BATCH : B 150
LESSON : STATISTICS-2
DATE : 22.06.2023
SUBJECT : Sample Distribution
Std. Error of Mean
Central Limit Theorem

techproeducation
techproeducation
techproeducation
techproeducation
techproedu

techproeducation.com info@techproeducation.com +1 (917) 768-7466

STATISTICS - 2

Data Science Program

Statistics Sessions -7

Interaction

Önceki derste anlamadığım konu yoktur

Session - 7 Content

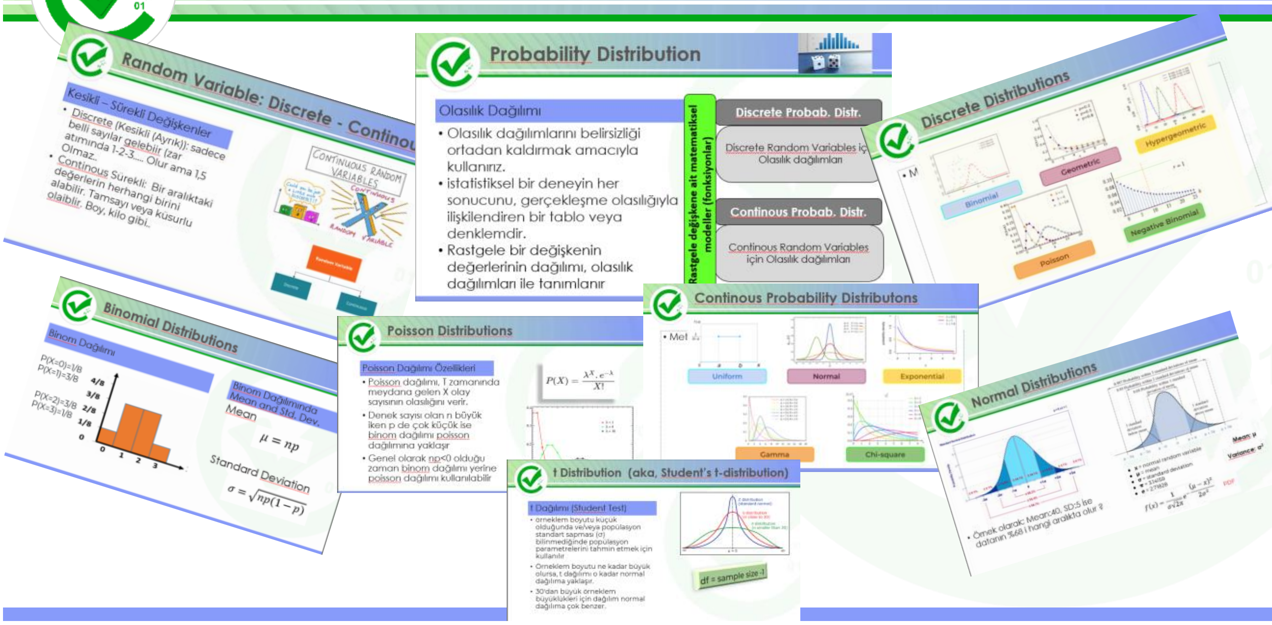
Content

- Sample Distribution
- Simple Random Sampling
- Standard Error of the Mean
- Central Limit Theorem
- Confidence Interval

RECAP

Herkes önceki dersten hatırladığı 1 cümle yazabilir mi?

Recap – Previous Lesson



SAMPLE DISTRIBUTIONS

Örneklem Dağılımları

Sample Distribution

- Hatırlayalım:
- İstatistik bir örneklemin ortalama, std. Sapma gibi değerlerini veriyordu – Popülasyon – Sample kavramları
- Descriptive Statistics – Inferential Statistics kavramları
- İstatistikler random variable lara bağlı olarak hesaplandığı için bu istatistiklerin kendisi de random variable dır.
- Istatistiklerin de kendi olasılıksal dağılımları vardır. Bunlara 'Sample Distribution' denir. Popülasyonun dağılımı değil de örneklemin dağılımı diye adlandırıyoruz.
- 100.000 nüfuslu bir üniversite öğrencisi olan popülasyondan 500 kişilik bir sample in kendine has bir distribution u vardır.

