14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Bahar 2009

14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Ders Notları 6

Konrad Menzel

24 Şubat 2009

1. Örnekler

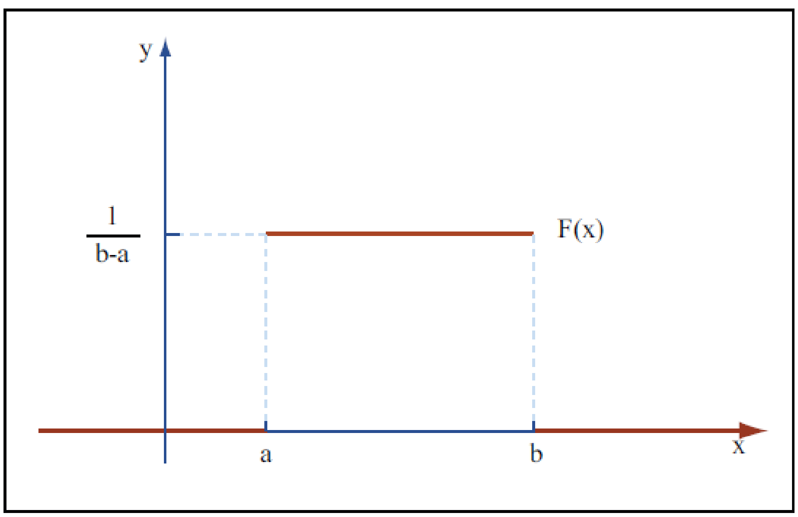
Bir rasgele değişkenin, reel eksen üzerindeki bazı [a, b] aralıklarında yer aldığını varsayalım, X'in bazı [a',b'] (burada a ≤ a' ≤ b'≤ b) alt gruplarına ait olma olasılığı, alt aralığın uzunluğu ile orantılıdır. .

Tanım 1. *Eğer aşağıdaki gibi bir olasılık yoğunluk fonksiyonuna sahip ise, X rasgele değişkeni [a, b], a < b, aralığında uniform dağılır.*

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

Sembolik olarak şöyle yazarız:

$$X \sim U[a, b]$$



Kaynak: MIT OpenCourseWare

Şekil 4. Uniform Rasgele Değişken için p.d.f., *X ~ [a, b]*

Örneğin, eğer *X ~ U[0, 10]* ise, o zaman

$$P(3 < X < 4) = \int_3^4 f(t)dt = \int_3^4 \frac{1}{10}dt = \frac{1}{10}$$

P(3 ≤ X ≤ 4) nedir? Olasılık P(X = 3) = 0 = P(X = 4) olduğu için, bu P(3 < X < 4)'ün aynısıdır.

Örnek 1. *Varsayalım ki X'in p.d.f.'si şöyledir:*

$$f_X(x) = \begin{cases} ax^2 & eğer\ 0 < x < 3\ ise \\ 0 & diğer\ bütün\ durumlarda \end{cases}$$

a ne olmak zorundadır? P(X ∈ ℝ) = 1 olduğu için, yoğunluk 1'e entegre olmalı, böylece a

$$1 = \int_{-\infty}^{\infty} f_X(t)dt = \int_0^3 at^2 dt = \left[\frac{ax^3}{3}\right]_0^3 = \frac{27}{3}a - 0 = 9a$$

olmak zorundadır. Böylece a = $\frac{1}{9}$.

P(1 < X < 2) nedir? Aşağıdaki integrali hesaplayalım

$$P(1 < X < 2) = \int_1^2 \frac{t^2}{9}dt = \frac{2^3}{9 \cdot 3} - \frac{1^3}{9 \cdot 3} = \frac{7}{27}$$

P(1 < X) nedir?

$$P(1 < X) = \int_1^{\infty} f_X(t)dt = \int_1^3 \frac{t^2}{9}dt = \frac{27-1}{27} = \frac{26}{27}$$

