14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Bahar 2009

14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Ders Notları 7

Konrad Menzel

26 Şubat 2009

1. X,Y gibi 2 Rasgele Değişkenin Birleşik Dağılımı (devamı)

X ve Y aynı S örneklem uzayında tanımlanmış iki sürekli rasgele değişkendir. (X,Y)'nin bileşik p.d.f.'si, $f_{XY}(x,y)$, (x,y) düzleminin herhangi bir alt kümesi olan A için aşağıdaki ifadeyi sağlayan bir fonksiyondur.

$$P((X,Y) \in A) = \int\int_A f_{XY}(x,y)dxdy$$

Tek-değişkenli durumda olduğu gibi, bu yoğunluk aşağıdaki koşulları sağlamak zorundadır:

her bir $(x, y) \in \mathbb{R}^2$ için $f_{XY}(x,y) \geq 0$

ve

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dxdy = 1$$

Böylece,

- herhangi bir nokta sıfır olasılıklıdır
- düzlemdeki herhangi bir tek-boyutlu eğri sıfır olasılıklıdır

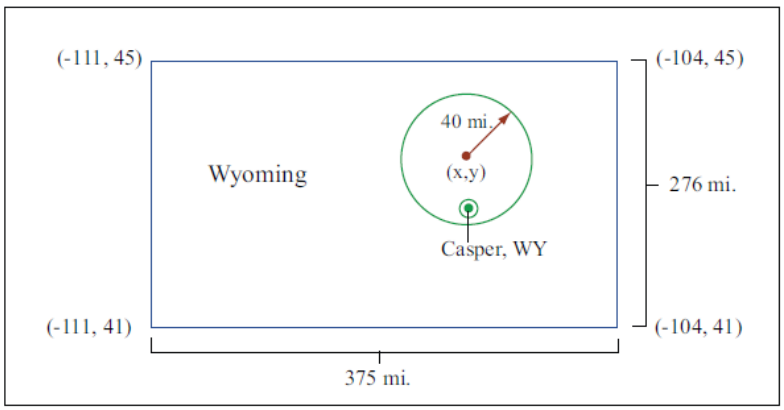
Örnek 1. *Wyoming üzerinde rasgele bir yerde bir UFO görülür. Dünyanın eğimini görmezlikten gelirsek, Wyoming gerçekçi bir şekilde 276 x 375 mil'lik bir dikdörtgen olarak tarif edilebilir. UFO'nun konumu bütün eyalet üzerinde uniform olarak dağılmıştır ve rasgele boylam değeri X (111-104 derece arası) ve rasgele enlem değeri Y (41-45 arası) olarak ifade edilsin.*

Bu, koordinatların birleşik yoğunluğunun aşağıdaki ifade ile verildiği anlamına gelir.

$$f_{XY}(x,y) = \begin{cases} \frac{1}{28} & \text{eğer} \quad -111 \le x \le -104 \quad \text{ve} \quad 41 \le y \le 45 \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

Eğer bir UFO 40 mil'lik bir uzaklıktan görülebilirse, eyaletin aşağı yukarı tam ortasında yer alan Casper'dan görülme olasılığı nedir?

Şimdi soruna grafik yardımıyla ile bakalım: Casper'dan UFO'nun görülebileceği yerlerin kümesi, Casper etrafında 40 Mil yarıçaplı bir daire ile ifade edilebilir. Aynı zamanda, uniform yoğunluğa göre, UFO'nun eyaletin A yerinden görülebilme olasılığı (yani A üzerinden sabit yoğunluğun integrali) A'nın kapladığı alan ile orantılıdır. Dolaysıyla, herhangi bir integral almak zorunda değiliz, çünkü olasılığı bulmak tamamen geometrik bir uygulamaya indirgenebilir.



