14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Bahar 2009

14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Ders Notları 14

Konrad Menzel

2 Nisan 2009

1. Koşullu Beklenen Değerler

Örnek 1. *Her sene, bir firmanın AR-GE bölümü rasgele bir süreç sonucunda X kadar buluş üretmektedir, burada E|X| = 2 ve Var(X) = 2'dir. Her buluş p = 0.2 olasılıkla ticari bir başarı gösterecektir (bağımsızlık varsayalım). Bir yıl içerisindeki ticari başarı sayısını S ile gösterelim. Bir yıl içerisinde X = x buluş sayısına koşullanmış S'nin ortalamasının S ~ B(x, p) = xp olduğunu bildiğimiz için, buluşların ortalama olarak xp kadarının başarılı olması gerekir.*

Y veriyken X'in koşullu beklenen değeri koşullu p.d.f'den elde edilen X'in beklenen değeridir:

Tanım 1.

$$\mathbb{E}[Y|X] = \begin{cases} \sum_y y f_{Y|X}(y|X) & \text{eğer Y kesikli ise} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|X)dy & \text{eğer Y sürekli ise} \end{cases}$$

$f_{Y|X}(y|X)$ *rasgele değişken* X'i kendi argümanı olarak taşıdığı için, koşullu beklenen değerin de aynı zamanda rasgele bir değişken olduğunu not ediniz. Ancak, X'in belli bir değeri veriyken, Y'nin koşullu beklenen değerini de tanımayabiliriz

$$\mathbb{E}[Y|X=x] = \begin{cases} \sum_y y f_{Y|X}(y|\text{x}) & \text{eğer Y kesikli ise} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|\text{x})dy & \text{eğer Y sürekli ise} \end{cases}$$

burada koşullu yoğunluk tanımlandığı sürece bu herhangi bir verili *x* değeri için sadece bir sayıdır.

Hesaplamalar tam olarak önceki gibi yapıldığı için (tek fark şimdi *koşullu* dağılım üzerinden integral alıyoruz), sayısal bir örnek yapmayacağız (problem seti için sadece tanımı uygulayınız). Bunun yerine, koşullu ve koşulsuz örnekler arasındaki farkı göstermek için daha kalitatif örnekleri tartışalım.

Örnek 2 (“Limon” Piyasası). *Aşağıdaki ekonomist George Akerlof’un meşhur kullanılmış araba piyasası modelinin basitleştirilmiş bir versiyonudur. Varsayalım ki üç tür X kullanılmış araba vardır: mükemmel durumdaki arabalar (karpuzlar), orta kalite arabalar (istatistiksel manada kesinlik ifade eden “ortalama” değil), ve çok kötü durumdaki arabalar (limonlar). Her tür araba eşit frekansa sahiptir, yani*

P(“limon”) = P(“orta”) = P(“karpuz”) = 1/3

Satıcı ve bir alıcı her bir araba türü için, sırasıyla Y_S ve Y_B, kadar aşağıda verilen değerleri biçiyorlar:

Tür	Satıcı	Alıcı
Limon	5000	6000
Orta	6000	10000
Karpuz	10000	11000

İlk dikkat edilmesi gereken şey, her araba çeşidi için, alıcının biçtiği değerin satıcının biçtiğinden daha yüksek olduğudur, dolayısıyla her bir tür araba için, alış veriş alıcı ile satıcının biçtiği değerin arasındaki bir fiyattan gerçekleşmeli. Ancak, kullanılmış arabalar da, kalite ilk anda görülen şey değildir. Bu nedenle, eğer ne alıcı ne de satıcı söz konusu arabanın türü X’i bilmiyorsa, onların beklenen değeri yinelenen beklentiler kanununa göre aşağıdaki gibidir:

$$\begin{aligned}
\mathbb{E}[Y_S] &= \mathbb{E}[Y_S|\text{” limon ”}]P(\text{” limon ”}) + \mathbb{E}[Y_S|\text{” orta ”}]P(\text{” orta ”}) + \mathbb{E}[Y_S|\text{” karpuz ”}]P(\text{” karpuz ”}) \\
&= \frac{1}{3}(5,000 + 6,000 + 10,000) = 7,000 \\
\mathbb{E}[Y_S] &= \mathbb{E}[Y_B|\text{” limon ”}]P(\text{” limon ”}) + \mathbb{E}[Y_B|\text{” orta ”}]P(\text{” orta ”}) + \mathbb{E}[Y_B|\text{” karpuz ”}]P(\text{” karpuz ”}) \\
&= \frac{1}{3}(6,000 + 10,000 + 11,000) = 9,000
\end{aligned}$$

Dolayısıyla alış verişin gerçekleşmesi gerekir.

Daha gerçekçi bir düzenlemede, arabanın satıcısı arabanın kalitesini alıcıdan daha iyi bilir (tamirat geçmişini, kazalarını vs.) ve arabayı satmaya arzulayacağı fiyat belirler. Eğer satıcı üç araba türünü de mükemmel bir şekilde ayırt edebilirse, ki alıcı bunu yapamaz, alıcının satıcının belirtilen fiyattan arabayı satma arzusuna koşullanmış beklentiler oluşturması gerekir.

Eğer satıcı 6000 dolardan daha düşük bir fiyat belirtirse, alıcı kesin olarak arabanın “limon” olduğunu bilirdi, çünkü diğer durumlarda satıcı en az 6000 dolar talep ederdi, yani

$$\mathbb{E}[Y_B|Y_S < 6000] = \mathbb{E}[Y_B|\text{” limon ”}] = 6000$$

ve alış veriş gerçekleşirdi. Ancak, eğer araba "karpuz" ise, satıcı en az 10000 dolar talep ederdi, hâlbuki alıcı en fazla

$$\mathbb{E}[Y_B|Y_S \leq 10,000] = \mathbb{E}[Y_B] = 9,000 < 10,000$$

kadar ödeyecekti, bu nedenle de satıcı yüksek kalite bir arabayı makul bir fiyata satamayacaktı.

