

GEÇİCİ KAPAK

*Kapak tasarımı
devam ediyor.*

BİDGE Yayınları

Olasılıksal Modeller, Simülasyon Yöntemleri ve İstatistiksel Uygulamalar

Editör: ZELİHA KAYGISIZ ERTUĞ

ISBN: -

1. Baskı

Sayfa Düzeni: Gözde YÜCEL

Yayınlama Tarihi: -

BİDGE Yayınları

Bu eserin bütün hakları saklıdır. Kaynak gösterilerek tanıtım için yapılacak kısa alıntılar dışında yayıncının ve editörün yazılı izni olmaksızın hiçbir yolla çoğaltılamaz.

Sertifika No: 71374

Yayın hakları © BİDGE Yayınları

www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com

Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.

Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya /
Ankara



İÇİNDEKİLER

KUMARBAZIN YENİLGİSİ: MARKOV ZİNCİRİ VE MARTİNGALE TEMELLİ BİR SİMÜLASYON	1
<i>MURAT SAĞIR, VEDAT SAĞLAM</i>	
HETEROJENLİK, TERCİH POLİTİKASI VE DAĞILIMSAL ASİMETRİ: HETEROJEN BİR KUYRUK SİSTEMİNDE MÜŞTERİ SAYISINA AİT KİTLESEL MOMENTLERİN İNCELEMESİ	20
<i>MURAT SAĞIR, MÜJGAN ZOBU, VEDAT SAĞLAM</i>	
BİYOİSTATİSTİKTE SİMÜLASYON YAKLAŞIMLARI VE UYGULAMALARI	34
<i>HAKAN ÖZTÜRK</i>	

BÖLÜM 0

KUMARBAZIN YENİLGİSİ: MARKOV ZİNCİRİ VE MARTİNGALE TEMELLİ BİR SİMÜLASYON

MURAT SAĞIR¹
VEDAT SAĞLAM²

Giriş

Markov zincirleri, olasılık süreçlerinin uzun dönem limit davranışlarını incelemede vazgeçilemez bir öneme ve yaygın bir kullanıma sahiptir (Norris, 1998) . Bir Markov zincirlerinin geçiş olasılıkları çok önemli olsa da bu geçişler altında değişmezlik ve denge özellikleri taşıyan özel fonksiyonlar da bir o kadar önemlidir ve istisnai özellikler sergiler. Dolayısı ile harmonik fonksiyonlar, Markov zincirlerinin yapısını anlamada kilit önem sahiptir (Woess, 2000). Fakat bu vazgeçilemez öneme rağmen, harmonik fonksiyonlar hakkında bilgi sahibi olmak her zaman kolay değildir (Kelvin & Tait, 1867). Özellikle Laplace denklemlerinin çözümünün sağlayacağı yarar tartışılmaz ama analizin sunulduğu biçim, işin ehli matematikçiler tarafından itici, sıradan matematik öğrencileri için ise oldukça zor gelmiştir.

¹ Dr. Öğr. Üyesi, İskenderun Teknik Üniversitesi, İşletme ve Yönetim Bilimleri Fakültesi, Ekonomi Bölümü Orcid: 0000-0001-7567-9327

² Prof. Dr., Ondokuz Mayıs Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü Orcid: [0000-0002-8586-1373](https://orcid.org/0000-0002-8586-1373)

Bir Markov zinciri için harmonik fonksiyon kavramı, geçiş operatörü altında ortalama değer özelliği ile tanımlanır(Levin & Peres, 2017). Özellikle ayırık durum uzaylarında (örneğin rasgele yürüyüş sürecinde) harmonik fonksiyonlar, sınır değer problemleri ve durma zamanları ile yakından ilişkilidir. Harmonik fonksiyonların bu özellikleri Dirichlet probleminin stokastik yorumunda hayati bir öneme sahip olmalarına neden olur (Doob & Doob, 1984; Woess, 2000). Ek olarak Markov zincirleri ile martingale teorisi arasındaki ilişki harmonik fonksiyonlar ile mümkün olur (Durrett, 2019). Uygun koşullar altında, harmonik bir fonksiyonun zincir boyunca değerlendirilmesi bir martingale oluşturur.

Bu bölümde, Markov zincirleri üzerinde tanımlı martingale yapılar ile harmonik fonksiyonlar ve bunların arasındaki ilişkiler ele alınmaktadır. Özellikle harmonik fonksiyonların özelliklerinden olan maksimum ilkesi ve eşsizlik teoremiyle literatürde kumarbazın yenilgisi (Grimmett & Stirzaker, 2020) olarak bilenen problemin çözümü incelendi. Ayrıca teorik sonuçların bir uygulaması olarak sayısal simülasyonlara yer verildi.

Martingale

Martingale, ortalama değeri sabit kalan bir süreçtir. Martingalenin sahip olduğu bu sabitlik bir dengeleme özelliğidir. Martingalelerin tanımlanması, genellikle stokastik bir sürecin dönüşümünü anlamada önemli bir adımdır.

$(X_n)_{n \geq 0}$ bir Markov zinciri ve $\mathbb{Z}$ üzerindeki basit simetrik bir rastgele yürüyüş olsun.

Yürüyüşün ortalama değeri sabittir; aslında yürüyüşün gelecekteki bir zamandaki ortalama değerinin her zaman yalnızca mevcut değer olması gibi daha güçlü bir özelliğe sahiptir. Yani,

$$\mathbb{E}(X_n) = \mathbb{E}(X_0), \quad (1)$$

ve daha güçlü özellik şunu söylüyor: $n \geq m$ için,

$$\mathbb{E}(X_n - X_m \mid X_0 = i_0, \dots, X_m = i_m) = 0 \quad (2)$$

ise $(X_n)_{n \geq 0}$ bir martingaledir.

Kesinlik için bir Markov zinciri $(X_n)_{n \geq 0}$ olsun ve yalnızca $X_0, \dots, X_n$ 'ye bağlı tüm kümelerin koleksiyonu $\mathcal{F}_n$ olsun. $(\mathcal{F}_n)_{n \geq 0}$ dizisine $(X_n)_{n \geq 0}$ 'ın filtrelenmesi denir ve $\mathcal{F}_n$ n zamanına kadar zincirin bilgi durumunu veya geçmişini temsil eden bir dizi olarak kabul edilir. Bir $(M_n)_{n \geq 0}$ işlemine, eğer M_n yalnızca $X_0, \dots, X_n$ 'ye bağlıysa adapte edilmiş denir. Bir $(M_n)_{n \geq 0}$ işlemine, eğer tüm n değerleri için $\mathbb{E}|M_n| < \infty$ ise integral denir. Uyarlanmış bir integrallenebilir işlem $(M_n)_{n \geq 0}$, eğer

$$\mathbb{E}[(M_{n+1} - M_n)1_A] = 0 \quad (3)$$

tüm $A \in \mathcal{F}_n$ ve tüm n için. $\mathcal{F}_n$ koleksiyonu, aşağıdaki gibi temel olayların sayılabilir birleşimlerinden oluştuğundan

$$\{X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\}, \quad (4)$$

bu martingale özelliği tüm $i_0, \dots, i_n$ ve tüm n için aşağıdaki denkleme eşdeğerdir:

$$\mathbb{E}(M_{n+1} - M_n \mid X_0 = i_0, \dots, X_n = i_n) = 0 \quad (5)$$

Martingale özelliğinin üçüncü bir formülasyonu, koşullu beklenti kavramını içerir. Integrallenebilir bir rastgele değişken Y verildiğinde, aşağıdaki koşullu beklenen değer tanımlanabilir (Norris, 1998).

$$\mathbb{E}(Y \mid \mathcal{F}_n) = \sum_{i_0, \dots, i_n} \mathbb{E}(Y \mid X_0 = i_0, \dots, X_n = i_n) 1_{\{X_0=i_0, \dots, X_n=i_n\}}. \quad (6)$$

Y 'den $\mathbb{E}(Y \mid \mathcal{F}_n)$ 'ye geçerken, her bir temel olay olan $A \in \mathcal{F}_n$ 'de, Y rastgele değişkenini ortalama değeri $\mathbb{E}(Y \mid A)$ ile değiştirilir. Uyarlanmış bir integrallenebilir sürecin $(M_n)_{n \geq 0}$ bir martingale olup olmadığını kontrol etmek, ancak ve ancak bütün n 'ler için aşağıdaki koşul sağlandığında mümkündür:

$$\mathbb{E}(M_{n+1} \mid \mathcal{F}_n) = M_n \quad (7)$$

Koşullu beklenti kısmi bir ortalamadır, dolayısıyla işlemi tamamlanır ve koşullu beklenen değer ortalaması alınır ve tam beklenen değer elde edilmelidir.

$$\mathbb{E}(\mathbb{E}(Y \mid \mathcal{F}_n)) = \mathbb{E}(Y) \quad (8)$$

Tam beklenen değer formülünün geçerliliği martingaleler için kolayca gösterilebilir

$$\mathbb{E}(M_n) = \mathbb{E}(\mathbb{E}(M_{n+1} \mid \mathcal{F}_n)) = \mathbb{E}(M_{n+1}) \quad (9)$$

tümevarım yoluyla

$$\mathbb{E}(M_n) = \mathbb{E}(M_0) \quad (10)$$

olur. Bu durum, martingalenin orijinal tanımında $A = \Omega$ aldığımızda zaten açıktı.

Bir rastgele değişkenin

$$T: \Omega \rightarrow \{0, 1, 2, \dots\} \cup \{\infty\} \quad (11)$$

$n < \infty$ için $\{T = n\} \in \mathcal{F}_n$ ise bir durma zamanıdır. Eşdeğer bir koşul, $n < \infty$ için $\{T \leq n\} \in \mathcal{F}_n$ olmasıdır.

Teorem 1 (İsteğe bağlı durdurma teoremi). $(M_n)_{n \geq 0}$ bir martingale ve T bir durdurma zamanı olsun. Aşağıdaki koşullardan en az birinin geçerli olduğunu varsayalım:

1. $T \leq n$ hemen hemen her yerde olacak şekilde bir $n \in \mathbb{N}$ vardır

2. $T < \infty$ hemen hemen her yerde sabit bir $C > 0$ için

$$|M_n| \leq C \text{ her } n \leq T.$$

O zaman

$$\mathbb{E}(M_T) = \mathbb{E}(M_0) \quad (12)$$

olur (Williams, 1991)

İspat. (i)'nin geçerli olduğunu varsayalım. O zaman

$$\begin{aligned} M_T - M_0 &= (M_T - M_{T-1}) + \cdots + (M_1 - M_0) \\ &= \sum_{k=0}^{n-1} (M_{k+1} - M_k) 1_{k < T} \end{aligned} \quad (13)$$

Şimdi $\{k < T\} = \{T \leq k\}^c \in \mathcal{F}_k$ çünkü T bir durma süresidir ve bu nedenle

$$\mathbb{E}[(M_{k+1} - M_k) 1_{k < T}] = 0 \quad (14)$$

çünkü $(M_k)_{k \geq 0}$ bir martingale olduğundan. Dolayısıyla

$$\mathbb{E}(M_T) - \mathbb{E}(M_0) = \sum_{k=0}^{n-1} \mathbb{E}[(M_{k+1} - M_k) 1_{k < T}] = 0 \quad (15)$$

1'i değil de 2'yi varsayarsak, önceki argüman durma süresi $T \wedge n$ için geçerlidir, böylece $\mathbb{E}(M_{T \wedge n}) = \mathbb{E}(M_0)$ olur. O zaman

$$|\mathbb{E}M_T - \mathbb{E}M_0| = |\mathbb{E}M_T - \mathbb{E}M_{T \wedge n}| \quad (16)$$

işlemlere devam edilsin $|E(X)| \leq E|X|$ olduğundan

$$|\mathbb{E}M_T - \mathbb{E}M_{T \wedge n}| = |\mathbb{E}(M_T - M_{T \wedge n})| \leq \mathbb{E}|M_T - M_{T \wedge n}| \quad (17)$$

Ayrıca

$$\begin{aligned} |M_T - M_{T \wedge n}| &= |M_T - M_n| 1_{\{T > n\}} \leq (|M_T| + |M_n|) 1_{\{T > n\}} \\ &\leq 2C 1_{\{T > n\}}. \end{aligned} \quad (18)$$

olur. Beklenen değer alındığında tüm n için:

$$\begin{aligned} |\mathbb{E}M_T - \mathbb{E}M_{T \wedge n}| &\leq \mathbb{E}|M_T - M_{T \wedge n}| \leq \mathbb{E}[2C 1_{\{T > n\}}] \\ &= 2CP(T > n). \end{aligned} \quad (19)$$

Ancak $n \rightarrow \infty$ iken $P(T > n) \rightarrow 0$ olduğundan

$$|\mathbb{E}M_T - \mathbb{E}M_0| \leq 2CP(T > n) = 0 \quad (20)$$

olur. Böylece $\mathbb{E}M_T = \mathbb{E}M_0$

Basit simetrik rastgele yürüyüş $(X_n)_{n \geq 0}$ için, $X_0 = 0$ olduğunu varsayalım ve aşağıdaki tanım verilsin:

$$T = \inf\{n \geq 0: X_n = -a \text{ veya } X_n = b\}$$

Burada $a, b \in \mathbb{N}$ birer parametredir ve verilmelidir. O zaman T bir durma süresidir ve sonlu kapalı sınıfların tekrarı nedeniyle $T < \infty$ 'dir. Dolayısıyla, isteğe bağlı durdurma teoreminin (ii) koşulu, $M_n = X_n$ ve $C = a \vee b$ için geçerlidir. $\mathbb{E}X_T = \mathbb{E}X_0 = 0$ olduğunu sonucuna ulaşılır.

$$p = \mathbb{P}(X_n \text{ } b \text{ den önce } -a \text{ ya çarpar}) \quad (21)$$

Olasılık p olan $X_T = -a$ ve olasılık $1 - p$ olan $X_T = b$ biliniyor, bu nedenle

$$0 = \mathbb{E}(X_T) = p(-a) + (1 - p)b \quad (22)$$

olur. Buradan p olasılığı aşağıdaki gibi olmalıdır.

$$p = \frac{b}{a + b}.$$

p 'yi hesaplamamanın tamamen farklı, Markovyen bir yolu vardır. Ancak $\mathbb{E}X_T = 0$ sonucunun ardındaki sezgi çok açıktır: Adil bir oyun oynayan bir kumarbaz, kayıpları a 'ya veya kazançları b 'ye ulaştığında, hangisi daha erken gerçekleşirse, kumarhaneden ayrılır; oyun adil olduğundan, ortalama kazanç sıfır olmalıdır.

