a mean length of 90 cm. The customer will reject the part if it is either less than 88 cm or more than 92 cm. Does the 95% confidence interval for the true mean length of all the components produced ensure acceptance by the customer?

Solution:

Here μ is the mean length of the components in the population.

The formula for the confidence interval is

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Here $\sigma = 1.6$, $Z_{\alpha/2} = 1.96$, $\bar{x} = 90$ and $n = 64$

Then $S.E. = \frac{\sigma}{\sqrt{n}} = \frac{1.6}{\sqrt{64}} = 0.2$

Critical Values Z_α	Level of significance (α)			
	1%	2%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert$=2.58	$\lvert Z_\alpha \rvert$=2.33	$\lvert Z_\alpha \rvert$=1.96	$\lvert Z_\alpha \rvert$=1.645
Right tailed test	Z_α=2.33	Z_α=2.055	Z_α=1.645	Z_α=1.28
Left tailed test	Z_α = -2.33	Z_α = –2.055	Z_α = –1.645	Z_α = –1.28

Therefore, $90 - (1.96 \times 0.2) \leq \mu \leq 90 + (1.96 \times 0.2)$

$(89.61 \leq \mu \leq 90.39)$

This implies that the probability that the true value of the population mean length of the components will fall in this interval (89.61,90.39) at 95% . Hence we concluded that 95% confidence interval ensures acceptance of the component by the consumer.

A sample of 100 measurements at breaking strength of cotton thread gave a mean of 7.4 and a standard deviation of 1.2 gms. Find 95% confidence limits for the mean breaking strength of cotton thread.

A sample of 100 measurements at breaking strength of cotton thread gave a mean of 7.4 and a standard deviation of 1.2 gms. Find 95% confidence limits for the mean breaking strength of cotton thread.

A sample of 100 measurements at breaking strength of cotton thread gave a mean of 7.4 and a standard deviation of 1.2 gms. Find 95% confidence limits for the mean breaking strength of cotton thread.

Solution:

Given, sample size = 100, $\bar{x} = 7.4$, since σ is unknown but $s = 1.2$ is known.

In this problem, we consider $\check{\sigma} = s$, $Z_{\alpha/2} = 1.96$

$$S.E. = \frac{\check{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{1.2}{\sqrt{100}} = 0.12$$

Hence 95% confidence limits for the population mean are

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$7.4 - (1.96 \times 0.12) \leq \mu \leq 7.4 + (1.96 \times 0.12)$$

$$7.4 - 0.2352 \leq \mu \leq 7.4 + 0.2352$$

$$7.165 \leq \mu \leq 7.635$$

Critical Values Z_α	Level of significance (α)			
	1%	2%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 2.33$	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 1.645$
Right tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 2.055$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -2.055$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

This implies that the probability that the true value of the population mean breaking strength of the cotton threads will fall in this interval (7.165,7.635) at 95% .

The mean life time of a sample of 169 light bulbs manufactured by a company is found to be 1350 hours with a standard deviation of 100 hours. Establish 90% confidence limits within which the mean life time of light bulbs is expected to lie.

The mean life time of a sample of 169 light bulbs manufactured by a company is found to be 1350 hours with a standard deviation of 100 hours. Establish 90% confidence limits within which the mean life time of light bulbs is expected to lie.

The mean life time of a sample of 169 light bulbs manufactured by a company is found to be 1350 hours with a standard deviation of 100 hours. Establish 90% confidence limits within which the mean life time of light bulbs is expected to lie.

Solution:

Given: n = 169, $\bar{x}$ = 1350 hours, s = 100 hours, since the level of significance is (100-90)% =10% thus a is 0.1, hence the significant value at 10% is $Z_{a/2} = 1.645$

$$S.E. = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{169}} = 7.69$$

Hence 90% confidence limits for the population mean are

$$\bar{x} - Z_{\alpha/2} SE < \mu < \bar{x} + Z_{\alpha/2} SE$$

$$1350 - (1.645 \times 7.69) \leq \mu$$

$$1337.35 \leq \mu \leq 1362.65$$

Critical Values Z_α	Level of significance (α)			
	1%	2%	5%	10%
Two-tailed test	$\lvert Z_\alpha \rvert$=2.58	$\lvert Z_\alpha \rvert$=2.33	$\lvert Z_\alpha \rvert$=1.96	$\lvert Z_\alpha \rvert$=1.645
Right tailed test	Z_α=2.33	Z_α=2.055	Z_α=1.645	Z_α=1.28
Left tailed test	Z_α = -2.33	Z_α = −2.055	Z_α = −1.645	Z_α = −1.28

Hence the mean life time of light bulbs is expected to lie between the interval (1337.35, 1362.65)

A tree consists of hundreds of apples. 46 apples are randomly chosen. The mean and standard deviation of this instance is found to be 86 and 6.2 respectively. Determine whether the apples are big enough or not.

$$CI = \hat{X} \pm Z \times \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$CI = 86 \pm 1.960 \times \left(\frac{6.2}{\sqrt{46}}\right)$$

$$CI = 86 \pm 1.79$$

The margin error in this problem is 1.79.

All the hundreds of apples are therefore likely to be in the range of 86 + 1.79 and 86 - 1.79

i.e. in the range of 84.21 and 87.79

Module 3: Class Test

Find the mean, median, mode, range, standard deviation and variance of the following data:
(a) 9, 7, 11, 13, 2, 4, 5, 5

Calculate correlation coefficient for x = 100, 106, 112, 98, 87, 77, 67, 66, 49 and y = 28, 33, 26, 27, 24, 24, 21, 26, 22.

Calculate the Spearman correlation coefficient

History	Geography
35	30
23	33
47	45
17	23
10	8
43	49
9	12
6	4
28	31

The average heights of a random sample of 400 people from a city is 1.75 m. It is known that the heights of the population are random variables that follow a normal distribution with a variance of 0.16.

1. Determine the interval of 95% confidence for the average heights of the population.

Estimation Using Python

- import numpy as np
- import scipy.stats as stats

- population_ages1 = stats.poisson.rvs(loc=18,mu=35,size=150000)
- population_ages2 = stats.poisson.rvs(loc=18,mu=35,size=100000)

- population_ages=np.concatenate((population_ages1,population_ages2))
- print("The population mean",population_ages.mean())

- samples_ages=np.random.choice(a=population_ages,size=500)
- print("The Sample mean",samples_ages.mean())

```
#Sampling distribution of sample mean
point_estimates = []
for x in range(200):
    sample=np.random.choice(a=population_ages,size=500)
    point_estimates.append(sample.mean())
print(np.array(point_estimates))
len(point_estimates)
print("The sampling mean",np.array(point_estimates).mean())
print("The popuation mean",population_ages.mean())
```

```
#BootStrapping methods
m_avg=[]
def meanBootstrap(X, numberb):
    x = [0]*numberb
    for i in range(numberb):
        sample=[X[j]
                for j
                in np.random.randint(len(X),size=len(X))
                ]
        m_avg.append(np.mean(sample))
    return x
m = meanBootstrap(population_ages,200)
print("The Bootstraping mean",np.array(m_avg).mean())
print("Population mean",population_ages.mean())
```