1.1. Karışık Rasgele Değişkenler/Dağılımlar

Esas itibariyle iki farklı sebepten ötürü birçok gerçek-dünya verisi bazı değerler için nokta etrafında yığılma gösterir:

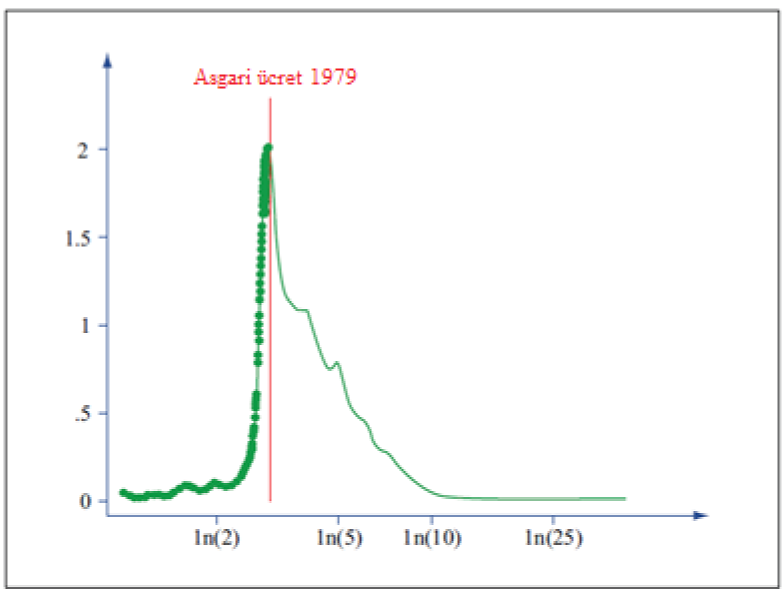
- bazı sonuçlar mekanik olarak bazı değerler ile sınırlandırılmıştır, böylece bir çok olasılık yığını rasgele değişkenin genişliğinin tam köşelerinde birikir, örneğin günlük yağış miktarı herhangi bir reel pozitif değer alır, fakat yağış miktarının sıfır olduğu bir çok gün vardır.

- ekonomik kararlar alan kişiler kendilerini kırılganlıklar ve kesintilere göre konumlandırarak belli kurumsal kuralara tepki gösterirler. Örneğin sosyal güvenlik kurumlarına ve gelirler genel müdürlüğüne rapor edilen gelirlere baktığımızda, vergi aralıklarının üst sınırlarında "yığılma" olduğunu görürüz (söz konusu kişiler için, gelirdeki en ufak artış vergi oranında büyük artış anlamına gelir).

Açık konuşmak gerekirse, ilgili dağılımlar sürekli değildir, çünkü gerçekleşme her ne kadar reel-değerli bir sayı olsa da, daha önceki bölümde yaptığımız gibi bir olasılık yoğunluk fonksiyonunu tanımlayamayız. Ancak nokta etrafında yığılma ile ayrıca ilgileneceğiz. Bunun bir kısmı ekonometri dersinizde karşınıza çıkacak, bu nedenle bunun üzerinde şimdi çok fazla durmayacağız ve sadece bir örneğe bakacağız.

Örnek 2. *İzleyen grafik 1979 yılına ait Current Population Survey (CPS) verisi kullanılarak oluşturulmuştur*[1].

Yazar, grafik için geliri çok düşük bir alt grup seçmiştir, böylece asgari ücretin sınırlandırmasının aksine örneklem nispeten daha büyük olmuştur. 1979 yılının asgari ücret değerinin solunda kalan bazı kişiler var. Bu, muhtemelen kısmı olarak asgari ücret kanunu kapsamı dışında kalan sektörleri yansıtmaktadır (örneğin, çiftçilik, genç işçiler gibi).



Kaynak: MIT OpenCourseWare

Şekil 2. 1979'da Lise'den Terk Kadınların Logaritmik Ücretleri

[1] DiNardo, J., N. Fortin ve T. Lemieux. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." Econometrica 64, no. 5 (1996): 1001- 1044'teki Şekil 3b.