Örneğin sonsuz zengin bir kumarhanede adil bir oyun oynamaya devam eden ve kesin olarak mahvolma sonucunun ortaya çıktığı bir kumarbazın mantığa aykırı durumunu ele alalım. Bu oyun, sonlu bir bitiş süresinde sona erer.

$$T = \inf\{n \geq 0: X_n = -a\} \quad (23)$$

Burada a , kumarbazın başlangıçtaki servetidir. $X_T = -a$ olduğundan,

$$\mathbb{E}(X_T) = -a \neq 0 = \mathbb{E}X_0$$

Ancak bu, isteğe bağlı durdurma teoremine aykırı değildir çünkü ne (i) koşulu ne de (ii) koşulu karşılanmaktadır. Dolayısıyla, sezgi $\mathbb{E}X_T = \mathbb{E}X_0$ 'ın oldukça açık olduğunu gösterse de her zaman doğru olmadığı için biraz dikkatli olmak gerekir.

Az önce ele alınan örnek, $(X_n)_{n \geq 0}$ zincirinin kendisinin bir martingale olması bakımından oldukça özeldi. Açıkçası, bu genel olarak doğru değildir; aslında bir martingale zorunlu olarak gerçel (reel) değerlidir ve genel olarak durum uzayı I 'in $\mathbb{R}$ 'de bulunması konusunda bir zorluk yoktur. Yine de, her Markov zincirine bir dizi martingale eşlik eder ve bu martingaleler zinciri karakterize eder. Bu, martingaleler ve Markov zincirleri arasındaki derin bağlantının temelidir.

$f: I \rightarrow \mathbb{R}$ fonksiyonu ve geçiş matrisi P olan bir Markov zinciri $(X_n)_{n \geq 0}$ verildiğinde, şunu elde ettiğimizi hatırlayalım:

$$(P^n f)(i) = \sum_{j \in I} p_{ij}^{(n)} f_j = \mathbb{E}_i(f(X_n)) \quad (24)$$

Ayrıca $n = 1$ için

$$(Pf)(i) = \sum_j P(i, j) f(j), \quad (P - I)f = Pf - f \quad (25)$$

Teorem 2. $(X_n)_{n \geq 0}$, değerleri I 'de olan rastgele bir süreç ve P stokastik bir matris olsun. $(X_n)_{n \geq 0}$ 'ın filtrelenmesi için $(\mathcal{F}_n)_{n \geq 0}$ $(\mathcal{F}_n = \sigma(X_0, \dots, X_n))$ yazın. Bu durumda aşağıdakiler eşdeğerdir:

$(X_n)_{n \geq 0}$ geçiş matrisi P olan bir Markov zinciridir; tüm sınırlı fonksiyonlar $f: S \rightarrow \mathbb{R}$ için aşağıdaki süreç bir martingale'dir (Ethier & Kurtz, 2009)

$$M_n^f = f(X_n) - f(X_0) - \sum_{m=0}^{n-1} (P - I)f(X_m) \quad (26)$$

İspat. İlk durumun geçerli olduğunu varsayalım. f sınırlı bir fonksiyon olsun. Yani $|f| \leq C$ diyelim. O zaman

$$|(Pf)(i)| = \left| \sum_{j \in I} p_{ij} f_j \right| \leq \sup_j |f_j| = C \quad (27)$$

ve

$$\begin{aligned} |(P - I)f(i)| &= |Pf(i) - f(i)| \leq |Pf(i)| + |f(i)| \leq C + C \\ &= 2C \end{aligned} \quad (28)$$

olur. Böylece

$$\begin{aligned} |M_n^f| &\leq |f(X_n)| + |f(X_0)| + \sum_{m=0}^{n-1} |(P - I)f(X_m)| \\ &= 2(n + 1) \sup_j |f_j| \leq 2(n + 1)C < \infty \end{aligned} \quad (29)$$

Yani M_n^f her n için integrallenebilir. $A = \{X_0 = i_0, \dots, X_n = i_n\}$ olsun.

$$\mathbb{E}(M_{n+1}^f - M_n^f \mid A) = \mathbb{E}[f(X_{n+1}) - (Pf)(X_n) \mid A] = 0 \quad (30)$$

olduğunu gösterilmelidir. Fark açık yazılırsa:

$$\begin{aligned} M_{n+1}^f - M_n^f &= \left(f(X_{n+1}) - f(X_0) - \sum_{m=0}^n (P - I)f(X_m) \right) \\ &\quad - \left(f(X_n) - f(X_0) - \sum_{m=0}^{n-1} (P - I)f(X_m) \right) \quad (31) \\ &= f(X_{n+1}) - f(X_n) - (P - I)f(X_n) \\ &= f(X_{n+1}) - Pf(X_n). \end{aligned}$$

Beklenen değer alındığında

$$\mathbb{E}[M_{n+1}^f - M_n^f \mid A] = \mathbb{E}[f(X_{n+1}) \mid A] - Pf(X_n) \quad (32)$$

yazılır. Markov özelliğinden

$$\mathbb{E}[f(X_{n+1}) \mid A] = \mathbb{E}[f(X_{n+1}) \mid X_n] = \text{Pf}(X_n) \quad (33)$$

elde edilir. Böylece

$$\begin{aligned} \mathbb{E}[M_{n+1}^f - M_n^f \mid A] &= \mathbb{E}[f(X_{n+1}) - (Pf)(X_n) \mid A] \\ &= \text{Pf}(X_n) - \text{Pf}(X_n) = 0 \end{aligned} \quad (40)$$

ve bu nedenle $(M_n^f)_{n \geq 0}$ bir martingale'dir. Öte yandan, eğer ikinci durum geçerliyse, o zaman bütün sınırlı f fonksiyonlar için aşağıdaki eşitlik geçerlidir.

$$\mathbb{E}[f(X_{n+1}) - (Pf)(X_n) \mid X_0 = i_0, \dots, X_n = i_n] = 0 \quad (41)$$

$f = 1_{\{i_{n+1}\}}$ alınsın. O zaman

$$f(X_{n+1}) = \mathbf{1}_{\{X_{n+1}=j\}} \quad (42)$$

olur.

$$\mathbb{E}[f(X_{n+1}) \mid A] = \mathbb{E}[f(X_{n+1}) \mid X_n] = \text{Pf}(X_n) \quad (43)$$

yukarıdaki denklemden

$$\begin{aligned} \mathbb{E}[\mathbf{1}_{\{X_{n+1}=i_{n+1}\}} \mid A] &= \text{Pf}(X_n) = \sum_k P(X_n, k) \mathbf{1}_{\{k=i_{n+1}\}} \\ &= P(X_n, i_{n+1}) \end{aligned} \quad (44)$$

elde edilir. Sol taraf, tanım gereği

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid A) \quad (45)$$

dir. Dolayısıyla

$$\mathbb{P}(X_{n+1} = i_{n+1} \mid A) = P(X_n, i_{n+1}) \quad (46)$$

elde edilir. Bu da tam Markov özelliğinden:

$$\begin{aligned} \mathbb{P}(X_{n+1} = i_{n+1} \mid X_0 = i_0, \dots, X_n = i_n) \\ = \mathbb{P}(X_{n+1} = i_{n+1} \mid X_n = i_n) = p_{i_n i_{n+1}} \end{aligned} \quad (47)$$

yazılır. Böylece $(X_n)_{n \geq 0}$ süreci, geçiş matrisi P olan bir Markov zinciridir.

Harmonik Fonksiyonlar

Harmonik fonksiyonlar, gerçek Öklid uzaylarının açık alt kümelerinde yer alırlar. Bu bölüm boyunca n , 1'den büyük sabit bir pozitif tam sayıyı, Ω ise $\mathbf{R}^n$ 'nin açık, boş olmayan bir alt kümesini ifade edecektir. Ω üzerinde tanımlanan, iki kez sürekli türevlenebilir, kompleks değerler alabilen bir u fonksiyonu, aşağıdakileri sağlandığı takdirde Ω üzerinde harmoniktir:

$$\Delta u \equiv 0 \quad (48)$$

Burada $\Delta = D_1^2 + \dots + D_n^2$ ve D_j^2 , j 'inci koordinat değişkenine göre ikinci kısmi türevi ifade eder. Δ operatörüne Laplasyen, $\Delta u \equiv 0$ denkleminde ise Laplace denklemi denir. $E \subset \mathbf{R}^n$ kümesi üzerinde (mutlaka açık olmayan) tanımlanmış bir u fonksiyonunun E üzerinde harmonik olduğunu, u 'nun E 'yi içeren açık bir küme üzerinde harmonik bir fonksiyona genişletilebilmesi durumunda söyleriz.

$x = (x_1, \dots, x_n)$ ifadesinin $\mathbf{R}^n$ 'deki tipik bir noktayı ve $|x| = (x_1^2 + \dots + x_n^2)^{\frac{1}{2}}$ ifadesinin x 'in Öklit normunu gösterdiğini varsayalım. En basit sabit olmayan harmonik fonksiyonlar koordinat fonksiyonlarıdır; örneğin, $u(x) = x_1$.

Düzgün fonksiyonlar için Laplas denklemi herhangi bir kısmi türevle değişmektedir. Genel olarak her harmonik fonksiyon sonsuz derecede türevlenebilirdir ve dolayısıyla bir harmonik fonksiyonun her kısmi türevi harmoniktir.

Harmonik Fonksiyonların Markov Zincirlerindeki Uzantısı

Eğer aşağıdaki özellik sağlanırsa bir $h: V \mapsto R$ fonksiyonu, belirli bir Markov zincirine göre harmonik denir (Axler vd., 2013).

$$h(i) = \sum_{j|\{i,j\} \in E} P_{ij}h(j) \quad (49)$$

Sezgisel olarak, köşeler üzerinde tanımlanan bir fonksiyon, tüm komşu köşelerin ağırlıklı ortalaması ise (ağırlıklar o komşuya geçme olasılığına karşılık gelir) harmonik dağılım gösterir.

Şimdi, V kümesinin bir alt kümesi olan $B \subset V$ (bazı durumlarda bu alt küme sınır olarak kabul edilir) ve bu alt küme üzerinde tanımlanmış bir $h_B: B \mapsto R$ fonksiyonuna sahip olduğumuzu varsayalım. O halde, h_B 'nin harmonik bir uzantısı şu şekilde tanımlanan bir fonksiyondur:

1. $h(x) = h_B(x) \forall x \in B$
2. h , tüm $x \notin B$ değerlerinde harmoniktir.

Dolayısıyla, rastgele bir köşe alt kümesi ve bu alt küme üzerinde rastgele bir fonksiyon verildiğinde, harmonik bir uzantı bu alt kümedeki fonksiyonla eşleşir ve bu alt kümede olmayan köşeler için fonksiyon değerlerini harmonik bir şekilde "yayar".

Ortalamada Değişmezlik Özelliği

Bir Markov zinciri $\{X_n\}_{n \geq 0}$ için tanımlı bir fonksiyon f geçiş operatörü P altında

$$Pf = f \quad (50)$$

koşulunu sağlıyorsa harmonik olarak adlandırılır (Norris, 1998). Bu koşul, fonksiyonun Markov zincirinin dinamiği altında koşullu beklenen değer anlamında değişmez olduğunu ifade eder. Sonuç olarak

$$\mathbb{E}_x[f(X_{n+1}) | X_n] = f(X_n) \quad (51)$$

eşitliği sağlanır. Özetle bir f fonksiyonunun harmonik olması $f(X_n)$ 'in bir martingale olması için gerek ve yeter koşuldur. Değişmezlik yerel bir özellik olmakla birlikte Markov zinciri

boyunca beklenen değerin korunmasını (değişmemesini) ifade eder. Bir Markov zincirinde birden fazla harmonik fonksiyon tanımlanabilir ve her biri değişmezlik özelliğini muhafaza edebilmektedir.

Maksimum İlkesi

Ortalama değer özelliğinin ayrık bir sonucu olarak, harmonik fonksiyonlar için maksimum ilkesi elde edilebilir. Bu ilke için aşağıdaki teorem verilmektedir.

Teorem 2 (Maksimum İlkesi): Durum uzayı bağlantılı olan bir Markov zincirinde tanımlı ve reel değerli bir u fonksiyonu harmonik olsun. Bu fonksiyon eğer iç noktada maksimuma veya minimuma ulaşıyor ise sabit olmak zorundadır.

İspat: Burada ispatı yaparken kitap bölümünün kapsamı da göz önüne alınarak klasik analitik ispat yapılmayacaktır. Bunun yerine martingale özelliğinden yararlanılarak kısa bir ispat aşağıdaki gibidir.

Varsayalım ki u fonksiyonu bir iç noktada (x) maksimum değere ulaşsın. u harmonik olduğu için

$$u(x) = \mathbb{E}_x[u(X_1)] \quad (52)$$

yazılabilir. X_1 ise zincirin bir adım sonraki durumu olsun. $u(X_1) \leq u(x)$ olduğu açıktır. Aynı zamanda yukarıdaki denklemin sağlanabilmesi için $u(X_1) = u(x)$ olması gerekmektedir. Dolayısı ile x 'e komşu her durum u fonksiyonu aynı değeri alır. Durum uzayının bağlantılı olmasından olayı bu durum tüm durum uzayı için geçerli olur ve böylece u sabit olmak zorundadır.

Teorem 3 (Dirichlet Problemi için Eşsizlik). $\{X_n\}_{n \geq 0}$ indirgenemez bir Markov zinciri, $B \subset V$ ve $h_B: B \mapsto R$ verildiğinde, köşeler üzerindeki fonksiyon $x \in V \setminus B$ için

$$h(x) = \mathbb{E}_x[h_B(X_{T_B})] \quad (53)$$

fonksiyonu h_B 'nin benzersiz harmonik uzantısıdır (Lawler ve Limic 2010). Yani bir x köşesinden başlayıp B 'deki ilk köşeye ulaşana kadar rastgele bir yürüyüş gerçekleştirilirse, T_B , B 'ye ilk ulaştığınız zamanı gösterir:

$$T_B = \inf\{n \geq 0: X_n \in B\} \quad (54)$$

X_{T_B} , B 'de ulaştığınız ilk köşeyi gösterir. Ve $h_B(X_{T_B})$, o köşedeki fonksiyon değeridir. Dolayısıyla, bu fonksiyonun beklenen değerinin h_B 'nin harmonik bir uzantısıdır.

$h_B(x)$ 'in tüm $x \in B$ için aynı v değeri olduğu durumu düşünmek faydalı olabilir. Bu durumda, $E[h_B(X_{T_B})] = v$ de olur, çünkü B 'ye her ulaştığımızda her zaman v değerini alırız. Dolayısıyla, ilginç olan tek durum, $x \in B$ 'deki farklı köşelerin farklı $h_B(x)$ değerlerine sahip olmasıdır.

