

ÖRNEK DAĞILIMININ ORTALAMASI VE STANDART SAPMASI

Birçok çalışmada, büyük veri gruplarının içinden daha küçük veri grupları seçilerek büyük veri gruplarının hakkında bilgi edinmek amaçlanır. Bunun nedeni

daha az zaman kullanmak,

tüm veriye ulaşamamak

maliyetinin çok fazla olmasıdır.

Örneğin,

Tüm gelecek seçimdeki seçmenler,

Herhangi bir ürünün üretilmesinde Pazar araştırmaları,

Herhangi bir hastalık üzerine yapılan çalışmalar,

İnternetteki tıklanma konuları

BAZI TANIMLAR

VERİ

Bir araştırma da, bir tartışmada bir akıl yürütmede sonuca ulaşabilmek için gözlemlenmiş ilk bilgiye veri denir. Veri; ölçüm, deney, gözlem, sayım ya da araştırma yolu ile elde edilir. İngilizce olarak data anlamına gelmektedir. Bir konu ile ilgili araştırma, tartışma, bilgi edinme ve akıl yürütme sonucunda oluşur. İşlenmemiş, yorum yapmaya imkan vermeyecek düzeyde sistemleştirilmiş ham bilginin derlenmiş ve formüle edilmiş şekline veri denir.

İSTATİSTİK

Belirli bir amaç için veri toplama, sonuçları yorumlama, sonuçları açıklama, örneklerden elde edilen sonuçları anakütle için genelleme, özellikler arasındaki ilişkiyi araştırma, çeşitli konularda geleceğe ilişkin tahmin yapma, deney düzenleme ve gözlem ilkelerini kapsayan sonuçta verilerden gidilerek bulunan olasılık deyimleri ile objektif karar vermede önemli rol oynayan bir bilimdir.

İstatistiksel Yöntemler •

- Tanımlayıcı İstatistikler

Verilerin toplanması, tanımlanması ve temel analizler

- Yorumlayıcı İstatistikler

Örnekten elde edilen istatistikler kullanılarak, bilinmeyen parametreler hakkında tahminde bulunma, karar verme

Anakütle (Populasyon)

ilgilenilen tüm elemanları içeren , bir veya daha fazla özellik (DEĞİŞKEN) açısından araştırma yapılmak istenen tüm elemanların içinde bulunduğu kümedir.

Örneğin,

Tüm mümkün seçmenler,

Bir haftada üretilen bir fabrikadaki tüm ürünler

Örnekleme(Örnek)

Bir ana kütlein bir alt kümesidir. Anakütleden seçilen ve ilgilenilen değişken açısından anakütlenin özelliklerini yansıtmaya özelliğine sahip alt kümedir.

Seçmenlerden rasgele seçilen 2000 kişi,

Bir hafta da üretilen ürünlerden 10 ürünü rasgele seçmek

Örneklemin amacı;

Zaman kaybını azaltmak

Maliyeti Minimum yapmaktır.

Örneğin özellikleri

Örnek veri rasgele toplanmalıdır.

Anakütleyi en iyi şekilde temsil etmelidir.

Parametre

Parametre, ilgili anakütle için, değişmeyen sayısal olarak ölçülebilen herhangi bir özelliği o anakütlenin parametresi olup bir sabittir.

