Marmara Üniversitesi

İstatistik Bölümü

Örnekleme Teorisi (6)

Doç. Dr. Atıf Evren

Farkların Örnekleme Dağılımı

Bir ana kütleden seçmiş olduğumuz n_1 hacimli bir örneğe ait S_1 istatistiğini ve diğer bir ana kütleden çekilmiş n_2 hacimli bir örneğe ait S_2 istatistiğini hesapladığımızı düşünelim. Bu işlemi çok sayıda örnek için yaptığımız takdirde $S_1 - S_2$' lerden meydana gelen bir bölünme elde ederiz ki bu bölünme ana kütlenin bölünmesi normal değilse dahi örneklerin mevcutları 30 veya daha büyük olduğu halde normale yaklaşacaktır. Bölünmenin ortalaması $\mu_{S_1-S_2}$, standart sapması $\sigma_{S_1-S_2}$ olarak tanımlanmakta ve bu istatistiklerin değerleri aşağıdaki formüller yardımı ile hesaplanmaktadır.

$$\mu_{S_1-S_2}=\mu_{S_1} - \mu_{S_2}$$

$$\sigma_{S_1-S_2}=\sqrt{(\sigma_{S_1})^2 + (\sigma_{S_2})^2}$$

Ortalamaların Farkının Dağılımı

Elimizde iki ayrı ana kütleden seçilmiş $\overline{X_1}$ ve $\overline{X_2}$ gibi örnek ortalamaları varsa bunların aritmetik ortalaması ve standart sapması yukarıdakilerine benzer şekilde

$$\mu_{\overline{X_1}-\overline{X_2}}=\mu_{\overline{X_1}} - \mu_{\overline{X_2}}$$

$\sigma_{\bar{X}_1-\bar{X}_2}=\sqrt{(\sigma_{\bar{X}_1})^2 + (\sigma_{\bar{X}_2})^2}$ olarak belirlenecektir. Fakat $\bar{X}$'lerin dağılımının varyansı

$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ olduğuna göre ortalama farklarının bölünmesinin standart hatasının formülü

$\sigma_{\bar{X}_1-\bar{X}_2}=\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ olur.

Örnek: iki imalatçı A ve B pil imal etmektedirler. Belirli voltaja sahip A marka piller devamlı kullanıldıkları halde ortalama 79 saat , aynı voltaja sahip B marka piller ise aynı şartlar altında 70 saat dayanmaktadırlar. Standart sapmalar sırasıyla 6 saat ve 3 saattir A marka piller arasından tesadüfi olarak 50 birimlik ve B marka piller arasından 80 birimlik örnekler seçerek test edersek, A marka pillerin B marka pillerden en az 11 saat fazla dayanması olasılığını hesaplayınız.

Çözüm:

$$P(\overline{X_A} - \overline{X_B} \geq 11) = P(\frac{\overline{X_A}-\overline{X_B}-(\mu_A-\mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}} \geq \frac{11-(79-70)}{\sqrt{\frac{6^2}{50}+\frac{3^2}{80}}})$$

$$P(Z \geq 2.19) = 0,0143$$

Bu sonuçlar sonlu bir ana kütleden iadeli olarak veya sonsuz bir ana kütleden seçim yapıldığı takdirde geçerlidir. Sonlu kütlelerden seçim yapıldığı veya $\frac{n}{N} \geq 0,05$ olduğu durumda standart hatanın

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ formülü ile hesaplanması gerekir.

Oranların Farkının Dağılımı

Bölünmeleri binoma uyan iki ayrı ana kütleden elde edilmiş örnek oranlarının farklarına ait ortalama ve standart sapma aşağıdaki şekilde hesaplanır.

$$\mu_{\hat{p}_1-\hat{p}_2}=\mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$$

$$\sigma_{\hat{p}_1-\hat{p}_2}=\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Örnek: Bir şehirdeki ailelerin % 70'inin televizyon sahibi olduğu bilinmektedir. 150'şer ailelik iki tesadüfi şekilde seçilmiş örneğin oranları arasında % 15 veya daha fazla fark bulunma olasılığını hesaplayınız.