İspat 1: Öncelikle $x \in B$ durumunu ele alalım. Bu durumda $T_B = 0$ olur.

$$h(x) = \mathbb{E}[h_B(x)] = h_B(x) \quad (55)$$

Böylece $x \in B$ için harmonik genişleme tanımı gösterimiş oldu. Diğer taraftan $x \notin B$ olduğu durumda rastgele yürüyüşün tanımını gereği aşağıdaki eşitlik yazılabilir.

$$h(x) = \mathbb{E}[h_B(X_{T_B})] = \sum_{y|\{x,y\} \in E} P_{xy} \mathbb{E}[h_B | x_1 = y] \quad (56)$$

Ancak zincir Markov olduğundan ve dolayısıyla zamana bağlı homojen (yani hafızasız) olduğundan:

$$E[h_B | x_1 = y] = E[h_B | x_0 = y] = h(y) \quad (57)$$

elde edilir. Bu nedenle, harmonik özellik için gerekli olan

$$h(x) = \sum_{(y|\{x,y\} \in E)} P_{xy} h(y) \quad (58)$$

sağlanmış olur.

Dirichlet Problemi için eşsizlik teoreminin ispatı maksimum ilkesi teoremi ile de gösterilebilir.

İspat 2: Fonksiyonumuzun benzersizliğine dair sezgisel bakış açısı, harmonik genişlemeyi tümevarımsal olarak düşünmekten gelir; burada B temel durumdur ve tümevarımsal adım, değeri bir köşeden komşularına yayarak tek bir çözüme ulaşır.

Matematik dili ile , iki harmonik genişlememiz olduğunu varsayalım: h ve u . Şimdi, tüm $x \in B$ için $h(x) = h_B(x) = u(x)$ 'dur, Dolayısı ile

$$h(x) - u(x) = 0 \quad (59)$$

yazılabilir. Öyleyse, $h(x) - u(x)$ 'in maksimum olduğu $x_i \notin B$ köşesini ele alınsın. Hem h hem de u harmonik olduğundan, $h(x_i) - u(x_i)$ değeri, her $j \mid \{i, j\} \in E$ için $h(x_j) - u(x_j)$ 'nin ağırlıklı ortalamasıdır. $h(x_i) - u(x_i)$ global maksimum olduğundan, ortalamanın doğru sonuç vermesinin tek yolu $h(x_j) - u(x_j) = v = h(x_i) - u(x_i)$ olmasıdır (eğer biri daha yüksek olsaydı, maksimum olmazdı; eğer biri daha düşük olsaydı, ağırlıklı ortalama işe yaramazdı). Bu argümanı, her j 'nin tüm komşularının da v değerine sahip olması gerektiğini söylemek için uygulamaya benzer şekilde devam edebiliriz. Ancak x_i 'den bir $b \in B$ 'ye giden bir yol olduğundan, b 'nin de v değerine sahip olması gerekir, ancak $v = h(b) - u(b) = 0$ olduğunu bildiğimiz için, değerlerdeki maksimum fark sıfırdır.

Tamamlayıcı olması açısından, minimum değerın sıfır olduğunun da gösterilmesi gerekmektedir. $h(x) - u(x)$ 'in minimum olduğu $x_i \notin B$ köşesini ele alalım. Hem h hem de u harmonik olduğundan, $h(x_i) - u(x_i)$ değeri, her $j \mid \{i, j\} \in E$ için $h(x_j) - u(x_j)$ 'nin ağırlıklı ortalamasıdır. $h(x_i) - u(x_i)$ küresel minimum

olduğundan, ortalamanın doğru olması için tek yol $h(x_j) - u(x_j) = v = h(x_i) - u(x_i)$ olmasıdır. Bu argümanı, her j 'nin tüm komşularının da v değerine sahip olması gerektiğini söylemek için uygulamaya devam edebiliriz. Ancak x_i 'den bir $b \in B$ 'ye giden bir yol olduğundan, b 'nin de v değerine sahip olması gerekir, ancak $v = h(b) - v(b) = 0$ olduğunu bildiğimiz için, değerlerdeki minimum fark sıfırdır. Dolayısıyla, zincirimizdeki tüm düğümler x için $g(x) = u(x)$ olduğu sonucu çıkar. Bu nedenle, iki genişletme fonksiyonumuz h ve u eşit olmalıdır ve bu nedenle herhangi bir B kümesi ve h_B fonksiyonu için yalnızca bir, benzersiz harmonik genişletme olabilir (Vöcking, 2002).

Uygulama: Kumarbazın Yenilgisi

Bölüme ait uygulamada yukarıda anlatılan martingale ve durdurma zamanları ve farmonik fonksiyonların maksimum ve eşsizlik özelliklerinin hepsinin harmanlayacak bir örnek olarak kumarbazın yenilgisi (Gambler's Ruin) örneğini simüle edeceğiz.

Bir kumarbazın sermayesini $(X_n)_{n \geq 0}$ ile gösterelim. Her oyunda (adımda) kubarbaz p olasılığı ile bir TL (burada başka bir miktar ya da başka bir para birimi de seçilebilir) kazanmakta ve $1 - p = q$ olasılığı ile bir TL kaybetmektedir. Dolayısı ile bu rasgele yürüyüş süreci

$$X_{n+1} = \begin{cases} X_n + 1, & \text{olasılık } p \\ X_n - 1, & \text{olasılık } q \end{cases} \quad (60)$$

şeklindedir. $(X_n)_{n \geq 0}$ durum uzayı

$$S = \{0, 1, 2, \dots, N\} \quad (61)$$

olan bir Markov zinciridir. Bu Markov zincirinde 0 (iflas) ve N (hedeflenen ve kumarbazın ulaştığında oyunu sonlandıracağı kazanç) yutan elemanlardır. Dolayısı ile sıfır (iflas) mecburi durdurma ve N isteğe bağlı durdurma olmak üzere, durdurma zamanı aşağıdaki gibi tanımlanabilir.

$$T = \inf\{n \geq 0: X_n \in \{0, N\}\} \quad (62)$$

Kumarbazın aklındaki o paraya (N) ulaşma olasılığı

$$h(x) = \mathbb{P}_x(X_T = N) \quad (63)$$

olarak tanımlansın. Markov özelliğinden $x \in \{1, \dots, N-1\}$ için

$$h(x) = ph(x+1) + qh(x-1) \quad (64)$$

olur. Dolayısıyla h fonksiyonu, iç noktalarda harmonik bir fonksiyondur ve sınır koşulları

$$h(0) = 0, \quad h(N) = 1$$

şeklinde. Bu, klasik bir Dirichlet problemidir. $n \geq 0$ için

$$M_n = h(X_n) \quad (65)$$

olarak tanımlansın. h fonksiyonu harmonik olduğundan $(M_n)_{n \geq 0}$ zinciri denklem (62)'deki T durdurma zamanı dikkate alındığında ortalama değişmezlik özelliğinden bir sınırlı bir martingale oluşturur.

Adil oyun varsayımında $p = q = 1/2$ 'dir. Bu varsayın altında çözüm doğrusal olup

$$h(x) = \frac{x}{N} \quad (66)$$

olur. Eğer $p \neq q$ olursa (adil olmayan oyun) o zaman

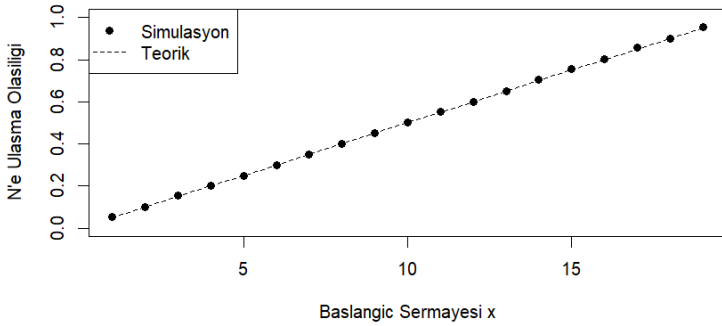
$$h(x) = \frac{1 - (q/p)^x}{1 - (q/p)^N} \quad (67)$$

olarak elde edilir. Harmonik fonksiyonların maksimum ilkesi ve Dirichlet problemini çözümünün eşsizlik teoremi gereğince, kumarbazın kazanma olasılığını tanımlayan fonksiyon tekdir.

Teorik olarak elde edilen ve harmonik fonksiyon olarak tanımlanan kazanma olasılığı, başlangıç sermayesi x için, 10 bin bağımsız rastgele yürüyüş adil bir oyun ($p = 0,5$) için simüle

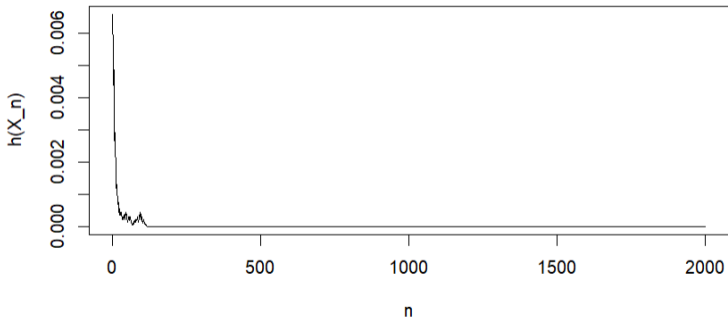
edilmiş ve kumarbazın aklında kararlaştırdığı ve ulaştığında oyundan çekileceği değer olan N 'e ulaşma olasılığı hesaplanmıştır. Şekil 1, simülasyon sonuçlarının teorik olarak elde edilen harmonik $h(x) = x/N$ fonksiyonu ile uyumlu olduğunu göstermektedir.

Şekil 1: Adil bir oyun için kazanma olasılığının teorik ve simülasyon değerlerinin grafiği



Adil olmayan bir oyunda($p \neq 0,5$) $h(X_n)$ sürecinin zaman grafiği aşağıdaki gibidir.

Şekil 2. $h(X_n)$ 'in zaman grafiği



Şekil 2 incelendiğinde Markov zincirinin yutulma durumlarına yaklaştığında $(0, N)$ martingale değerinin sıfıra yakınsadığı gözlenmektedir. Bu davranış, isteğe bağlı durma teoremi altında teorik olarak beklenen sonucu desteklemektedir.

Sonuç

Bu kitap bölümünde, ayrık zamanlı Markov zincirleri ekseninde harmonik fonksiyonlar ve martingale yapıları birlikte incelendi. Özetle harmonik fonksiyonların değişmezlik özellikleri bu fonksiyonların martingale özellikler göstermesini sağlamaktadır. Böylece Markov zincirlerinin uzun dönem davranışları martingale teorisi bağlamında incelenebilir.

Harmonik fonksiyonlar maksimum ilkesi ve dirichlet probleminin çözümünde teklik özelliklerine sahiptir. Bununla birlikte martingale teorisinde yer alan durma zamanları etkin bir şekilde gerçek hayatın modellenmesinde önemli argümanlar sergiler. Durma zamanları, klasik sınır değer problemlerinde stokastik bir vizyon sağlar.

Kumarbazın yenilgisi problemi, martingale ve harmonik fonksiyonların Markov zincirlerine uygulanması olarak sunulan teorik çerçevenin bir uygulaması olarak düşünüldü; harmonik fonksiyonların yardımıyla kazanma olasılıkları türetilmiş ve elde edilen sonuçlar martingale temelli simülasyonlarla desteklendi. Simülasyon sonuçlarının, teorik sonuçları doğruladığı gözlemlendi.

Gelecekte, kitap bölümünde yer alan ayrık zamanlı (kesikli zamanlı) teoremin sürekli zamanlı Markov süreçlerine genişletilmesi irdelenebilir. Ayrıca teorik çerçevenin enerji temelli maliyet fonksiyonları içeren durma problemlerine genişletilmesi literatüre önemli katkılar sağlayacaktır.

Kaynakça/References

Axler, S., Bourdon, P., & Wade, R. (2013). *Harmonic function theory* (C. 137). Springer Science & Business Media.

Doob, J. L., & Doob, J. I. (1984). *Classical potential theory and its probabilistic counterpart* (C. 262). Springer.

Durrett, R. (2019). *Probability: theory and examples* (C. 49). Cambridge university press.

Ethier, S. N., & Kurtz, T. G. (2009). *Markov processes: characterization and convergence*. John Wiley & Sons.

Grimmett, G., & Stirzaker, D. (2020). *Probability and random processes*. Oxford university press.

Kelvin, W. T. B., & Tait, P. G. (1867). *Treatise on natural philosophy* (C. 1). Claredon Press.

Levin, D. A., & Peres, Y. (2017). *Markov chains and mixing times* (C. 107). American Mathematical Soc.

Norris, J. R. (1998). *Markov chains* (Sayı 2). Cambridge university press.

Vöcking, B. (2002). Randomized algorithms & probabilistic analysis. *Lecture notes, Universität Dortmund, 14*.

Williams, D. (1991). *Probability with martingales*. Cambridge university press.

Woess, W. (2000). *Random Walks on Infinite Graphs and Groups, Cambridge Univ.* Cambridge.

BÖLÜM 0

HETEROJENLİK, TERCİH POLİTİKASI VE DAĞILIMSAL ASİMETRİ: HETEROJEN BİR KUYRUK SİSTEMİNDE MÜŞTERİ SAYISINA AİT KİTLESEL MOMENTLERİN İNCELEMESİ

Murat SAĞIR¹

Müjgan ZOBU²

Vedat SAĞLAM³

Giriş

Kuyruk modelleri, telekomünikasyon, sağlık hizmetleri, üretim sistemleri, bilgisayar ağları ve ulaşım gibi birçok alanda, sistem kapasitesinin etkin biçimde planlanabilmesinde kullanılmaktadır. (Gross ve Harris, 1998). Bu modeller arasında bekleme alanı olmayan kuyruk sistemleri, gerçek hayattaki birçok durumu oldukça gerçekçi biçimde temsil etmeleri nedeniyle özel bir öneme sahiptir.

Bu kitapta ele alınan farklı tercih olasılıklı heterojen kuyruk sistemi, Poisson süreci ile gelen müşterilerin, hizmet süreleri üstel

¹ Dr. Öğr. Üyesi, İskenderun Teknik Üniversitesi, İşletme ve Yönetim Bilimleri Fakültesi, Ekonomi Bölümü, Orcid: 0000-0001-7567-9327

² Doç. Dr., Amasya Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Orcid: 0000-0001-9182-8927

³ Prof. Dr., Ondokuz Mayıs Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, Orcid: 0000-0002-8586-1373

dağılıma sahip iki heterojen paralel sunucu tarafından hizmet gördüğü ve bekleme alanının bulunmadığı bir sistemdir. Sisteme gelen bir müşteri, her iki sunucunun da dolu olması durumunda hizmet alamadan sistemi terk eder. Bu özellik, modeli “kayıp sistemleri” sınıfına dahil etmektedir (Ross, 2014).

