Örnekleme İstatistiğinin Dağılımı

IST 108 Olasılık ve İstatistik
Bahar 2016

Yrd. Doç. Dr. Ferhat Dikbıyık

Örnekleme Ortalaması

- Her biri bir değerle eşleştirilmiş, bir elemanlar nüfusu (çoğunluğu) düşünelim (yani bir veri kümesi).
- Örneğin belirli bir gruptaki yetişkinler elemanlarımız olsun ve her bir yetişkinin yıllık geliri de eşleştirilen değer olabilir.
- Her bir eleman ile eşleştirilen her bir değer, bir rastgele değişken ile ifade edilebilir. Bu rastgele değişkene ait beklenti μ ile ve varyans ise σ^2 ile gösterilsin. Bu değerlere *nüfus ortalaması* ve *nüfus varyansı* denir.

Örnekleme Ortalaması

- $X_1, X_2, \ldots, X_n$ n adet örnekten oluşan bu nüfusa ait değerleri göstersin. Bu durumda örnekleme ortalaması

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- Örnekleme ortalaması rastgele değişkenlerin toplamının belirli bir değere bölünmesinden oluşan bir rastgele değişken olarak düşünülebilir. Bu durumda bu rastgele değişkene ait beklenti ve varyans da hesap edilebilir.

Örnekleme Ortalamasının beklentisi ve varyansı

$$E[\overline{X}] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{1}{n}\left(E[X_1] + E[X_2] + \ldots + E[X_n]\right)$$

$$= \frac{1}{n} n\mu = \mu$$

$$Var(\overline{X}) = Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n^2}\left(Var(X_1) + Var(X_2) + \ldots + Var(X_n)\right)$$

$$= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Merkezi Limit Teoremi

- Merkezi Limit Teoremi, kabaca, büyük sayıda birbirlerinden bağımsız rastgele değişkenlerin toplamının yaklaşık olarak normal dağılımı izlemesi olarak ifade edilebilir.
- $X_1, X_2, \ldots, X_n$, birbirinden bağımsız ve beklentisi μ ve varyansı ise σ^2 olan aynı dağılımlara sahip rastgele değişkenler olsun. Eğer *n* çok büyük ise, $X_1+X_2+ \ldots+X_n$ rastgele değişkenine ait dağılım, beklentisi $n\mu$ ve varyans ise $n\sigma^2$ olan bir normal dağılıma yakınsanabilir.

Merkezi Limit Teoremi

- Merkezi Limit Teoremine göre aşağıdaki rastgele değişken standart normal dağılım izler

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

- Bu durumda, Z bir standart normal dağılımlı rastgele değişkeni gösteriyorsa,

$$P\left\{\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P\{Z < x\}$$

Örnek Soru 1

- Bir sigorta şirketi, 25.000 araç-poliçesi yaptıran müşteriye sahiptir. Poliçe sahiplerinin yıllık sigorta talepleri eğer ortalaması 320 ve standart sapması 540 olan bir rastgele değişken ile ifade edilebiliyorsa, yıllık sigorta taleplerinin 8,3 milyonu aşması olasılığını hesaplayınız.

Örnek Soru 1

- Bir sigorta şirketi 25.000 araç poliçesi yaptıran müşteriye sahiptir. Poliçe sahipleri yıllık sigorta talepleri eğer ortalaması 320 ve standart sapması 540 olan bir rastgele değişken ile ifade edilebiliyorsa, yıllık sigorta taleplerinin 8,3 milyonu aşması olasılığını hesaplayınız.
- X, yıllık talebi göstersin. Poliçe sahiplerini numaralandıralım ve X_i i. poliçe sahibini yıllık sigorta talebi olsun. n = 25.000 ile merkezi limit teoremine göre X = $\sum_{i=1}^{n} X_i$ beklentisi 320x25.000 = 8 milyon olan ve standart sapması 540$\sqrt{25.000}$ = 85831 olan bir normal dağılıma yakınsanabilir.