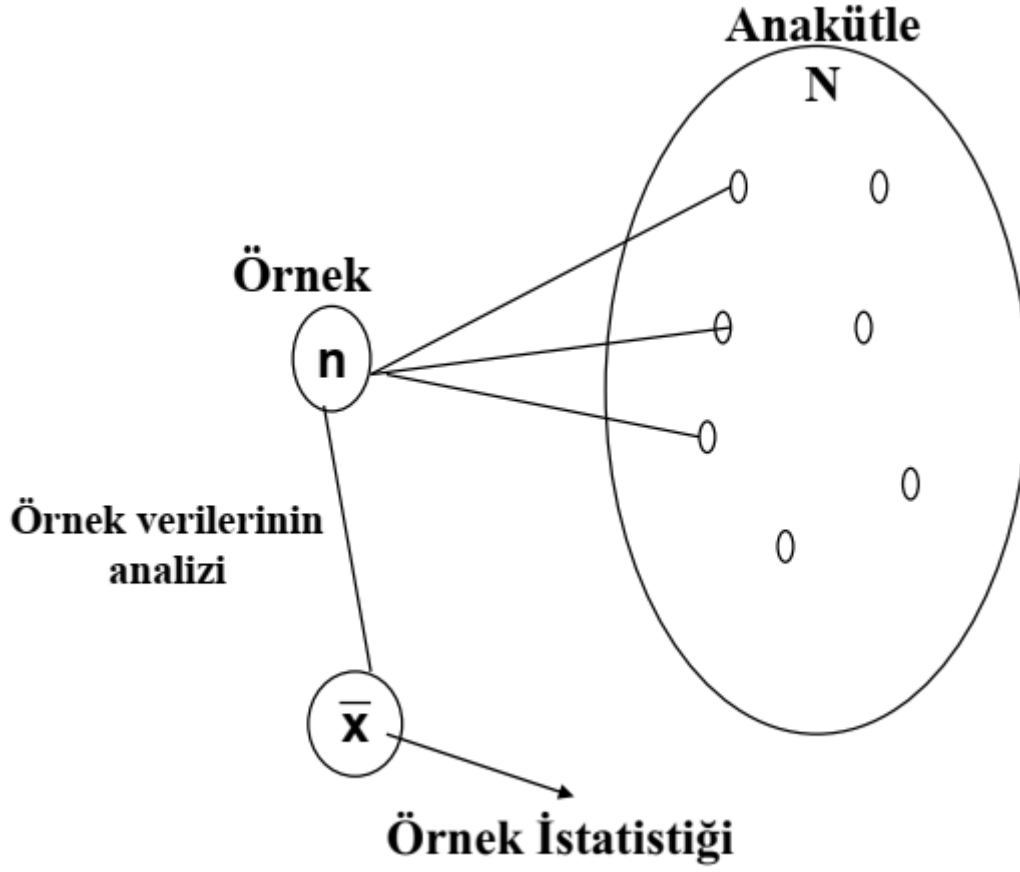
Bir fabrikada bir günde üretilen ürünlerin kusurlu olma oranı

Örnek İstatistiği

Anakütlenin belirli bir parametresinin hesaplanmasının zorluğundan dolayı alınan örnek yardımıyla bulunan parametre tahminine örnek istatistiği (istatistik / tahminleyici) adı verilir.

Örnek: Bir Ülkede tekstil de çalışan işçilerin aylık kazançlarının ortalamasını tahmin etmek amacıyla 100 işçilik bir örnek alınarak aylık kazanç miktarlarının ortalamasının bulunması istenebilir.

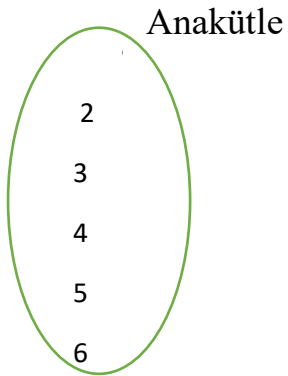
Anakütle-Örnek İlişkisi



Örneklerin Ortalamasının dağılımı

Örneklerin dağılımı, bir anakütledeki alabileceği her olası sonucu ve bu sonucun sıklığını gösterir.

ÖRNEK



olsun. İadeli çekilişlerin tüm ikili örnekleri oluşturalım.