Çözüm:

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 = 0$$

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0,70*0,30}{150} + \frac{0,70*0,30}{150}} = 0,053$$

$$P(\widehat{p_1} - \widehat{p_2} \geq 0,15) = P\left(\frac{\widehat{p_1} - \widehat{p_2} - (p_1 - p_2)}{\sigma_{\hat{p}_1-\hat{p}_2}} \geq \frac{0,15}{0,053}\right)$$

$$= P(Z \geq 2,83) = 0,0023$$

Standart Hata Kavramı ve Özellikleri

İstatistikte önem taşıyan başlıcs iki hata vardır: **Sistematik** ve **tesadüfi (rassal)** hatalar. Örneklemede rassallık sağlanamadığı zaman **sistematik** ya da tek yönde işleyen ve etkileri birbirini ortadan kaldırmayan hatalar yapılmış olur ki bunlar örneğin temsili olmasını önler. Örneğe dayanarak yapılan tahminler de hatanın yönüne göre gerçek değerin altında veya üstünde olacaklardır. Örneğin ana kütle içine bütün birimler dahil edilmemişse bir sistematik hata payı olacaktır.

Bir ana kütle içinden çok sayıda örnek seçtiğimiz ve bunların aritmetik ortalamalarını hesapladığımız zaman ortalamaların birbirlerinden ve ana kütle ortalamasından farklı olduklarını görürüz. Bu farklar ya da hatalar **tesadüfi hatalardır** ve her iki yönde etkili olduklarından birbirlerinin etkilerini yok edebilirler.

Bir istatistiğin örnekleme bölünmesinin standart sapması olarak tanımlamış olduğumuz **standart hata** bu tesadüfi hatanın bir ölçütüdür. Standart hatanın ana kütle standart sapması bilinmediği hallerde örnek standart sapmasından yararlanılarak nasıl bulunabileceği konusu üzerinde duralım.

Ana kütlenin tamamı hakkında bilgi toplanmadıkça genellikle kütlenin standart sapmasını bilmek mümkün değildir. Buna karşılık örneğe ait standart sapma hesaplanabilir. Ana kütle standart sapmasının bilinmediği hallerde

i) $n \geq 30$ ise s doğrudan doğruya σ yerine konularak standart hata hesaplanır, veya

ii) düzeltilmiş standart sapma $\hat{s} = \sqrt{n/(n-1)}s$ formülü ile daha yakın bir tahmin yapılmaya çalışılmaktadır. Bu durumda σ yerine s konularak standart hata

$\widehat{\sigma_{\bar{X}}}=\frac{\hat{s}}{\sqrt{n}}$ formülü ile aşağıdaki gibi hesaplanır:

$$\widehat{\sigma_{\bar{X}}}=\frac{\sqrt{n/(n-1)}s}{\sqrt{n}}$$

Örnekleme oranı $n/N > 0{,}05$ ise

$\widehat{\sigma_{\bar{X}}}=\frac{\sqrt{n/(n-1)}s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ formülü kullanılacaktır.

Genellikle nüfus, ücretler, kalite kontrolü gibi örnekleme araştırmalarında örnek mevcudunun ana kütle mevcuduna oranı % 5'in altında olacağından düzeltme faktörüne gerek yoktur.σ yerine s'in konması ile yapılacak tahmin tam ve kesin sonuç vermese de içerdiği hata payı $n \geq 30$ olması durumunda çok küçük olacağından bu yönted pratik öneminden bir şey kaybetmez.

Örnek: Bir mahallede oturan 620 aile arasından tesadüfi olarak 120 birimlik bir örnek seçilerek bu mahalledeki ailelerin aylık gelirleri ile ilgili bir araştırma yapılacaktır. Seçilen 120 ailenin aylık gelir ortalaması 6,5 milyon TL ve standart sapma 1,2 milyon TL'dir. Ortalamaların örnekleme bölünmesinin standart hatası nedir?

Çözüm:

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}} = \frac{1{,}2}{\sqrt{120-1}}\sqrt{\frac{620-120}{620-1}} = 0{,}09885M\ TL = 98850\ TL$$

Standart Hatayı Etkileyen Faktörler

Standart hatayı örneklere ait istatistiklerin ana kütle parametrelerine olan uzaklıklarının bir ölçüsü olarak da tanımlayabiliriz. Bu durumda örneğe ait bir değerin anlamlı olabilmesi standart hatanın da bilinmesine bağlıdır. $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ olarak tanımlanan standart hatanın büyüklüğü iki etkene bağlıdır:

i)ana kütlenin veya örneğin değişkenliği,

ii)örneklem hacmi.