2. Birikimli Dağılım Fonksiyonu (c.d.f)

Tanım 2. *Bir rasgele değişken X'in Birikimli Dağılım Fonksiyonu (c.d.f.) F_X her bir reel sayı için aşağıdaki gibi tanımlanır.*

$$F_X(x) = P(X \leq x)$$

Dikkat edilecek olursa, bu tanım kesikli, sürekli ve karışık rasgele değişkenler için aynıdır. Özelikle, X'in kesikli olmasına olanak verdiğimizden ve *$P(X \leq x)$'in $P(X < x)$'* den farklı olduğunu akılda tutmak koşuluyla, ilintili olayları birbirinden ayırt etmek önemli olmaktadır. C.d.f.'nin tanımında, X'i her zaman x'ten "küçük veya eşittir" şeklinde kullanacağız.

C.d.f. bir olasılık olduğu için, olasılık fonksiyonunun bütün özeliklerini içinde barındırır. Özelikle,

Özelik 1. *C.d.f sadece 0 ile 1 arasındaki değerleri alır.*

$$\text{Bütün } x \in \mathbb{R} \text{ için } 0 \leq F_X(x) \leq 1$$

Ayrıca, $x_1 < x_2$ için $X \leq x_1$ olayı $X \leq x_2$'nin *içinde* yer aldığından Özelik 2'yi elde ederiz.

Özelik 2. *F_x x'in azalmayan bir değeridir, yani*

$$x_1 < x_2 \text{ için } F_x(x_1) \leq F_x(x_2)$$

Eğer $x \to -\infty$ ise, olasılığın gerçekleşmesi bakımında ($X \leq x$) olayı imkânsız olaya "yakın" olur (burada bunun ne anlama geldiği konusunda emin değilim), halbuki eğer $x \to \infty$ ise ($X \leq x$) olayı neredeyse kesindir ve böylece Özelik 3 elde edilir.

Özelik 3.

$$\begin{aligned} \lim_{x \to -\infty} F(x) &= 0 \\ \lim_{x \to \infty} F(x) &= 1 \end{aligned}$$

Burada şuna dikkat etmek gerekiyor, eğer *sol limiti*

$$F(x^-) = \lim_{h>0, h \to 0} F(x-h)$$

olarak ve *sağ limiti* de

$$F(x^+) = \lim_{h>0, h \to 0} F(x+h)$$

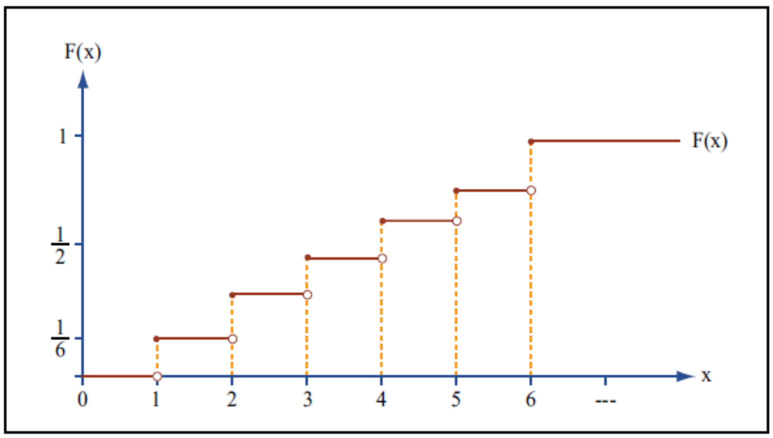
olarak tanımlarsak, C.d.f. her zaman sürekli olmak zorunda değildir.

x'te sürekli olabilmek için, F(x), $F(x^-) = F(x^+)$'i sağlamak zorundadır. Aşağıdaki örnekte görüldüğü gibi, genel olarak bu doğru değildir.