Örnek Soru 1

- Bir sigorta şirketi 25.000 araç-poliçesi yaptıran müşteriye sahiptir. Poliçe sahiplerinin yıllık sigorta talepleri eğer ortalaması 320 ve standart sapması 540 olan bir rastgele değişken ile ifade edilebiliyorsa, yıllık sigorta taleplerinin 8,3 milyonu aşması olasılığını hesaplayınız.
- Beklenti = 8 milyon olan ve standart sapma = 85831
- P{X > 8.3 milyon} = P{(X – 8 milyon)/85831 > (8.3 milyon – 8 milyon)/85831 } = P{Z > 3,51} = 0,00023

Örnek Soru 2

- İnşaat mühendislerinin hesabına göre, yapısal bir zarar oluşturmadan bir köprünün taşıyabileceği yük miktarı, W (ton olarak) beklentisi 400 ve standart sapması 40 olan bir normal dağılımla dağıtılmıştır. Bir arabanın ağırlığı (ton olarak) beklentisi 3 ve standart sapması 0,3 olan bir normal dağılımla dağıtılmış olsun. Yapısal zarar ihtimalinin 0,1'i geçmesi için köprü üzerinde kaç araç olmalıdır?

Örnek Soru 2

- İnşaat mühendislerinin hesabına göre, yapısal bir zarar oluşturmadan bir köprünün taşıyabileceği yük miktarı W (ton olarak) beklentisi 400 ve standart sapması 40 olan bir normal dağılımla dağıtılmıştır. Bir arabanın ağırlığı (ton olarak) beklentisi 3 ve standart sapması 0,3 olan bir normal dağılımla dağıtılmış olsun. Yapısal zarar ihtimalinin 0,1'i geçmesi için köprü üzerinde kaç araç olmalıdır?
- P_n, köprü üzerinde n tane araç olduğunda yapısal zarar ihtimali olsun. Bu durumda soruda asıl sorulan $P_n > 0,1$ olduğunda n değerinin kaç olduğudur.

Örnek Soru 2

- P_n, köprü üzerinde n tane araç olduğunda yapısal zarar ihtimali olsun. Bu durumda soruda sorulan $P_n >$ 0,1 olduğunda n kaç olur?
- X_i, köprü üzerindeki i. aracın ağırlığı olsun.
- $P_n = \{X_1 + X_2 + \cdots + X_n \geq W\}$
  $$= P\{X_1 + X_2 + \cdots + X_n - W \geq 0\}$$
- Merkezi Limit Teoremine göre araçların toplam ağırlığı $X = X_1 + X_2 + \cdots + X_n$ beklentisi 3n ve standart sapması 0,09n olan bir normal dağılım izler.
- X ve W birbirlerinden bağımsızdır.
- E[X – W] = 3n – 400
- Var (X – W) = Var(X) + Var(W) = 0,09n + 1.600

Örnek Soru 2

- P_n, köprü üzerinde n tane araç olduğunda yapısal zarar ihtimali olsun. Bu durumda soruda sorulan $P_n > 0,1$ olduğunda n kaç olur?
- $P_n = P\{X - W \geq 0\} > 0,1$
- $E[X - W] = 3n - 400$ ve $Var(X - W) = 0,09n + 1.600$
- $P\{X - W \geq 0\} = P\left\{\frac{X-W-(3n-400)}{\sqrt{0,09n+1600}} \geq \frac{0-(3n-400)}{\sqrt{0,09n+1600}}\right\}$

  $= P\left\{Z \geq \frac{-(3n-400)}{\sqrt{0,09n+1600}}\right\} > 0,1$
- $P\{Z \geq 1,28\} \approx 0,1$ olduğundan
- $\frac{-(3n-400)}{\sqrt{0,09n+1600}} \leq 1,28 \rightarrow n \geq 117$