Standart hata örnek büyüklüğü ile teers orantılı olduğuna göre standart hatanın küçültülebilmesi örnek hacmini **arttırmakla** olasıdır. Formüle göre standart sapmanın küçültülmesi ile de standart hata küçülecektir. Ancak gerek an kütle gerekse örnek standart sapması araştırmacı için veri olduklarından **değiştirilemezler.**

Az önceki örnekte örnek büyüklüğü 120 ve standart hata 98850 TL idi. Standart hatayı 50 bin TL'ye düşürebilmek için

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n-1}}\sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{\bar{X}}{}^2 = \frac{s^2}{n-1}\frac{N-n}{N-1}$$

$$(0{,}05)^2 = \frac{1{,}2^2}{n-1}\frac{6200-n}{6200-1} \Rightarrow n = 299{,}36 \cong 299$$

Görüldüğü gibi standart hatayı yarıya indirebilmek için örnek mevcudunun yaklaşık 2,5 katına çıkarılması gerekmektedir. Örnek mevcudunun arttırılması araştırmanın maliyetini yükselteceğinden araştırmacı için optimal kabul edilecek bir standart hata değeri ve örnek hacmi üzerinde karara varmak bir sorun oluşturmaktadır.

Örneğin ana kütleye oranla küçük olduğu hallerde istenilen standart hatayı verecek örneklem büyüklüğü

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \Longrightarrow n = \frac{\sigma^2}{{\sigma_{\bar{X}}}^2}$$

Örneğin ana kütleye oranı büyük olduğu hallerde ise

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \Longrightarrow n = \frac{N\sigma^2}{{\sigma_{\bar{X}}}^2(N-1)+\sigma^2}$ formülü ile örneklem hacmi belirlenir.

Aralık Tahminleri

Bir parametrenin aralık tahmini, örnekten yararlanılarak bulunan iki değer arasındaki aralığı ifade etmekte ve ana kütle parametresinin bu iki değer arasında bulunması beklenmektedir. Bu aralık, tahmini istenilen parametrelerin o aralıkta bulunması olasılığına göre belirlenmiştir. Aralık tahmininin sağladığı en önemli avantaj parametrenin ne ölçüde kesinlikle tahmin edildiğini belirtebilmesidir. Örneğin aralık dar ise kesinlik veya isabet derecesi yüksek olmaktadır. Bu şelilde belirtilen aralık tahminleri güven aralığı olarak da tanımlanmaktadır.

Ana Kütle Ortalamasının %100(1-α)'lık Güven Aralığı Tahmini

$X_1, X_2, \dots, X_n$ ortalaması μ, standart sapması σ olan bir dağılmdan çekilen bağımsız bir örnek olsun. **Büyük örnekler için ($n \geq 30$ için)** ana kütle ortalamsı için %100(1-α)'lık güven aralığı

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

ile verilir. Burada ana kütle parametresi **σ'nın bilinmemesi halinde**, eldeki örneklemin **büyük** olması varsayımından hareketle anakütle standart sapması yerine örnek standart sapması kullanılabilir. Bu durumda aralık tahmini

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

şeklinde gerçekleştirilir.

Örneklerin **bağımlı olduğu halde** (iadesiz örnekleme) aralık tahmini (**Büyük örnekler için ($n \geq 30$ için)**)

$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha$ biçiminde gerçekleştirilir.

Buna ek olarak anakütle dağılımının **normal ya da yaklaşık normal olduğu** durumlarda örneklem hacminin ne olduğundan bağımsız olarak örneklemenin iadeli ya da iadesiz olmasına göre

$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$ (iadeli seçim)

$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha$ (iadesiz seçim)

formülleri kullanılır.

Son olarak ana kütle dağılımının **normal** ya da yaklaşık normal olduğu ama **örneklem hacminin küçük** olduğu $(n < 30)$ ve anakütle standart sapmasının **bilinmediği** durumlarda

$P\left(\bar{X} - T_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \bar{X} + T_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$ formülü kullanılır. Burada dikkat edilecek olursa standart normal dağılımın güven katsayıları n-1 serbetlik dereceli student-T dağılımının güven katsayıları kullanılmaktadır. Bu aralık literatürdeki bazı çalışmalarda

$P\left(\bar{X} - T_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}} < \mu < \bar{X} + T_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}}\right) = 1 - \alpha$ şeklinde de ifade edilmektedir.