Heterojen kuyruk modeli, özellikle kapasitenin sınırlı, beklemenin mümkün olmadığı ve sunucuların hizmet hızlarının farklı olduğu durumların modellenmesinde kullanılmaktadır. Sunucular arasındaki heterojenlik, pratikte sıklıkla karşılaşılan bir durumdur; çünkü çalışanların deneyim düzeyleri, makinelerin teknik kapasiteleri veya hizmet kanallarının fiziksel özellikleri çoğu zaman birbirinden farklıdır (Kleinrock, 1975). Heterojen kuyruk sistemlerinin en yaygın modelleme aracı olarak kullanıldığı alanlar arasında sağlık hizmetleri gelmektedir. Çünkü acil servisler, yoğun bakım üniteleri, ameliyathaneler gibi ortamlarda, hasta kabul kapasitesi kısıtlıdır ve acil durumlarda beklemek genellikle imkansızdır. Özellikle farklı uzmanlık alanlarına sahip iki doktorun ya da farklı donanımlara sahip iki ameliyathanenin olduğu durumlar için, heterojen kuyruk modeli hasta kayıp olasılığını ve sistem doluluk oranını analiz etmek için yeterli matematiksel donanımı sağlamaktadır. (Green, 2006).

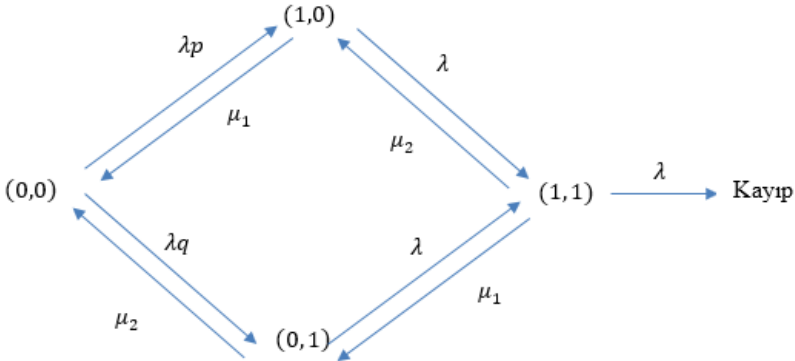
Benzer şekilde, telekomünikasyon sistemlerinde sınırlı sayıda kanalın bulunduğu ve çağrılar bekletilmeden reddedildiği durumlar klasik kayıp sistemleriyle modellenmektedir. Farklı bant genişliğine veya işlem hızına sahip iki sunucunun yer aldığı hücresel ağlar ya da anahtarlama sistemleri, heterojen sunuculu kayıp kuyruklarının önemli uygulama alanları arasındadır (Kelly, 2011). Benzer şekilde üretim ve imalat sistemlerindeki farklı makine kullanımı da heterojen kuyruk modelleri ile modellenenabilir (Buzacott ve Shanthikumar, 1993). Buna ek olarak, bilgisayar ve bilişim sistemlerinde, işlemci çekirdekleri veya sunucular farklı hızlara sahip olması ve tampon belleğin bulunmaması da heterojen kuyruk modelleri ile modellenenilmektedir (Trivedi, 2002).

Bu kitapta bir heterojen kuyruk modeli ayrıntılı biçimde ele alınacak; denge durum denklemleri ve sistemdeki kişi sayısının dağılımı elde edilecektir. Ayrıca sistemdeki kişi sayısına ait

momentler kapalı formda elde edilecektir. Elde edilen momentler ile kitlesel konum, dağılım ve yayılım ölçüleri elde edilecektir. Ayrıca sistemi oluşturan parametrelerden tercih olasılığının (tercih politikasının) ve hizmet oranlarının (μ_1, μ_2) heterojenliğinin (Gini indeks ile elde edilmektedir) kitlesel konum, dağılım ve yayılım ölçüleri üzerindeki etkisi sayısal analizler ile sezgisel olarak incelenecektir.

Farklı Tercih Olasılıklı Heterojen Kuyruk Sistemi

Müşteriler sisteme λ parametrelili Poisson akımı gelir. Her bir müşterinin k . kanalda hizmet süresi $1/\mu_k$ ($k = 1,2$) ortalaması üstel dağılıma sahiptir. Geliş anında müşteri kanalların her ikisi boş ise p ve $q(1 - p)$ olasılığı ile sırayla birinci ve ikinci kanalda hizmet alır. Yalnız bir kanal boş ise hizmetini bu kanalda alır. Sisteme gelen müşteri kanalların her ikisini de dolu bulursa hizmet almadan müşteri sistemi terk eder. Bu modele ait geçiş diyagramı aşağıdaki gibidir.



Şekil 1: Modele ait geçiş diyagramı

Limit Olasılıkları

$\{\xi_t = i, \eta_t = j; t \geq 0\}$ iki boyutlu sürekli parametrelili Markov zinciri dikkate alınıyor. Burada $i = j = 0,1,2$ dir. Bu bağlamda $P(\xi_t = i, \eta_t = j) = P_{ij}(t)$ olasılıklarının bulunması gerekiyor. $P_{ij}(t)$ t süresinde birinci kanalda i tane , ikinci kanalda

j tane müşterinin olması olasılığını gösterelim. $h \rightarrow 0$ için $P_{ij}(t)$ olasılıklarının Kolmogorov diferansiyel denklemler şöyledir.

$$P'_{00}(t) = -\lambda P_{00}(t) + \mu_1 P_{10}(t) + \mu_2 P_{01}(t) \quad (1)$$

$$P'_{10}(t) = -(\lambda + \mu_1)P_{10}(t) + p\lambda P_{00}(t) + \mu_2 P_{11}(t) \quad (2)$$

$$P'_{01}(t) = -(\lambda + \mu_2)P_{01}(t) + q\lambda P_{00}(t) + \mu_1 P_{11}(t) \quad (3)$$

$$P'_{11}(t) = -(\mu_1 + \mu_2)P_{11}(t) + \lambda(P_{01}(t) + P_{10}(t)) \quad (4)$$

Bu diferansiyel denklemlere ait sabit durum (steady-state) olasılıkları aşağıdaki gibidir.

$$0 = -\lambda P_{00} + \mu_1 P_{10} + \mu_2 P_{01} \quad (5)$$

$$0 = -(\lambda + \mu_1)P_{10} + p\lambda P_{00} + \mu_2 P_{11} \quad (6)$$

$$0 = -(\lambda + \mu_2)P_{01} + q\lambda P_{00} + \mu_1 P_{11} \quad (7)$$

$$0 = -(\mu_1 + \mu_2)P_{11} + \lambda(P_{01} + P_{10}) \quad (8)$$

X sistemdeki müşteri sayısını gösteren bir tesadüfi değişken olmak üzere olasılık fonksiyonu $p_k = P(X = k)$ olsun. Ayrıca:

$$p_0 = P_{00}, \quad p_1 = P_{01} + P_{10}, \quad p_2 = P_{11} \quad (9)$$

olarak alınıyor. Burada p_0 sistemin boş olması, p_1 kanallardan birisinin boş olması ve p_2 'de her iki kanalın dolu olması olasılığıdır. p_2 aynı zamanda sisteme gelen müşterinin hizmet almadan ayrılması (kaybolması) olasılığını da ifade eder. Böylece $P(X = k)$ aşağıdaki gibi tanımlanır.

$$P(X = k) = \begin{cases} \frac{a}{a + b + c}, & k = 0 \\ \frac{b}{a + b + c}, & k = 1 \\ \frac{c}{a + b + c}, & k = 2 \\ 0, & dd \end{cases} \quad (10)$$

Burada $a = \mu_1 \mu_2 (2\lambda + \mu_1 + \mu_2)$, $b = \lambda(\mu_1 + \mu_2)(\lambda + p\mu_2 + q\mu_1)$, $c = \lambda^2(\lambda + p\mu_2 + q\mu_1)$ olarak tanımlanmaktadır

(Saglam ve Torun, 2005) Dolayısı ile sisteme etkin geliş oranı aşağıdaki gibidir.

$$\lambda_{eff} = \lambda(1 - p_2) \quad (11)$$

Bu kitapta tanıtılan heterojen kuyruk sisteminde tercih olasılığı $p = 1/2$ alındığında sistem $M/M_i/2/0$ 'a dönüşmektedir.

Momentler

Bu alt başlıkta müşteri sayısına ait sıfır etrafındaki momentler elde edilecektir. Momentleri bulmadan önce olasılık üreten fonksiyonu elde edelim. Sistemdeki müşteri sayısına ait olasılık üreten fonksiyon aşağıdaki gibidir.

$$\begin{aligned} \eta_X(s) = E(s^X) &= \sum_{k=0}^{\infty} s^k P(X = k) \\ &= \frac{1}{a + b + c} (a + bs + cs^2) \end{aligned} \quad (12)$$

olarak elde edilir. Olasılık üreten fonksiyon yardımı ile sıfır etrafındaki ilk dört moment aşağıdaki gibi elde edilir.

$$E(X) = m_1 = \left. \frac{d}{ds} \eta_X(s) \right|_{s=1} = \frac{b + 2c}{(a + b + c)} \quad (13)$$

$$E(X(X - 1)) = \left. \frac{d^2}{ds^2} \eta_X(s) \right|_{s=1} = \frac{2c}{(a + b + c)} \quad (14)$$

Dolayısı ile

$$E(X^2) - E(X) = \frac{2c}{(a + b + c)} \quad (15)$$

$$E(X^2) = m_2 = \frac{b + 4c}{(a + b + c)} \quad (16)$$

$n \geq 3$ için

$$\left. \frac{d^n}{ds^n} \eta_X(s) \right|_{s=1} = 0 \quad (17)$$

olur. Bunun sebebi sistemde en fazla iki kişinin olabileceği gerçeğidir. Dolayısı ile

$$E(X(X-1)(X-3)) = 0 \quad (18)$$

ve

$$E(X(X-1)(X-3)(x-4)) = 0 \quad (19)$$

eşitlikleri geçerlidir. Bu eşitliklerden ve 1. ve 2. sıfır etrafındaki momentlerden yararlanarak sıfır etrafındaki 3. ve 4. momentler aşağıdaki gibi elde edilir.

$$E(X^3) = m_3 = \frac{b + 8c}{(a + b + c)} \quad (20)$$

$$E(X^4) = m_4 = \frac{b + 16c}{(a + b + c)} \quad (21)$$

Bu dağılımın momentleri kullanılarak sırasıyla Beklenen değer, Varyans, Değişim Katsayısı, Çarpıklık ve Basıklık katsayısı hesaplanabilir (Sağlam vd., 2018; Stewart, 2009)

$$E(X) = m_1 \quad (22)$$

$$Var(X) = \sigma^2 = m_2 - m_1^2 \quad (23)$$

$$\begin{aligned} \text{ÇK} = \alpha_3 &= \frac{E(X - E(X))^3}{\sigma^3} \\ &= \frac{m_3 - 3m_1m_2 + 2m_1^3}{(m_2 - m_1^2)^{3/2}} \end{aligned} \quad (24)$$

$$\begin{aligned} BS = \alpha_4 &= \frac{E(X - E(X))^4}{\sigma^4} \\ &= \frac{m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4}{(m_2 - m_1^2)^2} \end{aligned} \quad (25)$$

Uygulama

Bu bölümde yukarıda tanıtılan heterojen kuyruk modelinin nümerik örnekleri verilecektir. Özellikle kanallar arasındaki

heterojenliğin büyük olduğu durumlarda sistemdeki kişi sayısının dağılımlı hakkında konum ve dağılım ölçülerinin nasıl etkilendiği ve bunun yanında çarpıklık ve basıklık katsayısının nasıl etkilendiği araştırılacaktır. Modeldeki kişi sayılarına ait varyansın tercih olasılığı sayesinde azaltılabilir olup olmadığı araştırılacaktır. Özel olarak sistemdeki kişi sayısının az olması âtil zaman oluşumu ile verimliliği düşürürken, sistemim dolu olması (bu modelde sistemde iki kişinin olması) yeni gelen müşterilerin kaybolması maliyetini doğuracaktır. Bu durumda heterojen ortalamalara sahip kanallardan hızlı olanın seçilmesi politikası benimsenebilir. Bu tercih olasılığı yük dengeleme parametresi olarak kullanılabilir. Bu tercih olasılığının yarı yarıya durumundan (%50 olasılığından) fazla olmalı gerekir. Fakat optimal tercih olasılığının %50 ile %100 arasından nasıl belirlenebilir olduğu sadece sistemdeki kişi sayısının ortalaması ile üst mertebeden momentleri ile elde edilen kitlesel göstergeler ile karar verilmelidir. Aslında heterojenlik ve yönlendirme olasılığı, yalnızca ortalamaı değil dağılımın şeklini de belirgin biçimde etkilemektedir. İlk olarak heterojenliğin bir ölçüsü olarak gini indeks değerleri hesaplanmaktadır.

Bu kitap bölümünde ele alınan modelde, paralel kanalların heterojenliği Gini indeks ile ölçülebilir. Modelde yer alan paralel kanallar için hizmet kapasitesinin sabit ve sınırlı olduğu düşünüldüğünde bu kapasitenin kanallar arasında nasıl dağıldığını gini indeks ile ölçebiliriz (Yitzhaki, 1979). Gini indeks 0 (kanallar homojen) ile 1 (kanallar tamamen heterojen) arasında değerler alır (Alves vd, 2011). Gini indeksi, sürekli rassal değişkenler için aşağıdaki biçimde tanımlanabilir (Cowell, 2011):

$$G = \frac{1}{2\mu} \int_a^b \int_a^b |x - y| f(x) f(y) dx dy \quad (26)$$

Bizim modelimizde olduğu gibi iki noktalı ayrık dağılımlar için bu yukarıdaki formül daha basit bir hal alır. Özetle bu çalışmada ele alınan iki sunuculu sistem için Gini indeksi,

$$G = \frac{\frac{1}{2} |\mu_1 - \mu_2|}{2 \cdot \frac{\mu_1 + \mu_2}{2}} = \frac{|\mu_1 - \mu_2|}{\mu_1 + \mu_2} \quad (27)$$

ile hesaplanır. Modelde genelliği bozmaksınız $\mu_1 \geq \mu_2$ varsayıldığında gini indeks

$$G = \frac{\mu_1 - \mu_2}{\mu_1 + \mu_2} \quad (28)$$

olarak belirlenir. Gini indeks değerleri göz önüne alınarak seçili bazı durumlar için nümerik sonuçlar aşağıdaki gibi elde edildi.

