Ege Üniv. Ziraat Fak. Derg., 2003, **40** (3):89-96
ISSN 1018-8851

Bootstrap Metodu ve Uygulanışı Üzerine Bir Çalışma 1. Olasılık ve Bootstrap Metodu

Çiğdem TAKMA[1] **Hülya ATIL**[2]

Summary

A study on Bootstrap Method and It's Application I. Probability and Bootstrap Method

In statistics, the observations can be obtained for related topics to have information about population and parameters of population can be estimated by using this knowledge. Accuracy of parameter estimation has direct connection with sample. For that reason, more and large samples are necessary. Sampling from large population will cause loss of the time and costs. So, separate sets in the different sizes and quantities can be constituted by replacement with the samples in the observed data set, derived from population. In applied statistics this technique is called as Bootstrap Method. This study was conducted to give information about Bootstrap Method and its application on probability by using the macro of Resampling Excel-add-in, in Excel program.
Keywords: Bootstrap, Resampling, Resampling Excel-add-in

Giriş

İstatistikte populasyon parametrelerini tahminlemede o populasyona ait gözlemler kullanılmaktadır. Gözlemlerin tümünün tahminlemede kullanılmak istenmesi ise zaman kaybı ve masrafa yol açmaktadır. Bu nedenle, populasyonu iyi temsil eden örneklere ihtiyaç duyulmaktadır.

Herhangi büyüklükte bir veri setinde gözlemlerin şansa bağlı olarak yer değiştirilmesi ile yeniden örneklenerek çeşitli miktarda ve büyüklükte veri setleri oluşturulabilmektedir. Böylece, mevcut veri setinden mümkün olabildiğince fazla miktarda bilgi alınabilmektedir. Söz konusu metot,

[1] Arş. Gör. E.Ü. Ziraat Fakültesi, Zootekni Bölümü, Biyometri & Genetik A.B.D., 35100 Bornova-İZMİR (cigdem@ziraat.ege.edu.tr)
[2] Doç. Dr. E.Ü. Ziraat Fakültesi, Zootekni Bölümü, Biyometri ve Genetik A.B.D., 35100 Bornova-İZMİR

1979 yılında Bradley Efron tarafından, geliştirilmiş ve *Bootstrap (Resampling) Metodu* olarak adlandırılmıştır (5). İstatistiksel hesaplamalardaki modern gelişmelere parallel olarak, Bootstrap metodunda da ilerlemeler sağlanmış ve uygulamalı istatistik alanında bu metodun kullanımı giderek artmıştır.

Bu çalışmada, Excel programı altında çalışan bir makro (1) kullanılarak Bootstrap metodunun tanıtımı ve olasılıkta uygulanışı hakkında bilgi verilmesi amaçlanmıştır.

Bootstrap Metodu

Bootstrap metodu parametrik ve parametrik olmayan istatistik analizlerde kullanılabilen basit ve güvenilir bir metoddur (2).

Herhangi bir *S*(x) istatistiği N adet gözlemden oluşan bir veri seti üzerinde bu metot kullanılarak kolaylıkla açıklanabilir: Orijinal veri setinde ($x=(x_1,x_2,x_3,x_4,..,x_N)$) gözlemlerin yer değiştirilip, bu veri setinden 1/N kadar olasılıkla şansa bağlı iadeli seçim yapılarak elde edilecek yeni örnek veri seti $x_i^* = (x_1, x_2, x_3, \ldots, x_N)$ olacaktır. Bu şekilde oluşturulan veri seti Bootstrap örnek veri setidir (3). Bu işlem istenildiği kadar yinelenerek birbirinden farklı B adet Bootstrap veri setleri oluşturulmaktadır. İlgili istatistik, bu yeni veri setleri kullanılarak hesaplanmaktadır.

Örneğin, Bootstrap metoduna göre *S*(x) istatistiğinin standart sapmasının hesaplanmasında aşağıda verilen işlem basamakları gerçekleştirilmektedir.

1) x veri setinden yer değiştirme yapılarak N bireylik B adet Bootstrap örnek veri setleri $(x_1^*, x_2^*, x_3^*, \ldots., x_B^*)$ oluşturulur.

2) Her bir Bootstrap örneğinde söz konusu istatistik hesaplanır.