Örnek: Bir yüksek okuldaki kız öğrenciler arasından rassal seçilmiş 64 öğrencinin boy uzunluklarının ortalaması 162 cm. ve standart sapma 4 cm. dir. Okuldaki kızların boy uzunluklarının ortalaması için % 95'lik güven aralığı nedir?

Çözüm:

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(162 - 1{,}96\frac{4}{\sqrt{64}} < \mu < 162 + 1{,}96\frac{4}{\sqrt{64}}\right) = 0{,}95$$

$$P(161{,}02 < \mu < 162{,}98) = 0{,}95$$

Aşağıdaki tabloda belirli güven düzeylerine karşılık geken Z değerleri verilmektedir:

Güven Düzeyi	0,99	0,98	0,95	0,90	0,80
Z değeri	2,58	2,33	1,96	1,645	1,28

Örnek: 200 parçalık bir yığın içinden tesadüfi olarak seçilmiş 49 parçalık bir örneğe ait ağırlık ortalaması 70 gram ve standart sapma 3 gramdır.

a)200 parçalık yığının ortalaması için % 98'lik güven sınırları nelerdir?

b)Ana kütle ortalamasının $70 \pm 0{,}75$ arasında kalabileceğini hangi güven düzeyi için söyleyebiliriz?

Çözüm:

a)Örnek hacmi ana kütle hacmine oranla büyük olduğundan standart hatanın hesaplanması için $\sqrt{\frac{N-n}{N-1}}$ düzeltme faktörünü kullanmak gerekmektedir.

$$P\left(\bar{X} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right) = 1 - \alpha$$

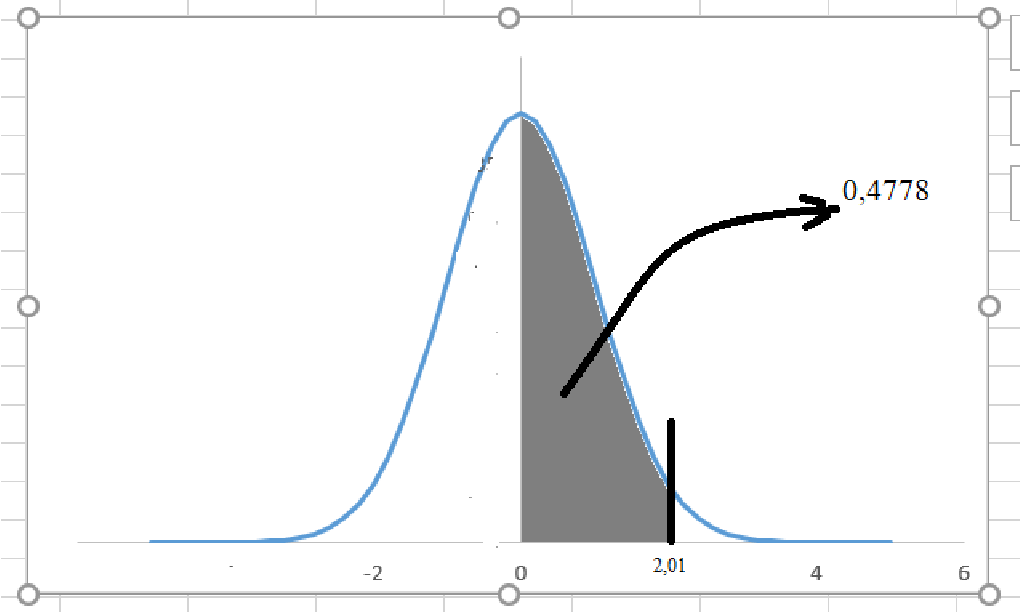
$$P\left(70 - 2{,}33\frac{3}{\sqrt{49}}\sqrt{\frac{200-49}{200-1}} < \mu < 70 + 2{,}33\frac{3}{\sqrt{49}}\sqrt{\frac{200-49}{200-1}}\right) = 0{,}98$$

