14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Bahar 2009

14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Ders Notları 19

Konrad Menzel

28 Nisan 2009

1. Maksimum Olabilirlik Tahmin: İlave Örnekler

Örnek 1. *Varsayalım ki* $X \sim N(\mu_0, \sigma_0^2)$*'dir ve bir i.i.d. örneklem* $X_1, \ldots, X_n$*'den* μ *ve* σ^2 *parametrelerini tahmin etmek istiyoruz. Olabilirlik fonksiyonu şöyledir:*

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

Log-olabilirliği maksimize etmenin daha kolay olduğunu ortaya koyabiliriz,

$$\begin{aligned}
\log \mathcal{L}(\theta) &= \sum_{i=1}^{n} \log \left\{ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \right\} \\
&= \sum_{i=1}^{n} \left\{ \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(X_i-\mu)^2}{2\sigma^2} \right\} \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i-\mu)^2
\end{aligned}$$

Maksimumu bulmak için, μ *ve* σ^2*'ye göre türevleri alıp sıfıra eşitleriz:*

$$0 = \frac{1}{2\widehat{\sigma^2}} \sum_{i=1}^{n} 2(X_i - \hat{\mu}) \Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Aynı şekilde,

$$0 = -\frac{n}{2} \frac{2\pi}{2\pi\widehat{\sigma^2}} + \frac{1}{2\left(\widehat{\sigma^2}\right)^2} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \Leftrightarrow \widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Hâlihazırda, bu tahmin edicinin σ_0^2 için sapmasız olmadığını gösterdiğimizi hatırlayınız, bu nedenle genel olarak Maksimum Olabilirlik Tahmin Edicileri sapmasız olmak zorunda değildir.

Örnek 2. *Uniform dağılımlı örneğe geri dönelim: varsayalım ki $X_i \sim U[0, \theta]$'dır ve θ'nın tahmini ile ilgileniyoruz. Momentler yöntemi tahmin edicisi için aşağıdakini görebilirsiniz,*

$$\mu_1(\theta) = \mathbb{E}_\theta[X] = \frac{\theta}{2}$$

böylece bunu örneklem ortalamasına eşitleyerek aşağıdakini elde ederiz:

$$\hat{\theta}_{MoM} = 2\bar{X}_n$$

Maksimum olabilirlik tahmin edicisi nedir? Açıkçası, biz herhangi bir $\hat{\theta} \leq \max\{X_1, \ldots, X_n\}$ almayacağız çünkü $\hat{\theta}$'dan büyük gerçekleşmiş bir örneklemin $\hat{\theta}$ altında sıfır olasılığı vardır. Biçimsel olarak, olabilirlik

$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{eğer} \quad 0 \leq X_i \leq \theta \quad \text{için} \quad i = 1, \ldots, n \quad \text{ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

$\theta \leq \max\{X_1, \ldots, X_n\}$'nin herhangi bir değeri maksimumu olamaz çünkü bütün o noktalarda $L(\theta)$'in sıfır olduğunu görebiliriz. Aynı zamanda, $\theta \geq \max\{X_1, \ldots, X_n\}$ için olabilirlik fonksiyonu θ'da kesin azalandır ve bu nedenle aşağıda ifade edildiği gibi maksimumdur

$$\hat{\theta}_{MLE} = \max\{X_1, \ldots, X_n\}$$

1 olasılıkla $X_i < \theta_0$ olduğu için, maksimum olabilirlik tahmin edicisi de 1 olasılıkla θ_0'dan düşük olacaktır, böylece sapmasız değildir. Daha da açık olmak gerekirse, $X_{(n)}$'in p.d.f.si aşağıdaki gibi verilir:

$$f_{X_{(n)}}(y) = n[F_X(y)]^{n-1} f_X(y) = \begin{cases} \frac{n}{\theta_0}\left(\frac{1}{\theta_0}\frac{y}{\theta_0}\right)^{n-1} & \text{eğer} \quad 0 \leq y \leq \theta_0 \quad \text{ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

Böylece,

$$\mathbb{E}[X_{(n)}] = \int_{-\infty}^{\infty} y f_{X_{(n)}}(y) dy = \int_0^{\theta_0} n\left(\frac{y}{\theta_0}\right)^n dy = \frac{n}{n+1}\theta_0$$

Çok kolay bir şekilde bir sapmasız tahmin edici, $\hat{\theta} = \frac{n+1}{n} X_{(n)}$*, oluşturabiliriz*.