Kaynak: MIT OpenCourseWare

Şekil 1. Casper, WY'de görülebilen (x, y)'deki UFO

Olasılığı şöyle hesaplayabiliriz:

$$P(Casper'dan\ itibaren\ 40\ mil'den\ az) = \frac{Alan(Casper'dan\ itibaren\ 40\ mil'den\ az)}{Alan(Bütün\ Wyoming)}$$

$$= \frac{40^2\pi}{375x276} \approx \%4.9$$

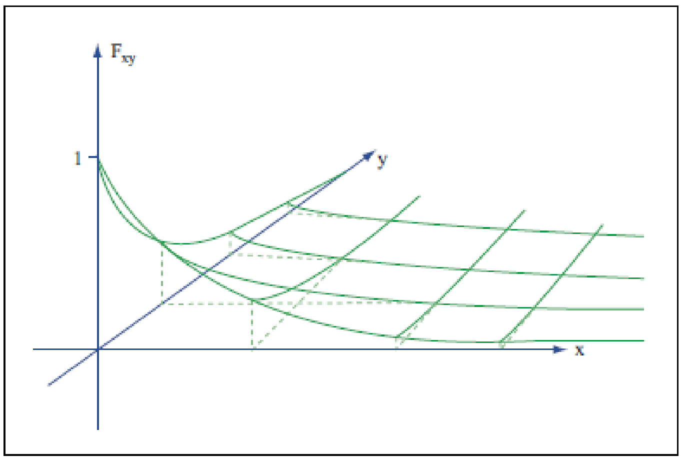
Uniform dağılımı için genellikle karmaşık integral almaya ihtiyaç duyulmadığına dikkat etmelisiniz, çünkü her şeyi salt geometrik olarak işleyebilirsiniz.

Son örnektekinden farklı olarak, olasılıkları elde etmek için yoğunluk fonksiyonun integralini almanın yolu yoktur, çünkü herhangi bir sabit olmayan yoğunluk, olasılık yığılması açısından farklı bölgeleri yeniden ağırlıklandırır. Bunu, açık ve sistematik bir şekilde izleyen örnekte göreceğiz:

Örnek 2. *Çim biçme makinenizde 2 buji olduğunu varsayalım, ve X 1'nci bujinin ömrünü Y'de 2'nci bujinin ömrünü temsil etsin. Varsayalım ki dağılımı aşağıdaki gibi ifade edebiliyoruz.*

$$f_{XY}(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & \text{eğer} \quad x \geq 0 \quad \text{ve} \quad y \geq 0 \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

Sayfa 4'teki Şekil 2 birleşik yoğunluğun nasıl görüneceğini göstermektedir.

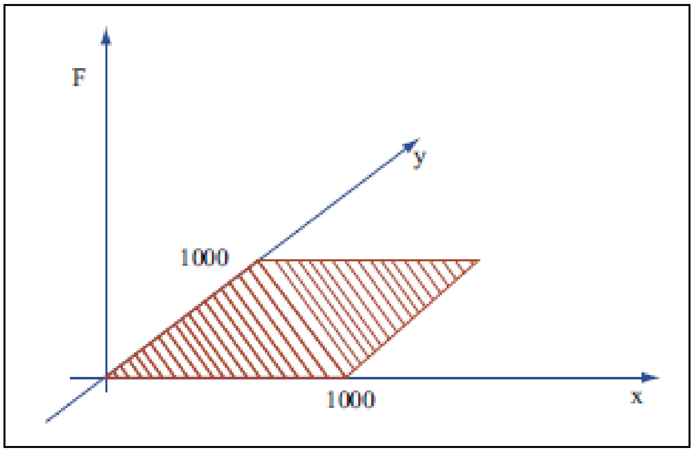


Kaynak: MIT OpenCourseWare

Şekil 2: 1. ile 2. Bujinin X ve Y Ömürlerinin Birleşik Yoğunluğu

Doğrusu, bu yoğunluk bujilerin birbirinden bağımsız olarak zaman içerisinde değişmeyen sabit bir λ oranında bozulduğu varsayımından türetilebilir.

Eğer her iki buji çalıştığı sürece çim biçme makinesi de çalışacaksa, çim biçme makinesinin 1000 saat içerisinde bozulma olasılığı nedir?