Merkezi Limit Teoreminin Binomial Rastgele Değişkene uygulanması

- Merkezi Limit Teoreminin en önemli uygulamalarından biri Binomial rastgele değişkene uygulanmasıdır. Binomial rastgele değişken X, her birinin başarılı olma ihtimali *p* olan *n* adet bağımsız deneyden başarılı olanların sayısını ifade ettiğinden, biz her bir deneyi bir Bernoulli rastgele değişken ile ifade edebiliriz. Bu durumda; $X = X_1 + X_2 + \ldots + X_n$

$$X_i = \begin{cases} 1 & i.\ deney\ basarili \\ 0 & aksi\ halde \end{cases} \qquad E[X_i] = p \qquad Var(X_i) = p(1-p)$$

$$\boxed{\frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)}$$

Örnek Soru 3

- Sakarya Üniversitesi Bilgisayar Mühendisliğinde birinci-sınıflar için ideal sınıf boyutunu 150 kişilik olarak belirlemiştir. Bölüm, geçmiş tecrübelerinden biliyor ki kabul alanların sadece %30'u derslere katılmaktadır. Bu nedenle 450 öğrenci kabul etmektedir. 150'den fazla öğrencinin derslere katılması ihtimalini hesaplayınız.

Örnek Soru 3

- Sakarya Üniversitesi Bilgisayar Mühendisliğinde birinci-sınıflar için ideal sınıf boyutunu 150 kişilik olarak belirlemiştir. Bölüm, geçmiş tecrübelerinden biliyor ki kabul alanların sadece %30'u derslere katılmaktadır. Bu nedenle 450 öğrenci kabul etmektedir. 150'den fazla öğrencinin derslere katılması ihtimalini hesaplayınız.
- X: derse katılacak öğrenci sayısı olsun. Her öğrencinin derse katılma ihtimali birbirinden bağımsız olduğu için X için Binomial rastgele değişken diyebiliriz. Burada n=450 ve p = 0,3 olur.
- Binomial kesikli, Normal dağılım ise sürekli olduğundan, normal yakınsamayı uygularken P{X=i} değerini hesaplamak için P{i-0,5 < X < i + 0,5}'i hesaplamak daha iyidir. Buna süreklilik düzeltmesi denir.
- $P\{X > 150{,}5\} = P\left\{\frac{X-(450)(0{,}3)}{\sqrt{450(0{,}3)(0{,}7)}} > \frac{150{,}5-(450)(0{,}3)}{\sqrt{450(0{,}3)(0{,}7)}}\right\} = P\{Z > 1{,}59\} \approx 0{,}06$

Merkezi Limit Teoreminin Binomial Rastgele Değişkene uygulanması

- Binomial rastgele değişkene ait iki yakınsama metodu olduğunu öğrendik.
  - Poisson: *n* çok büyük ve *p* çok küçük olduğunda iyi bir yakınsama sunar.
  - Merkezi Limit Teoremi: *np(1-p)* büyük olduğunda (örneğin ≥10) iyi bir yakınsana sunar.

Örnekleme Ortalamasının Yaklaşık Dağılımı

- Merkezi Limit Teoremi örnekleme ortalamasının dağılımını yakınsamak için de kullanılabilir.

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

$$\boxed{\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)}$$

Örneklemin ne kadar büyük olması gerekir?

- Merkezi limit teoreminin geçerli olabilmesi için örnekleme boyutu n ne olmalıdır?
- Eğer her bir değere ait dağılım normal dağılım ise, bu değerlerin toplamına ait dağılım, n değerine bağlı olmaksızın normal olacaktır.
- Genel bir kural olarak eğer n en az 30 ise normal dağılıma yaklaştırmak uygun olacaktır.
- Ama bir çok durumda daha düşük boyutlar için bile merkezi limit teoremi çok iyi bir yakınsama sunacaktır.

Örnek Soru 4

- Bir işçi popülasyonunda işçilerin ağırlıklarına ait beklenti167 ve standart sapma ise 27'dir.
- 36 tane işçi bu popülasyondan rastgele seçilirse, örnekleme ortalamalarının 163 ile 170 arasında olmasını yaklaşık olarak hesaplayınız.
- 144 işçi seçimi için cevabınız ne olurdu?