1.1. MLE'nin Özellikleri

Aşağıdakiler sadece MLE için elde edilen temel teorik sonuçların özetidir(bu aşamada ispatları yapmayacağız):

- Tutarlı tahmin ediciler grubunda etkin bir tahmin edici varsa, MLE onu oluşturur.
- Belli düzenleyici koşullar altında, MLE asimptotik olarak normal dağılım olabilir (bu esas itibariyle Merkezi Limit Teoreminin bir uygulamasından gelmektedir).

Maksimum olabilirlik her zaman yapılması gereken en iyi şey mi? Hayır

- sapmalı olabilir
- genellikle hesaplanması zordur
- ilgili dağılım ile ilgili yanlış varsayımlara karşı çok hassas olabilir.

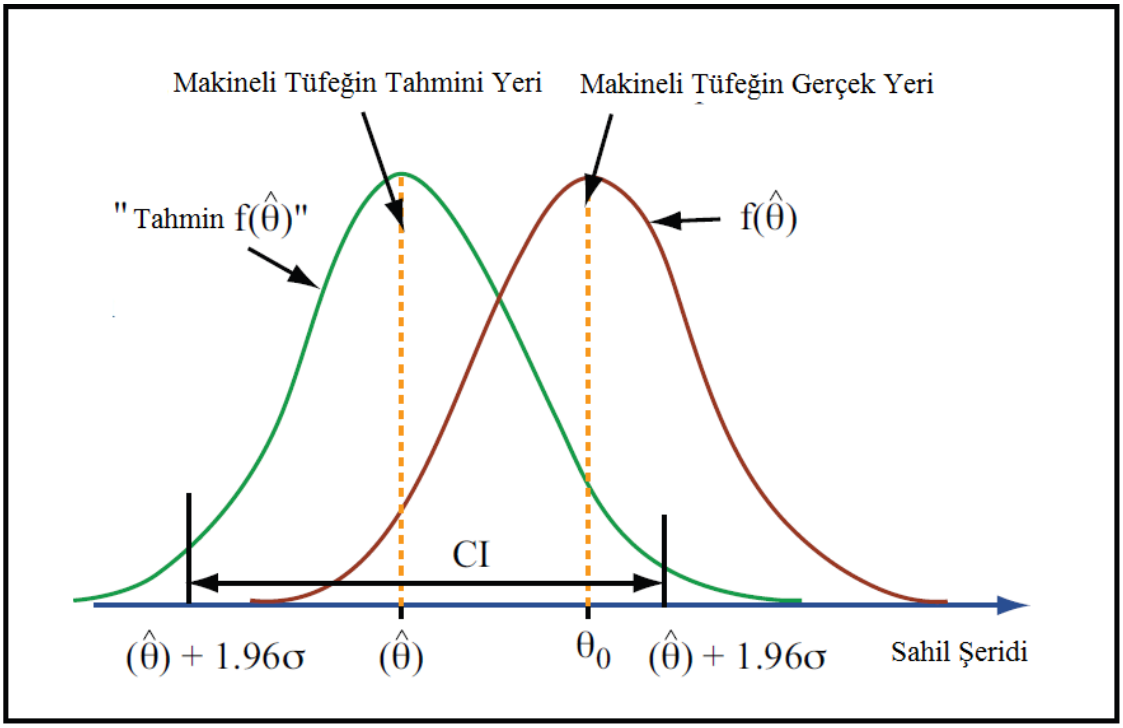
2. Güven Aralığı

Tahminin değeri ve onun doğruluğu(standart hatası tarafından verildiği gibi) hakkındaki bilgileri birleştirmek için, genellikle yapılan, bir tahminin etrafında muhtemelen gerçek değeri içeren bir aralık belirlemektir.