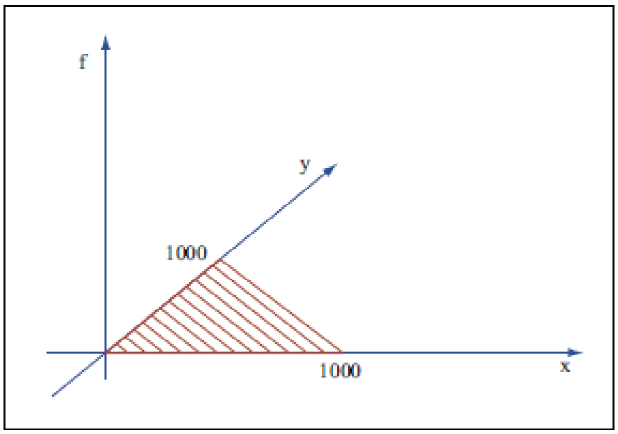
Kaynak: MIT OpenCourseWare

Şekil 3. “Çim Biçme Makinesi 1000 Saatten Önce Bozulur” Olayının İlk Durumu

$$
\begin{aligned}
P(X \leq 1000, Y \leq 1000) &= \int_0^{1000} \int_0^{1000} \lambda^2 e^{-\lambda(x+y)} dy dx \\
&= \int_0^{1000} \int_0^{1000} \lambda^2 e^{-\lambda x} e^{-\lambda y} dy dx \\
&= \int_0^{1000} \lambda e^{-\lambda x} \left(\int_0^{1000} \lambda e^{-\lambda y} dy \right) dx \\
&= \int_0^{1000} \lambda e^{-\lambda x} \left(1 - e^{-1000\lambda}\right) dx \\
&= \left(1 - e^{-1000\lambda}\right)^2
\end{aligned}
$$

Söz konusu olasılık, birinci bujinin bozulması halinde sadece ikinci bujinin kullanılması durumunda nedir? Yani $P(X + Y \leq 1000)$'i nasıl hesaplarız? Bunun sadece ilgilendiğimiz “olayı” değiştirdiğine dikkat ediniz, yani integralini hesapladığımız $\mathbb{R}^2$ alanı değişti, halbuki hala aynı yoğunluğun integralini alıyoruz.

$$
\begin{aligned}
P(X+Y \leq 1000) &= \int_0^{1000} \left[\int_0^{1000-x} \lambda^2 e^{-\lambda x} e^{-\lambda y} dy\right] dx \\
&= \int_0^{1000} \lambda e^{-\lambda x} \left[\int_0^{1000-x} \lambda e^{-\lambda y} dy\right] dx \\
&= \int_0^{1000} \lambda e^{-\lambda x} \left[1 - e^{-\lambda(1000-x)}\right] dx \\
&= \int_0^{1000} \lambda \left[e^{-\lambda x} - e^{-1000\lambda}\right] dx \\
&= 1 - e^{-1000\lambda} - 1000\lambda e^{-1000\lambda} = 1 - (1+1000\lambda)e^{-1000\lambda}
\end{aligned}
$$



Kaynak: MIT OpenCourseWare

Şekil 4: "Çim Biçme Makinesi 1000 Saatten Önce Bozulur" Olayının İkinci Durumu

Tekrar etmek gerekirse, sürekli iki-değişkenli rasgele değişkenler ile alakalı olaylar düzlemdeki alanlara tekabül eder. Ve bu alanlar üzerinden yoğunluğun integralini alarak olasılığı buluruz.

2. X,Y gibi 2 Rasgele Değişkenin Birleşik c.d.f.'si

Sadece tanımları vereceğim. Bu derste bunu çok fazla kullanmayacağız, fakat bunu görmeniz gerekiyor.

Tanım 1. *Rasgele değişkenler (X, Y) için birleşik c.d.f., $(x, y) \in \mathbb{R}^2$ için, $F_{XY}(x,y)$ fonksiyonu olarak tanımlanır.*

$$F_{XY}(x,y) = P(X \leq x, Y \leq y)$$

Birleşik c.d.f.'den olasılıkları aşağıdaki gibi hesaplarız.

$$P(a \leq X \leq b, c \leq Y \leq d) = F(b,d) - F(a,d) - F(b,c) + F(a,c)$$

Son terimi eklemek zorundayız çünkü daha önce iki kere çıkarıldı.