Piyasanın "karpuz" için çalışmamasının(break down) nedeni bu modelde satıcının alıcıya arabanın kalitesinin düşük olmadığı konusunda kabul edilebilir bir garanti verememesidir, bu nedenle alıcı alış verişinde kötü bir araba alma ihtimalini hesaba katar.

Örnek 3. *Bu örnekte, insanların gelecekteki politik olaylar üzerine bahse girdiği bir internet platformu olan IEM Political Markets'ın 2008 başkanlık adayları ile ilgili verisine bakacağız(veri için bkz.:http://www.biz.uiowa.edu/iem/markets/data_nomination08.html)*

Piyasa şöyle çalışıyor: Her bir politik aday i için, katılımcılar aşağıdaki getiriyi veren kontratlar satın alıyorlar

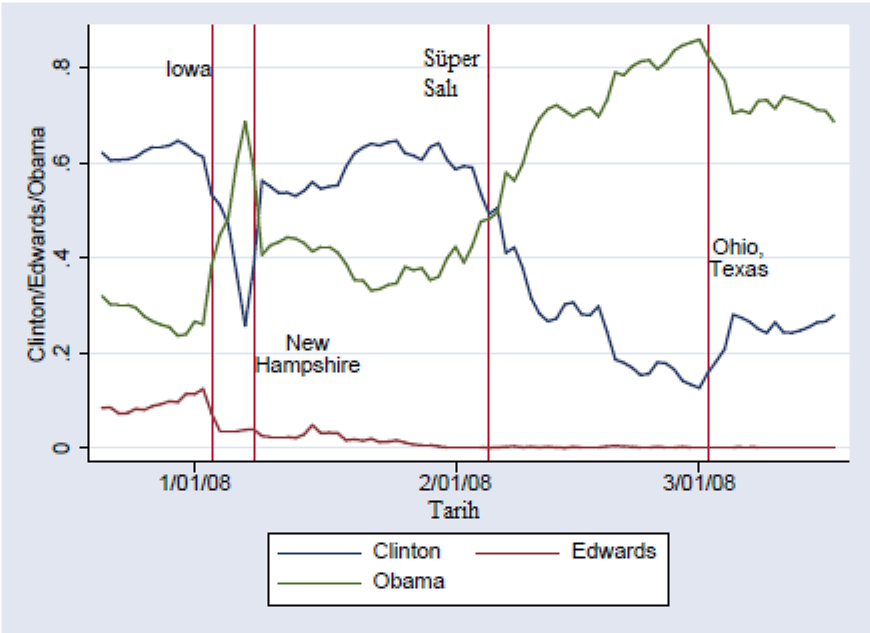
$$Y_i = \begin{cases} 1\ dolar & eğer\ aday\ i\ adaylığı\ kazanırsa \\ 0\ dolar & diğer\ bütün\ durumlarda \end{cases}$$

Verili bir t zamanda, piyasadaki katılımcılar X_t olarak adlandıracağımız dışarıdan gelen ilave bilgiye sahip olurlar, örneğin o ana kadar kazanılan delege sayısı, adayın propagandasının "momentumu", ya da adayın propaganda stratejisi hakkında görevlilerin demeçleri gibi.

Söz konusu ilave bilgi verilmişken, kontratın beklenen değeri

$$\mathbb{E}[Y_i|X_t = x] = \sum_{y_i} y_i f_{Y_i|X_t}(y_i|x) = 1 \cdot P(Y_i = 1|X_t = x) + 0 \cdot P(Y_i = 0|X_t = x) = P(Y_i|X_t)$$

Diğer bir ifadeyle, katılımcıların aday i'nin kontratı için ödemeyi arzuladıkları dolar miktarı t zamanda verili bilgiye göre i 'nin kendi parti adaylığını kazanma olasılığına eşittir. Son üç aydaki Demokrat Parti'nin ana adaylarının kontratlarının fiyatlarına bakalım: Demokratik adayların kazanma ihtimali hakkında önemli bilgileri ortaya çıkaran 3 olay için üç dik doğru çizdim:



- *Barrack Obama'nın Hillary Clinton'a karşı ezici bir farkla kazandığı Iowa parti kongresi,*
- *Iowa'daki yenilgiden sonra Hillary Clinton'un geri dönüşü olarak görülen New Hampshire ön seçimi,*
- *Ohio ve Taksas'ın ön seçimleri; bu iki önemli eyaletin ön seçimlerini Hillary Clinton kazandı*

Bu olayların her birisinden sonra koşullu beklenen değerlerde çok önemli değişiklikler olduğunu görebiliriz. Bu, adayların partilerinin adaylığını sağlama alma şansı hakkında piyasanın "inançlarının" nasıl değiştiğini göstermektedir.

Finansal Ekonomi'de, özel bir durumun gerçekleşmesi halinde 1 dolar ödeyen bu tür kontratlara Arrow-Debrue menkul kıymeti *de denir.*

Koşullu ve koşulsuz beklenen değerler arasındaki önemli bir ilişki Yinelenen Beklentiler Kanunu'dur(bu daha önce bu derste gördüğümüz Toplam Olasılık Kanunun yakındır).

Önerme 1 (Yinelenen Beklentiler Kanunu).

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y]$$

İSPAT: g(x) = $\mathbb{E}$[Y|X = x] olsun. g(x) x'in bir fonksiyonudur. Beklenen değeri şimdi hesaplayabiliriz:

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_{-\infty}^{\infty} \mathbb{E}[Y|X=x] f_X(x) dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y \frac{f_{XY}(x,y)}{f_X(x)} dy \right) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x,y) dy dx \\
&= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{XY}(x,y) dx \right) dy \\
&= \int_{-\infty}^{\infty} y f_Y(y) dy = \mathbb{E}[Y]
\end{aligned}$$

Önerme 2 ***(Koşullu Varyans / Toplam Varyans Kanunu).***

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathrm{Var}(Y|X)]$$

Bu sonuç aynı zamanda ANOVA eşitliği olarak bilinir. Burada ANOVA Varyans Analizi'dir. Özelikle, Var(Y|X) ≥ 0 olduğu için, aşağıdakine ulaşılır

$$\mathrm{Var}(Y) \geq \mathbb{E}[\mathrm{Var}(Y|X)]$$

Bu, kabaca söylemek gerekirse, "X'i bilmek Y'nin varyansını düşürür" şeklinde okunabilir.