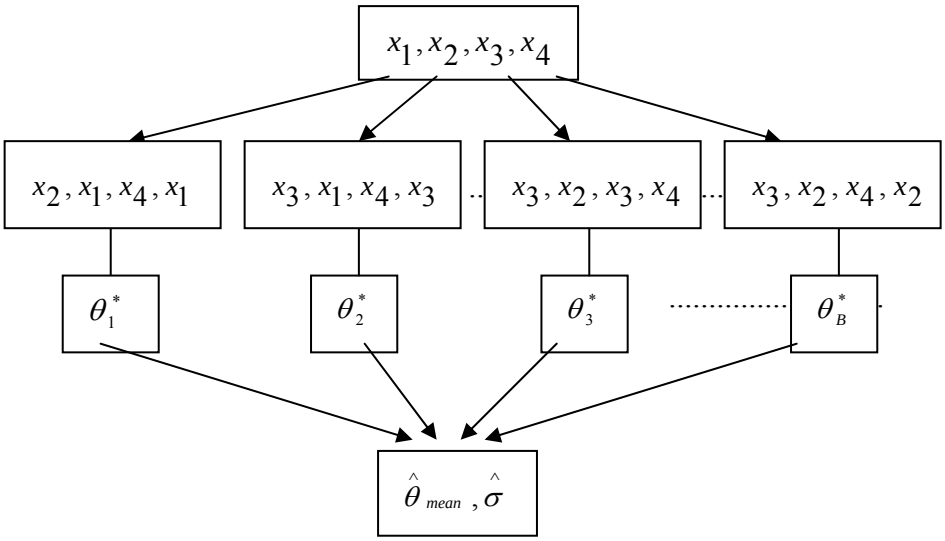
$$\hat{\theta}^* i = S(x_i^*) \qquad i = 1,2,....,B$$

3) Aşağıdaki formül ile standart sapma hesaplanır.

$$\sigma^* = \left[\frac{1}{B-1} \sum_{i=1}^{B} (\theta_i^* - \langle \theta^* \rangle)^2 \right]^{1/2}, \quad \langle \theta \rangle = \sum_{i=1}^{B} \theta_i^* / B$$

Şekil 1'de Bootstrap metodu şematik olarak gösterilmektedir. Orijinal verilerden yer değiştirmeyle şansa bağlı olarak seçilen dört

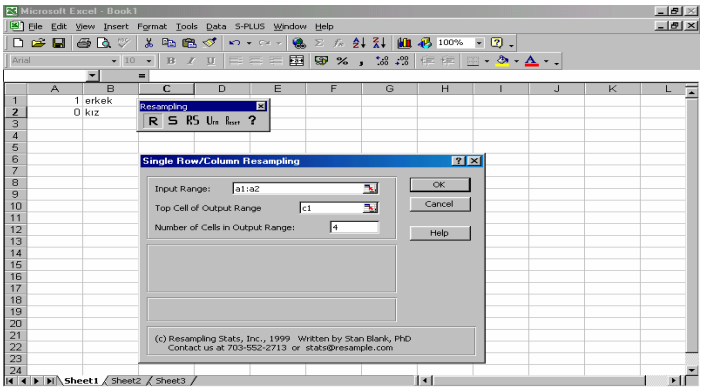
elemanlık B adet Bootstrap örneğinden *S*(x) istatistiğine ait Bootstrap tahminleri $(\theta_i, i = 1, B)$ elde edilmektedir. Bu tahminler daha sonra ortalama ve varyans hesabında kullanılmaktadır.



Şekil 1. Bootstrap metodunun şematik gösterimi (6).

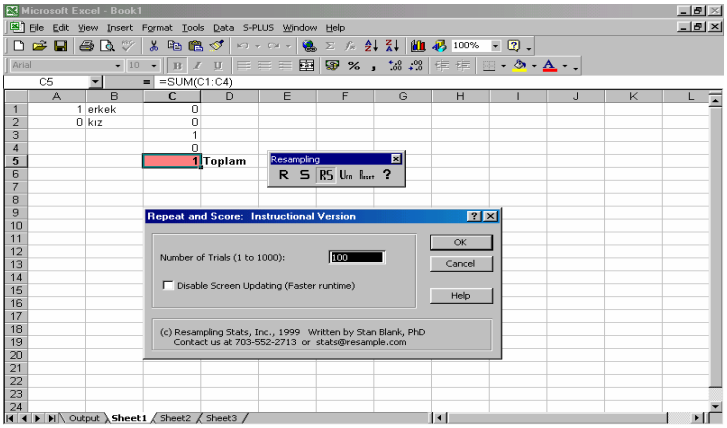
Olasılık ve Bootstrap Metodu

Olasılık hesaplamalarında Bootstrap metodunun kullanımı bir örnek üzerinde açıklanabilir: Bir çocuğun cinsiyetinin kız ya da erkek olma olasılığının birbirine eşit olduğu ve bu olasılığın bir önceki çocuğun cinsiyetinden bağımsız olduğu durumda, 4 çocuklu ailelerde 3 çocuğun da erkek olma olasığı araştırılmak istensin. Bunun Bootstrap metodu ile çözümünde Excel-Add in makrosunda yer alan "Resampling" menüsündeki "single row/column resampling" penceresinde veri seti tanımlanmaktadır. Burada gözlem değerleri erkekler için "1" kız çocuklar için "0" biçiminde kodlanarak 4 çocuklu aileler için yine toplam 4 gözlemlik bir veri seti tanımlanmıştır (Şekil 2).