Örnek 3. *Varsayalım ki deniz kuvvetlerinin bir topçekerinin(bot) kaptanı kıyı şeridi boyunca bir sahil koruma hattı oluşturacaktır, fakat ondan önce denizden doğrudan görülmeyen sahildeki bir makineli tüfeğin yok edilmesi veya en azından ağır tahribata uğratılması gerekecektir.*

Bota halihazırda sahilden birkaç kere atış açılır ve mermilerin geldiği yöne dayanarak, kaptan silahın konumu hakkında bir tahmin $\hat{\theta}$ *oluşturur. Tahmin, gerçek* θ_0 *konumun etrafında* σ_β^2 *varyanslı bir normal dağılımdır.*

Kaptan, sahillin bir aralığına füzelerle yaylım ateşinde bulunarak o alandaki her şeyi yok edebilir. Kaptan, sahillin hangi aralığına ateş edeceğini nasıl belirleyebilir ki %95 olasılıkla makineli tüfeğin orada olup tahrip olacağından emin olabilir ve böylece birlikleri güvenli bir şekilde sahile çıkarabilir?



Kaynak: MIT OpenCourseWare

Normal dağılım için, olasılık yığının %95'inin ortalamanın her iki tarafındaki 1.96 standart sapmalık aralığın içinde olduğunu biliyoruz. Böylece, eğer kaptan CI = [$\hat{\theta}$ – 1.96σ, $\hat{\theta}$ + 1.96σ] aralığı için ateş emri verirse, $\hat{\theta}$'in $\theta_0 \in$ CI olma olasılığı %95'tir.

Daha önce sadece gerçek parametre θ_0 değerine yakın değer veren tek fonksiyon $\hat{\theta}(X_1, \ldots, X_n)$ arıyor iken, şimdi belli bir değere eşit veya daha yüksek olasılıkla gerçek parametre değerini kapsayan (içeren) iki fonksiyon, $A(X_1, \ldots, X_n) < B(X_1, \ldots, X_n)$, oluşturmaya çalışacağız.

Tanım1. *Parametre θ_0 için bir 1-α'lık güven aralığı A(.) ve B(.) gibi veriye-dayalı iki fonksiyona bağlı bir aralıktır [$A(X_1, \ldots, X_n)$, $B(X_1, \ldots, X_n)$]. Yani,*

$$P_{\theta_0}\left(A(X_1,\ldots,X_n) \le \theta_0 \le B(X_1,\ldots,X_n)\right) = 1-\alpha$$

Bu fonksiyonlar benzersiz değildir fakat teamüllere bağlı olarak, A ve B'yi $\alpha/2$ olasılığı aralığın her iki tarafına eşit düşecek şekilde seçiyoruz.

Bir güven aralığının, $[A(x_1, \ldots, x_n), B(x_1, \ldots, x_n)]$, gerçekleşmesi için, $P(A(x_1, \ldots, x_n) \le \theta_0 \le B(x_1, \ldots, x_n) = 1 - \alpha$ olduğunu söylemek mantıklı değil, çünkü aralığın limitleri ve gerçek parametre şimdi reel sayılardır, böylece örneklemin gerçekleşmesi veri iken,

tahmin edilen aralık ya θ_0 kapsar(1 olasılıkla) ya da kapsamaz. Gerçek parametre veri iken rasgele olan güven aralığıdır, yoksa θ_0 değil.

Aşağıdaki kendisi için güven aralığı oluşturmak istediğimiz en yaygın durumdur.