İSPAT:

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathrm{Var}(Y|X)] &= \left(\mathbb{E}[\mathbb{E}[Y|X]^2] - (\mathbb{E}[\mathbb{E}[Y|X]])^2\right) + \left(\mathbb{E}\left[\mathbb{E}[Y^2|X]\right] - \mathbb{E}\left[\mathbb{E}[Y|X]^2\right]\right) \\
&= \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2 + \mathbb{E}[Y^2] - \mathbb{E}[\mathbb{E}[Y|X]^2] \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \mathrm{Var}(Y)
\end{aligned}$$

burada birinci eşitlik VarX = $\mathbb{E}[X^2]$ - $\mathbb{E}[X]^2$ özelliğini kullanır, ikinci adım yinelenen beklentiler kanununu kullanır, ve son adımda birinci ve ikinci terimler birbirini götürünce ispat tamamlanır.

Örnek 4. *Her sene, bir firmanın AR-GE bölümü rasgele bir süreç sonucunda X kadar buluş üretmektedir, burada* $\mathbb{E}[X]$ *= 2 ve Var(X) = 2'dir. Her buluş p = 0.2 olasılıkla ticari bir başarıdır (bağımsızlık varsayalım). Bir yıl içerisindeki ticari başarı sayısını S ile gösterelim.*

(a) *Varsayalım ki bu sene 5 yeni buluşumuz var. Onlardan S tanesinin ticari başarı gösterme olasılığı nedir? X = 5 veri iken, S'nin koşullu p.d.f.si bir binomdur, bu nedenle, örneğin,*

$$P(S=2|X=x)=\left(\begin{array}{c}5\\3\end{array}\right)(0.2)^3(1-0.2)^2\approx 0.05$$

(b) *5 buluş veri iken, beklenen başarı sayısı nedir? S|X = 5 ∽ B(5,0.2) olduğu için, son dersteki sonuçları kullanabiliriz,*

$$\mathbb{E}[S|X=5]=\mathbb{E}[B(5,0.2)]=0.2\cdot 5=1$$

(c) *Buluşların koşulsuz beklenen değeri nedir? Yinelenen beklentiler kanunu ile şunu buluruz,*

$$\mathbb{E}[S]=\mathbb{E}[\mathbb{E}[S|X]]=\mathbb{E}[pX]=0.2\mathbb{E}[X]=0.2\cdot 2=0.4$$

burada E[X] = 2 varsayıyoruz.

(d) *S'in koşulsuz varyansı nedir? Toplam varyans kanunu hatırlayınız,*

$$\begin{aligned}\mathrm{Var}(S) &= \mathrm{Var}(\mathbb{E}[S|X])+\mathbb{E}[\mathrm{Var}(S|X)]\\ &= \mathrm{Var}(Xp)+\mathbb{E}[p(1-p)X]=p^2\mathrm{Var}(X)+p(1-p)\mathbb{E}[X]=0.04\cdot 2+0.16\cdot 2=0.4\end{aligned}$$

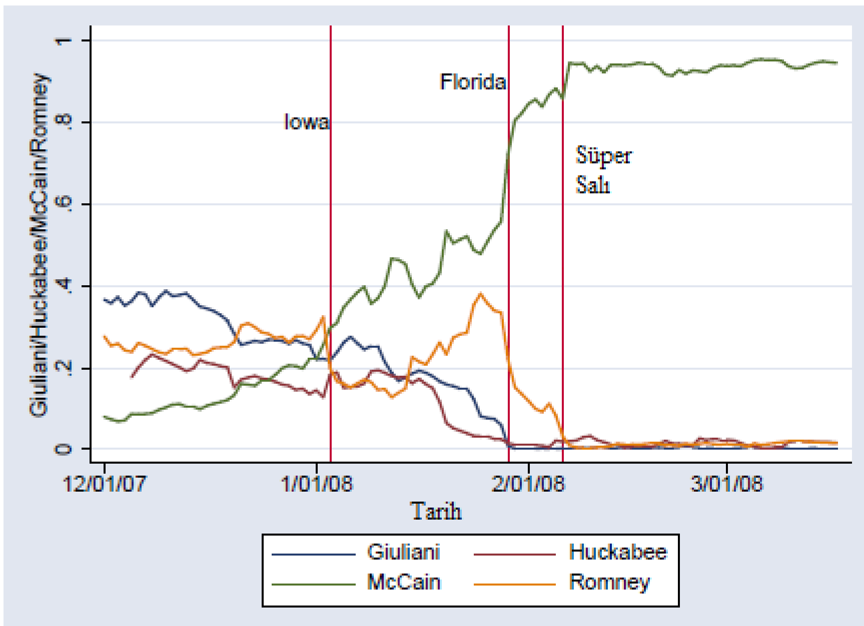
Bu binomların karışımının bir örneğidir, yani X'e koşullanmış S için binom dağılımımız var. O zaman yinelenen beklentiler kanununu kullanarak başarıların toplam sayısını elde edebiliriz.