Örnekler	örneklerin ortalaması				
(2,2), (2,3), (2,4), (2,5), (2,6)	2	2.5	3	3.5	4
(3,2), (3,3), (3,4), (3,5), (3,6)	2.5	3	3.5	4	4.5
(4,2), (4,3), (4,4), (4,5), (4,6)	3	3.5	4	4.5	5
(5,2), (5,3), (5,4), (5,5), (5,6)	3.5	4	4.5	5	5.5
(6,2), (6,3), (6,4), (6,5), (6,6)	4	4.5	5	5.5	6

$$\mu_{\bar{x}} = \frac{2 + 2.5 + 3 + 3.5 + 4 + 2.5 + 3 + 3.5 + 4 + 4.5 + 3 + 3.5 + 4 + 4.5 + 3 + 3.5 + 4 + 4.5 + 5 + 3.5 + 4 + 4.5 + 5 + 5.5 + 4 + 4.5 + 5 + 5.5 + 6}{25}$$

$$\mu_{\bar{x}} = 100 / 25 = 4$$

$$\sigma_{\bar{x}}^2 = ((2-4)^2 + (2.5-4)^2 + (3-4)^2 + (3.5-4)^2 + (4-4)^2 + (2.5-4)^2 + (3-4)^2 + (3.5-4)^2 + (4-4)^2 + (4.5-4)^2 + (3-4)^2 + (3.5-4)^2 + (4-4)^2 + \dots + (5.5-4)^2 + (6-4)^2) / 25$$

$$= 1$$

Anakütlenin ortalaması

$$\mu = \frac{2 + 3 + 4 + 5 + 6}{5} = 4 = \mu_{\bar{x}}$$

Anakütlenin standart sapmasının karesi

$$\sigma^2 = \frac{(2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2}{5} = \frac{10}{5} = 2$$

$$\mu = 4 \quad \mu_{\bar{x}} = 4 \quad \sigma^2 = 2 \quad \sigma_{\bar{x}}^2 = 1$$

Buradan

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Eğer örnekler iadesiz çekilişler ile yapılıyorsa bu durumda tüm ikili örnekler

(2,3), (2,4), (2,5), (2,6)	2.5	3	3.5	4
(3,4), (3,5), (3,6)		3.5	4	4.5
(4,5), (4,6)			4.5	5
(5,6)				5.5

$$\mu_{\bar{x}} = \frac{2.5+3+3.5+4+3.5+4+4.5+4.5+5+5.5}{10} = 4$$

$$\sigma_{\bar{x}}^2 = \frac{(2.5-4)^2 + (3-4)^2 + (3.5-4)^2 + (4-4)^2 + (3.5-4)^2 + (4-4)^2 + (4.5-4)^2 + (4.5-4)^2 + (5-4)^2 + (5.5-4)^2}{10} = \frac{7.5}{10} = 0.75$$

İadesiz çekilişler için;

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

N: Ana kütledeki eleman sayısı

n: örneğin eleman sayısı

ortalaması μ , standart sapması σ ve eleman sayısı N olan bir anakütleden iadesiz çekilişlerle oluşan tüm n elemenlı örneklerin ortalamasının dağılımının ortalaması $\mu_{\bar{x}}$ ve standart sapması $\sigma_{\bar{x}}$ ile gösterilirse;

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Çekilişler iadeli ve anakütlenin eleman sayısı ∞ (iadesiz çekilişler içinde) ise

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

dir $\frac{\sigma}{\sqrt{n}}$ standart hatayı verir. Anakütlenin eleman sayısı ∞ olduğunda yada çok çok büyük olduğunda tüm n li örneklerin alınması mümkün değildir. Bu durumda rasgele n li örnek alınır. Bu örneklerin dağılımının parametresi $(\mu_{\bar{x}}, \sigma_{\bar{x}})$ dir. Bu durumda bu parametrelerin değerleri yaklaşıktır ve hata payı içerir.

Anakütlenin eleman sayısı sonsuz ise iadesiz çekilişlerde de

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

kullanılır.

İstatistiksel çıkarsama: TAHMİN

Örneklerden alınan bilgidan faydalanarak anakütle parametrelerine ilişkin tahminde bulunmaktır. Anakütle parametresinin tahmini örneklerin dağılımından tek bir sayısal değerle ifade edilebiliyorsa bu değer nokta tahmini denir.