Sample Distribution

Örneklem Dağılımı

- Örneklem dağılımları, hipotez testi yapmak için gerekli bilgiyi sağlar.
- Standard Error Kavramı
- Random Statistics

kavramı

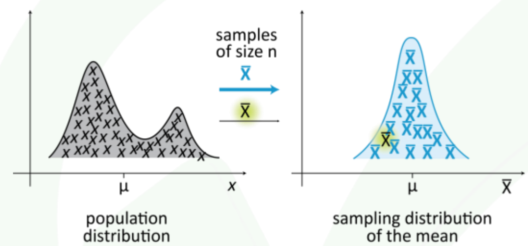
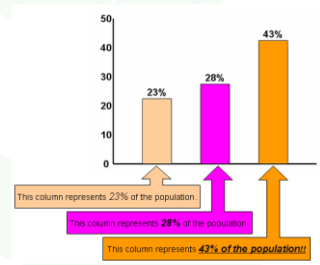


Table 5.1 Samples from the Weibull Distribution of Example 5.19

Sample	1	2	3	4	5	6
1	6.1171	5.07611	3.46710	1.55601	3.12372	8.93795
2	4.1600	6.79279	2.71938	4.56941	6.09685	3.92487
3	3.1950	4.43259	5.88129	4.79870	3.41181	8.76202
4	0.6694	8.55752	5.14915	2.49759	1.65409	7.05569
5	1.8552	6.82487	4.99635	2.33267	2.29512	2.30932
6	5.2316	7.39958	5.86887	4.01295	2.12583	5.94195
7	2.7609	2.14755	6.05918	9.08845	3.20938	6.74166
8	10.2185	8.50628	1.80119	3.25728	3.23209	1.75468
9	5.2438	5.49510	4.21994	3.70132	6.84426	4.91827
10	4.5590	4.04525	2.12934	5.50134	4.20694	7.26081
$\bar{x}$	4.401	5.928	4.229	4.132	3.620	5.761
$\tilde{x}$	4.[illegible]60	6.144	4.608	3.857	3.221	6.342
s	2.642	2.062	1.611	2.124	1.678	2.496

Note that the sample means, medians, and standard deviations are all different – randomness!

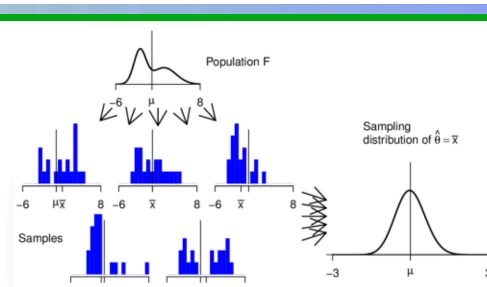
7

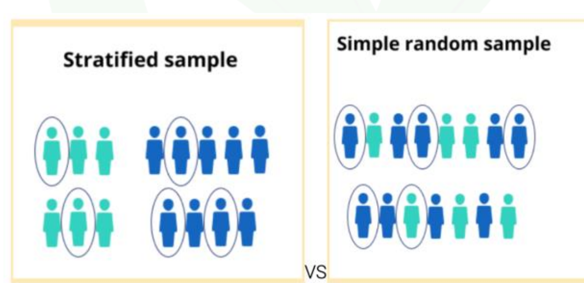


Simple Random Sampling (SRS)

Örneklem Dağılımı

- Bir istatistik örneklem dağılımı:
  - Popülasyon dağılım türüne
  - Örneklem büyüklüğüne
  - Örneklem seçme yöntemine bağlıdır
- Mevcut örnekleme yöntemlerinden Simple Random Sampling kullanılacak





2. What is sampling? How many sampling methods do you know?

"Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined." *Read the full answer* here.