$P(69{,}134 < \mu < 70{,}866) = 0{,}98$

b) $Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 0{,}75 \Longrightarrow Z_{\frac{\alpha}{2}}\frac{3}{\sqrt{49}}\sqrt{\frac{200-49}{200-1}} \Longrightarrow Z_{\frac{\alpha}{2}} = 2{,}01$

Bu değer tabloda alanın 0,4778'ine karşılık geldiği için istenilen güven düzeyi

2*0,4778=0,956=% 95,6 olur. Bu durum aşağıdaki grafik yardımı ile de izlenebilir:



Ana Kütle Oranının %100(1-α)'lık Güven Aralığı Tahmini

Bilindiği gibi örnek büyüklüğünün yeterince büyük olduğu durumlarda, örnek oranının örnekleme dağılımı normale yakın olmaktadır. Bu dağılımın standar hatası

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ formülü ile hesaplanmaktadır. Bu durumda ana kütle oranının

%100(1-α)'lık güven aralığı tahmini

$P\left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$ biçiminde olacaktır. Fakat ana kütle oranı p bilinmediği için örnek oranıi ana kütle oranının kestiricisi olarak kullanılacak ve elde edilecek yaklaşık güven aralığı

$P\left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$ biçiminde ifade edilecektir.

Bu formül iadesiz örnekleme için örnekleme oranı (n/N), % 5'den büyük ise

$$P\left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}}\right) = 1-\alpha$$

halini alacaktır.

Örnek: Bir bölgeden rastgele seçilen 144 seçmenin % 20'sinin X partisini desteklediği gözlenmiştir. Bu bölgedeki X partisini destekleyen seçmen oranı için % 95'lik güven aralığı nedir?

Çözüm:

$$P\left(\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1-\alpha$$

$$P\left(0{,}20 - 1{,}96\sqrt{\frac{0{,}20 * 0{,}80}{144}} < p < 0{,}20 + 1{,}96\sqrt{\frac{0{,}20 * 0{,}80}{144}}\right) = 0{,}95$$

$P(0{,}136 < p < 0{,}264) = 0{,}95$ bulunur.

Ortalama Farkları için Güven Aralığı Tahminleri

Ortalamaları μ_1, μ_2, varyansları σ_1^2 , σ_2^2 olan iki ana kütleden çekilen n_1 ve n_2 büyüklüğünde iki bağımsız örneğin ortalamaları $\overline{X_1}$ ve $\overline{X_2}$ olsun. n_1 ve n_2 yeterince büyük ise iki ortalama farkının %100(1-α)'lık güven aralığı tahmini

$$P\left(\overline{X_1} - \overline{X_2} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{X_1} - \overline{X_2} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha$$

olur. σ_1^2 ve σ_2^2'nin bilinmemesi halinde n_1 ve n_2 yeterince büyük olduğu için yerlerine örnek varyansları s_1^2 ve s_2^2 konularak yaklaşık bir aralık tahminine gidilebilir.

$$P\left(\overline{X_1} - \overline{X_2} - Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{X_1} - \overline{X_2} + Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right) \cong 1-\alpha$$

Örnek: Aynı ürünü elde eden A ve B fabrikalarından yapılan imalattan 200 ve 150 birimlik örnekler seçilmiş ve bu örneklerin ömürlerinin sırası ile 118 ve 105 saat olduğu görülmüştür. Birinci örneğin standart sapması 12 saat, ikinci örneğin standart sapması da 6 saat olarak hesaplanmıştır. Bu verilere göre iki ana kütle ortalaması arasındaki farkın % 99'luk güven aralığını hesap ediniz.

Çözüm:

$$P\left(\overline{X_1}-\overline{X_2}-Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}<\mu_1-\mu_2<\overline{X_1}-\overline{X_2}+Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)\cong 1-\alpha$$

$$P\left(118-105-2{,}58\sqrt{\frac{12^2}{200}+\frac{6^2}{150}}<\mu_1-\mu_2<118-105+2{,}58\sqrt{\frac{12^2}{200}+\frac{6^2}{150}}\right)\cong 0{,}99$$