Örnek 3. *Rasgele değişken X'in zar atma sayısı ile ilişkilendirildiği, zar atma örneğini tekrar düşünün. O durumda X'in c.d.f.'si aşağıdaki gibi verilir.*

$$F_X(x) = \begin{cases} 0 & \text{eğer} \quad x < 1 \\ \frac{1}{6} & \text{eğer} \quad 1 \leq x < 2 \\ \cdots & \quad\quad \cdots \\ \frac{5}{6} & \text{eğer} \quad 5 \leq x < 6 \\ 1 & \text{eğer} \quad x \geq 6 \end{cases}$$

Burada 1, 2, …, 6 sayılarında süreksiz atlamalar vardır.



Kaynak: MIT OpenCourseWare

Şekil 3: Zar atmanın c.d.f.'si

Ancak, reel analizlerin bir sonucu olarak, bir monoton fonksiyon (burada özelikle c.d.f F_X) sadece *sayılabilir süreksiz birçok nokta*ya sahip olabilir.

Daha da iler gitmek gerekirse, her zaman Özelik 4'e sahibiz.

Özelik 4. *Herhangi bir c.d.f. sağ-süreklidir, yani*

$$F(x) = F(x^+)$$

C.d.f'nin daha fazla özeliğini gösterebilmek için şimdi olasılıkla ilgili bilgimizi kullanabiliriz.

Önerme 1. Verilen herhangi bir x için,

$$P(X > x) = 1 - F_X(x)$$

İSPAT: Olasılığın özeliklerinden hareketle,

$$P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)$$

Aynı şekilde,

Önerme 2. *İki reel sayı için* $x_1 < x_2$

$$P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$$

Önerme 3. *Herhangi bir x için,*

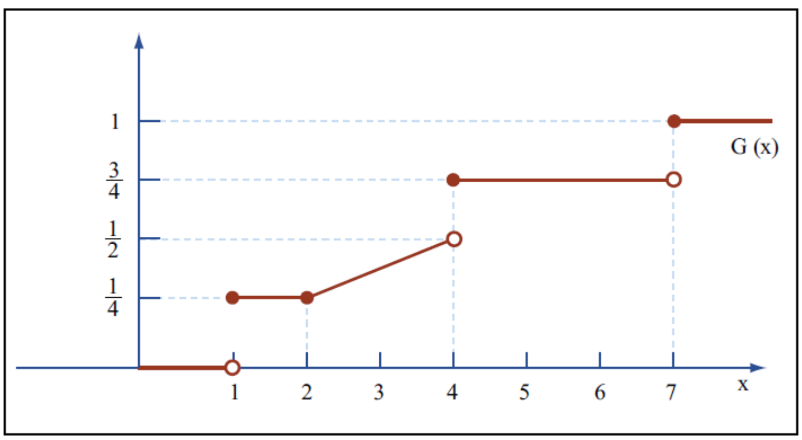
$$P(X < x) = F(x^-)$$

Önerme 4. *Herhangi bir x için,*

$$P(X = x) = F(x^+) - F(x^-)$$

Bu son sonuçlar sürekli değişkenler için bütün x değerleri için özelikle (P(X = x) = 0 anlamına gelir.

Örnek 4. *Aşağıdaki grafikte* $G_x(x)$ *fonksiyonun bir c.d.f. olup olmadığını kontrol edelim.*