Örnek 5. *(IEM Politik Piyasa, devam) Cumhuriyetçilerin geçen seneki ön seçimlerine bakacak olursak, belirsizliklerin çoğunun Süper Salı'da çözüldüğünü görebiliriz. Diyelim ki, koşul değişkeni X_t t tarihindeki yeminli delegasyonun sayısı olsun. Toplam varyans kanunu ışığında Iowa önseçimlerinden önceki "koşulsuz" ortalamalar ile Süper Salı'dan sonraki "koşullu" ortalamaları karşılaştırabiliriz.*

$$\mathrm{Var}(Y_i)=\mathbb{E}[\mathrm{Var}(Y_i|X_t)]+\mathrm{Var}(\mathbb{E}[Y_i|X_t])$$

Iowa seçimlerinde önce, ana adaylarin $\mathbb{E}[Y_i]$ 'si büyük dalgalanmalar ile %10 ile %40 gibi orta düzey bir aralıkta yer alıyorlardı. Ancak, Süper Salı'dan sonra, fiyatlar (yani $\mathbb{E}[Y_i|X_t]$) 0 veya 1'e yaklaştı, ve "oynamalar" çok ufak olmaya başladı. Bu nedenle, koşullu varyans formülüne göre, Süper Salı'dan sonra koşullu ortalama hakkındaki belirsizlik, Var($\mathbb{E}[Y_i|X_t]$), gelişmelerden sonraki varyansın, Var(Y_i), en büyük bölümünü oluşturuyordu, hâlbuki koşullu varyansın, Var($Y_i|X_t$) , katkısı nispeten daha düşük görünüyordu.

Eğer bunu Demokratların yarışının grafiğiyle karşılaştırırsak, Süper Salı'dan sonra demokratlar için hala büyük hareketlenme olduğu görülebilir, bu nedenle Süper Salıdan ötürü yeminli delege sayısının üzerine koşul koymak varyansın önemli bir bölümünü yok etmemektedir, yani varyans, Var($Y_i|X_t$), hala çok yüksektir. Diğer taraftan, Cumhuriyetçilerin yarışının çok çabuk bitmesinin sıkça belirtilen nedeni, her bir eyalette, Cumhuriyetçilerin ön seçimlerinde delege sayısının adayların oy oranına göre değil (bu Demokratların çoğu önseçiminde bir kuraldır) kazanan hepsini alır kuralına göre belirlendiği içindir. Bu nedenle, en ufak bir oy farkında bile kazanan aday diğer adaylara fark atabilir ve rakiplerin arayı kapatması oldukça zorlaşır.



2. Özel Dağılımlar

Bu derste, şimdiye kadar yaygın olarak kullanılan üç dağılım gördük, binom, uniform ve üstel. Bundan sonraki iki derste, bu listeyi birkaç önemli örnek ile genişleteceğiz ve onların en sık kullanılanı ile başlayacağız, yani normal dağılımla.

2.1 Hatırlatma: Derste şimdiye kadar gördüğümüz dağılımlar

Tanım 2. *Eğer p.d.f.si aşağıdaki gibiyse, X değişkeni, X ∽ B(n, p), (n, p) parametreleriyle bir binom dağılımdır:*

$$f_X(x) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} p^x(1-p)^{n-x} & \text{eğer} \quad x \in \{0, 1, \ldots, n\} \text{ ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

X ∽ B(n, p) için daha önce aşağıdaki ilişkileri göstermiştik.

$$\begin{aligned} \mathbb{E}[X] &= np \\ \mathrm{Var}(X) &= np(1-p) \end{aligned}$$

Tanım 3. *Eğer p.d.f.si aşağıdaki gibiyse, X değişkeni [a, b] aralığında bir uniform dağılımdır, X ~ U[a, b]:*

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{eğer} \quad a \leq x \leq b \text{ ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

Tanım 4. *Eğer p.d.f.si aşağıdaki gibiyse, X değişkeni λ parametresiyle bir üstel dağılımdır:*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{eğer} \quad x \geq 0 \quad \text{ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

2.2 Standartlaştırılmış Rasgele Değişken

Bazen, rasgele değişken X'in standardizasyonu olan aşağıdaki Z'ye bakmakta fayda vardır.

$$Z := \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}$$

Son birkaç derste türetilen beklenen değer ve varyans kurallarını kullanarak aşağıdakileri elde ederiz:

$$\mathbb{E}[Z] = \frac{\mathbb{E}[X - \mathbb{E}[X]]}{\sqrt{\mathrm{Var}(X)}} = 0$$

ve

$$\mathrm{Var}(X) = \frac{\mathrm{Var}(X - \mathbb{E}[X])}{\mathrm{Var}(X)} = \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} = 1$$

Eğer rasgele değişkenleri bu şekilde normalize edersek, ölçek ve konumdan bağımsız olarak farklı dağılımların şeklini karşılaştırmak daha kolay olur.

2.3 Normal Dağılım

Normal dağılım sürekli rasgele değişken ile ilintilidir. Çok sayıda ki istatistiki deneyin sonuçlarının en iyi tahminini verdiği ortaya çıkmıştır (biraz sonra bunun için bir örnek göreceğiz, daha fazlasını Merkezi Limit Teoremini işledikten sonra göreceğiz).

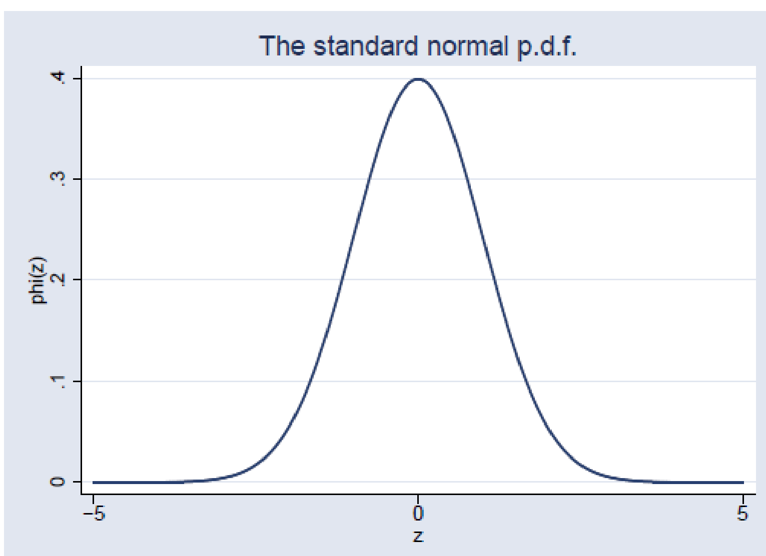
Tanım 5. *Eğer herhangi bir $z \in \mathbb{R}$ için p.d.f. aşağıdaki gibiyse, rasgele bir Z değişkeni normal dağılımlıdır – sembolik olarak $Z \sim N(0, 1)$,*

