İstatistik:
Teori, Metodoloji ve Uygulama

Editör
Dr. Öğr. Üyesi Bayezid Gülcan

DOI: 10.5281/zenodo.17415782

İstatistik:

Teori, Metodoloji ve Uygulama

Editör

Dr. Öğr. Üyesi Bayezid Gülcan

İmtiyaz Sahibi
Platanus Publishing®

Editör
Dr. Öğr. Üyesi Bayezid Gülcan
Kapak & Mizanpaj & Sosyal Medya
Platanus Yayın Grubu

Birinci Basım
Ekim, 2025

Yayımcı Sertifika No
45813

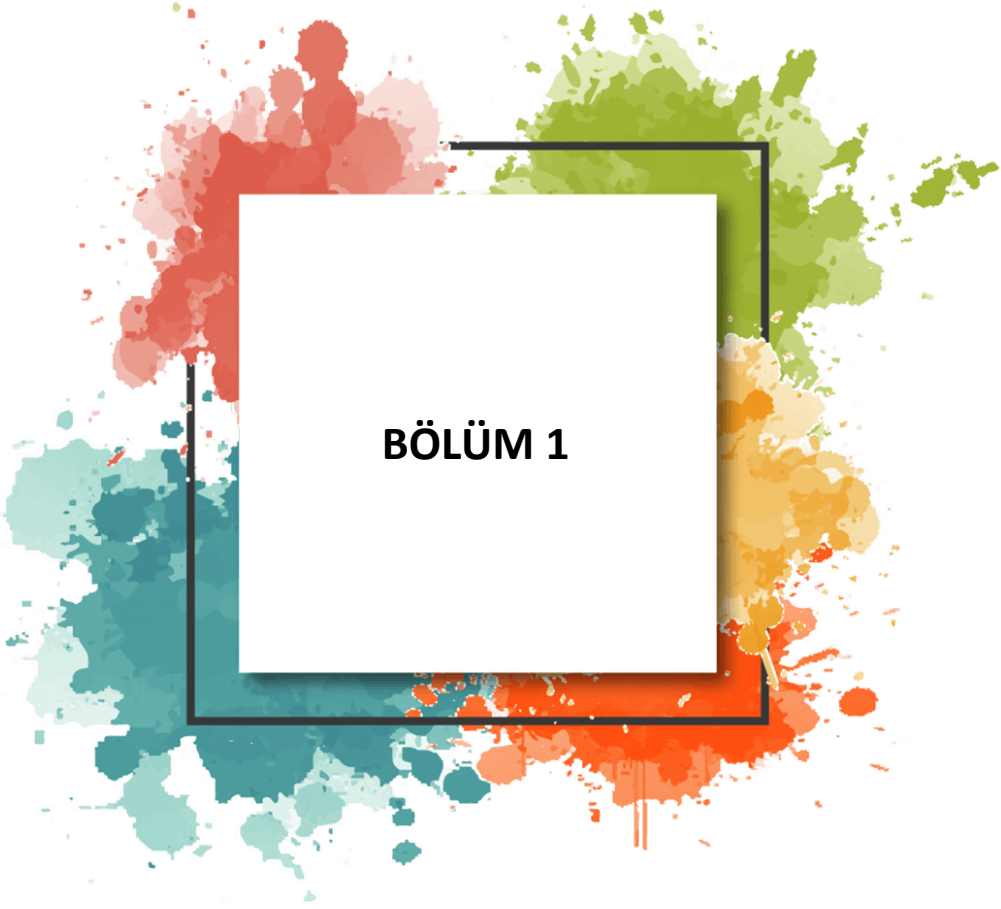
ISBN
978-625-7374-99-6

Adres: Natoyolu Cad. Fahri Korutürk Mah. 157/B, 06480, Mamak, Ankara, Türkiye.

Telefon: +90 312 390 1 118
web: www.platanuspublishing.com
e-mail: platanuskitap@gmail.com

Platanus Publishing®

İÇİNDEKİLER


BÖLÜM 1

Sayma Verileri İçin Simülasyon ve Performans Değerlendirmesi

Barış Ergül[1] & Arzu Altın Yavuz[2]