Birleşik c.d.f.'ler p.d.f'ler ile aşağıdaki şekilde ilişkilidir: sürekli rasgele değişkenler için

$$F_{XY}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{XY}(u,v)\,du\,dv$$

$$f_{XY}(x,y) = \frac{\partial^2}{\partial y \partial x} F_{XY}(x,y)$$

Kesikli durumda,

$$F_{XY}(x,y) = \sum_{u \leq x}\sum_{v \leq y} f_{XY}(u,v)$$

3. Marjinal p.d.f.ler

Eğer birleşik dağılımlarımız varsa, tek değişken X'in dağılımını yeniden elde etmek isteyebiliriz. Eğer *X* ve *Y* bileşik p.d.f.'si f_{XY} olan *kesikli* rasgele değişkenler ise, o zaman

$$f_X(x) = \sum_{\text{Tüm } y} f_{XY}(x,y)$$

$$f_Y(y) = \sum_{\text{Tüm } x} f_{XY}(x,y)$$

Eğer X ve Y sürekli ise, esas itibariyle toplamı integral ile değiştirmemiz gerekir, böylece

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx$$

Örnek 3. *Bu örnek 1977'de[1] Redbook dergisi tarafından toplanan gerçek evlilik-dışı ilişkiler verisine dayanmaktadır. Ankette, kişilerden evliliklerini 1'den (mutsuz) 5'e (mutlu) ölçeklemeleri ve evlilik-dışı ilişkilerinin sayısı bölü yıllar itibariyle evlilik süresini belirtmeleri istenmiş. Şimdilik önce "evlilik kalitesi" X ile yıllar itibariyle evlilik süresi Y'nin*

[1]Verinin mevcut olduğu adres: http://pages.stern.nyu.edu/wgreene/Text/Edition6/tablelist6.htm

bileşik dağılımına bakalım. P.d.f.'den elde edilen "hücre" olasılıklarından başlayabiliriz, ve daha sonra tablonun solunda ve altında bulunan marjinal p.d.f.'leri doldurabiliriz:

İlginç bir şekilde, her ne kadar marjinal dağılımlar eşit dağılmışsa da, birleşik dağılımlar tablonun alt sol ve üst sağ köşelerinde, birleşik p.d.f.lerde daha düşük değerler alarak üst sol ve alt sağ köşelerinde yoğunlaşmış gibi görünüyor.

	f_{XY}	Y = 1	Y = 8	Y = 12	f_X
	1	4.66%	11.48%	12.98%	29.12%
X	2	5.16%	14.81%	12.31%	32.28%
	3	13.48%	16.47%	8.65%	38.60%
	f_Y	23.30%	42.76%	33.94%	100.00%

Örnek 4. Geçen seferki iki bujili örneğini hatırlayınız. Birleşik p.d.f aşağıdaki gibiydi:

$$f_{XY}(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & \text{eğer } x \geq 0 \text{ ve } y \geq 0 \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

X'in marjinal yoğunluğu şöyle olur:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \int_0^{\infty} \lambda^2 e^{-\lambda(x+y)} dy \\ &= \lambda e^{-\lambda x} \int_0^{\infty} \lambda e^{-\lambda y} dy = \lambda e^{-\lambda x}[1-0] = \lambda e^{-\lambda x} \end{aligned}$$

Benzer şekilde,

$$f_Y(y) = \lambda e^{-\lambda y}$$

4. Bağımsızlık

Eğer P(AB) = P(A)P(B) ise, A ve B olaylarının bağımsız olduğunu söylemiştik. Şimdi benzer bir kavramı rasgele değişkenler için tanımlayacağız.