$$f_Z(z) = \varphi(z) := \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

C.d.f.si ise şöyle gösterilir:

$$\Phi(z) := P(Z \leq z) = \int_{-\infty}^{z} \varphi(s) ds$$

Bir standart normal rasgele değişkenin c.d.f.si kapalı-form ifadeye sahip değildir(fakat tablo değerlerine ya da istatistiksel yazılım paketlerine bakılabilir). P.d.f *$\varphi(z)$* çan eğrisi ve sıfır etrafında simetrik özeliklere sahiptir:



2.3.1. Standart Normal Dağılımın Önemli Özelikleri

Özelik 1. *Bir standart normal rasgele değişken Z için*

$$\mathbb{E}[Z] = 0$$

ve

$$\mathrm{Var}(Z) = 1$$

Normal dağılımın neden yararlı olduğuna dair vurgulanması gereken ilk önemli nokta, çok sayıda n deneme için Binom rasgele değişkenlerin normal dağılım ile tahmin edilebildiğinin ortaya çıkmasıdır.

Teorem 1 *(DeMoivre-Laplace Teoremi)*. *Eğer X ~ B(n, p) bir binom rasgele değişken ise, o zaman c ≤ d gibi herhangi bir sayı için şu ifade yazılabilir,*

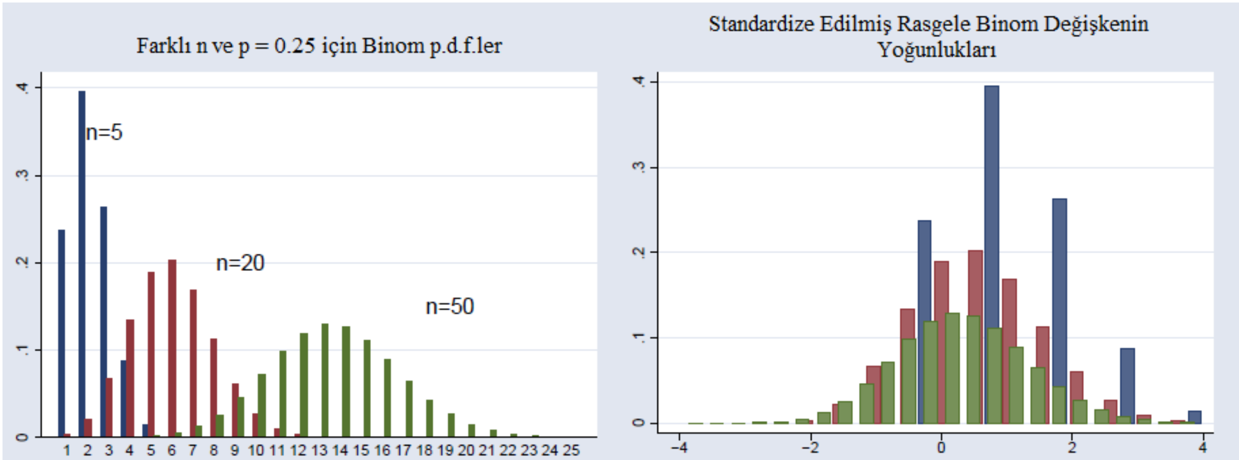
$$\lim_{n\to\infty} P\left(c \leq \frac{X - np}{\sqrt{np(1-p)}} < d\right) = \lim_{n\to\infty} P\left(c < \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} \leq d\right) = \int_c^d \varphi_Z(z)dz$$

Binom değişkenin aşağıdaki ifadeye dönüştürüldüğüne dikkat ediniz.

$$Z = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}}$$

Bu ifade gerçekte standardizasyondur. Bu sonuçlar, çok sayıda *n* için standardize edilmiş binom rasgele değişken X'in (c, d| aralığına düşme olasılığının aşağı yukarı normal değişkenin aynısı olduğunu ifade etmektedir. Örnek olsun diye sadece yükselen n değerleri için binom p.d.f.'leri grafik üzerinde gösterdim ve daha sonra normalizasyon uyguladım.

n = 50 için, bar grafiğinin şekli nispeten normal yoğunluğun çan eğrisine benzemektedir. Özelikle dikkat edilecek olursa, az sayıdaki n için dağılımın çarpıklığı nerdeyse tamamen yok olmuştur ("başarı"nın düşük olasılığından ötürü, p = 1/4).



Şekil 1. DeMoivre-Laplace Teoreminin Gösterimi

Örnek 6. *Uygulama açısından bu tür tahminlerin neden gerçekten yararlı olduğunu görmek için ardışık n denemenin olasılığını hesaplamak istiyoruz. Diyelim ki p = 1/5'dir bu durumda en az %25'lik başarı vardır.*

Eğer n = 5 ise, 25%'ten daha fazla başarıya sahip olmama olasılığı aşağıdaki gibi hesaplanır:

$$P(X \leq 1.25) = P(X=0) + P(X=1) = \left(\begin{array}{c} 5 \\ 0 \end{array}\right)(1-p)^5 + \left(\begin{array}{c} 5 \\ 1 \end{array}\right)p(1-p)^4 = \frac{4^5}{5^5} + 5\frac{4^4}{5^5} = \frac{1280}{3125} \approx 40.96\%$$