GİRİŞ

Sosyal bilimlerden sağlığa, ekonomiden mühendisliğe kadar pek çok alanda karşımıza çıkan veri türlerinden biri de sayılabilir verilerdir (count data). Bu veriler, bir olayın ya da davranışın belirli bir zaman dilimi veya mekanda kaç kez gerçekleştiğini göstermektedir ve yalnızca sıfır ya da pozitif tam sayılarla ifade edilen kesikli (discrete) verilerdir. Bu tür verilerin analizi, sürekli değişkenler için geliştirilmiş klasik regresyon modellerinin varsayımlarına uymadığı için özel istatistiksel yöntemler gerektirmektedir. Bu ihtiyaca karşılık olarak Poisson regresyonu, negatif binom regresyonu ve sıfır-şişkin modeller gibi sayılabilir veriye özel modeller geliştirilmiştir ve zamanla farklı alanlarda yaygın biçimde kullanılmaya başlanmıştır. Sayılabilir veri modelleri, olayların ne sıklıkla meydana geldiğini anlamaya yönelik kuramsal temele dayanmaktadır. Temel varsayım, olayların birbirinden bağımsız olarak ve sabit bir ortalama hızla gerçekleştiğidir. Bu amaçla, ilk başvurulan model Poisson regresyon modelidir. Poisson regresyonu, olay sayılarının poisson dağılımına uyduğunu ve ortalama ile varyansın birbirine eşit olduğunu kabul etmektedir. Model, bir logaritmik bağlantı kullanmaktadır ve bu bağlantı fonksiyonu sayesinde, model tarafından tahmin edilen olay sayısı, her zaman pozitif kalmaktadır. Ancak bu modelin temel bir sınırlaması vardır; varyansın sabit, yani ortalama ile eşit olduğu varsayımı çoğu gerçek veri setinde karşılanmaz. Gerçek hayattaki veriler genellikle beklenenden daha fazla değişkenlik gösterir. Bu duruma aşırı saçılma (overdispersion) denir. Aşırı saçılma, Poisson modelinin standart hataları olduğundan küçük tahmin etmesine ve bu nedenle sonuçların istatistiksel anlamlılık açısından yanıltıcı olmasına yol açmaktadır (Cameron & Trivedi, 2013).

Poisson modelinin dezavantajlarını aşmak için geliştirilen negatif binom regresyon modeli, bu yaklaşımın daha esnek bir versiyonudur. Bu model, varyansın ortalamadan büyük olmasına izin vererek Poisson modeline göre daha gerçekçi sonuçlar sunmaktadır. Negatif binom modelinde yer alan saçılma parametresi (α), verideki aşırı değişkenliği dengeleyerek modelin daha doğru

[1] Dr. Öğr. Üy., Eskişehir Osmangazi Üniversitesi Fen Fakültesi, İstatistik Bölümü, 0000-0002-1811-5143
[2] Prof. Dr., Eskişehir Osmangazi Üniversitesi Fen Fakültesi, İstatistik Bölümü, 0000-0002-3277-740X

tahminler yapmasını sağlamaktadır (Hilbe, 2011). Bununla birlikte, bazı veri setlerinde gözlemlerin çok büyük bir kısmı sıfır değerinden oluşur. Bu gibi durumlarda, klasik Poisson ya da negatif binom modelleri, sıfırların bu kadar yoğun oluşunu açıklamakta yetersiz kalmaktadır. Bu tür verilere sıfır-ağırlıklı ya da sıfır-şişkin (zero-inflated) veriler denir. Sıfır-şişkin Poisson (ZIP) ve sıfır-şişkin negatif binom (ZINB) modelleri, bu tür veri yapılarını analiz edebilmek için geliştirilmiştir. Bu modeller, verideki sıfırları iki ayrı sürece ayırarak açıklar; İlki, olayları yapısal olarak gerçekleşmesinin mümkün olmadığı durumlar olarak ele almaktadır. Diğeri, olayın gerçekleşmemiş olsa da gerçekleşme ihtimalinin bulunduğu, yani rastlantısal olarak sıfır kalan durumlardır. Bu modeller, ikili bir seçim modeli (genellikle logit ya da probit) ile sayım modelini birleştirerek daha esnek ve açıklayıcı analizler yapılmasına olanak tanımaktadırlar (Lambert, 1992).

Sayılabilir veri modellerinin çok disiplinli doğası, yalnızca istatistik ya da ekonometri gibi teknik alanlarda değil, aynı zamanda kamu sağlığı, eğitim, ulaşım ve enerji politikaları gibi uygulamalı alanlarda da etkili olmasına olanak tanımaktadır. Bu yönüyle, model seçimi yapılırken hem kullanılan verinin yapısal özellikleri hem de analizin hizmet ettiği kuramsal çerçeve göz önünde bulundurulmalıdır. Ek olarak, parametrik modellere alternatif olarak geliştirilen yarı-parametrik ve parametrik olmayan yöntemler de sayılabilir veri analizinde yeni ufuklar açmaktadır. Bu modeller, esnek yapıları sayesinde veri dağılımındaki varsayım ihlallerine karşı daha dayanıklı sonuçlar sunabilmektedir. Özellikle küçük örneklem büyüklüğü veya düzensiz dağılıma sahip veri setlerinde bu tür yöntemlerin kullanımı ön plana çıkmaktadır. Sayılabilir veri modelleri; hem teorik açıdan sağlam bir temel sunmakta, hem de pratik uygulamalarda esneklik sağlayarak araştırmacılara önemli analiz araçları kazandırmaktadır. Ancak modelleme sürecinde sadece teknik uygunluk değil, aynı zamanda, bütünlük ve yorumlanabilirlik de göz önünde bulundurulmalıdır. Bu doğrultuda, yeni yöntemlerin geliştirilmesi ve mevcut modellerin farklı alanlara uyarlanması, sayım verilerine dayalı araştırmaların bilimsel değerini arttıracağı da göz önüne alınmalıdır.