Bir anakütle parametresi tahmini tek bir değerle ifade etmek güvenilir olmayabilir. Herhangi bir güven aralığında belirli bir hata payı ile anakütle parametrelerini tahmin etmeye aralık tahmini denir.

Örnek 10000 kişinin katıldığı bir sınavda notların ortalaması 68 ve standart sapması 7.6 olup normal dağılım göstermektedir. Bu anakütleden iadesiz çekilişlerle $n=20$ kişilik 45 örnek alınmıştır.

- 1) Örneklerin ortalamalarının dağılımının parametrelerini bulunuz
- 2) Bu 45 örnekten kaç tanesinin ortalaması [65-70] aralığındadır.
- 3) Ortalaması 67 veya daha fazla kaç örnek vardır.

1) Dağılımın parametreleri

$$\mu_{\bar{x}} = \mu \quad \mu_{\bar{x}} = 68$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{7.6}{\sqrt{20}} \sqrt{\frac{10000-20}{10000-1}} = 1.6978$$

$$2) P\left(\frac{64.5-68}{1.6978} \leq z \leq \frac{70.5-68}{1.6978}\right) = P(-2.06 \leq z \leq 1.47) = 0.4803 + 0.4292 = 0.9095$$

$$45(0.9095) = 40.92 = 41 \text{ örnek}$$

$$3) P\left(z \geq \frac{66.5-68}{1.6978}\right) = P(z \geq -0.88) = 0.5 + 0.3106 = 0.8106$$

$$40(.8106) \approx 36$$

tane örnek

ÖRNEK:

50 adet erkek öğrencinin boy ortalaması 178 cm hesaplanmıştır. 17-22 yaş grubundaki erkeklerin boy ortalaması 177 cm ve standart sapması 6 olan bir normal dağılım göstermektedir. örneğin boy ortalamasının, bu anakütleden çekilecek 50'lik örneklerin ortalamalarına ait dağılımda olma olasılığını bulunuz..

Bu anakütleden çekilen örneklerin ortalaması, standart sapması, $\mu_{\bar{x}} = \mu = 177$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{50}} = 0.8485$$

olan bir dağılım gösterir. Bu dağılımda boy ortalamalarının 178'den daha fazla olma olasılığı,

$$P\left(\frac{177.5-177}{0.8485} \leq z \leq \frac{178.5-177}{0.8485}\right) = P(0.59 \leq z \leq 1.77) = 0.4616 - 0.2224 = 0.2392$$

ÖRNEK:

Bir fabrikada üretilen ampullerin ömürlerinin ortalaması 1700 saat ve standart sapması 175 saattir. Rasgele seçilen 25 ampullü bir örneğin parametrelerini bulunuz .Örneklerin ortalamalarının dağılımının normal dağılım olması için anakütleyle ilgili ne söylenebilir? Örneklerin dağılımındaki standart hatanın 20 birimin altında kalması için örneklerin boyutu(eleman sayısı) kaç olmalıdır?

$$\mu_{\bar{x}} = \mu = 1700 \text{ saat} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{175}{\sqrt{25}} = 35$$

Anakütledeki ampullerin ömürleri normal dağılıyorsa, örneklerin ortalamalarının dağılımının normal dağılır.

$$\frac{\sigma}{\sqrt{n}} = \frac{175}{\sqrt{n}} \leq 20 \quad 30625 \leq 400n \quad 76.77 \leq n$$

TOPLAMLARIN VE FARKLARIN DAĞILIMI

İki tane anakütle olsun. 1. Anakütleden n elemanı 2. Anakütleden m elemanlı bütün örnekler seçilsin. Rasgele 1. Anakütleden n elemanı, 2. Anakütleden m elemanlı birer örnek alınsın bu örneklerin parametreleri $\mu_{\bar{x}_1}, \sigma_{\bar{x}_1}$ ve $\mu_{\bar{x}_2}, \sigma_{\bar{x}_2}$ olsun.

$$\mu_{\bar{x}_1 + \bar{x}_2} = \mu_{\bar{x}_1} + \mu_{\bar{x}_2}, \quad \sigma_{\bar{x}_1 + \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}, \quad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Örnek,

A fabrikasında üretilen ampullerin ömürleri 2500 saat standart sapması 250 saattir. B fabrikasında üretilen ampullerin ömürleri 2100 saat standart sapması

150 saattır. A ve B fabrikalarından 50 ampulluk birer örnek alınmıştır. A fabrikasından alınan örneğin ortalamasının B fabrikasından alınan örneğin ortalamasından 175 saat fazla çıkması olasılığını hesaplayınız.

$$\mu_{\bar{x}_A - \bar{x}_B} = \mu_{\bar{x}_A} - \mu_{\bar{x}_B} = \mu_A - \mu_B = 2500 - 2100 = 400$$

$$\sigma_{\bar{x}_A - \bar{x}_B} = \sqrt{\sigma_{\bar{x}_A}^2 + \sigma_{\bar{x}_B}^2} = \sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}} = \sqrt{\frac{250^2}{50} + \frac{150^2}{50}} = 41.23$$

$$P(z \geq \frac{175 - 400}{41.23}) = P(z \geq -5.457) = 1$$

ÖRNEK

Araba yakıtlarına konan iki antifrizin yakıt tüketimine etkisi araştırılıyor. A antifrizinin denendiği 20 arabada ortalama 5.8 lt/100km, B antifrizinin denendiği 22 arabada ise ortalama tüketim 6.4 lt/100km bulunuyor. Bu örneklerin varyansı 1.44 olan bir anakütleden çekilen bizim örnek çiftimiz gibi örnek çiftlerinin ortalamaları arasındaki farkın örnekleme dağılımına dâhil olma olasılığı nedir?

$D = 6.4 - 5.8 = 0.6$ lt farkın ortalaması sıfır, varyansı da yukarıdaki eşitlikten

$$\sigma_D^2 = \sigma^2 \left(\frac{n_A + n_B}{n_A \cdot n_B} \right) = 1.44 \frac{42}{440} = 0.13745$$

olan örnekleme dağılımına dâhil olma olasılığı, yani bu dağılımda 0.6 ve daha büyük olan farkların nispi miktarı, ilgili örnekleme dağılımını standart normal dağılıma çevirerek bulunur:

$$z = \frac{D - 0}{\sigma_D} = \frac{0.6}{\sqrt{0.13745}} = \frac{0.6}{0.37075} \cong 1.62$$

$$P(z > 1.62) = 0.5 - P(0 < z < 1.62) = 0.5 - .4474 \\ = .0526$$

ÖRNEK

Bir anakütledeki

Örnek: Bir fabrikadaki işçilerin haftalık ücretlerinin standart sapma 40 ile normal dağılım göstermektedir. 10 işçinin toplam haftalık ücreti 4400 bulunduğuna göre kitle ortalaması (μ) için %90 güven aralığını bulunuz.

Haftalık ücretler $\sim N(\mu, 40^2)$

$n=10$ (kişi sayısı)

$$X_1 + X_2 + \dots + X_{10} = 4400 \quad \bar{x} = 4400/10 = 440$$

$$1 - \alpha = 0.90$$

$$\alpha = \%10 \quad \alpha/2 = 0.05$$

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

bulduğumuz değerleri

formülde yerine yazalım.

$$P(440 - z_{0.05} \cdot 40 / \sqrt{10} < \mu < 440 + z_{0.05} \cdot 40 / \sqrt{10}) = 0.90$$

$$P(419 < \mu < 460) = 0.90$$

yorum: %90 güvenle işçilerin haftalık ücretlerinin ortalaması (μ) 419 tl ile 460 tl arasındadır.