Örnek 4. *Varsayalım ki* $\hat{\theta} \sim N(\theta_0, \sigma^2)$*'dir ve bir 1-*$\alpha$*'lık bir güven aralığı oluşturmak istiyoruz. Eğer* $z_{1-\alpha/2}$ *standart normal dağılımın 1- (*α*/2) ondalığı ise yani* $\Phi(z_{1-\alpha/2}) = 1-(\alpha/2)$ *ise, o zaman aşağıdakinin*

$$CI = [\hat{\theta} - \sigma z_{1-\alpha/2}, \hat{\theta} + \sigma z_{1-\alpha/2}]$$

θ_0*'ı aşağıdaki olasılıkla kapsadığını kontrol edebiliriz,*

$$\begin{aligned} P_{\theta_0}\left(\hat{\theta} - \sigma z_{1-\alpha/2} \leq \theta_0 \hat{\theta} + \sigma z_{1-\alpha/2}\right) &= P_{\theta_0}\left(-z_{1-\alpha/2} \leq \frac{\theta_0 - \hat{\theta}}{\sigma} \leq z_{1-\alpha/2}\right) \\ &= \Phi(z_{1-\alpha/2}) - \Phi(-z_{1-\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \end{aligned}$$

burada $\frac{\sigma_0 - \hat{\sigma}}{\sigma}$ $\hat{\theta}$*'nın standardizasyonu olduğu için standart normal dağılımlıdır.*

Böylece eğer bir %95'lik güven aralığı istiyorsak, $z_{1-\alpha/2} = z_{0.975} = 1.96$*'dır, bu nedenle güven aralığı* $\hat{\theta} \pm 1.96\sigma$ *ile verilir.*

Bu güven aralığını elde etmenin en yaygın yoludur, bu nedenle bunun nasıl işlediğini anlamanız gerekiyor.

Örnek 5. *Anket sonuçları genellikle bir "hata payı" ile rapor edilir. Örneğin Gallup'un 18 Nisan raporuna göre seçmenlerin %46'sının McCain'e karşı Clinton'e, %44'ünün McCain'e oy vereceğini, %10'nun ise ya her ikisi için de oy kullanmayacağını ya da herhangi bir fikri olmadığını söylemiştir. Bu sonuçlar 4385 kişiyle yapılan görüşmeye dayanmaktadır ve rapor ayrıca "ulusal yetişkinlerin toplam örneklemine dayalı sonuçlar için, %95 güvenilirlikle maksimum örneklem hata payı yüzde iki olduğu söylenebilir" ifadesine yer vermiştir. Bu ne anlama gelmektedir? – Eğer bir adayın gerçek oy oranı p ise, n sayıdaki seçmen örnekleminde ortalama payın varyansı VAR(*$\overline{X}_n$*) =* $\frac{p(1-p)}{n}$ *'dir. Bu varyansın p = 0.5 için en yüksek olduğunu kendiniz de kontrol edebilirsiniz. Dolaysıyla 4385 görüşmeli bir örneklem için, maksimum standart sapma* $\sqrt{Var(\overline{X}n)} \leq \sqrt{\frac{0.5(0.5)}{4385}} \approx \%0.76$*'tir.*

Merkezi Limit Teoremine göre, $\overline{X}_n$ *yaklaşık olarak normal dağılımlıdır ve bir normal dağılım için %95'lik bir olasılık kütlesinin ortalamanın 1.96 standart sapmalık aralığında*

yer aldığını daha önce görmüştük. Bu nedenle, [$\overline{X}_n$ -1.96(0.76) , $\overline{X}_n$ + 1.96(0.76)] aralığı gerçek oy oranını %95'ten daha büyük bir olasılıkla içerecektir. Seçmenin daha küçük alt grupları için hata payı daha büyük olacaktır.

Örnek 6. *Bir laboratuvar bir davada kanıt olarak kullanılabilecek bir kan örneği üzerinde kimyasal analiz yapmaktadır. Kanıt olarak kabul edilebilmesi için, bazı maddelerin mevcudiyetinin %90'lık güven aralığında % 0.001 g/ml'den daha az olması gerekir. Analizler için kullanılan makine gerçek değer etrafında standart sapması σ = 0.05g/ml olan normal dağılımlı sonuçlar vermektedir. %90'lik güven aralığının 0.001 g/ml'den az olduğundan emin olmak için kaç tane sonuç almamız gerekir?*

%95'lik güven aralığının genişliği şöyledir:

$$l = 2\frac{\sigma}{\sqrt{n}}\Phi^{-1}(0.95) \approx 2\frac{0.005}{\sqrt{n}}1.645 = \frac{0.01645}{\sqrt{n}}$$

Dolayısıyla, $l \leq 0.001$ olması için, $n \geq 16.45^2 = 270.6025$'e ihtiyacımız var, bu nedenle de en az 271(bağımsız) sonuç almamız gerekir.