Q32. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Let's continue our Data Science Interview Questions blog with some more statistics questions.

INTERVIEW QUESTION

Sampling Distributions

Rastgele değişkene ait beklenen değer

Kesikli Rastgele Değişken

$$E(x) = \sum_{x} xp(x)$$

Sürekli Rastgele Değişken

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

Örnekleme Dağılımı

- Expected Value (Beklenen Değer) (E(x))
- Örnek:

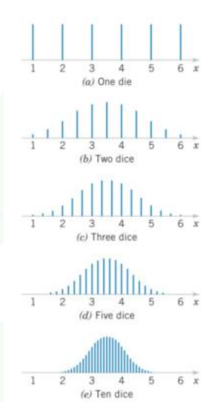
X, zar atışında bir zarın alacağı değerleri göstermektedir.
E(X) =?

$$p(x) = \frac{1}{6}$$

$$E(x) = \sum_{x} xp(x) = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = 3.5$$

$$\sigma^2 = \sum_{x}(x-\mu)^2 p(x) = (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + \ldots + (6-3.5)^2(1/6) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$



Bir Zar atıldığında böyle bir sayı ile karşılaşılabilir mi???

Sampling Distributions of $\overline{X}$

Örnekleme Dağılımı

- 2 zar atışının ortalaması, n=2

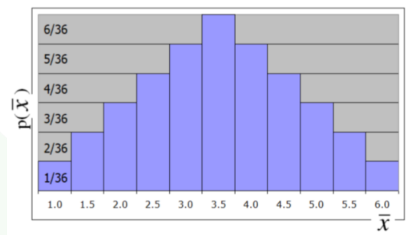
Sample		Sample		Sample	
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

$$\mu_{\bar{X}} = \sum_{\bar{x}} \bar{x} p(\bar{X}=\bar{x}) = 1(1/36) + 1.5(2/36) + \ldots + 6(1/36) = 3.5$$

$$\sigma_{\bar{x}}^2 = \sum_{\bar{x}} (\bar{x}-\mu)^2 p(\bar{X}=\bar{x}) = (1-3.5)^2(1/36) + (1.5-3.5)^2(2/36) + \ldots + (6-3.5)^2(1/36) = 1.46$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.46} = 1.21$$

$\bar{x}$	$p(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



$$\left.\begin{aligned} &1.\ E(\overline{X}) = \mu_{\overline{X}} = \mu \\ &2.\ V(\overline{X}) = \sigma_{\overline{X}}^2 = \sigma^2/n \\ &3.\ \sigma_{\overline{X}} = \sigma / \sqrt{n} \end{aligned}\right\}$$

Genel Kural

Standard Error of the Mean

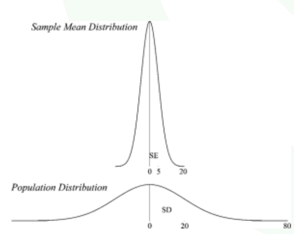
Ortalamanın Standard Hatası

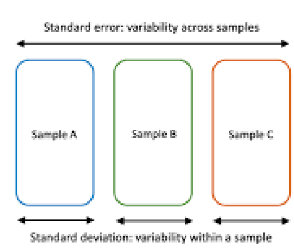
- Standart sapmalar popülasyon verilerini kullanırken, standart hata örnek verileri kullanır.
- Standart hata ne kadar küçük olursa örneklem istatistiği popülasyonun parametrelerine o kadar yakınlaşmış olur

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

Standard Deviation....

...Standard Error!!!