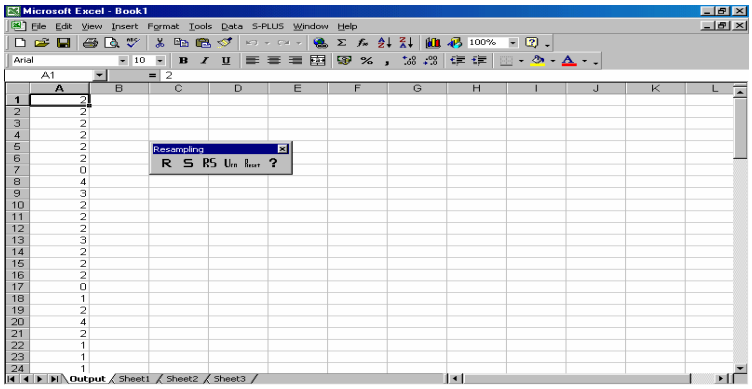
Şekil 2. Excel-Add in ile olasılık hesabı (veri tanımlaması).

Bu veri setindeki ilk 4 çocuk içinden erkeklerin toplamı alınarak bir hücreye yazdırılmıştır. Gözlemlerin yer değiştirilerek yeni veri setlerinin oluşturulması işlemi ise "Repeat and Score" menüsü kullanılarak yapılmıştır. Söz konusu örnek için, "Repeat and Score" menüsünde deneme sayısı 100 tanımlanmıştır (Şekil 3).



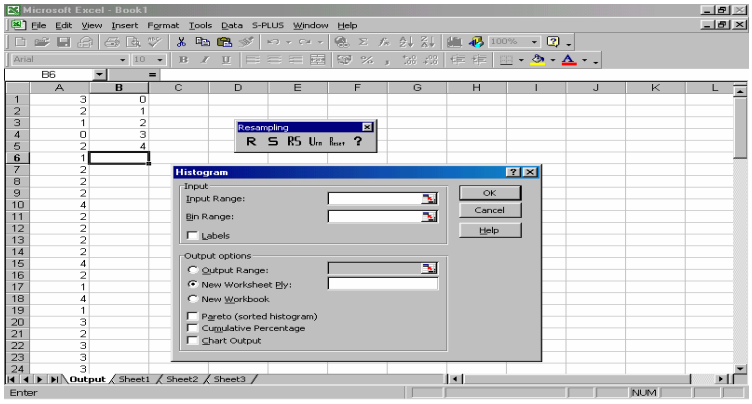
Şekil 3. "Repeat and Score" menüsünde deneme sayısının tanımlanması.

Buna göre Excel-Add in makrosunda dört çocuklu 100 ailedeki erkek çocuk sayılarına ilişkin toplamlar Şekil 4'deki gibi bir sonuç dosyasına yazdırılmıştır.



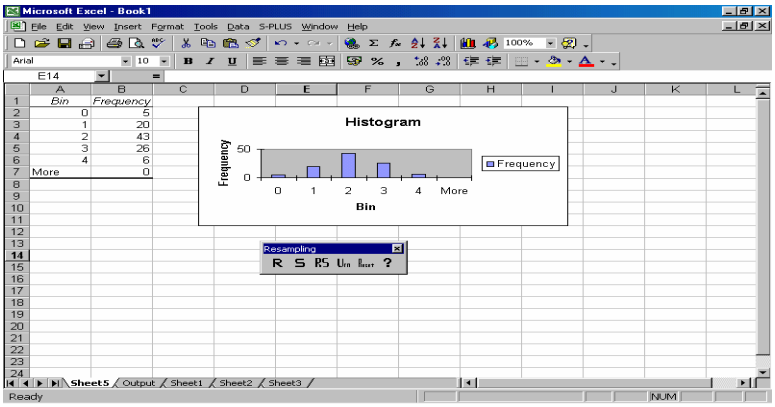
Şekil 4. Sonuç dosyası.