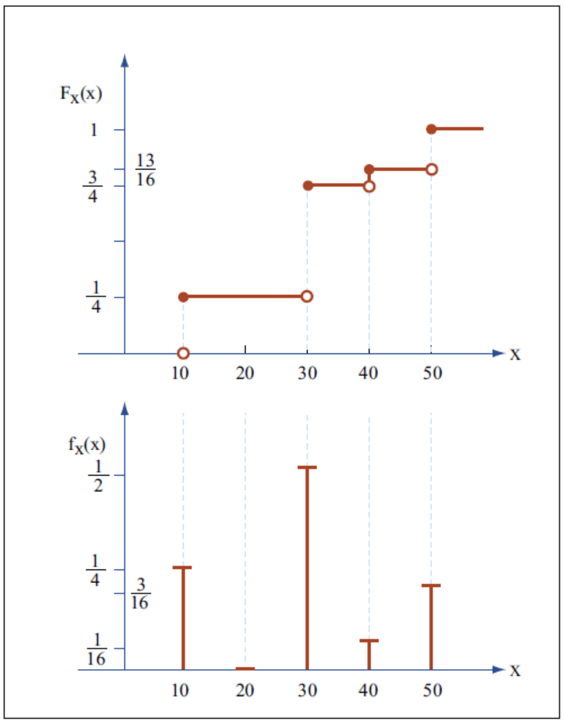


Kaynak: MIT OpenCourseWare

Fonksiyon 0 ile 1 arasındadır, monotonik artmaktadır ve sağ-süreklidir. Şimdi son 4 önermeyi bu örneğe uygulayalım(sadece grafikten doğrudan rakamlar elde edilecektir):

- $P(X > 4) = 1 - F(4) = 1 - \frac{3}{4} = \frac{1}{4}$
- $P(3 < X \leq 4) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}$
- $P(X < 4) = F(4^-) = \frac{1}{2}$
- $P(X = 4) = F(4^+) - F(4^-) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$

Örnek 5. *P.d.f ve c.d.f'yi bir birine bağlayan tek-doğru formülümüzün olduğu sürekli rasgele değişken durumundan farklı olarak, kesikli durumda daha yeni tartıştığımız c.d.f'den elde edilen olasılık sonuçlarını kullanmak zorundayız. Şimdi ilişkiye başka bir grafik örneğiyle bakalım:*



Kaynak: MIT OpenCourseWare

Şekil 4: Kesikli bir rasgele değişken için c.d.f ve p.d.f

2.1 Sürekli Rasgele Değişkenler için p.d.f ve c.d.f

Eğer X p.d.f *f(x)*'li ve *F(x)*'li sürekli bir dağılıma sahip ise (karışıklık olmadığı sürece bundan sonra X için altsimgeyi kullanmayacağım), o zaman

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Kalkülüsün temel teoreminden hareketle, c.d.f ve p.d.f. arasındaki ilişkiyi bu durumda aşağıdaki gibi yazabiliriz:

$$F'(x) = \frac{d}{dx}F(x) = f(x)$$

Örnek 6. *Aşağıdaki gibi bir fonksiyonumuz olsun,*

$$F(x) = \begin{cases} 0 & \text{eğer } x < 0 \\ \frac{x}{1+x} & \text{eğer } x \geq 0 \end{cases}$$

F(x) bir c.d.f mi? - Şimdi temel özelikleri kontrol edelim:

- $\lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to\infty} F(x) = 1$
- $F(\cdot)$ azalmayan fonksiyondur (aşağıdaki türetme kontorl edilebilir)

p.d.f. f(x) nedir?

$$f(x) = F'(x) = \begin{cases} 0 & \text{eğer } x < 0 \text{ ise} \\ \frac{1}{(1+x)^2} & \text{diğer bütün durumlarda} \end{cases}$$

f(x) bir p.d.f. mi? -Doğrusu, biz zaten temelde F(X)'in bir c.d.f. olduğunu göstermiştik. Aşağıdaki ifadeleri hemen görebiliriz