$$P(10{,}47<\mu_1-\mu_2<15{,}53)\cong 0{,}99$$

Oran Farkları için Güven Aralığı Tahminleri

Başarı oranları p_1 ve p_2 olan ve Bernoulli dağılan iki ana kütleden çekilen n_1 ve n_2 hacimli iki örneğin başarı oranları $\widehat{p_1}$ ve $\widehat{p_2}$ olsun. n_1 ve n_2'nin yeterince büyük olması durumunda iki ana kütle oranı farkı için %100(1-α)'lık güven aralığı tahmini

$$P(\widehat{p_1}-\widehat{p_2}-Z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}<p_1-p_2<\widehat{p_1}-\widehat{p_2}+Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}})=1-\alpha$$

olur. Fakat yukarıdaki formülasyonda $p_1 ve\ p_2$ bilinmemektedir. Dolayısıyla standart hata formülünde geçen bu parametrelerin örnek başarı oranları yardımı ile tahmin edilmesi ile aşağıdaki yaklaşık güven aralığı formülüne ulaşılır:

$$P(\widehat{p_1}-\widehat{p_2}-Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}<p_1-p_2<\widehat{p_1}-\widehat{p_2}+Z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}+\frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}})=1-\alpha$$

Benzer bir biçimde örnekleme yönteminin iadesiz ve $\frac{n_1}{N_1}$ ve $\frac{n_2}{N_2}$ örnekleme oranlarının % 5'den büyük olmaları durumunda bu formül

$$P\left(\widehat{p_1} - \widehat{p_2} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}\frac{(N_1-n_1)}{(N_1-1)} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}\frac{(N_2-n_2)}{(N_2-1)}} < p_1 - p_2 < \widehat{p_1} - \widehat{p_2} + Z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1}\frac{(N_1-n_1)}{(N_1-1)} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}\frac{(N_2-n_2)}{(N_2-1)}}\right)=1-\alpha$$

biçimini alır.

Örnek: Bir şehirde rastgele olarak seçilmiş 300 erkek arasından 150'si ve 250 kadın arasından 100'ü belirli bir gazeteyi okuduklarını bildirmişlerdir. Ana kütle oranları arasındaki fark için % 98'lik güven aralığını hesaplayınız.

Çözüm:

$\widehat{p_1} = \frac{150}{300} = 0,5;$ $\widehat{p_2} = \frac{100}{250} = 0,4$

$$P\left(\widehat{p_1} - \widehat{p_2} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} < p_1 - p_2 < \widehat{p_1} - \widehat{p_2} + Z_{\alpha/2}\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}\right)=1-\alpha$$

$$P\left(0,5 - 0,4 - 2,33\sqrt{\frac{0,5*0,5}{300} + \frac{0,4*0,6}{250}} < p_1 - p_2 < 0,5 - 0,4 + 2,33\sqrt{\frac{0,5*0,5}{300} + \frac{0,4*0,6}{250}}\right)=0,98$$

$P(0,004 < p_1 - p_2 < 0,196)=0,98$

Tahmin Hatasının Verilmesi Durumunda Örnek Büyüklüğünün Belirlenmesi

Daha önceki bölümlerde tartışıldığı gibi

$e = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ ve

$e = Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ formülleri yardımıyla örnek hacminin elde edilmesi mümkündür. Bu formüllerden verili bir hata miktarı için örneklem büyüklüğünün ne olması gerektiğini veren formüller elde edilebilir. Sözgelimi örnek ortalamasının dağılımı ile ilgili olarak

$$e = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Longrightarrow n = \frac{\left(Z_{\alpha/2}\right)^2\sigma^2}{e^2}$$

olur. Ana kütle varyansının bilinmediği durumlarda örnek varyansı ana kütle varyansının kestiriminde kullanılır. Bu durumda yukarıdaki formül

$n \cong \frac{\left(Z_{\alpha/2}\right)^2 s^2}{e^2}$ olur.

Örnek: Bir firmada çalışanların ortalama yılda kaç gün izin yaptıkları araştırılmaktadır. Daha önceden 100 kişilik bir örnekten hareketle örnek standart sapması 12 olarak belirlenmiştir. Bulunacak olan ortalamanın gerçek ana kütle ortalamasından en fazla 3 gün sapması istenilmekte ise % 95 güven düzeyi için kaç kişilik bir örnek seçilmelidir?

Çözüm:

$$n \cong \frac{\left(Z_{\alpha/2}\right)^2 s^2}{e^2}$$

$n \cong \frac{(1{,}96)^2 12^2}{3^2} = 61{,}46 \approx 61$ bulunur.