Tablo 1: $\lambda = 4,2$ müş/dk ve $\mu_1 + \mu_2 = 6$ müş/dk olmak farklı gini indeks ve tercih olasılıkları için konum ve yayılım ölçüleri

μ_1	μ_2	G	p	E(X)	Varyans	ÇK	BK
3,00	3,00	0,00	0,50	0,994	0,586	0,010	1,707
3,20	2,80	0,07	0,50	0,995	0,585	0,008	1,709
			0,60	0,994	0,586	0,011	1,707
			0,70	0,992	0,587	0,013	1,705
			0,80	0,990	0,587	0,016	1,703
			0,90	0,989	0,588	0,019	1,701
			1,00	0,987	0,589	0,022	1,699
4,00	2,00	0,33	0,50	1,028	0,571	-0,046	1,753
			0,60	1,020	0,575	-0,033	1,741
			0,70	1,012	0,578	-0,020	1,729
			0,80	1,003	0,582	-0,005	1,718
			0,90	0,994	0,586	0,010	1,707
			1,00	0,985	0,589	0,026	1,697

4,80	1,20	0,60	0,50	1,113	0,524	-0,173	1,918
			0,60	1,100	0,531	-0,156	1,889
			0,70	1,087	0,540	-0,137	1,860
			0,80	1,072	0,548	-0,116	1,831
			0,90	1,057	0,557	-0,092	1,800
			1,00	1,039	0,565	-0,064	1,770
5,20	0,80	0,73	0,50	1,182	0,474	-0,254	2,101
			0,60	1,170	0,484	-0,242	2,066
			0,70	1,156	0,494	-0,227	2,028
			0,80	1,141	0,505	-0,209	1,987
			0,90	1,123	0,517	-0,187	1,943
			1,00	1,103	0,530	-0,161	1,896
5,80	0,20	0,93	0,50	1,339	0,327	-0,174	2,312
			0,60	1,333	0,333	-0,192	2,333
			0,70	1,327	0,340	-0,211	2,351
			0,80	1,319	0,349	-0,230	2,364
			0,90	1,309	0,359	-0,248	2,371
			1,00	1,297	0,372	-0,266	2,368

Yukarıdaki tablo incelendiğinde gini indeksinin sıfır olduğu satır kanalların homojen olduğu durumu gösteriyor. Homojen duruma göre sistemdeki ortalama kişi sayısı yaklaşık bire eşit ve

çarpıklık katsayısı neredeyse sıfır olduğu için simetrik bir durum söz konusudur. Bu durum heterojen durumlar için bir referans noktası sağlamaktadır. Düşük heterojenlik durumunda ($G = 0,07$) olması durumunda yönlendirme olasılığının etkisi zayıf kalmaktadır. Yönlendirme olasılığı hızlı kanala doğru büyüdükçe ortalama çok az azalıyor, varyans çok az artıyor, çarpıklık simetriden sağa doğru artıyor, basıklık ise çok az düşüyor. Orta heterojenlikte ($G = 0,33$) ise yönlendirme olasılığı ile ortalama düşüyor (aslında birin yukarisından bire doğru toparlanıyor), fakat varyans düşük heterojenlikte olduğu gibi yine artıyor. Çarpıklık negatif değerlerden (sola çarpık durumdan) artarak önce sıfıra (simetrik durumu) daha sonra pozitifte doğru (sağa çarpık durum) artıyor. Dolayısı ile orta heterojenlikte yönlendirme politikası sistemdeki kişi sayısının dağılımını ve çarpıklığını etkilediği nümerik olarak gözlemlenmektedir. Yani rastgele yönlendirme (eşit olasılıklı seçim) altında dağılım sola çarpık iken, hızlı kanala yönlendirme olasılığı (p) büyüdükçe dağılım sağa çarpık bir yapıya dönüşüyor. Yüksek heterojenlik durumunda ($G \geq 0,6$) ortalama gini indeksi artıkça artıyor. Bu durum sistemin dolu olması anlamına geliyor ve âtil zamanın azaldığı ve sistemin verimli çalıştığı durumu ifade ediyor. Fakat dolu bir sistemde kayıpların olmasının kaçınılmaz olduğu da aşikâr. Varyans kademeli bir azalış gösteriyor. Çarpıklık bu durumda daima sola çarpık olarak heterojenliğe bağlı olarak artıyor. Tercih politikası bu sola çarpıklığın etkisini azaltsa da dağılımın simetrik duruma gelmesi için yeterli olmuyor. Basıklık katsayısı ise heterojenliğin artması ile düzgün dağılım basıklığından kalın kuyruklu normal dağılım basıklığına doğru bir yaklaşım sergiliyor. Bu durum, dağılımın basıklığının artmasına ve uç değerlerin daha baskın hâle gelmesine yol açmaktadır. Heterojenlik düzeyi arttıkça sistemdeki kişi sayısının dağılımı simetrik yapısını kaybetmekte ve belirgin biçimde asimetrik (sola çarpık) hâle gelmektedir. Yönlendirme politikası ile çarpıklığın işaret değiştirmesi, heterojenliğin yalnızca yayılımı değil, dağılımın yönünü de etkilediğini göstermektedir. Yönlendirme politikası, heterojenliğin neden olduğu dağılımsal etkileri tamamen ortadan kaldırmamakta; ancak dağılımın şeklini kısmen dengeleyici bir rol üstlenmektedir.

Sayısal sonuçlar, sunucuların heterojenlik düzeyinin sisteme ait ölçümlerde yalnızca ortalamalar için değil, dağılımın şekli (simetri ya da çarpık dağılım) içinde önemli etkileri olduğunu göstermektedir. Gini indeksi ile ölçülen heterojenlik arttıkça, sistemdeki kişi sayısının dağılımı daha asimetric (sola çarpık) ve kalın kuyruklu bir basıklık yapısına dönüşmektedir. Yönlendirme politikası bu etkiyi nispeten azaltsa da, yüksek heterojenlik durumlarında dağılımın kuyruk davranışı daha ağır basmaktadır. Bu bulgular, heterojen servis sistemlerinde yalnızca ortalama performans ölçütlerine odaklanmanın yetersiz olabileceğini ortaya koymaktadır.

Sonuçlar ve Gelecek Çalışmalar

Bu bölümde, önünde beklemeye müsaade edilmeyen iki kanallı heterojen bir servis sisteminde sunucu heterojenliğinin sistemdeki kişi sayısının dağılımı üzerindeki etkileri incelenmiştir. Sisteme ait kanallardaki hizmet oranlarının ortalamaları arasındaki farklılık, Gini indeksi yardımıyla nicel olarak ölçülmüş ve üç düzeyde sınıflandırılmıştır (düşük, orta ve yüksek). Ayrıca sistemdeki kişi sayısının dağılımından yararlanarak sistemdeki kişi sayısının olasılık üreten fonksiyon u elde edilmiştir. Daha sonra türev geçişleri ile sıfır etrafında ilk dört moment kapalı formda elde edilmiştir. Daha sonra bu momentler yardımı ile sistemdeki kişi sayısına ait kitlesel konum, yayılım ve dağılım ölçüleri elde edilmiştir.

Kapalı formda elde edilen kitlesel konum ve dağılım ölçüleri sistemin heterojenliği ile nümerik olarak incelenmiştir. Sonuçlar sunucuların heterojenliğinin yalnızca ortalama olarak değil, dağılımın şeklini de belirgin biçimde etkilediğini göstermektedir. Heterojenlik arttıkça sistemdeki kişi sayısının dağılımı simetric yapısını kaybetmekte, daha asimetric (sola çarpık) ve düzgün dağılım basıklığından kalın kuyruklu normal dağılım basıklığına doğru bir geçiş sergilemektedir. Bu durum, sistemin çoğu zaman düşük doluluk seviyelerinde kalmasına rağmen, nadir gerçekleşen yüksek doluluk durumlarının daha baskın hâle gelmesine yol açmaktadır. Yönlendirme politikası (hızlı kanalı tercih etme olasılığı) bu etkileri kısmen dengeleyebilmekte; ancak yüksek

heterojenlik düzeylerinde dağılımın kuyruk davranışı baskın gelmekte ve yönlendirme olasılığı baskınlık altında etkisiz kalmaktadır.

Gelecekte maliyet yapısı modele dahil edilerek genişletilebilir. Hizmet kanallarının çalıştırılmasından kaynaklanan işletme maliyetleri, müşterileri kaybından kaynaklanan maliyetler bir maliyet fonksiyonu olarak tasarlanabilir. Özellikle heterojenlik düzeyinin zaman içinde değiştiği dinamik kuyruk modellerinde, düşük ortalama doluluğa rağmen artan kuyruk kalınlığı, maliyet minimizasyonu açısından önemli çözümler sağlayabilir. Bu bağlamda, heterojenliğin yalnızca yapısal bir özellik değil, dinamik olarak yönetilebilen bir karar değişkeni olarak değerlendirilmesi hem teorik hem de uygulamalı çalışan araştırmacılar için ilgi çekici olabilir.

Sonuç olarak, bu çalışma heterojen servis sistemlerinde dağılım temelli performans analizinin önemini vurgulamakta ve maliyet temelli, dinamik optimizasyon yaklaşımlarına yönelik bir orijin noktası olmaktadır.

Kaynaklar

- Alves, F. S. Q., Yehia, H. C., Pedrosa, L. A. C., Cruz, F. R. B., & Kerbache, L. (2011). Upper bounds on performance measures of heterogeneous M/M/c queues. *Mathematical Problems in Engineering*, 2011(1), 702834.
- Buzacott, J. A., & Shanthikumar, J. G. (1993). *Stochastic models of manufacturing systems*. Prentice Hall.
- Cowell, F. (2011). *Measuring inequality*. Oxford University Press
- Green, L. V. (2006). Queueing analysis in healthcare. *Patient Flow: Reducing Delay in Healthcare Delivery*, 281–307.
- Gross, D., & Harris, C. M. (1998). *Fundamentals of queueing theory* (3rd ed.). Wiley.
- Kelly, F. P. (2011). *Reversibility and stochastic networks*. Cambridge University Press.
- Kleinrock, L. (1975). *Queueing systems, Volume I: Theory*. Wiley.
- Ross, S. M. (2014). *Introduction to probability models* (11th ed.). Academic Press.
- Saglam, V., & Torun, H. (2005). On optimization of stochastic service system with two heterogeneous channels. *International Journal of Applied Mathematics*, 17(1), 1.
- Sağlam, V. Sağır M., Yücesoy E., Olasılığa Giriş, Nobel Kitabevi, Ankara, 2023.
- Stewart, W. J. (2009). *Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling*. Princeton university press.

Trivedi, K. S. (2002). *Probability and statistics with reliability, queueing, and computer science applications* (2nd ed.). Wiley.

Yitzhaki, S. (1979). Relative deprivation and the Gini coefficient. *The quarterly journal of economics*, 93(2), 321-324.

BÖLÜM 0

BIYOİSTATİSTİKTE SİMÜLASYON YAKLAŞIMLARI VE UYGULAMALARI

HAKAN ÖZTÜRK¹

Giriş

Simülasyon, istatistiksel yöntemlerin performansını değerlendirmek amacıyla yapay veri üretimi ve bu veriler üzerinde sistematik analizler gerçekleştirmeyi sağlayan güçlü bir araçtır. Gerçek dünyadan elde edilen veriler; örneklem büyüklüğü, gözlem süresi, eksik veri oranı veya etki büyüklüğü gibi açılardan yetersiz veya ideal olmayan koşullara sahip olabilir. Bu durumda, araştırmacılar gerçekçi fakat kontrol edilebilir senaryolar altında yöntemlerin nasıl davrandığını görmek için simülasyonlara başvururlar (Burton, Altman, Royston & Holder, 2006).

Biyoistatistik alanında, özellikle yeni geliştirilmiş istatistiksel modellerin uygulanabilirliği, güç analizi, örneklem büyüklüğü hesaplaması, çoklu test düzeltmeleri ve eksik veri atama yöntemlerinin karşılaştırılması gibi konularda simülasyon çalışmaları kritik öneme sahiptir. Simülasyonlar sayesinde “gerçeklik” olarak tanımlanabilecek veri üretim mekanizması

¹ Araş. Gör. Dr., Aydın Adnan Menderes Üniversitesi, Biyoistatistik AD, Orcid: 0000-0001-8112-4934

bilinmekte ve buna göre test edilen yöntemin sapma, tutarlılık, girdi değişkenlerine duyarlılık gibi özellikleri incelenabilmektedir (Morris, White & Crowther, 2019).

Bu kitap bölümünde amaç, biyoistatistik bağlamında simülasyon yaklaşımlarının kuramsal temellerini, türlerini, uygulama alanlarını, yazılım altyapılarını ve güncel eğilimlerini sistematik şekilde ele almaktır. Özellikle sağlık ve tıp alanında, yöntem geliştirme sürecinde veya gözlemsel verilerin doğası gereği sınırlı olduğu durumlarda simülasyonun nasıl etkin kullanılabileceği üzerinde durulacaktır. Böylece hem araştırmacılar hem de uygulayıcılar için rehber niteliğinde bir kaynak sunulması hedeflenmektedir.

SİMÜLASYONUN KURAMSAL TEMELLERİ

İstatistiksel açıdan simülasyon, olasılıksal süreçlerin bilgisayar ortamında defalarca üretilmesi ve bu süreçlerden elde edilen sonuçların dağılım özelliklerinin incelenmesine dayanır. Temelinde, büyük sayılar yasası ve olasılık kuramının rastgelelik ilkesine dayalıdır. Bu yönüyle simülasyon, karmaşık veya analitik çözümü olmayan olasılık modellerini yaklaşık yöntemlerle analiz etmeye olanak sağlar (Law, 2015).

Biyoistatistikte kullanılan simülasyonların dayandığı en önemli kuramsal yapı Monte Carlo yöntemidir. Monte Carlo simülasyonları, rastgele sayılar kullanarak belirli bir istatistiğin (örneğin ortalama, varyans, p-değeri) dağılımını tahmin eder. Bu yaklaşım, özellikle karmaşık modellerin teorik olarak çözülemeyeceği durumlarda kullanılabilir bir alternatif sunar (Rubinstein & Kroese, 2017). Örneğin, klinik araştırmalarda bir testin Tip I hata oranı, örneklem büyüklüğü veya etki büyüklüğü değiştikçe nasıl değişir sorusu, binlerce rastgele veri kümesi üretilerek incelenebilir.

Simülasyon sürecinde tekrarlama kavramı kritik bir öneme sahiptir. Aynı senaryonun binlerce kez tekrar edilmesi, bir istatistiğin beklenen değerine yakınsama olanağı sağlar. Bu özellik, olasılık kuramının temel ilkelerinden biri olan yakınsama teoremi ile de uyumludur: örneklem sayısı arttıkça simülasyon ortalaması, teorik beklentiye yaklaşır (Morris et al., 2019). Bu nedenle simülasyon sayısının yeterince yüksek seçilmesi (genellikle ≥ 1000) elde edilen sonuçların güvenilirliği açısından önemlidir.

