14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Bahar 2009

# 14.30 Ekonomide İstatistiksel Yöntemlere Giriş
# Ders Notları 12

Konrad Menzel

Mart 19 2009

## 1. Medyan ve Yüzdeliklerin Özelikleri

Rasgele bir değişkenin medyanını şöyle tanımlarız

$$P(X > \text{medyan}(X)) = 1/2$$

X kesikli ve c.d.f'de sıçramalara neden olan nokta yığılmalarına sahip ise, bu tanım yararlı olamayabilir, bu nedenle daha genel durumda, medyanı aşağıdaki gibi tanımlarız

$$\text{Medyan}(X) := \min\left\{m \in \mathbb{R} : P(X \leq m) \geq \frac{1}{2}\right\}$$

Dar tanım ile ilgili değişiklik, c.d.f.'nin süreksizliğe sahip olması ve 1/2'nin üzerine çıkmasıyla, medyanı sadece süreksizlik noktasında aramaktır. X dağılımının diğer yüzdeliklerini de tanımlayabiliriz.

**Tanım 1**. Rasgele bir değişken X için, $\alpha$ quantile aşağıdaki ile verilir

$$q(X, \alpha) := \min\{q \in \mathbb{R} : P(X \leq q \geq \alpha\}$$

*Ayrıca q(X, p/100)'yi p'nci yüzdelik(percentile) olarak adlandırıyoruz.*

Bu tanımdan hareketle, medyanın 50'nci yüzdeliğe tekabül ettiğine dikkat ediniz. Diğer daha sık kullanılan ondalıklar (quantile) (p = 10, 20, 30, …, 90) ve çeyreklikler (quartile) (p = 25, 50, 75)'dır.

Beklenen değerlerde yaptığımız gibi ondalıkların özellikleri için çok zaman harcamayacağız, fakat medyanın beklenen değerden farklı davrandığı iki önemli noktaya temas etmek istiyorum: İlki için, Jensen'in Eşitsizliğinde bir u(X) fonksiyonu için beklenen değer, $\mathbb{E}$[u(X)], önemli oranda X'in olasılık yığının bulunduğu bölgenin eğriliği u(x)'e bağlı olduğunu gördük. Genel olarak, medyan *medyan(u(X))* u(medyan(X))'ten farklı da olacaktır, ancak bunun dikkat edilmesi gereken bir istisnası şudur:

**Önerme 1**. *X'i desteklemek için u(X)'in kesin artan olduğunu varsayalım. O zaman*

$$\textit{medyan}(u(X) = u(\textit{medyan}(X))$$

İSPAT: X'in medyanı P(X < medyan(X)) =1/2'yi sağlar. u(x) kesin artan olduğu için, herhangi bir sabit m değeri için olay X < m olay u(X) < u(m)'ye eşittir. Bu nedenle P(u(X) < u(medyan(X)) = P(X < medyan(X)) = 1/2'dir, böylece u(medyan(X)) gerçekten u(X)'in medyanıdır.

Sezgisel olarak, medyan kesin artan dönüşüm ile korunan rasgele değişkenin *ordinal* özeliğine bağlıdır.

Beklenen değerin, çoklu rasgele değişkenin doğrusal fonksiyonun beklenen değeri beklenen değerlerin aynı doğrusal fonksiyonuna eşit olduğu manasında, doğrusal *olduğunu gördük. İzleyen örnekte gösterildiği gibi medyan için bu doğru değildir:*

***Örnek 1**. Varsayalım ki $X_1$ ve $X_2$ aşağıdaki benzer marjinal dağılımdan elde edilen*

$$f_X(x) = \begin{cases} 0.6 & \text{eğer} \quad x = 0 \ \text{ ise} \\ 0.4 & \text{eğer} \quad x = 1 \ \text{ ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

*ve birbirinden bağımsız olan kesikli rasgele değişkenler olsun. O zaman $Y = X_1 + X_2$ 0,1 ve 2 değerlerini alabilir ve p.d.f.si*

$$f_Y(y) = \begin{cases} 0.36 & \text{eğer} \quad x = 0 \ \text{ ise} \\ 0.48 & \text{eğer} \quad y = 1 \ \text{ ise} \\ 0.16 & \text{eğer} \quad y = 2 \ \text{ ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

*$X_1$ ve $X_2$'nin medyanı sıfırdır, ancak medyan(Y) = 1 ≠ 0 + 0 = medyan($X_1$) + medyan($X_2$).*

Daha genel olarak, ortalamaların ondalıkları ondalıkların ortalamasından farklı olabilir. İzleyen örnek bu anlayışın oldukça pratik bir başka yorumunu vermektedir (aşağıdaki sayısal örneği verdiği için Aleksandr Tamarkin'e teşekkürler).