Örnek Soru 4

- Bir işçi popülasyonunda işçilerin ağırlıklarına ait beklenti167 ve standart sapması ise 27'dir.
- 36 tane işçi bu popülasyondan rastgele seçilirse, örnekleme ortalamalarının 163 ile 170 arasında olmasını yaklaşık olarak hesaplayınız.
- Merkezi Limit Teoremine göre örnekleme ortalaması beklentisi 167 olan ve standart sapması 27/6 = 4,5 olan bir normal dağılıma yakınsanabilir.
- $P\{163 < \bar{X} < 170\} = P\left\{\frac{163-167}{4,5} < \frac{\bar{X}-167}{4,5} < \frac{170-167}{4,5}\right\}$
  $= P\{-0,8889 < Z < 0,8889\} = 2P\{Z < 0,8889\} - 1$ =0,6259

Örnek Soru 4

- Bir işçi popülasyonunda işçilerin ağırlıklarına ait beklenti167 ve standart sapması ise 27'dir.
- 36 tane işçi bu popülasyondan rastgele seçilirse, örnekleme ortalamalarının 163 ile 170 arasında olmasını yaklaşık olarak hesaplayınız.
- 144 işçi seçimi için cevabınız ne olurdu?
- Bu durumda standart sapma 27/12 = 2,25 olur.
- $P\{163 < \bar{X} < 170\} = P\left\{\frac{163-167}{2,25} < \frac{\bar{X}-167}{2,25} < \frac{170-167}{2,25}\right\}$ $= P\{-1,7778 < Z < 1,7778\} = 2P\{Z < 1,7778\} -$ 1 =0,9246

Örnekleme Varyansı

- Örnekleme varyansı (ve standart sapması) da aynı örnekleme ortalaması gibi bir rastgele değişken olarak düşünülebilir.

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} \qquad S = \sqrt{S^2}$$

Örnekleme Varyansının beklentisi

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2]$$

$$(n-1)E[S^2] = nE[X_i^2] - nE[\overline{X}^2]$$

$$(n-1)E[S^2] = nVar(X_i) + n\left(E[X_i]\right)^2 - nVar(\overline{X}) + n\left(E[\overline{X}]\right)^2$$

$$(n-1)E[S^2] = n\sigma^2 + n\mu^2 - n\left(\sigma^2 / n\right) + n\mu^2$$

$$(n-1)E[S^2] = (n-1)\sigma^2$$

$$E[S^2] = \sigma^2$$

Sınırlı Popülasyondan Örnekleme

- *N* elemandan oluşan bir popülasyonda, popülasyon, *p* oranında belirli bir karakteristiği gösteriyor olsun.
- Yani *Np* eleman belirli bir karakteristiği gösterirken *N*(1-*p*) eleman bu karakteristiği göstermemektedir.
- Bu popülasyondan *n* boyutunda bir örnek seçilecek. Eğer popülasyonun *n* boyutundaki tüm alt kümelerinin örnek olma ihtimali eşit ise bu her bir örneğe *rastgele örnek* diyoruz.
- Örneğin, {a, b, c} elemanlarından oluşan bir popülasyon için, 2 elemanlı her bir alt küme {a, b}, {a, c} ve {b, c} aynı ihtimalle örnek oluşturabilirlerse bir rastgele örnek seçilebilir.

Sınırlı Popülasyondan Örnekleme

- *n* boyutunda bir rastgele örneğin *N* elemanlı bir popülasyondan seçildiğini varsayalım ve aşağıdaki tanımlamayı yapalım.

$$X_i = \begin{cases} 1 & \text{i. eleman belirli bir karakteristikte ise} \\ 0 & \text{aksi halde} \end{cases}$$

- Bu durumda belirli bir karakteristiği gösteren elemanların sayısı bu rastgele değişkenlerin toplamı olacaktır.

$$X = X_1 + X_2 + \ldots + X_n$$

- Bu durumda, bu değerlerin ortalaması, örnek içinde belirli bir karakteristiğe sahip elemanların oranı olacaktır.