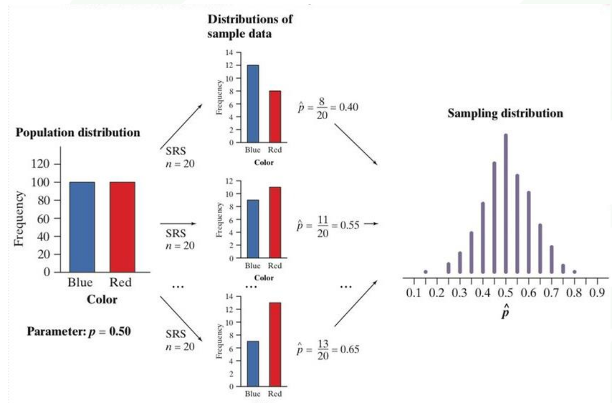




Sampling Distributions

Örnekleme Dağılım

- Popülasyon distr. İle bundan oluşan Sample Distr. Birbirinden farklı



Python Coding

Law of Large Numbers-lecturer.ipynb dosyasına bakalım..

- Bu notebook ta Law of Large Numbers için hesaplamalar bulunmaktadır.

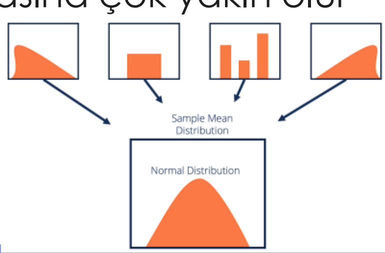
Central Limit Theorem

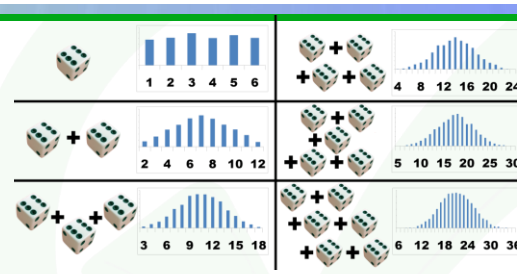
Merkezi Limit Teoremi

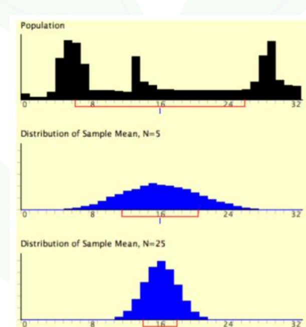
Central Limit Theorem

Merkezi Limit Teoremi

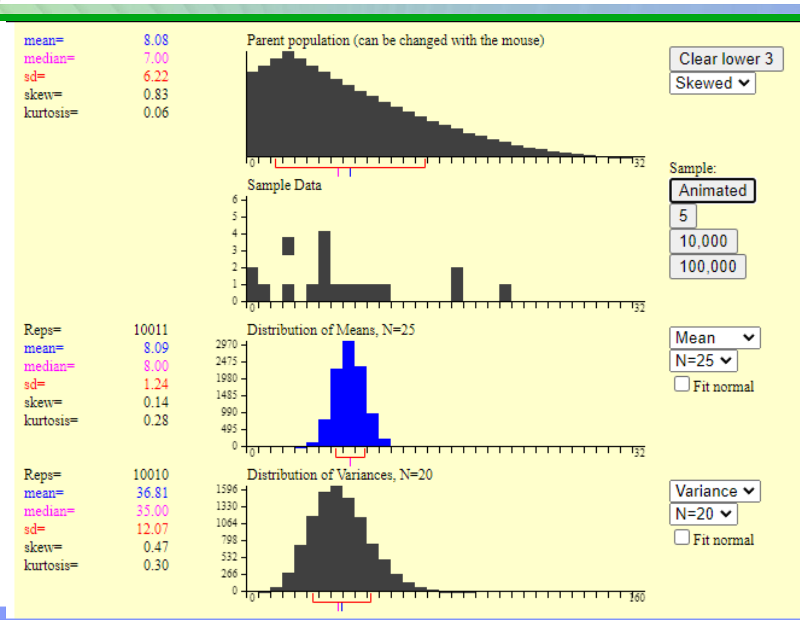
- Eğer sample size yeterince büyük olursa, örneklem ortalamasının dağılımı normal dağılıma yakınsar, ortalaması popülasyonun ortalamasına çok yakın olur