Simülasyon çalışmalarında rastgelelik kontrolü için seed (tohum) değeri kullanımı zorunludur. Seed değeri, kullanılan rastgele sayı üreteçlerinin aynı koşullarda tekrar çalıştırıldığında aynı sonuçları vermesini sağlar. Bu da bilimsel araştırmalarda “yeniden üretilebilirlik” ilkesinin temelini oluşturur (Burton et al., 2006). Özellikle metodolojik çalışmalarda, kullanılan seed değerinin, algoritma parametrelerinin ve yazılım sürümlerinin açık biçimde raporlanması gereklidir.

Simülasyonlar genellikle parametrik ve parametrik olmayan biçimlerde sınıflandırılır. Parametrik simülasyonlarda veri belirli dağılımlardan (örneğin normal, binom, Poisson) üretilirken, parametrik olmayan yaklaşımlarda gözlenen veriden doğrudan örnekleme yapılır. Bootstrap yöntemi bunun en bilinen örneğidir; gerçek veri kümesinden tekrar örnekleme yapılarak istatistiksel belirsizlik hesaplanır (Efron & Tibshirani, 1993).

Son yıllarda biyoistatistikte, klasik Monte Carlo simülasyonlarının ötesinde Bayesyen simülasyon teknikleri (örneğin Markov Chain Monte Carlo – MCMC) yaygınlaşmıştır. Bu yöntemlerde amaç, posterior olasılık dağılımlarından rastgele örnekler çekerek karmaşık modellerde parametre tahminleri elde etmektir (Gelman et al., 2014). Bu yaklaşım özellikle hiyerarşik modeller, çok düzeyli veriler ve eksik veri analizlerinde güçlü bir alternatif sunar.

Özetle, simülasyonun kuramsal temelleri olasılık teorisi, rastgelelik ve tekrar kavramları üzerine inşa edilmiştir. Bu temeller, biyoistatistikteki uygulamaların istatistiksel doğruluk, genellenebilirlik ve yöntem karşılaştırmaları açısından güvenilir biçimde yürütülmesine olanak tanır.

SİMÜLASYON TÜRLERİ

Biyoistatistikte simülasyonlar, amaçlarına ve veri üretim mekanizmalarına göre farklı biçimlerde sınıflandırılır. Her simülasyon türü, belirli istatistiksel problemlere odaklanır ve yöntemlerin farklı koşullar altındaki davranışlarını incelemeyi amaçlar. En yaygın kullanılan beş temel simülasyon yaklaşımı Tablo 1’de özetlenmiştir.

Tablo 1 Biyoistatistikte Temel Simülasyon Türleri ve Kullanım Alanları

Simülasyon Türü	Temel Özellik	Biyoistatistikte Kullanım Alanı
Monte Carlo	Rastgele veri üretimi, ortalama davranışın incelenmesi	Güç, Tip I hata, model doğruluğu
Bootstrap	Gerçek veriden tekrar örnekleme	Güven aralığı, yanlılık düzeltilmesi
Bayesyen	Posterior dağılımlardan örnekleme	Hiyerarşik modeller, eksik veri
Mikrosimülasyon	Birey düzeyinde süreç modelleme	Epidemiyoloji, sağlık ekonomisi
Agent-Based Model	Etkileşimli ajan temsili	Bulaşıcı hastalık yayılımı, davranış modelleri

Monte Carlo Simülasyonu

Monte Carlo simülasyonları, rastgele sayı üretimi yoluyla olasılık dağılımlarından tekrar tekrar örneklem alarak bir istatistiğin dağılımını tahmin etmeyi amaçlar. Bu yöntem, karmaşık veya analitik çözümü bulunmayan modellerin davranışını anlamaya olanak sağlar (Rubinstein & Kroese, 2017).

Biyoistatistikte, özellikle testlerin güç ve Tip I hata oranlarını incelemek, çoklu karşılaştırma düzeltmelerinin performansını değerlendirmek ve örneklem büyüklüğünün belirlenmesi için sıklıkla kullanılır. Örneğin, iki grup arasındaki ortalama farkını test eden bağımsız örneklemeler t-testinin farklı örneklem büyüklüklerinde gücü Monte Carlo yaklaşımıyla kolayca simüle edilebilir (Burton et al., 2006).

Bootstrap Simülasyonu

Bootstrap yöntemi, parametrik bir dağılım varsayımı yapmadan, mevcut veri kümesinden tekrar örnekleme yoluyla istatistiksel belirsizliğin tahmin edilmesini sağlar (Efron & Tibshirani, 1993).

Biyoistatistiksel uygulamalarda bootstrap simülasyonlar; güven aralıklarının oluşturulması, parametre tahminlerindeki yanlılığın (bias) düzeltilmesi ve küçük örneklemelerde yöntemlerin kararlılığının değerlendirilmesinde önemli bir rol oynar. Ayrıca, medyan ya da oran gibi istatistikler söz konusu olduğunda, parametrik varsayımların sağlanmadığı durumlarda bootstrap yaklaşımı sıklıkla tercih edilir.

Bayesyen Simülasyon

Bayesyen simülasyonlar, olasılık dağılımlarının doğrudan hesaplanamadığı durumlarda Markov Chain Monte Carlo (MCMC) algoritmalarını kullanarak posterior dağılımlardan örnekler üretir (Gelman et al., 2014).

Günümüzde hiyerarşik modeller, çok düzeyli veri yapıları ve eksik veri problemlerinin çözümünde en güçlü yöntemlerden biridir. Örneğin, bir klinik çalışmada hasta grupları arasındaki varyasyonun hem grup içi hem de gruplar arası belirsizliklerle birlikte modellenmesi MCMC simülasyonlarıyla gerçekleştirilebilir.

Mikrosimülasyon

Mikrosimülasyon veya birey temelli simülasyon, birey düzeyinde özelliklerin zaman içinde nasıl değiştiğini izler. Her bir bireyin belirli bir olasılıkla hastalanması, tedavi alması veya ölmesi gibi süreçler, rassal geçiş olasılıklarıyla modellenir (Karnon, 2003).

Bu yaklaşım özellikle epidemiyoloji, sağlık ekonomisi ve politika analizlerinde kullanılır. Örneğin, bir aşı programının toplum düzeyinde etkisi, bireylerin enfeksiyon geçirme olasılıklarına dayalı mikrosimülasyonlarla tahmin edilebilir.

Sistem ve Agent-Based Simülasyonlar

Sistem dinamiği ve agent-based modeller, makro düzeyde karmaşık sağlık sistemlerini veya bireyler arası etkileşimleri analiz etmek için kullanılır. Agent-based modeller, her bir bireyin “ajan” olarak davranış kurallarına sahip olduğu ve bu etkileşimlerin popülasyon düzeyinde sonuçlar ürettiği simülasyonlardır (Macal & North, 2010).

Biyoistatistikte bu yöntemler, bulaşıcı hastalık yayılım modelleri, davranışsal müdahalelerin etkisi ve sağlık hizmeti kullanım dinamikleri gibi konularda kullanılmaktadır.

SİMÜLASYON SÜRECİ

Simülasyon çalışmaları, yalnızca rastgele veri üretiminden ibaret olmayıp dikkatli bir planlama, açık senaryo tanımı ve sistematik raporlama gerektirir. Biyoistatistikte yürütülen bir simülasyon araştırmasının geçerliliği, sürecin her adımında yapılan tercihlerle doğrudan ilişkilidir (Morris, White & Crowther, 2019). Bu nedenle, tasarım ve yürütme aşamalarının belirli standartlara uygun biçimde gerçekleştirilmesi önerilir (Burton et al., 2006).

Aşağıda, simülasyon süreci altı temel adımda özetlenmiştir.

Amaç ve Tasarımın Belirlenmesi

Bir simülasyon çalışmasının ilk adımı, araştırma sorusunun ve amacın net biçimde tanımlanmasıdır. Araştırmacı hangi yöntemi, hangi koşullar altında test ettiğini ve hangi ölçütlerle performans değerlendirmesi yapacağını önceden belirtmelidir (White et al., 2011).

Örneğin, “farklı örneklem büyüklüklerinde t-testinin Tip I hata oranı değişir mi?” veya “Benjamini–Hochberg ve Holm yöntemleri arasında FDR (False Discovery Rate) kontrol performansı açısından fark var mı?” gibi açık sorular tanımlanmalıdır.

Bu aşamada ayrıca aşağıdaki tasarım bileşenleri belirlenir:

- Senaryoların sayısı ve kombinasyonları (ör. örneklem büyüklüğü, varyans, etki büyüklüğü)
- Tekrarlama sayısı — genellikle 1000 veya daha fazla önerilir
- Korelasyon yapısı ve dağılım varsayımları
- Modelin türü (ör. parametrik, parametrik olmayan, Bayesyen)

Senaryo ve Veri Üretim Mekanizması

Simülasyonun ikinci adımı, veri üretim mekanizmasının tanımlanmasıdır. Bu aşamada araştırmacı hangi değişkenlerin hangi dağılımlardan üretileceğini, aralarındaki ilişkileri ve olası hata yapısını açık biçimde belirtmelidir (Law, 2015).

Örneğin:

- Sürekli değişkenler için normal ($N(\mu, \sigma^2)$) veya log-normal dağılımlar,
- Kategorik değişkenler için Bernoulli veya multinomiyal dağılımlar,
- Korelasyonlu yapılar için çok değişkenli normal dağılımlar tercih edilebilir.

Ayrıca, eksik veri veya ölçüm hatası gibi durumlar incelenmek isteniyorsa, bunlar da veri üretim aşamasına dahil edilmelidir. Bu bağlamda, veri üretim modeli ile analiz modelinin birbirinden ayrılması, yöntemlerin yanlılık ve varyans açısından performanslarının daha sağlıklı biçimde değerlendirilmesine olanak sağlar (Morris et al., 2019).

Modelleme ve Analiz Aşaması

Üretilen her simülasyon tekrarında ilgili istatistiksel yöntem uygulanır. Bu adımda amaç, her senaryoda yöntemin ne ölçüde doğru sonuç verdiğini test etmektir.

Kullanılabilecek yöntemler arasında klasik parametrik testler (ör. t-testi, ANOVA), regresyon modelleri (ör. doğrusal, lojistik, Poisson) veya çoklu karşılaştırma düzeltmeleri (ör. BH, BY, Holm) bulunabilir.

Her bir yinelemede elde edilen sonuçlar (örneğin p-değeri, katsayı tahmini, model uyum ölçütü) kaydedilir. Böylece binlerce tekrar sonunda yöntemin genel performansına ilişkin istatistiksel özetler hesaplanabilir.

Performans Ölçütlerinin Hesaplanması

Simülasyonun başarısı, tanımlanmış performans ölçütleri üzerinden değerlendirilir. Bu ölçütler, simülasyonun amacına göre

değişmekle birlikte, genellikle aşağıdakileri kapsar (Burton et al., 2006; Morris et al., 2019):

- Tip I hata oranı: Yanlış pozitif sonuçların oranı
- Güç: Gerçek bir etkinin tespit edilme olasılığı
- Sapma: Tahminin gerçek değerden ortalama farkı
- Varyans veya RMSE: Tahminlerin yayılımı
- FDR / FWER: Çoklu karşılaştırmalarda hata kontrolü
- Kapsama olasılığı: Güven aralığının gerçek değeri kapsama oranı

Performans ölçütleri sadece tek bir senaryo için değil, farklı parametre kombinasyonları altında da karşılaştırılmalı; sonuçlar tablo veya grafik biçiminde özetlenmelidir.

Sonuçların Özetlenmesi ve Yorumlanması

Her simülasyon senaryosundan elde edilen sonuçlar birleştirilerek genel eğilimler belirlenir. Ortalama değerler, standart sapmalar, yüzdelikler veya oranlar kullanılabilir. Bulgular yorumlanırken;

- Simülasyon parametrelerindeki değişimlerin performans ölçütleri üzerindeki etkisi,
- Örneklem büyüklüğü veya varyans gibi faktörlerin güce etkisi,
- Modeller arası karşılaştırmalar açık biçimde sunulmalıdır.

Bu aşamada, görsel sunumlar (ör. boxplot, ısı haritaları, eğri grafikleri) okuyucunun senaryolar arasındaki farklılıkları kolayca görmesini sağlar.

Tekrarlanabilirlik ve Raporlama İlkeleri

Son adım, simülasyonun yeniden üretilebilir biçimde raporlanmasıdır. Morris ve ark. (2019), simülasyon çalışmalarında aşağıdaki bilgilerin mutlaka raporlanmasını önermektedir:

- Amaç ve senaryoların açık tanımı
- Veri üretim mekanizması ve dağılımlar
- Analiz yöntemi ve kullanılan yazılım
- Performans ölçütleri ve hesaplama biçimleri
- Tekrarlama sayısı ve seed değeri
- Sonuçların özetlenme biçimi

Bu ilkeler, “yeniden üretilebilir araştırma” yaklaşımının temelini oluşturur. Ayrıca, kullanılan kodların, parametrik değerlerin ve yazılım sürümlerinin açık erişimle paylaşılması, bilimsel şeffaflığı ve güvenilirliği artırır (Peng, 2011).

SİMÜLASYONUN BİYOİSTATİSTİKTE UYGULAMA ALANLARI

Simülasyon yöntemleri, biyoistatistikte hem metodolojik hem de uygulamalı araştırmaların temel araçlarından biri hâline gelmiştir. Gerçek dünyadan elde edilen sağlık verileri çoğu zaman eksik, dengesiz veya karmaşık yapıda olduğundan, araştırmacıların bu tür zorluklara karşı yöntemsel dayanıklılığı test etmeleri gerekir. Simülasyonlar, istatistiksel yöntemlerin farklı koşullar altında nasıl performans gösterdiğini değerlendirmek için kontrollü bir ortam sağlar (Burton et al., 2006; Morris, White & Crowther, 2019).

Aşağıda, biyoistatistikte yaygın olarak kullanılan başlıca uygulama alanları özetlenmiştir.

Klinik Deneme Tasarımları ve Güç Analizleri

Klinik arařtırmalarda istatistiksel güç, örneklem büyüklüğü ve hata oranı arasındaki ilişki çoęu zaman teorik olarak belirlenemeyecek kadar karmařık olabilir. Bu durumda, simülasyonlar farklı örneklem büyüklükleri, deęişkenlik düzeyleri ve etki büyüklükleri altında testin performansını deęerlendirmek için kullanılır (Vallat et al., 2021).

Örneęin, bir faz III klinik çalışmada iki tedavi grubunun ortalama iyileşme sürelerinin karşılaştırılacağı varsayılın. Arařtırmacı, etki büyüklüğünü (δ) ve varyansı (σ^2) deęiřtirerek 1000 tekrarlı simülasyonlar aracılığıyla testin gücünü ($1-\beta$) tahmin edebilir. Böylece örneklem büyüklüğünün klinik olarak anlamlı farkı tespit etmek için yeterli olup olmadığı belirlenir.