Sınırlı Popülasyondan Örnekleme

- *N* eleman arasındaki her bir elemanın örnekteki i. eleman olma ihtimali eşittir. Yani;

$$P\{X_i = 1\} = \frac{Np}{N} = p$$

$$P\{X_i = 0\} = 1 - P\{X_i = 1\} = 1 - p$$

Sınırlı Popülasyondan Örnekleme

- Fakat X_1, X_2, …, X_n değerleri birbirinden bağımsız değildir. Çünkü örnek için seçilen ikinci eleman, N elemandan herhangi biri olabileceğinden ikinci elemanın karakteristiğe sahip olması ihtimali *Np*/*p*=*p*'dir. $P\{X_2=1\}=p$.
- Fakat, birinci seçilen elemanın karakteristiğe sahip olup olmadığı verildiğinde

$$P\{X_2 = 1 \mid X_1 = 1\} = \frac{Np - 1}{N - 1}$$

$$P\{X_2 = 1 \mid X_1 = 0\} = \frac{Np}{N - 1}$$

- Yani birinci elemanın karakteristiğe sahip olup olmadığı bilgisi ikinci elemanı etkiler.

Sınırlı Popülasyondan Örnekleme

- Ama eğer *N* çok büyükse, bu etki çok küçük olacaktır. Örneğin, *N*=1.000 ve *p*=0,4 için bu koşullu olasılık, koşulsuz olasılığa çok yakın olur.

$$P\{X_2 = 1 \mid X_1 = 1\} = \frac{399}{999} = 0,3994$$

$$P\{X_2 = 1 \mid X_1 = 0\} = \frac{400}{999} = 0,4004$$

$$P\{X_2 = 1\} = 0,4$$

- Yani *N* eğer *n*'e göre çok büyük ise X_1, X_2, …, X_n değerleri yaklaşık olarak birbirinden bağımsızdır ve X toplamını *n* ve *p* parametrelerine sahip bir Binomial rastgele değişken gibi düşünebiliriz.

Örnek Soru 5

- Nüfusun gelecek seçimlerde %45 oranında belirli bir partiyi desteklediğini varsayın. Eğer nüfustan rastgele 200 kişi seçilirse,
- Bu partiyi destekleyeceklerin sayısına ait beklenti ve standart sapma nedir?
- Yarısından fazlasının bu partiyi desteklemesi ihtimali nedir?

Örnek Soru 5

- Nüfusun gelecek seçimlerde %45 oranında belirli bir partiyi desteklediğini varsayın. Eğer nüfustan rastgele 200 kişi seçilirse,
- Bu partiyi destekleyeceklerin sayısına ait beklenti ve standart sapma nedir?
  - $E[X] = 200 \times 0,45 = 90$ ve $SD(X) = \sqrt{(200 \times 0,45 \times 0,55)} = 7,0356$
- Yarısından fazlasının bu partiyi desteklemesi ihtimali nedir?

Örnek Soru 5

- Nüfusun gelecek seçimlerde %45 oranında belirli bir partiyi desteklediğini varsayın. Eğer nüfustan rastgele 200 kişi seçilirse,
- Bu partiyi destekleyeceklerin sayısına ait beklenti ve standart sapma nedir?
  - $E[X] = 90$ ve $SD(X) = 7{,}0356$
- Yarısından fazlasının bu partiyi desteklemesi ihtimali nedir?
  - Bir program vasıtası ile Binomial ile çözersek $P\{X \geq 101\} = 0{,}0681$.
  - Normal yakınsama ile çözersek:
  - $P\{X \geq 101\} = P\{X \geq 100{,}5\}$
    $= P\{(X-90)/7{,}0356 \geq (100{,}5-90)/7{,}0356\} = P\{Z \geq 1{,}4924\}$
    $= 0{,}0678$