Son yıllarda, makine öğrenmesi ve büyük veri analitiği gibi gelişmelerin etkisiyle, sayılabilir veri modellerine yönelik ilgi giderek artmaktadır. Poisson ve negatif binom modelleri, artık yalnızca klasik istatistiksel çerçevede değil, aynı zamanda modern hesaplamalı yöntemlerle bütünleşik olarak değerlendirilmektedir. Bu sayede hem model performansı artırılmakta hem de karmaşık veri yapıları daha etkin şekilde analiz edilebilmektedir.

Bu çalışmada, 2 farklı sayma verisi içeren (sıfır şişkinlik ve aykırı değer) 5 sayma modeli kullanılarak simülasyon yapılmış, hangi modelin seçileceği konusunda çeşitli çıkarsamalarda bulunulmuştur.

2.MATERYAL VE METOT

Poisson regresyon modeli, belirli bir zaman ya da mekân içinde olayların ne sıklıkla meydana geldiğini tahmin etmek için kullanılan bir istatistiksel yaklaşımdır. Özellikle sayım verileri üzerine geliştirilmiş olan bu yöntem, bağımlı değişkenin yalnızca sıfır ve pozitif tam sayılar aldığı durumlarda uygundur. Sayım verileri kesikli olduğu için, bu tür verilerin analizinde geleneksel doğrusal regresyon modelleri uygun değildir; çünkü bu modeller sürekli verilerle çalışmak üzere tasarlanmıştır (Cameron & Trivedi, 2013).

Modelin temel varsayımı, olayların sabit bir oranla ve birbirinden bağımsız biçimde gerçekleştiğidir. Poisson modelini diğerlerinden ayıran en önemli özellik ise, olayların ortalama sayısı ile varyansının eşit olmasıdır. Bu özellik eş-dağılım (equidispersion) olarak adlandırılır. Poisson dağılımı için olasılık yoğunluk fonksiyonu (1) eşitliğindeki gibi tanımlanır:

$$P(Y = y) = \frac{\lambda^y e^{-y}}{y!} \qquad (1)$$

Burada: Y; gözlemlenen sayım değişkenini, λ; olayların beklenen ortalamasını (ve aynı zamanda varyansı), y; 0, 1, 2, … gibi tam sayıları temsil eder.

Poisson regresyon modelinde, λ (beklenen olay sayısı), açıklayıcı değişkenlerle logaritmik bir bağlantı fonksiyonu üzerinden (2) no'lu eşitlikteki gibi ilişkilendirilir:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \qquad (2)$$

Buradan $\lambda_i = \exp(X_i\beta)$ formülü elde edilir. Bu matematiksel yapı sayesinde λ daima pozitif kalmaktadır ve bu sayede, modelin tahminleri anlamlı hale gelmektedir. Modeldeki katsayılar, açıklayıcı değişkenlerdeki bir birimlik artışın beklenen olay sayısını (λ) logaritmik olarak nasıl etkilediğini açıklamaktadır. Poisson regresyon modelleri için parametre tahminleri, genel olarak Maksimum Olabilirlik Tahmini (MLE) yöntemiyle çözülmektedir.

Poisson regresyon modelleri epidemiyoloji, ulaşım, sigortacılık, çevre bilimleri gibi çok çeşitli alanlarda kullanılır:

Gerçek hayattaki verilerde, genellikle varyans değeri ortalamadan daha büyük çıkmaktadır. Bu durum aşırı saçılma olarak bilinmektedir ve Poisson regresyon modelinin en büyük dezavantajlarından biridir. Aşırı saçılma durumu olduğunda, Poisson modelinden elde edilen tahmin değerleri tutarsız olmaktadır ve güvenilirliklerini yitirmektedirler. Bu problemin yaygın nedenleri arasında şunlar sayılabilmektedir; gözlemler arasındaki ilişki (bağımlılık), modelde yer almayan önemli değişkenlerin eksikliği ve verinin kendi içindeki çeşitlilik (heterojenlik). