


DT/NT : DATA SCIENCE
LESSON : STATISTICS -2
SUBJECT: SAMPLE DISTRIBUTION
STD. ERROR OF MEAN
CENTRAL LIMIT
THEOREM

BATCH: 223



TECHPRO
EDUCATION

techproeducation.com [+1 \(585\) 304 29 59](https://wa.me/915853042959)

[f](#) [X](#) [in](#) [v](#) [i](#) [@](#)

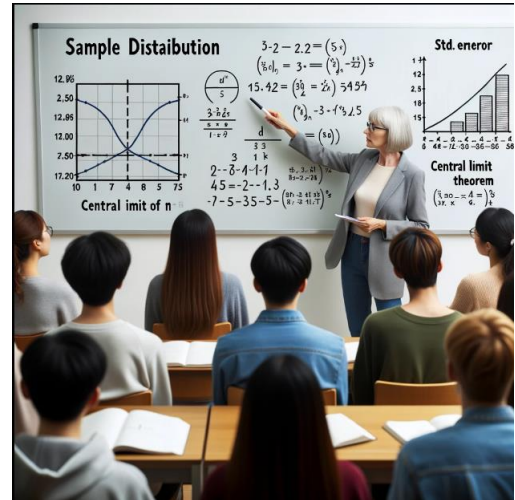
STATISTICS - 2

Data Science Program
Statistics Session -7

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

Random Variable: Discrete - Continuous

Kesikli – Sürekli Değişkenler

- Discrete (Kesikli) (Ayrık): sadece belli sayılar gelebilir (zar atımında 1-2-3... Olur ama 1.5 değeri her hangi birini alamaz. Tamsayı veya kusuru olabilir. Boy, kilo gibi).
- Continuous (Sürekli): Bir aralıktaki değerlerin herhangi birini alabilir. Tamsayı veya kusuru olabilir. Boy, kilo gibi).

Probability Distribution

Olasılık Dağılımı

- Olasılık dağılımını belirsizliği ortadan kaldırmak amacıyla kullanılır.
- İstatistiksel bir deneyin her sonucunu, gerçekleşme olasılığıyla ilişkilendiren bir tablo veya denklemdir.
- Rastgele bir değişkenin değerlerinin dağılımı, olasılık dağılımları ile tanımlanır

Discrete Probab. Distr.

- Discrete Random Variables için Olasılık dağılımları

Continuous Probab. Distr.

- Continuous Random Variables için Olasılık dağılımları

Rastgele değişkene ait matematiksel modeller (fonksiyonlar)

Discrete Distributions

Binomial Distributions

Binom Dağılımı

$P(X=0)=1/8$
 $P(X=1)=3/8$
 $P(X=2)=3/8$
 $P(X=3)=1/8$

Binom Dağılımında Mean and Std. Dev.

Mean: $\mu = np$

Standard Deviation: $\sigma = \sqrt{np(1-p)}$

Poisson Distributions

Poisson Dağılımı Özellikleri

- Poisson dağılımı, T zamanında meydana gelen X olay sayısının olasılığı verir.
- Denek sayısı olan n büyükken p de çok küçük ise binom dağılımı poisson dağılımına yaklaşıp
- Genel olarak $np < 10$ olduğu zaman binom dağılımı yerine poisson dağılımı kullanılabilir

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

Continuous Probability Distributions

• Most

Normal Distributions

• Örnek olarak: Mesurata 32.5 ile datanın 36.61 hangi aralıkta olur?

Mean: μ
 Variance: σ^2

$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

t Distribution (aka, Student's t-distribution)

t Dağılımı (Student Test)

- Örnekleme büyüklüğü küçük olduğunda ve/veya popülasyon standart sapması bilinmediğinde popülasyon parametrelerini tahmin etmek için kullanılır
- Örnekleme büyüklüğü ne kadar büyük olursa, t dağılımı o kadar normal dağılıma yaklaşıp
- 30'dan büyük örneklem büyüklükleri için dağılım normal dağılıma çok benzep

$n = \text{sample size}$

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.
- İstatistiklerin 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