Tanım 2. *Eğer herhangi bir A, B ⊂ ℝ bölgesi için aşağıdaki ilişki sağlanırsa rasgele değişkenler olan X ile Y'nin bağımsız olduğunu söyleyebiliriz,*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Bu gereklilik koşulunun katı olduğuna dikkat ediniz: X ∈ A ve Y ∈ B türü olaylara bakıyoruz ve o çiftlerin tümünün karşılıklı bağımsız olmasını istiyoruz.

Bu tanım kendi başına pratik değildir çünkü kontrol etmek zordur, ancak eğer X ile Y bağımsız ise, tanımdan hareketle şu ifadeyi elde ederiz:

$$F_{XY}(x,y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$$

Buradan, genellikle doğrulanması daha kolay olan aşağıdaki koşulu elde edebiliriz.

Önerme 1. *X ve Y bağımsızdır eğer sadece ve sadece bunların birleşik ve marjinal p.d.f.leri aşağıdaki ilişkiyi sağlarsa,*

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

İSPAT: Kesikli rasgele değişkenler için, bu doğrudan A = {x} ve B = {y} tanımları uygulanarak elde edilebilir. Sürekli değişkenler için, eğer X ile Y bağımsız ise

$$F_{XY}(x,y) = F_X(x)F_Y(y)$$

denkleminin her iki tarafının da türevini alarak aşağıdakine ulaşırız:

$$f_{XY}(x,y) = \frac{\partial^2}{\partial y \partial x} F_{XY}(x,y) = \frac{\partial^2}{\partial y \partial x}[F_X(x)F_Y(y)] = \frac{\partial}{\partial y} f_X(x)F_Y(y) = f_X(x)f_Y(y)$$

Bunun aksine, eğer marjinal p.d.f.lerin çarpımı birleşik p.d.f'lere eşit ise, o zaman integralini alırız

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B f_{XY}(x,y)dydx = \int_A \int_B f_X(x)f_Y(y)dydx \\ &= \left(\int_A f_X(x)dx\right)\left(\int_B f_Y(y)dy\right) \end{aligned}$$

Dolaysıyla marjinal üzerindeki koşul bağımsızlığı sağlar ve eşitliğin her iki yönünü de ispatlamış oluruz.

Örnek 5. *Evlilik Dışı ilişkiler örneğine geri dönecek olursak, "evlilik kalitesi" X ile yıllar itibariyle evlilik süresi Y'nin marjinal p.d.f.lerini rapor ettiğimizi hatırlayınız, yani*

$$f_X(1) = 29.12\%, \quad f_X(2) = 32.28\%, \quad f_X(3) = 38.60\%$$

ve

$$f_Y(1) = 23.30\%, \quad f_Y(8) = 42.76\%, \quad f_Y(12) = 33.94\%$$

Eğer iki rasgele değişken gerçekten de bağımsız ise birleşik dağılım nasıl görünmelidir? Burada şunu elde ederiz.

$$\tilde{f}_{XY}(3,1) = f_X(3)f_Y(1) = 38.60\% \cdot 23.30\% = 8.99\%$$

Birleşik p.d.f.'nin o noktadaki gerçek değeri 13.48'di ve açıkça görüldüğü gibi iki değişken bağımsız değildir. Şimdi bağımsızlık varsayımı altında tablonun geri kalanını doldurabiliriz: Bunu son tablomuzla karşılaştırınca bazı sistematik çelişkiler olduğunu görebiliriz, özelikle oluşturulan birleşik p.d.f. $\tilde{f}_{XY}$ belirgin bir şekilde diyagonalde yoğunlaşmamaktadırlar, halbuki bu gerçek birleşik p.d.f.lerin üzerinde durmaya değer özelliğiydi.

	$\tilde{f}_{XY}$	Y: 1	8	12	f_X
X	1	6.78%	12.45%	9.88%	29.12%
	2	7.52%	14.81%	10.96%	32.28%
	3	8.99%	16.50%	13.10%	38.60%
	f_Y	23.30%	42.76%	33.94%	100.00%

Fakat gerçekten bu X ile Y'nin bağımsız olmadığı anlamına mı gelir? Dikkat edilmesi gereken bir husus, olasılıkları, belirtilen dağılımdan "çekilen" örneklemden edindiğimiz birleşik p.d.flerden hesapladık, bu nedenle gerçek hücre olasılıklarını doğru bir şekilde ölçebilme konusunda bazı belirsizlikler var. Dersin son bölümünde, "oluşturulan" ve gerçek p.d.f.ler arasındaki farkın X ile Y'nin bağımsız olmadığını gösterecek kadar büyük olup olmadığını formal olarak öneren bir yöntem göreceğiz.