Peki, eğer $\hat{X} \sim$ *B(100, 1/5) ise, yani n'i 100'e yükseltirsek ne olur? Prensipte, aşağıdakini hesaplayabiliriz:*

$$P(\tilde{X} \leq 25) = P(X=0) + P(X=1) + \ldots + P(X=25)$$

Böylece, her birinin toplamını ayrı ayrı hesaplayabiliriz. Ancak onlardan çok fazla olduğu için, bu çok ağır bir yük getirecektir. Diğer bir seçenek olarak, DeMoivre-Laplace Teoremini kullanarak iyi bir tahminde bulunabilirdik. Standardize edilmiş $\hat{X}$ *şöyledir:*

$$Z = \frac{\tilde{X} - 100 \cdot \frac{1}{5}}{\sqrt{100 \cdot \frac{1}{5} \cdot \frac{4}{5}}} = \frac{\tilde{X} - 20}{4}$$

Bu nedenle,

$$P(\tilde{X} \leq 25) = P(\tilde{X} - 20 \leq 5) = P\left(Z \leq \frac{5}{4}\right) \approx \int_{-\infty}^{\frac{5}{4}} \phi(z) dz \approx 89.44\%$$

Bu tahmin ne kadar iyidir? Hesaplamaları yaptım ve $P(\tilde{X} \leq 25) \approx$ %91.25 elde ettim. Eğer aynı örneği n = 200 için tekrarlarsak, "tam olarak" $P(\tilde{X} \leq 50) \approx$ %96.55 binom olasılığını ve $P(\tilde{X} \leq 50) \approx$ %96.15 normal tahminini elde ettim.

$Z \sim N(0, 1)$ için, ayrıca herhangi bir rasgele değişkeni elde edebiliriz.

$$X = \mu + \sigma Z$$

Bu *normal* bir rasgele değişkendir, ortalaması μ ve varyansı σ^2'dir. Sembolik olarak

$$X \sim N(\mu, \sigma^2)$$

X'in p.d.f.si nedir? Daha önce derste gördüğümüz değişken değiştirme formülünü kullanabiliriz:

$$f_X(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Bir normal rasgele değişkenin doğrusal dönüşümünün yine normal bir X_1 değişkeni olduğunu göz önünde bulundurarak tartışmayı daha ileri bir aşamaya taşıyabiliriz.

Önerme 3. *Eğer $X_1 \sim N(\mu, \sigma^2)$ ve $X_2 = a + bX_1$ ise, o zaman*

$$X_2 \sim N(a + b\mu, b^2\sigma^2)$$

Bu sonucu yine değişken değiştirme formülünü kullanarak kontrol edebilirsiniz.

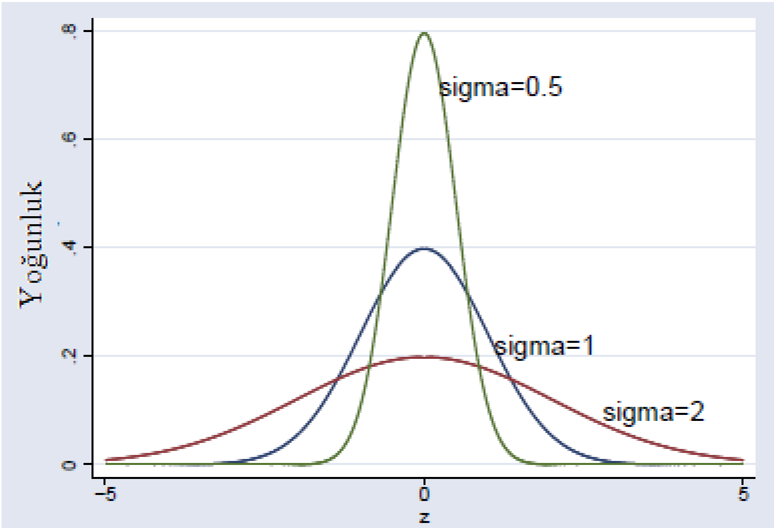
N sayıdaki $X_1, ..., X_n$ değişkenin toplamının beklenen değerinin, beklenen değerlerinin toplamı olduğunu ve n *bağımsız* ve rasgele değişkenin varyansının da varyanslarının toplamı olduğunu görmüştük. Eğer X_i'ler aynı zamanda normal ise, toplamları da normaldir:

Önerme 4. *Eğer $X_1, ..., X_n$ bağımsız normal rasgele değişkenler ve $X_i \sim N(\mu_i, \sigma_i^2)$ ise, o zaman,*

$$Y = \sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Genel olarak, birkaç hafta önce gördüğümüz bükülme formülünü kullanmak zorunda olabilirdik, ancak normallerin toplamı için, sadece toplamın beklenen değerini ve varyansını hesaplamamız yeterlidir. Bu durumda p.d.f.leri hemen bulabiliriz:

$$f_Y(y) = \frac{1}{\sqrt{2\pi \sum \sigma_i^2}} e^{-\frac{(y-\sum \mu_i)^2}{2\sum \sigma_i^2}}$$



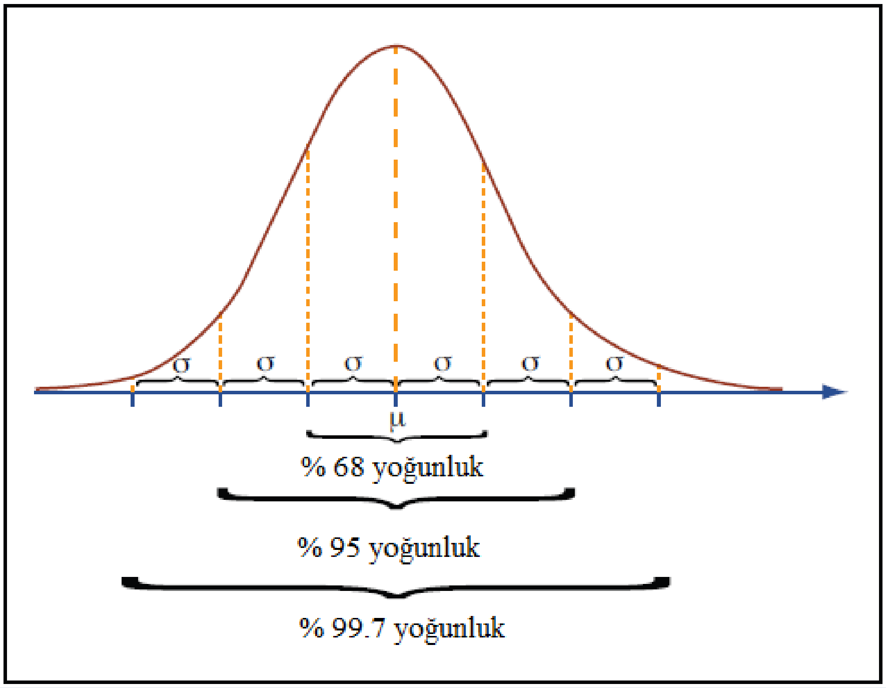
Şekil 2. σ'in Farklı Değerleri için Normal Yoğunluk