Sonraki örnek tahmin edicinin dağılımının normal olmadığı durumlarda güven aralığı oluşturmanın bir yolunu göstermektedir.

Örnek 7. *Varsayalım ki $X_1, \ldots, X_n$ i.i.d.'dir, dağılımı $X \sim U[0, \theta]$'dir ve θ_0 için %90'lik güven aralığı oluşturmak istiyoruz.*

$$\hat{\theta} = \max\{X_1, \ldots, X_n\} = X_{(n)}$$

Yukarıdaki ifade n'nci sıra istatistiği olsun (önceki derslerde gösterildiği üzere bu aynı zamanda bir maksimum olabilirlik tahmin edicisidir). Daha önce gördüğümüz gibi, $\hat{\theta}$ θ için sapmasız olmamasına rağmen, onu θ için bir güven aralığı oluşturmakta kullanabiliriz. Sıra istatistiğinin sonuçlarından gördük ki $\hat{\theta}$'nın c.d.f.sini veren $\hat{\theta}$'in c.d.f.si aşağıdaki gibi belirlenmektedir:

$$F_{\hat{\theta}}(\theta) = \begin{cases} 0 & \text{eğer } \theta \leq 0 \text{ ise} \\ \left(\frac{\theta}{\theta_0}\right)^n & \text{eğer } 0 < \theta \leq \theta_0 \text{ ise} \\ 1 & \text{eğer } \theta > \theta_0 \text{ ise} \end{cases}$$

burada U[0, θ_0] olan bir rasgele değişkenin c.d.f.sini, F(x) = x/ θ_0, yerine koyduk,

A ve B fonksiyonlarını elde etmek için, önce a ve b sabit değerlerini bulalım,

$$P_{\theta_0}(a \leq \hat{\theta} \leq b) = F_{\hat{\theta}}(b) - F_{\hat{\theta}}(b) = 0.95 - 0.05 = 0.9$$

a ve b değerlerini aşağıdakileri çözünce bulabiliriz

$$F_{\hat{\theta}}(a) = 0.05 \text{ and } F_{\hat{\theta}}(b) = 0.95$$

böylece a = $\sqrt[n]{0.05\theta_0}$ ve b = $\sqrt[n]{0.95\theta_0}$'i elde ederiz. Bu bize henüz bir güven aralığı vermemektedir, çünkü güven aralığının tanımına göre biz gerçek θ_0 değerini eşitsizliğin ortasında isteriz. Ve her iki tarafın fonksiyonları sadece veriye ve diğer bilinmeyen büyüklükler bağlıdır.

Ancak aşağıdakini yazabiliriz,

$$0.9 = P_{\theta_0}(a \leq \hat{\theta} \leq b) = P_{\theta_0}\left(\sqrt[n]{0.05}\theta_0 \leq \hat{\theta} \leq \sqrt[n]{0.95}\theta_0\right) = P_{\theta_0}\left(\frac{\hat{\theta}}{\sqrt[n]{0.95}} \leq \theta_0 \leq \frac{\hat{\theta}}{\sqrt[n]{0.05}}\right)$$

Bundan ötürü aşağıdaki θ_0 için bir %90'lik güven aralığıdır.

$$[A, B] = [A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)] = \left[\frac{\max\{X_1, \ldots, X_n\}}{\sqrt[n]{0.95}}, \frac{\max\{X_1, \ldots, X_n\}}{\sqrt[n]{0.05}}\right]$$

Bu durumda, aralığın sınırları sadece $\hat{\theta}(X_1, \ldots, X_n)$ tahmin ediciler aracılığıyla veriye bağlıdır. Bu genel olarak doğru olmak zorunda değildir.