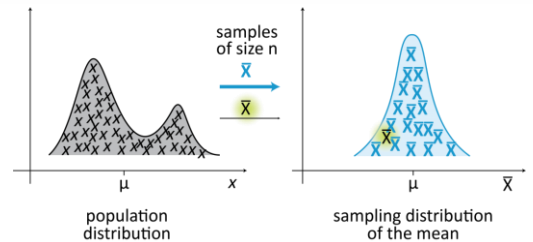
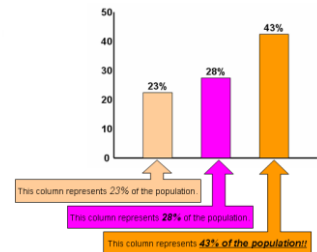


Table 5.1 Samples from the Weibull Distribution of Example 5.19

Sample	1	2	3	4	5	6
1	6.1171	5.0761	3.4671	1.5560	3.1237	8.9379
2	4.1600	6.7927	2.7198	4.5694	6.0985	3.9248
3	3.1950	4.4329	5.8812	4.7987	3.4118	8.7620
4	0.6694	8.5575	5.1491	2.4975	1.6549	7.0556
5	1.8552	6.8248	4.9963	2.3326	2.2912	2.3093
6	5.2316	7.3998	5.8687	4.0129	2.1283	5.9419
7	2.7609	2.1475	6.0918	9.0845	3.2093	6.7416
8	10.2185	8.5062	1.8019	3.2578	3.2309	1.7546
9	5.2438	5.4951	4.2194	3.7032	6.8426	4.9182
10	4.5590	4.0452	2.1294	5.5013	4.2064	7.2608
\bar{x}	4.401	5.928	4.229	4.132	3.620	5.761
s	4.260	6.144	4.608	3.857	3.221	6.342
s	2.642	2.062	1.611	2.124	1.678	2.496

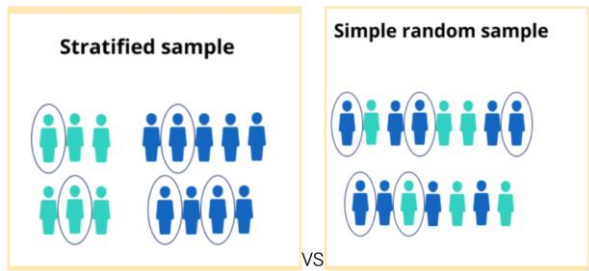
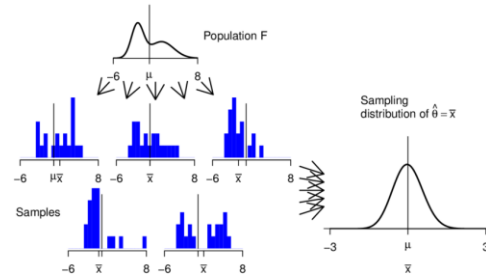
Note that the sample means, medians, and standard deviations are all different – randomness!



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.

Sampling Distributions

Rastgele değişkene ait beklenen değer

Örnekleme Dağılımı

- Expected Value
(Beklenen Değer) ($E(x)$)

Örnek:

X , zar atışında bir zarın alacağı değerleri göstermektedir.

$E(X) = ?$

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

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

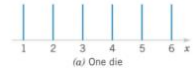
$$\sigma^2 = \sum_x (x - \mu)^2 p(x) = (1 - 3.5)^2 \left(\frac{1}{6}\right) + (2 - 3.5)^2 \left(\frac{1}{6}\right) + \dots + (6 - 3.5)^2 \left(\frac{1}{6}\right) = 2.92$$

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

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

Sürekli Rastgele Değişken

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



Sampling Distributions of \bar{X}

Örnekleme Dağılımı

- 2 zar atışının ortalaması, $n=2$

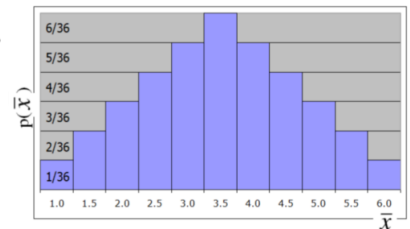
$$\mu_{\bar{x}} = \sum \bar{x} p(\bar{x}) = 1 \left(\frac{1}{36}\right) + 1.5 \left(\frac{2}{36}\right) + \dots + 6 \left(\frac{1}{36}\right) = 3.5$$

$$\sigma_{\bar{x}}^2 = \sum (\bar{x} - \mu)^2 p(\bar{x}) = (1 - 3.5)^2 \left(\frac{1}{36}\right) + (1.5 - 3.5)^2 \left(\frac{2}{36}\right) + \dots + (6 - 3.5)^2 \left(\frac{1}{36}\right) = 1.46$$

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

Sample	Sample	Sample
1, 1	3, 1	5, 1
1, 2	3, 2	5, 2
1, 3	3, 3	5, 3
1, 4	3, 4	5, 4
1, 5	3, 5	5, 5
1, 6	3, 6	5, 6
2, 1	4, 1	6, 1
2, 2	4, 2	6, 2
2, 3	4, 3	6, 3
2, 4	4, 4	6, 4
2, 5	4, 5	6, 5
2, 6	4, 6	6, 6

\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



$$1. E(\bar{X}) = \mu_{\bar{X}} = \mu$$

$$2. V(\bar{X}) = \sigma_{\bar{X}}^2 = \sigma^2 / n$$

$$3. \sigma_{\bar{X}} = \sigma / \sqrt{n}$$

Genel Kural



Standard Error of the Mean

Ortalamanın Standard Hatası

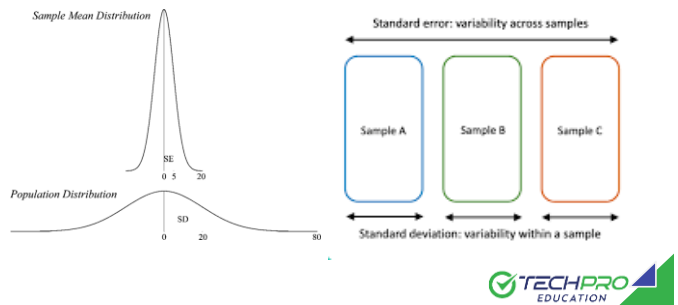
- Standart sapmalar popülasyon verilerini kullanırken, standart hata örnek verileri kullanır.
- Standart hata ne kadar küçük olursa örneklem istatistiği popülasyonun parametrelerine o kadar yaklaşmış olur

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

Standard Deviation....

••• Vs. •••

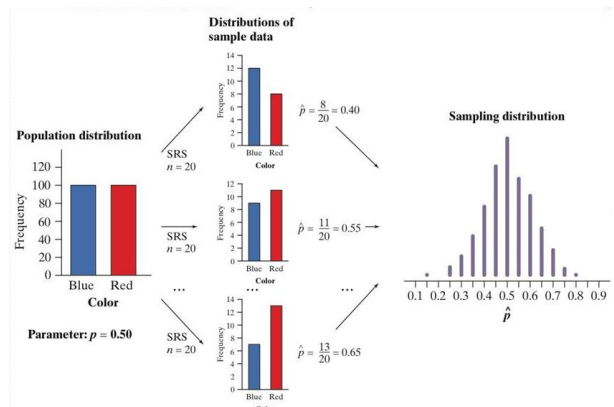
...Standard Error!!!



Sampling Distributions

Örnekleme Dağılımı

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



Python Coding

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

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

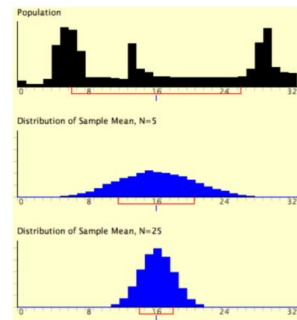
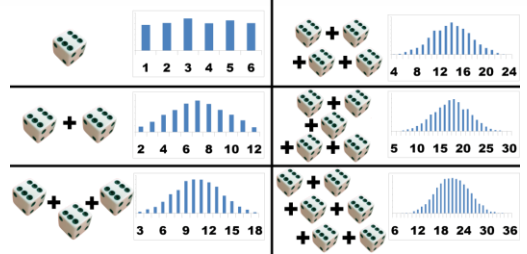
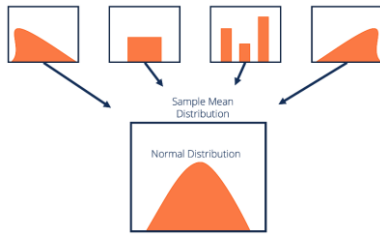
CENTRAL LIMIT THEOREM

Merkezi Limit Teoremi

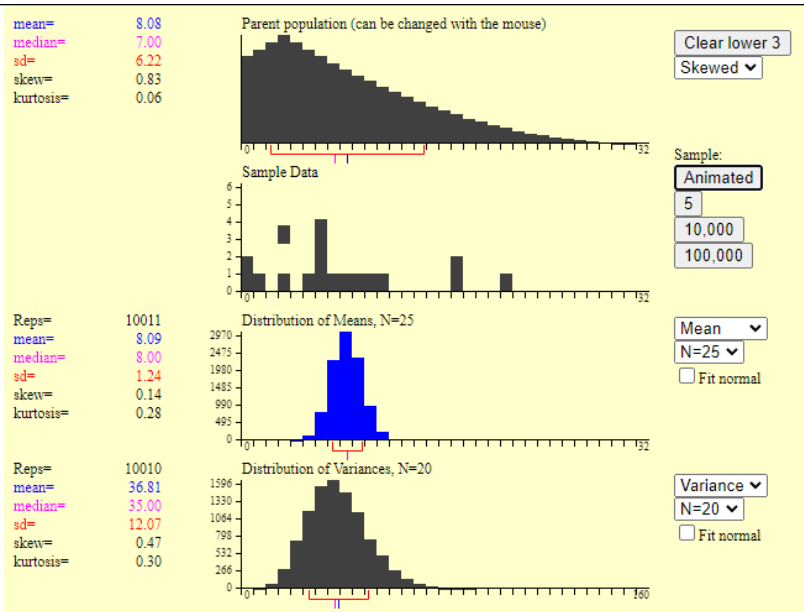
Central Limit Theorem

Merkezi Limit Teoremi

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

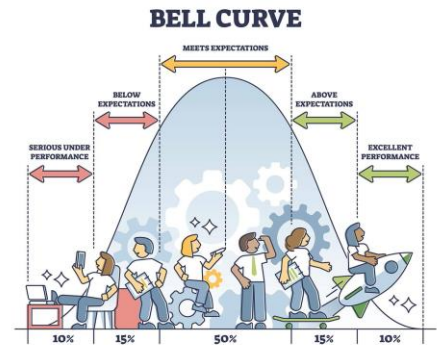
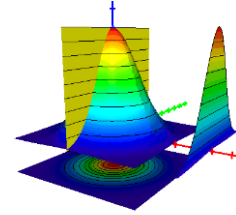


Central Limit Theorem



Normal Distribution Advantages

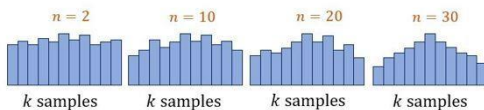
- Analiz ve Yorumlama Kolaylığı
- Parametrik Testlerin Kullanımı
- Merkezi Limit Teoremi
- Hata Terimlerinin Dağılımı
- Anormalliklerin ve Outlier Tespiti
- Öngörü ve Tahmin



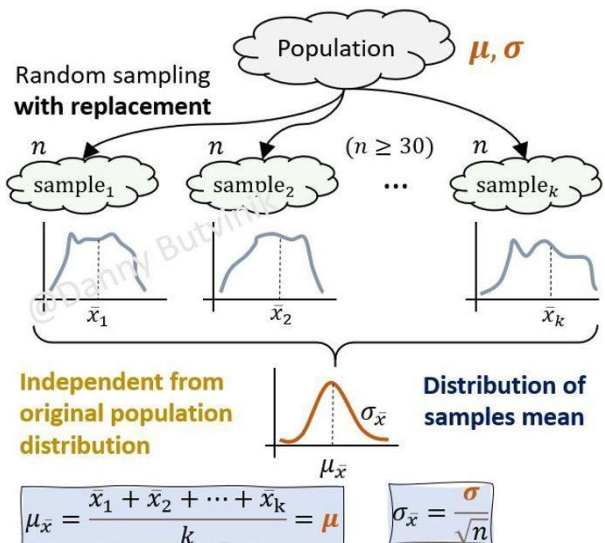
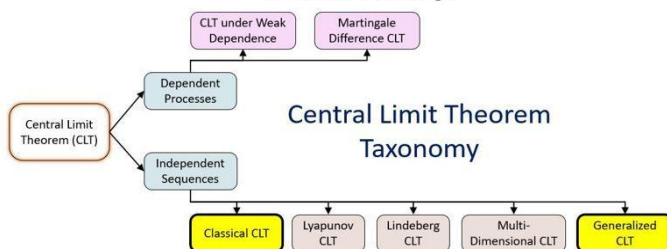
Central Limit Theorem

μ – mean of entire population
 σ – std of entire population
 n – sample size (preferably >30; equal among all samples)
 k – number of 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

Law of Large Numbers Principle



n = size of each sample



Question13: What is the Central Limit Theorem?

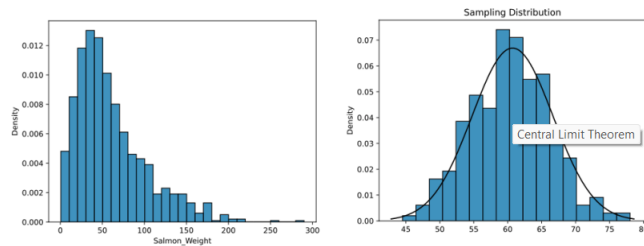
Central limit theorem states that, if you have a population mean (μ) and standard deviation (σ) and take large random samples from the population with replacement.

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?

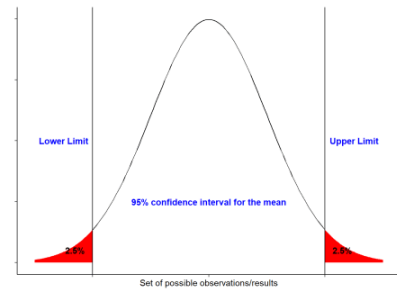
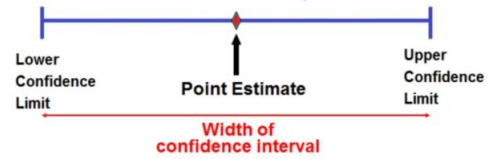
"Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can't obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about 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

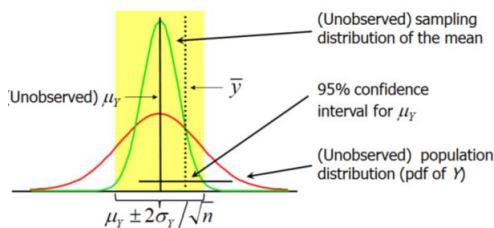
- Noktasal tahminler
- Aralık tahmini
- Güven aralığı kavramı ile konuşmak
- «Ortalamalar %90 güven ile Aralıktadır» deriz.
- $\alpha(\text{alfa})=0,05$ yani genelde %95 CI esas alınır



Confidence Interval (CI)

Güven Aralığı

- $1-\alpha(\text{alfa})=0,05$ yani genelde %95 CI esas alınır



CONFIDENCE INTERVAL ESTIMATES



$$\left[\begin{array}{l} \text{Point estimate} - \text{reliability factor} * \text{standard error} \\ \bar{X} - \text{reliability factor} * \frac{\sigma}{\sqrt{n}} \end{array} , \begin{array}{l} \text{Point estimate} + \text{reliability factor} * \text{standard error} \\ \bar{X} + \text{reliability factor} * \frac{\sigma}{\sqrt{n}} \end{array} \right]$$

$$\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

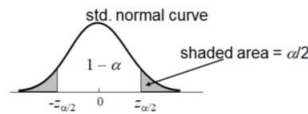
Lower confidence limit \downarrow Upper confidence limit \downarrow
 Unknown Target parameter Confidence level

Confidence Interval (CI)

Güven Aralığı

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

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



Level of Confidence (1-\alpha)	\alpha/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 Width =
2x Margin of Error**



Confidence Interval (CI)

Güven Aralığı

Örnek:

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

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

$$\sigma = 2.5$$

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

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

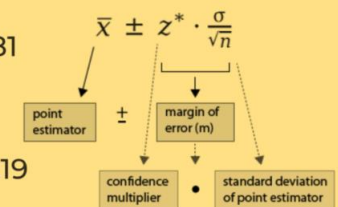
Lower limit

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

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 α is the level of significance.



Python Coding

Confidence Interval-CI.ipynb dosyasına bakalım..

- › Bu notebook ta Confidence Interval için hesaplamalar bulunmaktadır.