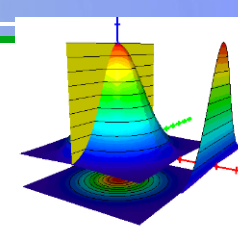
Central Limit Theorem

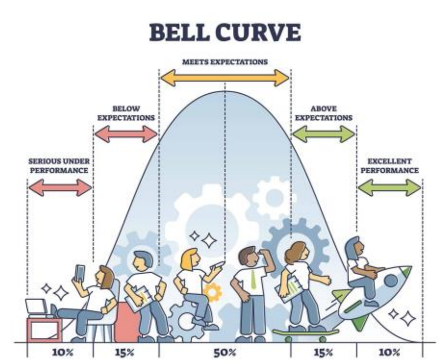


https://onlinestatbook.com/stat_sim/sampling_dist/index.html

Normal Distribution Advantages

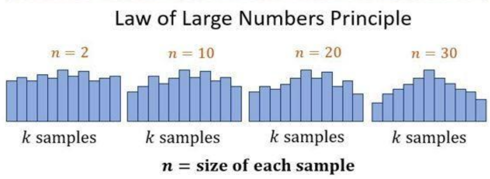
- Analiz ve Yorumlama Kolaylığı
- Parametrik Testlerin Kullanımı
- Merkezi Limit Teoremi
- Hata Terimlerinin Dağılımı
- Anormalliklerin ve Outlier Tespiti
- Öngörü ve Tahmin

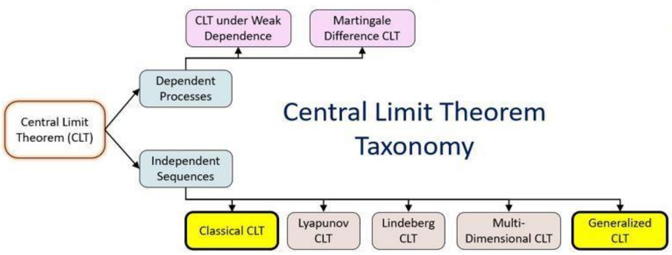


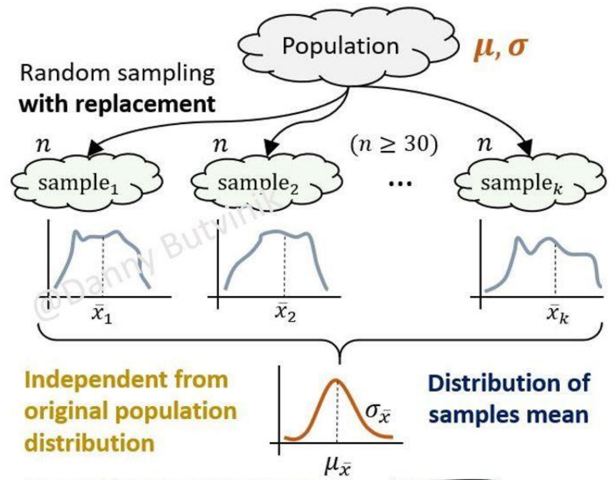


Central Limit Theorem

μ – mean of entire population
σ – std of entire population
n – sample size (preferably >30; equal among all samples)
k – number pf samples
$\bar{x}_i$ - mean of sample i
$\mu_{\bar{x}}$ - mean of means of the samples
$\sigma_{\bar{x}}$ - std of distribution of the means of samples







$$\mu_{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_k}{k} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Question13: What is the Central Limit Theorem?

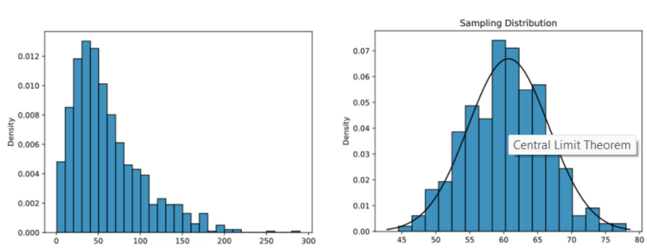
Central limit theorem states that, if you have a population mean (μ) and standard deviation (σ) and take large random samples from the population with replacement.