Şimdi güven aralığına nasıl ulaştığımızı tekrarlayalım:

1. önce $\hat{\theta}$ $(X_1, \ldots, X_n)$ tahmin edicileri ve $\hat{\theta}$'in dağılımını elde et,
2. aşağıdaki koşulu sağlayacak olan a(θ) ve b(θ)'yi bul

$$P(a(\theta) \leq \hat{\theta} \leq b(\theta)) = 1 - \alpha$$

3. θ'yı çözerek olayı yeniden yaz

$$P(A(X) \leq \theta \leq B(X)) = 1 - \alpha$$

4. A(X), B(X) değerlerini gözlemlenen örneklem $X_1, \ldots, X_n$'i kullanarak bul,
5. 1 – α'lık güven aralığı aşağıdaki ile verilir:

$$\widehat{CI} = [A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)]$$

2.1 Önemli Durumlar

1. $\hat{\theta}$ normal dağılımlıdır, Var($\hat{\theta}$)≡ $\sigma_{\theta 0}^2$ bilinmiyor: Güven aralığı aşağıdaki gibi oluşturulabilir

$$[A(X), B(X)] = \left[\hat{\theta} - \sqrt{\sigma_{\hat{\theta}}^2}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \hat{\theta} + \sqrt{\sigma_{\hat{\theta}}^2}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

2. $\hat{\theta}$ normal dağılımlıdır, Var($\hat{\theta}$) bilinmiyor fakat $\hat{S}^2$ = $\widehat{Var(\hat{\theta})}$ tahmin edicisi var: Güven aralığı aşağıdaki ile verilir

$$[A(X), B(X)] = \left[\hat{\theta} - \sqrt{\hat{S}^2}t_{n-1}\left(1 - \frac{\alpha}{2}\right), \hat{\theta} + \sqrt{\hat{S}^2}t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

   Burada $t_{n-1}(p)$*değeri n -1* serbestlik dereceli t-dağılımının *p*nci yüzdeliğidir.

3. $\hat{\theta}$ normal değil, fakat n > 30 veya daha fazla: öyle anlaşılıyor ki gördüğümüz bütün tahmin ediciler(unifom dağılım için örneklemin maksimumu hariç) merkezi limit teoremine göre asimptotik olarak normaldir (Merkezi Limit Teorisini nasıl uygulayacağımız konusu her zaman açık değil değildir). Bu durumda güven aralığını 2'deki gibi oluştururuz.

4. $\hat{\theta}$ normal değil, n küçük: eğer $\hat{\theta}$'in p.d.f.si biliniyor ise, 1'nci kullanılarak güven aralığı oluşturulabilir(son örnekteki gibi). Eğer p.d.f. bilinmiyor ise, yapabileceğimiz bir şey yok.

2nci durumda t dağılımını kullanmamızın nedeni şudur: $\hat{\theta} \sim$ N(μ, σ^2/n) olduğu için,

$$\frac{\hat{\theta} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Diğer taraftan, şunu kontrol edebiliriz

$$\frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi_{n-1}^2$$

burada $\hat{S}$ genellikle ortalaması sıfır ve varyansı σ^2 olan normal hataların karelerinin toplamı için yazılır. Dolaysıyla,

$$\frac{\hat{\theta} - \mu}{\sqrt{\hat{S}^2/n}} = \frac{\frac{\hat{\theta}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\hat{S}^2}{\sigma^2}/\sqrt{n-1}}} \sim \frac{N(0,1)}{\sqrt{\chi_{n-1}^2}} \sim t_{n-1}$$

Ayrıca 4'ün genel durumunda (ve uniform içeren son örnekte), $\hat{\theta}$ (X_1, …, X_n) istatistiğinin herhangi bir şeyin sapmasız ve tutarlı tahmin edicisi olmasını istemedik, fakat gerçek parametrede kesin monoton olmak zorundaydı. Ancak, normal durumlar($\hat{\theta}$'in varyansı hakkında bilgi sahip olsak ta olmasak ta) ve durum 3 için güven aralığını oluşturduğumuzda, tutarlı olmak zorundaydık.