Simülasyonlar ayrıca uyarlanabilir klinik denemelerin planlanmasında da kullanılır. Bu tür çalışmalarda, ara analizler sonrası randomizasyon oranı veya örneklem büyüklüğü yeniden düzenlenebilir. Simülasyonlar bu adaptif kararların Tip I hata oranına ve güce etkisini önceden deęerlendirmek için kritik bir rol oynar (Lee & Tsiatis, 2020).

Eksik Veri Analizleri

Eksik veri, biyoistatistikte en sık karşılaşılan metodolojik sorunlardan biridir. Eksik veri mekanizması; tamamen rastgele eksik (MCAR), rastgele eksik (MAR) veya rastgele olmayan eksik (MNAR) olabilir (Little & Rubin, 2019).

Simülasyonlar bu mekanizmaların yöntem performansına etkisini deęerlendirmek için kullanılır. Örneęin, çoklu atama yöntemleri (MICE, EM, veya missForest) farklı eksiklik oranlarında karşılaştırılabilir. Her senaryoda tahminlerin sapma, varyans ve kapsama olasılığı deęerleri hesaplanarak hangi yöntemin daha saęlam olduęu belirlenir (Morris et al., 2019).

Bu yaklaşım, özellikle klinik verilerin sıklıkla eksik olduğu uzunlamasına çalışmalarda büyük önem taşır. Simülasyon sonuçları, uygun atama yönteminin seçilmesine yardımcı olur.

Çoklu Karşılaştırma ve Hata Kontrolü

Çoklu hipotez testleri, özellikle genomik, proteomik veya sağlık bilişimi çalışmalarında yaygındır. Ancak çok sayıda test yapılması yanlış pozitif sonuçların artmasına neden olur.

Simülasyonlar, bu durumda kullanılan düzeltme yöntemlerinin (örneğin Bonferroni, Holm, Benjamini–Hochberg, Benjamini–Yekutieli) FDR ve FWER (Family-Wise Error Rate) kontrol performanslarını karşılaştırmak için kullanılır (Chen et al., 2017).

Örneğin, ilişkili değişkenlerden türetilen 10000 testin yer aldığı bir senaryo altında Benjamini–Hochberg (BH) ve Benjamini–Yekutieli (BY) yöntemleri simüle edilerek FDR kontrol düzeyi ve istatistiksel güç açısından farkları değerlendirilebilir. Bu tür simülasyonlar, özellikle yüksek boyutlu verilerde hangi düzeltme yönteminin tercih edilmesi gerektiğine dair pratik öneriler sunar.

Risk Skoru ve Nomogram Geliştirme

Klinik risk tahmin modellerinde (örneğin mortalite, komplikasyon veya depresyon riski), modelin güvenilirliği hem örneklem büyüklüğüne hem de değişken seçimi stratejisine bağlıdır.

Simülasyon çalışmaları, farklı örneklem büyüklükleri ve olay sıklıkları altında bir modelin kalibrasyon (tahmin edilen ve gözlenen olasılıklar arasındaki uyum) ve diskriminasyon (ayırma gücü, örn. AUC) performanslarını test etmek için kullanılır (Steyerberg, 2019).

Örneğin, lojistik regresyonla oluşturulan bir risk skoru modelinde, varyans büyüklüğü veya bağımsız değişken sayısı değiştirildiğinde AUC'nin nasıl etkilendiği simülasyonlarla

incelenebilir. Böylece, modelin örnekleme duyarlılığı ve genellenebilirliği değerlendirilir.

Yüksek Boyutlu Veri ve Makine Öğrenmesi Uygulamaları

Simülasyonlar, makine öğrenmesi (ML) algoritmalarının performanslarını karşılaştırmak için sıklıkla kullanılır (Boulesteix et al., 2020).

Gerçek veriler genellikle sınırlı veya dengesiz sınıf yapısına sahip olduğundan, simülasyonlar belirli korelasyon, etki büyüklüğü veya örneklem senaryolarında ML yöntemlerinin doğruluk, F1 skoru veya ROC-AUC gibi metriklerini inceleme olanağı sağlar.

Ayrıca, yüksek boyutlu veri ortamlarında ($p > n$) değişken seçimi algoritmalarının hata oranları, yanlış keşif ve tahmin gücü bakımından karşılaştırılmasında simülasyonlar önemli rol oynar. Bu tür çalışmalar, özellikle genomik veya metabolomik veri analizlerinde metodolojik güvenilirliğin sağlanması açısından değerlidir.

Sağlık Politikası ve Ekonomik Değerlendirmeler

Sağlık politikalarının etkilerini doğrudan gerçek verilerle değerlendirmek her zaman mümkün değildir. Bu nedenle mikrosimülasyon ve sistem dinamiği modelleri sağlık ekonomisi çalışmalarında sıkça kullanılır (Karnon, 2003; Brennan et al., 2006).

Bu tür simülasyonlarda bireylerin yaşam boyu maliyet, tedavi, hastalık geçirme ve ölüm olasılıkları modellenir. Örneğin, bir tarama programının maliyet-etkililiği farklı katılım oranları ve hastalık insidansları altında simüle edilerek değerlendirilebilir.

Eğitimsel ve Metodolojik Çalışmalar

Simülasyon, aynı zamanda biyoistatistik eğitiminde de güçlü bir öğretim aracıdır. Öğrenciler, istatistiksel kavramların soyut doğasını somutlaştırmak için simülasyonları kullanabilir; örneğin,

merkezi limit teoreminin veya örneklem dağılımlarının davranışı binlerce tekrarlı simülasyonla gösterilebilir (Kaplan & Haile, 2019).

Metodolojik açıdan ise simülasyonlar, yeni istatistiksel yöntemlerin performansını değerlendirmede yaygın ve güçlü bir araçtır.

YAZILIM VE KODLAMA ALTYAPISI

Biyoistatistikte simülasyon çalışmaları, günümüzde yoğun biçimde bilgisayar destekli ortamlarda yürütülmektedir. Simülasyonun temelinde rastgele veri üretimi, tekrarlamalı hesaplama ve sonuçların özetlenmesi bulunduğundan, kullanılan yazılım altyapısının esnek, tekrarlanabilir ve paralel işlemeye uygun olması büyük önem taşır (Burton et al., 2006; Morris, White & Crowther, 2019).

Simülasyonun başarılı yürütülebilmesi için yazılımın şu özellikleri sağlaması gerekir:

- Rastgele sayı üreticilerinin güvenilirliği (deterministik fakat tekrarlanabilir)
- Yüksek performanslı döngü ve vektörleştirilmiş hesaplama
- Paralel işlem desteği (çok çekirdekli CPU veya bulut tabanlı işlem)
- Sonuçların kolayca özetlenip görselleştirilebilmesi
- Reprodüksiyon ve açık kod paylaşımı olanakları

Bu çerçevede, biyoistatistikte en yaygın kullanılan yazılımlardan ikisi R ve Python'dur.

R Ortamında Simülasyon Altyapısı

R programlama dili, istatistiksel hesaplama ve görselleştirme olanakları açısından en köklü altyapılardan biridir (R Core Team, 2024). Özellikle simülasyon temelli çalışmalar için geliştirilmiş çok sayıda paket içerir (Tablo 2).

R ortamı, özellikle akademik ve metodolojik araştırmalarda tercih edilir; çünkü istatistiksel modelleme, veri manipülasyonu ve sonuç görselleştirmeyi tek çatı altında birleştirir.

Ayrıca, R Markdown ve Quarto altyapıları sayesinde simülasyon kodu, sonuç tablosu ve açıklama metinleri tek bir yeniden üretilebilir rapor hâlinde sunulabilir. Bu tür raporlamalarda `set.seed()` fonksiyonunun kullanılması da önemlidir; çünkü bu fonksiyon, tüm rastgele sayı üretim süreçlerinin aynı koşullar altında tekrarlandığında aynı sonuçları vermesini sağlar ve böylece araştırmanın yeniden üretilebilirliği güvence altına alınır (Xie, Allaire & Golemund, 2022; Peng, 2011).

Tablo 2. R Ortamında Simülasyon Çalışmalarında Yaygın Kullanılan Paketler

Paket	Temel İşlev	Öne Çıkan Kullanım Alanı
simstudy	Parametrik veri üretimi ve senaryo tanımı	Klinik deneme, epidemiyolojik simülasyonlar
MonteCarlo	Çoklu senaryolu Monte Carlo deneylerinin yönetimi	Güç analizi, hata oranı çalışmaları
mvtnorm	Çok değişkenli normal dağılımlardan örnekleme	Korelasyon yapıları veri üretimi
miceadds	Eksik veri simülasyonları ve çoklu atama	MCAR–MAR senaryoları, yöntem karşılaştırmaları
doParallel, foreach	Paralel simülasyon ve dağıtık işlem	Büyük tekrarlı çalışmalar
ggplot2, reshape2	Sonuçların görsel özeti	Boxplot, ısı haritası, dağılım grafikleri

Python Ortamında Simülasyon Altyapısı

Python, veri bilimi ve makine öğrenmesi alanlarında yaygın olarak kullanılan açık kaynaklı bir programlama dilidir.

Son yıllarda NumPy, SciPy ve PyMC gibi kütüphaneler sayesinde biyoistatistiksel simülasyon çalışmalarında da güçlü bir seçenek hâline gelmiştir (Van Rossum & Drake, 2023). Python'un simülasyon çalışmalarında en sık kullanılan kütüphaneleri Tablo 3'te özetlenmiştir.

Python'un önemli avantajları arasında hesaplama verimliliği, geniş ekosistemi ve makine öğrenmesi kütüphaneleriyle kolay entegrasyonu yer alır.rneğin, bir biyoistatistiksel modelden elde edilen çıktılar doğrudan TensorFlow veya scikit-learn ortamlarında kullanılarak modelleme ve karşılaştırma süreçlerine aktarılabilir. Bu özellik, özellikle simülasyon tabanlı makine öğrenmesi uygulamalarında (ör. model performansı karşılaştırmaları, hiperparametre duyarlılığı analizleri) Python'u öne çıkarır (Boulesteix et al., 2020).

Tablo 3. Python Ortamında Simülasyon Çalışmalarında Yaygın Kullanılan Kütüphaneler

Paket	Temel İşlev	Öne Çıkan Kullanım Alanı
numpy.random	Rastgele sayı üretimi (normal, binom, Poisson vb.)	Veri üretimi ve temel simülasyonlar
scipy.stats	Olasılık dağılımları ve istatistiksel testler	Parametrik–parametrik olmayan testler
pandas	Veri çerçevesi yapısı, özet istatistikler	Senaryo çıktılarının saklanması
joblib, multiprocessing	Paralel hesaplama	Çok tekrarlı simülasyonlar
PyMC	Bayesyen MCMC simülasyonları	Posterior tahmin, hiyerarşik modeller
Matplotlib, Seaborn	Görselleştirme	Dağılım grafikleri, korelasyon matrisleri

Yazılım Seçimi ve Entegrasyon İlkeleri

Her iki yazılım da güçlüdür; ancak tercih genellikle araştırmacının niteliğine bağlıdır.

- R, istatistiksel modelleme ve metodolojik araştırmalar için uygundur.
- Python, büyük veri, derin öğrenme veya yüksek performanslı hesaplama gerektiren durumlarda avantaj sağlar.

Güncel eğilim, bu iki ortamın entegrasyonuna yönelmiştir. R kullanıcıları “reticulate” paketiyle Python kodlarını doğrudan R ortamında çalıştırabilir; benzer şekilde Python kullanıcıları “rpy2” modülüyle R fonksiyonlarına erişebilir. Bu sayede simülasyon sürecinde istatistiksel analiz, görselleştirme ve makine öğrenmesi modülleri tek bir çalışma akışında birleştirilebilir (Tippmann, 2015). Ayrıca, GitHub, Zenodo veya Open Science Framework (OSF) gibi açık erişim platformlarında simülasyon kodlarının paylaşılması, hem bilimsel şeffaflığı hem de metodolojik etkiyi artırır.

SİMÜLASYON TASARIMI VE RAPORLAMA İLKELERİ

Simülasyon çalışmaları, yalnızca veri üretimi ve analizinden ibaret değildir; aynı zamanda sistematik bir araştırma tasarımı sürecini de içerir. Simülasyonun amacı, kapsamı, veri üretim modeli ve performans ölçütleri net olarak tanımlanmadığında elde edilen sonuçlar metodolojik olarak güvenilirliğini yitirir (Morris, White & Crowther, 2019). Bu nedenle, simülasyon araştırmalarında tasarım, yürütme ve raporlama aşamaları belirli standartlara uygun biçimde planlanmalıdır.

Simülasyon Tasarımında Temel İlkeler

İyi tasarlanmış bir simülasyon, hem istatistiksel hem de sayısal açıdan tutarlı sonuçlar üretir. Burton ve ark. (2006), biyoistatistikte etkili bir simülasyonun beş ana tasarım bileşenini şu şekilde tanımlamıştır:

- Amaç: Simülasyonun hangi istatistiksel problemi test ettiği açıkça belirtilmelidir. Örneğin, “Benjamini–Hochberg ve Holm yöntemleri arasında FDR kontrol düzeyinde fark var mı?” gibi doğrudan ölçülebilir bir araştırma sorusu formüle edilmelidir.
- Veri üretim mekanizması: Değişkenlerin dağılımı, korelasyon yapısı ve örneklem büyüklükleri tanımlanmalı; kullanılan rastgelelik kaynakları ve olasılık varsayımları raporlanmalıdır.
- Analiz yöntemi: Her simülasyon döngüsünde uygulanan istatistiksel test veya model açıkça belirtilmelidir.
- Performans ölçütleri: Güç, Tip I hata, sapma, RMSE, FDR veya kapsama oranı gibi ölçütlerin hesaplanma biçimi açıkça tanımlanmalıdır.
- Tekrarlama yapısı: Simülasyonun kaç kez tekrarlanacağı ve rastgele seed değerinin nasıl kontrol edildiği belirtilmelidir.

Bu unsurların açık biçimde tanımlanması hem sonuçların karşılaştırılabilirliğini hem de metodolojik geçerliliği artırır.

Simülasyon Çalışmalarında Raporlama Standartları

Simülasyon sonuçlarının bilimsel yayınlarda yer alabilmesi için şeffaf ve yeniden üretilebilir biçimde raporlanması gerekir. Morris ve ark. (2019) ile White ve ark. (2011) tarafından önerilen

raporlama çerçevesine göre bir simülasyon çalışması şu başlıkları içermelidir:

- Giriş: Simülasyonun amacı, ilgili literatürdeki boşluk ve araştırma gerekçesi.
- Yöntem: Veri üretim modeli (ör. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$).
 - Uygulanan istatistiksel yöntem(ler) ve parametre değerleri.
 - Simülasyon sayısı, örneklem büyüklüğü ve dağılım varsayımları.
 - Kullanılan yazılım, paket ve sürümler.
- Sonuçlar:
 - Performans ölçütlerinin tablo veya grafiklerle özeti.
 - Senaryolar arasındaki farkların yorumu.
 - Yöntemlerin göreceli üstünlükleri ve sınırlılıkları.
- Tartışma:
 - Bulguların genellenebilirliği, model varsayımlarının kısıtları ve olası yanlılık kaynakları.
 - Gelecekteki çalışmalar için öneriler.