**Örnek 2**. *$X_1$ sözel, $X_2$ analitik ve $X_3$ sayısal şeklinde üç bölümden oluşan standart bir test olan GRE sınavına girdiğinizi düşünün. Testin her bölümde puanınız yüzde 90'lık dilimin üzerindedir. Bu, genel puanda da yüzde 90'lık dilimin üzerinde olduğunuz anlamına mı gelir? Genel olarak cevap hayırdır?*

*Varsayalım ki, siz dâhil, sınava giren 100 kişi vardır ve puanların dağılımı şöyledir: 84 kişi hiçbir bölümden bir tek puan bile almaz, siz her bölümden 250 puan aldınız, ve bunun dışında sınava giren üç türlü kişi vardır ki her birisi sadece bir bölüm için her nasılsa dar görüşlü bir dâhiye benzeyen bir yeteneğe sahiptir. Daha açık olmak gerekirse, 5 kişi sözelde çok aşırı yeteneklidir ve sözel bölümden 800 diğer bölümlerden 0 alırlar. Diğer 5 kişi analitik bölümden 800, ve bir diğer 5 kişi ise sayısal* *bölümden 800 alırken öteki bölümlerden 0 alırlar. Toparlayacak olursak, puanların bileşik dağılımı şöyledir (bu tipik GRE puanların dağılımında oldukça farklıdır)*

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = \begin{cases} 0.84 & (x_1,x_2,x_3) = (0,0,0) \\ 0.01 & (x_1,x_2,x_3) = (250,250,250) \text{ (sen)} \\ 0.05 & (x_1,x_2,x_3) = (800,0,0) \\ 0.05 & (x_1,x_2,x_3) = (0,800,0) \\ 0.05 & (x_1,x_2,x_3) = (0,0,800) \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

*Dolayısıyla, siz her bölümde en az %95'lik dilimdesiniz, fakat 15 kişinin toplam skoru 800 iken sizin ki sadece 750'dir, bu nedenle her üç bölüme göre toplam puanların sadece %85'lik dilimindesiniz.*

## 2. Varyans

*Varyans* rasgele bir değişkenin yayılmasının ölçüsüdür

**Tanım 2**. *Rasgele bir değişkenin varyansı aşağıdaki ile verilir:*

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

*Bazen biz varyansı $\sigma^2$ = Var(X) ile ifade ederiz*.

**Özelik 1**. *Sadece ve sadece bazı sabit sayılar için P(X = c) = 1 ise Var(X) = 0'dır.*

**Özelik 2**. *Eğer Y = aX + b ise, o zaman*

$$\mathrm{Var}Y = a^2\mathrm{Var}(X)$$

İSPAT: Yine, sadece sürekli duruma bakalım. Beklenen değer için elde edilen önceki sonuçları kullanarak

$$\begin{aligned} \mathrm{Var}(Y) &= \int_{-\infty}^{\infty} (aX + b - \mathbb{E}[aX+b])^2 f_X(x)dx = \int_{-\infty}^{\infty} (ax - a\mathbb{E}[X])^2 f_X(x)dx \\ &= a^2 \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x)dx = a^2\mathrm{Var}(X) \end{aligned}$$

Yayılmanın ölçümü için rasgele değişken gibi bir çeşit birim kullanmak daha uygun düşer. Ancak, bu son sonuç Var(X)'in biriminin X'in biriminin karesi olabileceğini ima etmektedir. Bu nedenle, varyans yerine sıklıkla *standart sapma $\sigma$(X)*'i kullanırız:

$$\sigma(X) := \sqrt{\mathrm{Var}(X)}$$

**Özelik 3**.

$$\mathrm{Var}(X) = \mathbb{E}|X^2| - \mathbb{E}|X|^2$$

İSPAT:

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

**Özelik 4**. *Eğer*

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + b$$

*ve $X_1$, …, $X_n$ bağımsız ise, o zaman*

$$\mathrm{Var}(Y) = a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2) + \ldots + a_n^2\mathrm{Var}(X_n)$$

**Örnek 3**. *Varsayalım ki X kesikli bir rasgele değişkendir ve p.d.f.si şöyle olsun:*

$$f_X(x) = \begin{cases} \frac{1}{5} & \text{eğer } x \in \{-2, 0, 1, 3, 4\} \text{ ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

*Eğer Y = 4X – 7 ise, Y'nin varyansı nedir?*

$$\mathrm{Var}(Y) = 4^2\mathrm{Var}(X) = 16(\mathbb{E}[X^2] - \mathbb{E}[X]^2)$$

*Şimdi aşağıdakini hesaplayabiliriz.*

$$\mathbb{E}[X] = (1/5)(-2 + 0 + 1 + 3 + 4) = 6/5$$

*ve*

$$\mathbb{E}[X^2] = \frac{1}{5}((-2)^2 + 0^2 + 1^2 + 3^2 + 4^2) = \frac{30}{5} = 6$$

*Bundan ötürü*,

$$\mathrm{Var}(Y) = 16\left[6 - \left(\frac{6}{5}\right)^2\right] = 16\frac{150-36}{25} = \frac{1824}{25} \approx 73$$

**Örnek 4**. Varsayalım ki *Y ~ B(n, p). Y n* sayıda bağımsız denemenin sonuçlarının toplamı olarak yazılabilir

$$Y = X_1 + \ldots + X_n, \text{ burada } X_i = \begin{cases} 1 & \text{p olasılığı ile} \\ 0 & \text{1 - p olasılığı ile} \end{cases}$$

*Şunu hesaplayabiliriz*

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$$

*O halde $X_i$'nin varyansı nedir? Açıkçası*

$$\mathbb{E}[X_i] = p\text{'dir}$$

*aynı zamanda,*

$$\mathbb{E}[X_i^2] = 1 \cdot p + 0 \cdot (1-p) = p$$

*Dolaysıyla, Özellik 3'e göre,*

$$\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = p - p^2 = p(1-p)$$

*Bundan ötürü,*

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = np(1-p)$$

Varyans bir beklenen değer olduğu için, rasgele değişkenin fonksiyonunun beklenen değerini doğrudan rasgele değişkenin fonksiyonunun varyansına uygulayabiliriz: eğer Y = r(X) ise,

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \mathbb{E}[r(X)^2] - \mathbb{E}[r(X)]^2 = \int_{-\infty}^{\infty} r(t)^2 f_X(t)dt - \left[\int_{-\infty}^{\infty} r(t)dt\right]^2$$

## 2.1 Daha Yüksek Dereceli Momentler

Beklenen değerin ilgili dağılımın *konumunun* ölçüsü olduğunu görmüştük, ama öte yandan varyans *yayılmayı* ölçer. Dağılımı karakterize etmek için rasgele değişkenin diğer *momentlerine* bakabiliriz, örneğin simetrik mi? Kalın kuyruklu mu? vs.

**Tanım 3**. X'in *r*'nci momenti aşağıdaki ile verilir,

$$\mu'_r = \mathbb{E}[X^r]$$

ve *r*'nci merkezi momenti şöyle tanımlanır:

$$\mu_r = \mathbb{E}[(X - \mathbb{E}[X])^r]$$

Beklenen değer, dolaysıyla, X dağılımının *birinci momenti*, varyans ise *ikinci merkezi momenti* olarak da ifade edilir.

Bir dağılımın sık sık kullanılan diğer özelikleri şunlardır:

$$\mu_3 = \mathbb{E}[(X - \mathbb{E}[X])^3]$$

bu dağılımın *çarpıklığı* olarak adlandırılır, ve aşağıdaki de X'in *basıklığıdır*.

$$\mu_4 = \mathbb{E}[(X - \mathbb{E}[X])^4]$$

Yüksek *basıklık* olasılık yığılmasının kuyruklarda yoğunlaştığını ifade eder.