2.3.2. Standart Normal'in Tablo Değerlerini Kullanma

Eğer X ~ N(μ, σ) ise, X'in kendi ortalamasından bir veya iki standart sapmadan daha daha uzakta olmama olasılığının kabaca bir tahminini verebiliriz:

$$\begin{aligned} P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) &= \Phi(1) - \Phi(-1) \quad \approx \quad 68\% \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= \Phi(2) - \Phi(-2) \quad \approx \quad 95\% \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &= \Phi(3) - \Phi(-3) \quad \approx \quad 99.7\% \end{aligned}$$

Yani dağılım yığınının çoğu ortalamadan itibaren bir veya iki standart sapma aralığının içindedir. Eğer elinizin altında c.d.f. tablosu yoksa normal olasılığın kabaca bir tahmini elde etmek için bu üç nicel değeri hatırlamakta fayda vardır.



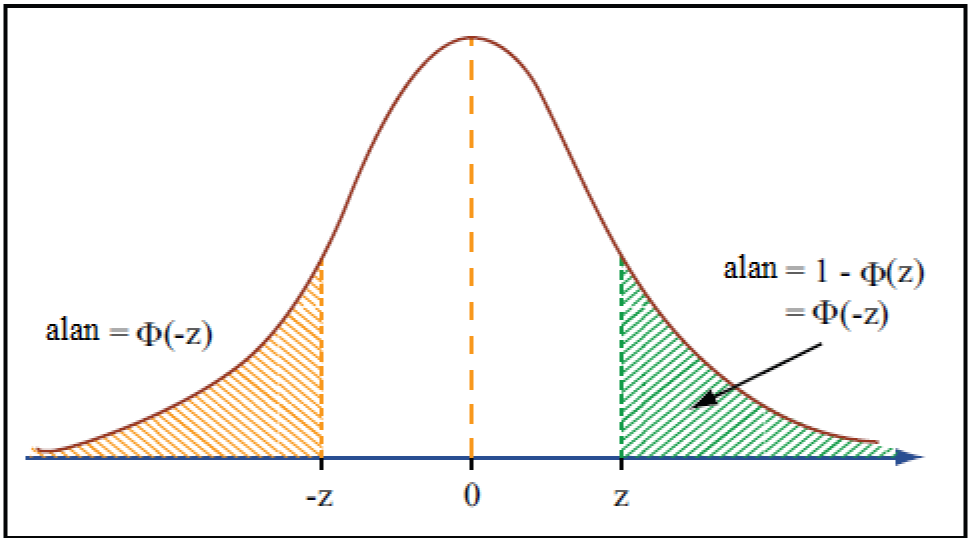
Kaynak: MIT OpenCourseWare

Standart normal dağılım yaygın olarak kullanıldığı için, Φ(z)'in c.d.f.sinin tablolaştırılmış değerlerini herhangi bir istatistik kitabında bulabilirsiniz.

Çoğunlukla söz konusu tablolar sadece $z \leq 0$ değerlerini içerirler, fakat dağılımın simetrisini kullanarak c.d.f.nin $\tilde{z} > 0$ değerlerini elde edebilirsiniz.

$$\Phi(\tilde{z}) = 1 - \Phi(-\tilde{z})$$

Örneğin, eğer $P(Z \leq 1.95)$'i bilmek istiyorsak, $P(Z \leq -1.95)$'i arayabilir ve $P(Z \leq 1.95) = 1 - P(Z \leq -1.95) = 0.9744$'ü hesaplayabiliriz.



Kaynak: MIT OpenCourseWare

Genel olarak, eğer X ~ N(μ, σ) ise, a ≤ b için P(a ≤ X ≤ b) türü olasılıkları aşağıdaki adımları takip ederek elde edebiliriz:

1. değişkeni standardize et, yani bir *standart* normal rasgele Z değişkeni için olayı aşağıdaki gibi tekrar yaz:

$$P(a \leq X \leq b) = P\left(a \leq \mu + \sigma Z \leq b\right) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

2. standart normal c.d.f., Φ(.), cinsinden olasılığı yeniden ifade et:

$$P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

3. olasılığı elde etmek için yukarıda hesaplanan değerlerin standart normal c.d.f.si için tablodan değerleri bul.

2.4 Bir Parantez Açmak: Standart Normal Rasgele Değişkenin Çizimi

Integral dönüştürme kullanarak uniform rasgele çekilişleri diğer herhangi bir sürekli dağılıma dönüştürmenin mümkün olduğunu daha önce görmüştük(rasgele değişkenlerin dönüştürülmesi üzerine olan ders notlarına bakınız). Eğer bir bilgisayarınız yoksa ne yaparsınız? 1900'lü yıllarda, ünlü istatistikçi Francis Galton zar kullanarak normal rasgele bir değişkeni taklit eden akılı bir alet yaptı[1].