Bu çerçeve, simülasyon çalışmalarında şeffaflık, karşılaştırılabilirlik ve yeniden üretilebilirlik ilkelerinin korunmasını sağlar (Peng, 2011).

Yeniden Üretilbilirlik ve Şeffaflık

Simülasyonların doğası gereği rastgelelik içerdiğinden, her çalışmada kullanılan seed değerleri, paket sürümleri ve kod akışı mutlaka raporlanmalıdır.

Araştırmanın yeniden üretilebilirliğini güvence altına almak için aşağıdaki uygulamalar önerilmektedir (Sandve et al., 2013; Peng, 2011):

- Kodların GitHub, Zenodo veya Open Science Framework (OSF) gibi açık erişim platformlarında paylaşılması,
- Kullanılan yazılımların sürümlerinin ve işletim ortamının (ör. R 4.5.2, Python 3.13) belirtilmesi,
- Tüm analizlerin R Markdown, Quarto veya Jupyter Notebook biçiminde tek bir belge altında sunulması,
- Simülasyonların tamamının belirli bir random seed ile çalıştırılması,
- Sonuçların elde edilmesinde kullanılan tüm fonksiyonların ve parametrelerin ek materyal olarak sunulması.

Bu uygulamalar hem araştırma şeffaflığını hem de metodolojik güvenilirliği artırır. Özellikle metodolojik biyoistatistik çalışmalarında, kod paylaşımı ve yeniden üretilebilirlik raporlaması editoryal bir gereklilik hâline gelmiştir (Stodden et al., 2016).

Görselleştirme ve Bulguların Sunumu

Simülasyon sonuçlarının yalnızca tablo biçiminde değil, aynı zamanda görsel olarak da sunulması önerilir. Boxplot'lar, güç eğrileri, FDR-FWER ısı haritaları ve hata-varyans dağılımları yöntemlerin karşılaştırılmasında oldukça etkilidir (Burton et al., 2006).

Ayrıca, sonuç grafiklerinde kullanılan renk paletleri ve eksen aralıkları standartlaştırılmalı; görseller, senaryolar arası farkları sezgisel biçimde yansıtmalıdır.

Bu tür grafikler, hem istatistiksel yorumlamayı kolaylaştırır hem de okuyucunun farklı yöntemlerin performansını hızla karşılaştırmasına olanak tanır.

Etik İlkeler ve Yayın Standartları

Simülasyon çalışmaları doğrudan insan verisi içermese de, bilimsel dürüstlük ve şeffaflık açısından etik ilkelere tabidir.

- Kullanılan veri setleri (örneğin açık erişimli klinik veriler) etik izin gerektiriyorsa kaynak belirtilmelidir.
- Simülasyon sonuçları, gerçek veriyle birebir örtüşecek şekilde manipüle edilmemelidir.
- Model sınırlılıkları ve potansiyel yanlılıklar dürüst biçimde raporlanmalıdır.

Bu ilkeler, biyoistatistikte metodolojik araştırmaların güvenilirliğini ve akademik bütünlüğünü korur (Stodden et al., 2016).

Sonuç

Simülasyon tasarımı ve raporlama ilkeleri, biyoistatistik araştırmalarının metodolojik kalitesini belirleyen kritik unsurlardır.

Açıkça tanımlanmış senaryolar, şeffaf veri üretim süreçleri ve yeniden üretilebilir raporlama yaklaşımları, yalnızca bir simülasyonun doğruluğunu değil, aynı zamanda bilimsel topluluğa olan katkı değerini de artırır.

Günümüzde bu ilkelerin sistematik biçimde uygulanması, metodolojik dergilerde yayın olasılığını artırmakta ve açık bilim

standartlarıyla uyumlu bir araştırma kültürünün gelişimini desteklemektedir.

GÜNCEL YÖNTEMLER VE ARAŞTIRMA EĞİLİMLERİ

Biyoistatistikte simülasyon yöntemleri, klasik Monte Carlo ve bootstrap yaklaşımlarının ötesine geçerek giderek daha karmaşık modelleme teknikleriyle bütünleşmektedir. Bilgisayar gücündeki artış, açık kaynak yazılımların gelişimi ve yüksek boyutlu verinin yaygınlaşması, simülasyon tabanlı yöntemleri hem teorik hem de uygulamalı araştırmalarda vazgeçilmez hâle getirmiştir (Morris, White & Crowther, 2019). Güncel eğilimler, yalnızca doğruluk ölçümüyle sınırlı olmayan, aynı zamanda sağlamlık (robustluk), nedensellik, yapay zekâ destekli simülasyon ve yüksek boyutlu veri modelleme alanlarında yeni olanaklar sunmaktadır.

Robust ve Kantil Tabanlı Simülasyon Yaklaşımları

Klasik istatistiksel modeller genellikle normal dağılım varsayımına dayanır; ancak sağlık verilerinde sıklıkla ağır kuyruklu, asimetrik veya uç değer içeren dağılımlar gözlenir. Bu durum, geleneksel yöntemlerin hatalı sonuçlar üretmesine yol açabilir.

Son yıllarda, bu tür durumlara dayanıklı robust ve kantil tabanlı simülasyonlar geliştirilmiştir (Koenker, 2022). Bu yaklaşımlar, sadece ortalama etkiyi değil, aynı zamanda dağılımın farklı kantillerinde (örneğin 0.25, 0.5, 0.75) parametre davranışını analiz etmeyi mümkün kılar.

Örneğin, klinik çalışmalarda tedavi etkisinin medyan veya üst kantillerde daha belirgin olup olmadığı, kantil regresyon tabanlı simülasyonlarla değerlendirilebilir. Bu yöntemler, özellikle heterojen popülasyon yapısına sahip epidemiyolojik veya pediatrik verilerde kullanıldığında yüksek duyarlılık sağlar.

Bayesyen ve Yarı-Parametrik Simülasyon Yaklaşımları

Bayesyen yöntemlerin popülaritesi, bilgisayar tabanlı MCMC algoritmalarının erişilebilir hâle gelmesiyle önemli ölçüde artmıştır (Gelman et al., 2014). Güncel eğilimler, klasik Monte Carlo simülasyonlarının ötesine geçerek Bayesyen Monte Carlo (BMC) ve Targeted Maximum Likelihood Estimation (TMLE) gibi yarı-parametrik modellerin simülasyon tabanlı değerlendirilmesine odaklanmaktadır (van der Laan & Rose, 2018).

Bu tür modellerde, önsel dağılımlar belirsiz veya zayıf bilgiye dayalı olsa bile, simülasyon yoluyla posterior dağılımın tahmin edilmesi sağlanır.

Örneğin, bir tedavinin klinik etkinliği hakkındaki ön bilgilerin posterior dağılım üzerindeki etkisi, farklı prior senaryolarının simüle edilmesiyle test edilebilir.

Bu yaklaşım, özellikle adaptif klinik denemelerde karar verme süreçlerinin optimizasyonunda kullanılmaktadır.

Yapay Zekâ Destekli Simülasyonlar

YZ ve ML yöntemlerinin biyoistatistikle bütünleşmesi, simülasyon temelli araştırmalarda yeni bir paradigma yaratmıştır (Boulesteix et al., 2020).

YZ destekli simülasyonlarda amaç, sadece önceden belirlenmiş parametrelerle veri üretmek değil, aynı zamanda verinin yapısını öğrenebilen ve gerçek veriye benzer sentetik veri setleri üretebilen modeller geliştirmektir.

Bu alandaki en yeni yöntemlerden biri Generative Adversarial Networks (GANs) temelli simülasyonlardır. GAN'ler, özellikle nadir olayların modellenmesinde (ör. nadir görülen hastalıklar, komplikasyonlar) kullanılarak küçük örneklemelerden gerçekçi veri üretmeyi sağlar (Xu et al., 2019).

Ayrıca, simülasyon tabanlı model değerlendirme yaklaşımıyla YZ modellerinin yanlılık ve genellenebilirlik performansları test edilebilmektedir (Cranmer et al., 2020).

Yüksek Boyutlu Veri Simülasyonları

Modern biyomedikal araştırmalarda veriler genellikle çok boyutludur; örneğin gen ekspresyonu, epigenetik, proteomik veya metabolomik veri kümelerinde değişken sayısı (p) gözlem sayısından (n) fazladır. Bu durumda klasik yöntemler performans kaybına uğradığından, yüksek boyutlu simülasyonlar özel önem taşır.

Güncel araştırmalar, Lasso, Elastic Net, Random Forest ve XGBoost gibi regularization ve ensemble tabanlı yöntemlerin yüksek boyutlu senaryolardaki yanlılık, varyans ve FDR kontrol performansını simülasyonlarla değerlendirmektedir (Hastie, Tibshirani & Friedman, 2021).

Örneğin, $p=5000$ değişken ve $n=300$ gözlem içeren bir genomik veri simülasyonunda, değişken seçimi algoritmalarının duyarlılık ve özgüllük düzeyleri farklı korelasyon yapıları altında incelenebilir.

Bu alanda kullanılan bir diğer yöntem, veri bağımlı (data-driven) simülasyon yaklaşımıdır. Burada gerçek veri setinden elde edilen özet istatistikler (ör. korelasyon matrisi, varyans yapısı) kullanılarak sentetik veri üretilir; böylece simülasyon sonuçları daha gerçekçi olur (Leek & Peng, 2015).

Nedensel Çıkarım ve Simülasyon Tabanlı Değerlendirme

Son yıllarda nedensellik analizleri, gözlemsel verilerden güvenilir sonuçlar elde etmede kritik bir yöntem hâline gelmiştir. Ancak bu modellerin performansını analitik olarak test etmek genellikle zordur.

Bu nedenle, nedensel simülasyon çalışmaları, farklı karıştırıcı (confounder) yapıları veya örneklem dengesizlikleri altında yöntemin doğru çıkarım yapma olasılığını ölçmek için kullanılmaktadır (Hernán & Robins, 2020).

Örneğin, Propensity Score Matching (PSM), Inverse Probability Weighting (IPW) veya Doubly Robust Estimation (DRE) gibi nedensel yöntemler, farklı karıştırıcı yoğunlukları altında simüle edilerek karşılaştırılabilir. Bu tür simülasyonlar, gerçek dünyadaki klinik veya epidemiyolojik verilerin hangi koşullarda güvenilir sonuçlar verebileceğine dair önemli çıkarımlar sağlar.

Geleceğe Yönelik Araştırma Eğilimleri

Biyoistatistikte simülasyon tabanlı araştırmaların önümüzdeki yıllarda aşağıdaki yönlerde evrilmesi beklenmektedir:

- Açık bilim ilkeleri doğrultusunda yeniden üretilebilir simülasyon çerçeveleri,
- Bulut tabanlı paralel hesaplama (ör. Google Colab, AWS, HPC kümeleri) ile yüksek hacimli senaryoların verimli çalıştırılması,
- Veri gizliliğini koruyan simülasyon teknikleri (ör. farklılaştırılmış gizlilik – differential privacy),
- Karma (hibrit) yaklaşımlar: Gerçek ve sentetik verinin birleştirildiği yarı-empirik simülasyonlar,
- Eğitim amaçlı interaktif simülasyon platformları, özellikle tıp ve sağlık bilimleri lisansüstü eğitiminde kullanılabilecek dinamik ortamlar.

Bu eğilimler, biyoistatistikte simülasyonun hem metodolojik hem de pedagojik işlevinin önümüzdeki dönemde daha da güçleneceğine işaret etmektedir.

Kaynakça

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279-4292. <https://doi.org/10.1002/sim.2673>

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074-2102.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). McGraw-Hill Education.

Rubinstein, R. Y., & Kroese, D. P. (2017). *Simulation and the Monte Carlo method* (3rd ed.). John Wiley & Sons.

Karnon, J. (2003). Alternative decision modelling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. *Health Economics*, 12(10), 837-848.

Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modeling and simulation. *Journal of Simulation*, 4(3), 151-162.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226-1227. [.org/10.1126/science.1213847](https://doi.org/10.1126/science.1213847)

White, I. R., Morris, T. P., Crowther, M. J., & Carpenter, J. R. (2011). Simulations for statistical methods: Design, implementation and reporting. *Statistical Methods in Medical Research*, 20(1), 3-22.

- Boulesteix, A. L., Lauer, S., Eugster, M. J. A., & Binder, H. (2020). Simulation studies in machine learning: The case of random forests for high-dimensional data. *Computational Statistics & Data Analysis*, 147, 106941. <https://doi.org/10.1016/j.csda.2020.106941>
- Brennan, A., Chick, S. E., & Davies, R. (2006). A taxonomy of model structures for economic evaluation of health technologies. *Health Economics*, 15(12), 1295–1310.
- Chen, X., Li, Y., & Bai, X. (2017). Multiple testing in omics research: Challenges and solutions. *Frontiers in Genetics*, 8, 59. <https://doi.org/10.3389/fgene.2017.00059>
- Kaplan, D., & Haile, M. K. (2019). Teaching statistical concepts through Monte Carlo simulation. *Journal of Statistics Education*, 27(3), 179–188.
- Lee, J. J., & Tsiatis, A. A. (2020). Adaptive clinical trial designs and their implementation. *Biostatistics*, 21(2), 339–351. <https://doi.org/10.1093/biostatistics/kxy058>
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.
- Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer.
- Vallat, R., Walker, M. P., & Aeschbach, D. (2021). Simulation-based sample size estimation for clinical studies. *PLOS ONE*, 16(8), e0255620.
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. [.R-project.org](https://www.R-project.org)

Tippmann, S. (2015). Programming tools: Adventures with R and Python. *Nature*, 517(7532), 109–110. <https://doi.org/10.1038/517109a>

Van Rossum, G., & Drake, F. L. (2023). The Python language reference manual (3rd ed.). Python Software Foundation.

Xie, Y., Allaire, J. J., & Golemund, G. (2022). R Markdown: The definitive guide (2nd ed.). Chapman & Hall/CRC.

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology*, 9(10), e1003285.

Stodden, V., Seiler, J., & Ma, Z. (2016). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11), 2584–2589.

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48), 30055–30062.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. Chapman & Hall/CRC.

Koenker, R. (2022). Quantile regression (2nd ed.). Cambridge University Press.

Leek, J. T., & Peng, R. D. (2015). Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6), 1645–1646.

van der Laan, M. J., & Rose, S. (2018). Targeted learning in data science: Causal inference for complex longitudinal data. Springer.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32, 7333–7343.

GEÇİCİ KAPAK

*Kapak tasarımı
devam ediyor.*