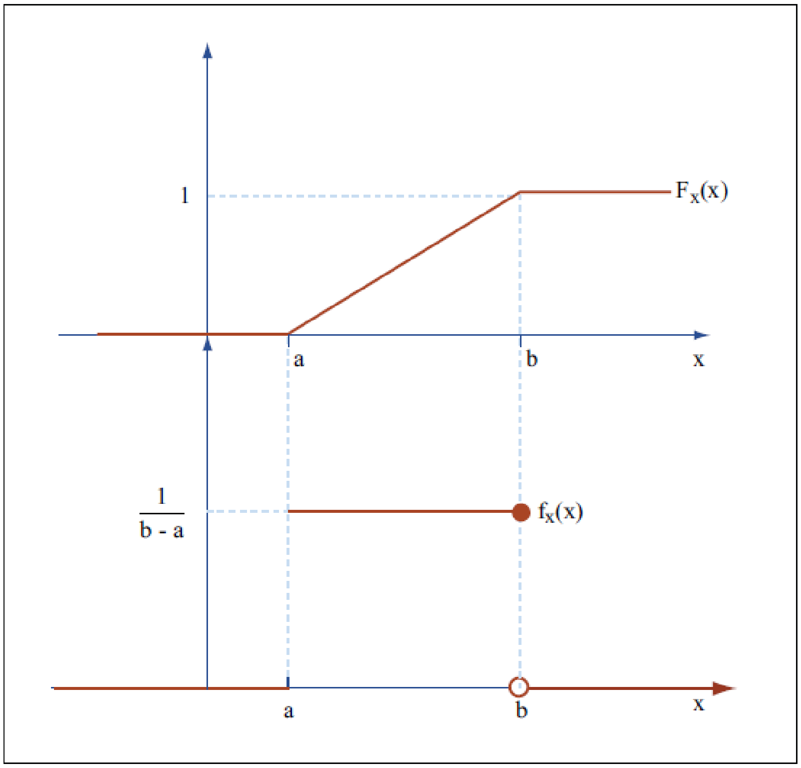
bütün x değerleri için, f(x) ≥ 0

ve aynı zamanda,

$$\int_{-\infty}^{\infty} f(t)dt = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1 - 0 = 0$$

Örnek 7. *Eğer X ~ U[0, 1] ise, o zaman c.d.f. aşağıdaki gibidir:*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & \text{eğer } x < 0 \\ x & \text{eğer } 0 \leq x < 1 \\ 1 & \text{eğer } x \geq 1 \end{cases}$$



Kaynak: MIT OpenCourseWare

Şekil 5: X ~ U[a, b] için p.d.f. ve c.d.f

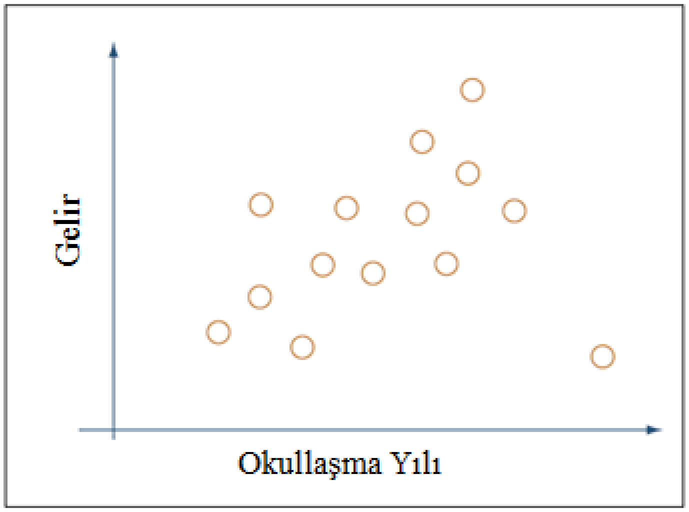
3. X, Y gibi 2 Rasgele Değişkenin Birleşik Dağılımı

Birçok durumda, sadece bir tek rasgele değişken ile değil, bir veya daha fazla değişken arasındaki ilişkiyle ilgileniriz, mesela bir sürecin sonucunun bir diğerinin sonucunu etkileyip etkilemediği gibi.

Örneğin aşağıdaki ilişkilere bakabiliriz:

- Tek yumurta ikizlerin IQ'sü - yani *X* çocuklardan birisinin, *Y* ise diğerinin IQ'sü olabilir.

- Eğitime katılım X ile gelir Y: Eğitim ile gelirin dağılımına ayrı ayrı bakabileceğimiz gibi, bir veri tabanından elde edilen verileri kullanarak iki değişkeni bir grafikte de gösterebiliriz. Doğrusu, grafikte iki değişken arasında şüphesiz bir ilişki varmış gibi görünüyor.