Örnek 6. *Daha önceki iki bujili örneğini hatırlayınız. Birleşik p.d.f. aşağıdaki gibiydi*

$$f_{XY}(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & \text{eğer } x \geq 0 \text{ ve } y \geq 0 \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

ve bir önceki bölümde de marjinal p.d.f.leri elde etmiştik

$$\begin{aligned} f_X(x) &= \lambda e^{-\lambda x} \\ f_Y(y) &= \lambda e^{-\lambda y} \end{aligned}$$

Dolayısıyla, bunların çarpımı şöyledir:

$$f_X(x) f_Y(y) = \lambda^2 e^{-\lambda x} e^{-\lambda y} = f_{XY}(x, y)$$

Böylece 1'nci ve 2'nci bujinin yaşam süreleri bağımsızdır.

Açıklama 1: *Sürekli rasgele değişkenler için, bileşik ve marjinal yoğunlukların bağımsızlığının koşulu aşağıdaki gibi yeniden belirtilebilir: Her ne zaman bileşik p.d.f'nin faktörünü alırsak, yani,*

$$f_{XY}(x, y) = g(x)h(y)$$

o zaman X ile Y bağımsızdır. Burada g(.) sadece x'e ve h(.) sadece y'ye bağlıdır. Özelikle belirtmek gerekirse, marjinal yoğunlukları doğrudan hesaplamak zorunda değiliz.

Örnek 7. *Diyelim ki, aşağıdaki gibi bir birleşik p.d.f'miz var*

$$f_{XY}(x, y) = \begin{cases} ce^{-(x+2y)} & \text{eğer} \quad x \geq 0, y \geq 0 \text{ ise} \\ 0 & \text{diğer bütün değerler için} \end{cases}$$

O zaman, örneğin $g(x) = ce^{-x}$ ve $h(y) = e^{-2y}$ 'dir diyebiliriz. Her ne kadar bunlar uygun yoğunluklar değilse de, X ile Y'nin bağımsız olduğunu göstermek için yeterlidir.

Örnek 8. *Varsayalım ki aşağıdaki gibi bir birleşik p.d.f.miz var.*

$$f_{XY}(x, y) = \begin{cases} cx^2 y & \text{eğer} \quad x^2 \leq y \leq 1 \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

X ve Y bağımsız olabilir mi?

Her iki durumda (yani $x^2 \leq y \leq 1$ sağlanır mı yoksa sağlanmaz mı?) p.d.f. x ve y'nin fonksiyonlarına ayrıştırılırsa (faktör, ÇN) (sıfır bölümü için doğruluğu şüphelidir) bile, X'in desteğinin Y'ye bağımlı olduğunu görebiliriz ve dolaysıyla X ve Y bağımsız olamaz-örneğin eğer $X \geq 1/2$ ise, $Y \geq 1/4$ olmak zorundadır. Böylece

$$P\left(X \geq \frac{1}{2}, Y \leq \frac{1}{4}\right) = 0 < P\left(X \geq \frac{1}{2}\right) P\left(Y \leq \frac{1}{4}\right)$$

X ve Y'nin bağımsız olabilmesi için iki rasgele değişkenin bileşik desteğinin dikdörtgen (büyük ihtimalle bütün $\mathbb{R}^2$) olmak zorunda olduğuna dikkat ediniz. Eğer değilse, gerçekleşen bazı X değerleri için, belli Y değerleri gerçekleşebilecekken, gerçekleşemez. Fakat eğer doğruysa, yani X'in Y hakkında bilgi verdiğini bilirsek, o zaman onlar bağımsız olamaz. Destek için bu koşul tek başına bağımsızlığı sağlar.