Erkek çocuk olma olasılıklarına ilişkin frekans dağılışlarını görmek için toplamlar küçükten büyüğe sıralanmış ve "Data" menüsünde, "Data Analysis" penceresinde "Histogram" seçeneği kullanılmıştır (Şekil 5-6).



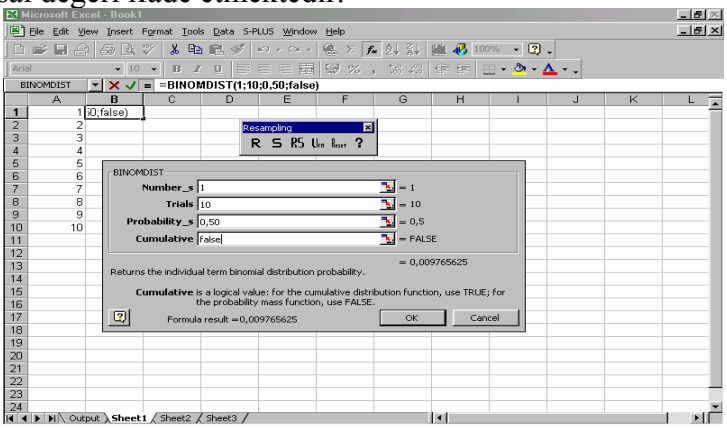
Şekil 5. Excel-Add-in "histogram" penceresi.

Histogram seçeneğinde histograma ilişkin aralık "input range", çizilecek histogramın biçimi ise "output options" seçenekleri ile belirlenmektedir. Bu dağılışın sınıf değerlerini tanımlamak için "bin" seçeneği kullanılmaktadır.



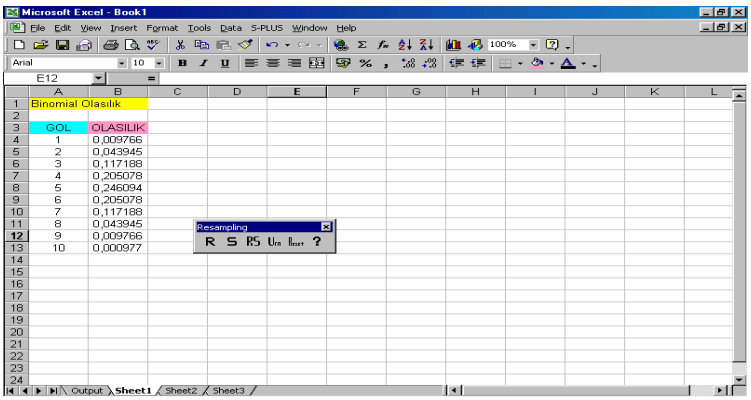
Şekil 6. Excel-Add-in ve Histogram.

Bootstrap metodu ile binom dağılışında olasılıklar da kolaylıkla hesaplanabilmektedir. Binom olasılık dağılışının Bootstrap metodu ile belirlenmesinde bir futbolcunun yaptığı atışlardaki başarı olasılığını gözönüne alınsın. Bu futbolcunun her penaltı vuruşunun ½ olasılıkla isabetli olduğu varsayılırsa, on atışından dördünün gol olma olasılığı araştırılsın. Böyle bir problemin çözümü için Excel programında binom olasılık dağılışının tanımlanacağı "BINOMDIST" fonksiyonundan yararlanılmaktadır (Şekil 7). Bu fonksiyonun komut tanımlamalarında Number_s: Denemedeki başarı sayısını, Trials: Bağımsız deneme sayısını, Probability_s: Başarı olasılığını, Cumulative: Fonksiyondan tanımlanan mantıksal değeri ifade etmektedir.



Şekil 7. BINOMDIST fonksiyon menüsünde binom olasılık probleminin tanımlanması.

Eğer “cumulative” seçeneği için gerçek (TRUE) tanımlaması yapılırsa, BINOMDIST fonksiyonu en fazla başarı olasılığını ifade eden kümülatif dağılış fonksiyon değerini vermektedir. Bu değer için yanlış (FALSE) tanımlaması yapıldığında başarı sayısına ait olasılığı veren olasılık fonksiyon değeri elde edilmektedir. Örnekte, “cumulative” seçeneğine FALSE tanımlaması yapılarak futbolcunun başarılı atışlarına ilişkin olasılık değerlerine ulaşılmıştır (Şekil 7-8).



Şekil 8. Binom dağılışı için olasılık probleminin çözümü.