Kaynak: MIT OpenCourseWare

Şekil 6. Okullaşma ve Gelir

- Relaps (nüks) süresi: Kanseri ameliyat ile yok etmek her zaman mümkün olmadığı için, tıbbi prosedürlerin etkinliğini değerlendirmek isteyebiliriz. Bunu (a) yeni bir ameliyatın gerekli olduğunun anlaşılması için ne kadar zaman geçtiğine (X) ve (b) ne kadar süre sonra hastanın öldüğüne (Y) bakarak yaparız. Her ne kadar iki sonuçla ilgilensek de, her iki olay da bağımsızıdır: eğer yeni bir ameliyattan önce hasta ölürse, onun ölmemesi durumunda ne zaman yeniden ameliyat olmak zorunda kalacağını gözlemleyemeyiz.

Dersin bu bölümünde, iki (veya daha fazla) rasgele değişkenin, aralarındaki ilişki dahil, aynı anda özeliklerini ele alacağız. Aynı zamanda "bağımsızlık" ve "koşullu olasılık" olayları ile benzeşen kavramları da tanıtacağız. (X, Y) (ortaklaşa) aynı değerleri (x, y) alan iki rasgele değişken olsun. Her iki değişken sürekli, kesikli veya karışık olabilir.

3.1. Kesikli Rasgele Değişken

Kesikli durumda, herhangi bir (X, y) $\in \mathbb{R}^2$ için birleşik p.d.f. aşağıdaki gibidir:

$$f_{XY}(x, y) = P(X = x, Y = y)$$

Eğer $\{(x_1, y_2), \ldots, (x_n, y_n)\}$ (X,Y)'nin mümkün olan bütün değerlerini içerirse, o zaman

$$\sum_{i=1}^{n} f_{XY}(x_i, y_i) = 1$$

Herhangi bir alt küme $A \subset \mathbb{R}^2$ için,

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

Örnek 8. *Bir süpermarkette, X sıradan bir kasa sırasında bulunan müşterilerin sayısı, Y'de ekspres kasada bulunan müşterilerin sayısı olsun. Bu durumda X ile Y'nin ortak p.d.f'si aşağıdaki gibi görünebilir: Bu yapıdaki bir tablo, (X, Y)'nin birleşik p.d.f'lerinden elde edilen hücre-olasılıklarını özetleyen ve marjinal olasılıklarını yanda gösteren, ihtimal tablosu olarak adlandırılır. Daha öncede tartışıldığı gibi, tablodaki olasılıkların toplamı 1 olmalı ve nitekim öyleler.*

f_{XY}		Y					
		0	1	2	3	4	Toplam
	0	0.1	0.05	0.05	0	0	0.2
X	1	0.05	0.2	0.2	0.05	0	0.5
	2	0	0	0.1	0.1	0.05	0.25
	3	0	0	0	0	0.05	0.05
		0.15	0.25	0.35	0.15	0.1	1

Tablo değerlerinden, iki değişken arasında bir çeşit ilişki varmış gibi göründüğünü görebiliyoruz: Sıradan kasada ödeme yapan kişi sayısı yüksek olduğu zaman, ekspres kasadaki kişi sayısı da yüksek olma eğilimdedir.

Tabloda verilen p.d.f. lere dayanarak farklı olaylar için de olasılıklar hesaplayabiliriz:

$$\begin{aligned}
P(X=2) &= 0+0+0.1+0.1+0.05 = 0.25 \\
P(X \geq 2, Y \geq 2) &= \sum_{x=2}^{3}\sum_{y=2}^{4} f_{XY}(x,y) = 0.1+0.1+0.05+0+0+0.05 = 0.3 \\
P(|X-Y| \leq 1) &= P(X=Y) + P(|X-Y|=1) \\
&= 0.1+0.2+0.1+0+0.05+0.05+0.2+0+0.1+0+0.05 = 0.85
\end{aligned}$$