Eğer sonlu bir ana kütleden örnek seçimi yapılacak olursa standart hatanın belirlenmesinde

$\sqrt{\frac{N-n}{N-1}}$ düzeltme faktörünün kullanılması gerektiğinden örneklem büyüklüğü

$n = \frac{N\left(Z_{\alpha/2}\right)^2 \sigma^2}{(N-1)e^2+(Z_{\frac{\alpha}{2}})^2\sigma^2}$ olur. Yine ana kütle varyansının bilinmediği ve örneklem büyüklüğünün yeterli olduğu durumlarda ana kütle varyansının kestirimi olarak örnek varyansı kullanılır:

$$n \cong \frac{N\left(Z_{\alpha/2}\right)^2 s^2}{(N-1)e^2 + (Z_{\frac{\alpha}{2}})^2 s^2}$$

Örnek: Az önceki örnekte firmada çalışanların toplam sayısının 800 olduğu varsayılsın. Bu durumdaki örnek hacmi

$$n \cong \frac{N\left(Z_{\alpha/2}\right)^2 s^2}{(N-1)e^2 + (Z_{\frac{\alpha}{2}})^2 s^2}$$

$$n \cong \frac{800(1{,}96)^2 12^2}{(800-1)3^2+(1{,}96)^2 12^2} = 57{,}14 \approx 57$$ bulunur.

Bernoulli dağılımı ile ilgili olarak da

$$e = Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \Longrightarrow n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 p(1-p)}{e^2}$$

formülü kullanılır. Benzer bir biçimde anakütle oranı bilinmediğinde ise uygun bir ana kütle oranı tahmini mevcut ise kullanılarak

$$n \cong \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 \hat{p}(1-\hat{p})}{e^2}$$ formülü kullanılabilir.

Ana kütle oranının önsel olarak tahmin edilmesinin mümkün olmadığı durumda ise örneklem hacmini maksimum kılacak p değeri seçilir. Başka bir deyişle

$$Maks\, n = Maks\, \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 p(1-p)}{e^2}$$

$$\frac{d}{dp}\left[\frac{\left(Z_{\frac{\alpha}{2}}\right)^2 p(1-p)}{e^2}\right] = 0 \Longrightarrow \frac{d(p(1-p)}{dp} = \frac{d\left(p-p^2\right)}{dp} = 0 \Longrightarrow p = \frac{1}{2}$$ bulunur.

Bu değer n$= \frac{\left(Z_{\frac{\alpha}{2}}\right)^2 p(1-p)}{e^2}$ formülünde yerine konacak olursa

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2}{4e^2}$$ bulunur.

Örnek: Bir ürünün kullanım oranını belirlemek için örnekleme yapılacaktır. Eğer ana kütle oranı ile ilgili herhangi bir bilgi yoksa ve tahmin hatasının en fazla % 2 olması isteniyorsa % 90 güven düzeyin için kaç kişilik örnek oluşturulmalıdır?

Çözüm:

$$n = \frac{\left(Z_{\frac{\alpha}{2}}\right)^2}{4e^2}$$

$$n = \frac{(1{,}645)^2}{4(0{,}02)^2} = 1691$$

Sonlu ana kütleden seçim yapılması veya ana kütlenin örneğe göre büyük olmaması halinde örneklem hacmi

$$n \cong \frac{N\left(Z_{\frac{\alpha}{2}}\right)^2 p(1-p)}{(N-1)e^2+\left(Z_{\frac{\alpha}{2}}\right)^2 p(1-p)}$$ formülü kullanılır.

Yine ana kütle oranının bilinmediği durumda p değeri yerine ½ değeri konacak olursa

$$n \cong \frac{N\left(Z_{\frac{\alpha}{2}}\right)^2}{4(N-1)e^2+\left(Z_{\frac{\alpha}{2}}\right)^2}$$ formülü elde edilir.

Örnek: Az önceki örnekte potansiyel kullanıcı sayısının 6000 olduğunun bilinmesi halinde

$$n \cong \frac{6000(1{,}645)^2}{4(6000-1)0{,}02^2+(1{,}645)^2} = 1319{,}54 \approx 1320$$ bulunur.