Tartışma

Son yıllarda çeşitli metot ve yazılımların ortaya çıkışı ile istatistiksel analizlerde hızlı gelişmeler kaydedilmiştir. Bu gelişmelere dayalı olarak Bootstrap metodu da çeşitli istatistiksel analizlerde kullanılmaya başlanmıştır. Bootstrap metodunun bilinen klasik istatistik metodlarına göre bir çok avantajı bulunmaktadır (7). Klasik yaklaşımda verilerin dağılışı hakkında gerekli olan varsayımların geçerliliğinden şüphe edildiğinde, alınan sonuçlara güvenilememektedir. Bootstrap metodu verilerin dağılışı hakkında herhangi bir varsayım taşımaması nedeniyle**,** diğer metodların kullanımının uygun olmadığı ya da bilinen varsayımların geçersiz olduğu durumlarda tercih edilebilmektedir (8**)**. Ayrıca, büyük veri setlerinde Bootstrap metoduna göre alınan sonuçlar ile klasik istatistik metodlardan alınan sonuçlar benzerlik göstermektedir. Bunun yanısıra, çok küçük veri setlerinde de güvenilir sonuçlara ulaşılmaktadır. Diğer taraftan, Bootstrap metodunun bilgisayar destekli olarak kullanımı hesaplanması zor ve uzun zaman alan problemlerde kolaylık sağlamıştır. Bootstrap metodu

bir çok avantaj taşımasına rağmen, yaygın biçimde kullanılmama sebebi, bu yaklaşımı kullanan paket programların yakın zamanda yazılmış olmasıdır. Tüm bu nedenlere bağlı olarak Bootstrap metodu üzerinde daha detaylı çalışmaya ve farklı Bootstrap fonksiyonlarını içeren daha kapsamlı paket programlara ihtiyaç olduğu anlaşılmaktadır.

Bu çalışmada Excel-add in makrosu yardımıyla Bootstrap metodu ve olasılıkta uygulanışı hakkında bilgi verilmiştir. Güven aralıklarının tahminlenmesi, hipotez testi ve regresyon analizinde Bootstrap metodunun kullanılışına bir başka çalışmada yer verilmesi uygun görülmüştür.

Özet

İstatistikte populasyon hakkında bilgi edinmek için, incelenen konu ile ilgili olarak gözlemler yapılmakta ve bunlardan yararlanılarak populasyon parametreleri tahminlenmektedir. Güvenilir bir tahminlemenin yapılabilmesi ise, alınan örneğin populasyonu iyi temsil etmesi ile doğrudan ilişkilidir. Bunun için çok sayıda ve büyük veri setlerinden oluşan örneklere ihtiyaç duyulmaktadır. Populasyonlara ait örnek büyüklüğünün artırılması çok fazla zaman kaybına ve masrafa neden olmaktadır. Bu nedenle, populasyondan alınmış mevcut veri setinde gözlemlerin yer değiştirilerek yeniden örneklenmesi ile çeşitli miktarda ve büyüklükte veri setleri oluşturulmaktadır. Uygulamalı istatistikte kullanımı giderek artan bu teknik *Bootstrap Metodu* olarak anılmaktadır. Bu çalışmada Bootstrap Metodu ve kullanım alanları hakkında genel bir tanıtım yapılmıştır. Metodun olasılıkta uygulanışı, Excel programı altında yazılmış Resampling Excel-add-in makrosu yardımıyla yapılmıştır.

Anahtar Kelimeler*:* Bootstrap, Resampling, Resampling Excel-add-in

Kaynaklar

1. Blank, S. 1999. Resampling Stats Excel Add-in, Online Trial Version, Resampling Stats, Inc. Erişim:[http://**www.resample.com**, Ağustos, 2001].
2. Efron, B. 1979. Bootstrap methods: Another look at the jackknife**.** Ann.Statist. (1):1-26.
3. Efron B. and Tibshirani R. 1993. An introduction to the Bootstrap.Chapman and Hall. New York.
4. Hamajima N., Yuasa H. and Matsuo K. 1999. Methods for statistical inferences. Biotherapy, Vol:13(6), 739-744.
5. Peterson, I. 1991. Pick a sample. Science News. Vol:140, 56-57.
6. Sacchi, M.D., 1998. A bootstrap procedure for high-resolution velocity analysis. Geophysics, Vol:63(5).
7. Visscher P.M., Thompson R. and Haley C.S. 1996. Confidence intervals in QTL mapping by bootstrapping. Genetics, Vol:143 (2), 1013-1020.
8. Wehrens R., Putter H. and Buydens L.M.C. 2000. The bootstrap: A tutorial. Chemometrics and Intelligent Laboratory Systems, Vol:54, 35-52.