Şekil 3'te görülen üç farklı zar peşi sıra atılmıştır. Bu yapılırken, deneyi yapan kişi tesadüfi çekilişlerin listesini tabloya aşağıdaki gibi aktarır: Birinci zar gerçek değeri verir(her zaman size doğru olan yüzeyin altındaki rakam okursunuz). İlk zamanda yıldızlar boş bırakılırken, fakat daha sonra ikinci zarın atılışıyla doldurulurlar. En sonunda, üçüncü zar ilk iki zar ile oluşturduğunuz çekilişlerin önüne konulacak artı ve eksilerin sırasını verir.

Sonucun gerçekten de standart normal rasgele değişkene benzemesi için zar üzerindeki rakamlar özelikle normal dağılımın pozitif yarısının eşit aralıklı yüzdeleri olarak seçildi.

2.5 Standart Normallerin Fonksiyonları: Ki-kare, t- ve F-dağılımları

Tahmin ve test etmek için standart normal dağılımın öneminden ötürü, standart normal rasgele değişkenlerin bazı fonksiyonları da önemli rol oynarlar ve sık sık istatistik testlerde tablo haline getirilirler. Bütünlük açısından şimdilik sadece tanımları vereceğiz, fakat, dersin sonlarına doğru (son üçte birinde) tekrar uygulamalara döneceğiz. İlgili p.d.f.leri vermeyeceğim çünkü onların kullanımı pratik değildir.

Tanım 6. *Eğer Z_1, Z_2, …, Z_k bağımsız ve dağılımı $Z_i \sim N(0,1)$ ise, $Y = \sum_{i=1}^{k} Z_i^2$ 'in k serbestlik derecesiyle ki-kare dağılımlı olduğu söylenir. Semboller ile ifade edecek olursak,*

$$Y \sim \chi_k^2$$

Burada "serbestlik derecesi" karesi alınmış ve toplanmış bağımsız çekilişlerin sayısını ifade etmektedir. Ki-karenin beklenen değeri serbestlik derecesi ile verilir,

$$Y \sim \chi_k^2 \Rightarrow \mathbb{E}[Y] = \sum_{i=1}^{k} \mathbb{E}[X_i^2] = k$$

Tanım 7. *Eğer $X \sim N(0,1)$ ve $Y \sim \chi_k^2$ ise, o zaman*

[1] bkz. Stigler, S. (2002): Statistics on the Table

$$T = \frac{Z}{\sqrt{Y}} \sim t_k$$

ilişkisinin k serbestlik derecesiyle (öğrenci) t-dağılımı olduğu ifade edilir.

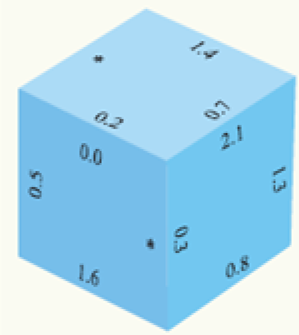
Serbestlik derecesi k'nin büyük değerleri için, t- dağılımı standart normal dağılımı ile çok doğru bir şekilde tahmin edilir.

Tanım 8. *Eğer* $Y_1 \sim \chi^2_{k_1}$ *ve* $Y_2 \sim \chi^2_{k_2}$ *ise, o zaman*

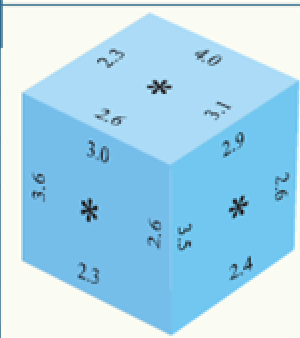
$$F = \frac{Y_1/k_1}{Y_2/k_2} \sim F(k_1, k_2)$$

ilişkisinin (k_1, k_2) serbestlik derecesiyle F- dağılımlı olduğu ifade edilir.

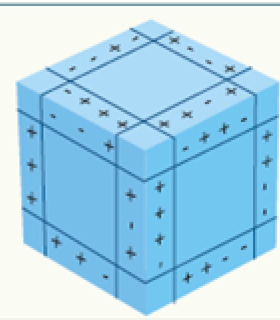
Zar I'in Değerleri			
0.03	0.51	1.04	1.78
0.11	0.59	1.14	1.95
0.19	0.67	1.25	2.15
0.27	0.76	1.37	(2.40)
0.35	0.85	1.50	(2.75)
0.43	0.94	1.63	(3.60)



Zar II'nin Değerleri			
2.29	2.51	2.77	3.25
2.32	2.55	2.83	3.36
2.35	2.59	2.90	3.49
2.39	2.64	2.98	3.65
2.43	2.68	3.06	4.00
2.47	2.72	3.15	4.55

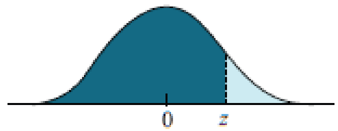


Zar III'ün Değerleri			
+ + + +	+ − − +	− − + +	+ − +
+ + + −	+ − − −	− − + −	+ − −
+ + − +	− + + +	− − − +	− + +
+ + − −	− + + −	− − − −	− + −
+ − + +	− + − +	+ + +	− − +
+ − + −	− + − −	+ + −	− − −



Şekil 3. Galton’nun Zarının Üç Çeşidi. Zarlar 1890’dan kalma 1.25 inçlik küptürler ve normal dağılımlı rasgele değişkeni taklit etmek için kullanılırlar. Stigler, S. (2002): Statistics on the Table: The History of Statistical Concepts and Methods’tan uyarlanmıştır.

Standart Normal Dağılım Altındaki Birikimli Alan



(Devam)

z	0	1	2	3	4	5	6	7	8	9
-3	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0238	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0300	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0570	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3112
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Kaynak: MIT OpenCourseWare

Standart Normal Dağılım Altındaki Birikimli Alan

(Devam)

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2..3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	09988	0.9989	0.9989	0.9989	0.9990	0.9990

Source: B. W. Lindgren, *Statistical Theory* (New York: Macmillan. 1962), pp. 392-393.

Kaynak: MIT OpenCourseWare