Then the distribution of the sample means will be approximately normally distributed regardless of whether the population is normal or skewed.

Provided that the sample size is sufficiently large (n > 30).

Question35: What is the Central Limit Theorem?

The Central Limit Theorem (CLT) states that, given a sufficiently large sample size from a population with a finite level of variance, the sampling distribution of the mean will be normally distributed regardless of if the population is normally distributed.



Question36: What general conditions must be satisfied for the central limit theorem to hold?

The central limit theorem states that the sampling distribution of the mean will always follow a normal distribution under the following conditions:

The sample size is sufficiently large (i.e., the sample size is $n \geq 30$).

The samples are independent and identically distributed random variables.

The population's distribution has finite variance.

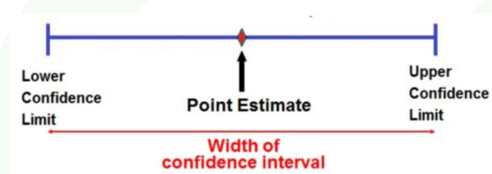
1. What is the Central Limit Theorem and why is it important?

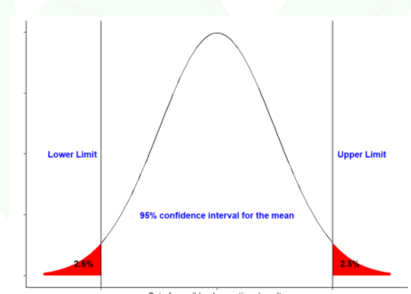
"Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can't obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly." *Read more* here.

Confidence Interval (CI)

Güven Aralığı

- Noktasal tahminler
- Aralık tahmini
- Güven aralığı kavramı ile konuşmak
- «Ortalamalar %90 güven ile .... Aralıktadır» deriz.
- a(alfa)=0,05 yani genelde %95 CI esas alınır

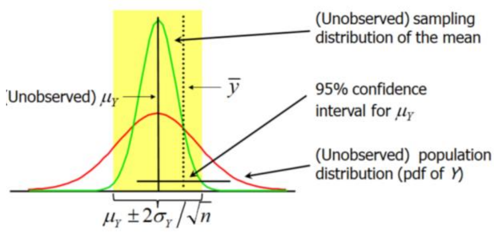


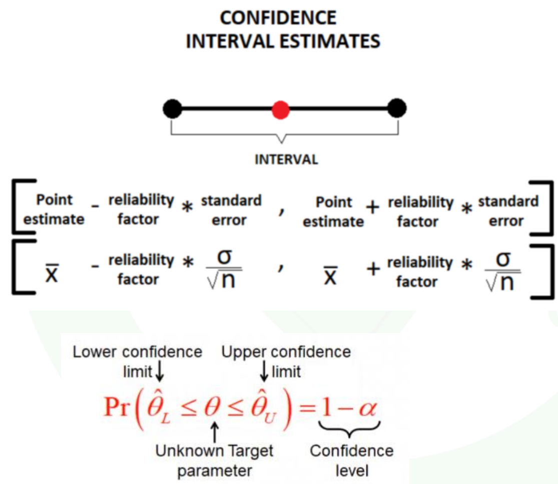


Confidence Interval (CI)

Güven Aralığı

- 1-a(alfa)=0,05 yani genelde %95 CI esas alınır



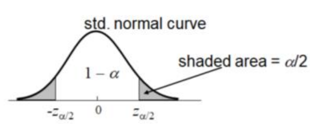


Confidence Interval (CI)

Güven Aralığı

- Normal dağılıma uygun ortalaması 0, std. Sapması 1 olan bir sample dan hareketle CI

100(1 – α)% confidence interval = $\bar{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

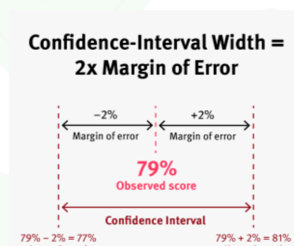


Level of Confidence (1-α)	α /2	$z_{\alpha/2}$
.90	.05	1.645
.95	.025	1.96
.99	.005	2.58

$$P\left\{\left(\bar{Y}-1.96\frac{\sigma}{\sqrt{n}}\right) < \mu < \left(\bar{Y}+1.96\frac{\sigma}{\sqrt{n}}\right)\right\} = .95$$

$$\bar{y} \pm 1.96\frac{\sigma_Y}{\sqrt{n}}$$

%95 CI formülü



Confidence Interval (CI)

Güven Aralığı

Örnek:

- Data seti: 2,3,5,6,9
- Hesaplamalar yanda görülmektedir.
- CI en altta alt-üst limitlerle gösterilir.

$$\bar{x} = \frac{2+3+5+6+9}{5} = 5$$

$$\sigma = 2.5$$

$$\sigma_{\bar{x}} = \frac{2.5}{\sqrt{5}} = 1.118$$

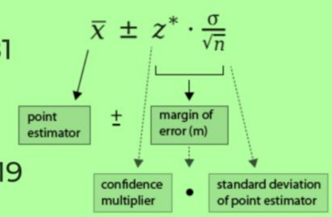
Margin of error = $z \times \frac{\sigma}{\sqrt{n}}$

Lower limit

- = 5 - (1.96)(1.118) = 2.81

Upper limit

- = 5 + (1.96)(1.118) = 7.19



CI = [2.81, 7.19]

Q8. What is the difference between Point Estimates and Confidence Interval?

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

A confidence interval gives us a range of values which is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by 1 — alpha, where alpha is the level of significance.

INTERVIEW QUESTION

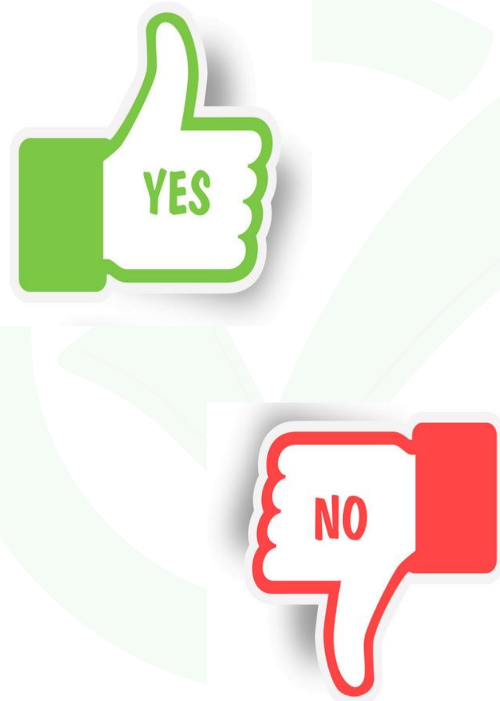
Python Coding

Confidence Interval-CI.ipynb dosyasına bakalım..

- Bu notebook ta Confidence Intervaliçin hesaplamalar bulunmaktadır.

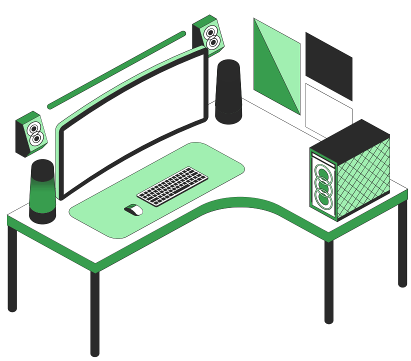
Kahoot Uygulaması

Final Question



Bu dersi tamamen anladım..

FINISH



Do you have any questions?

Send it to us! We hope you learned something new.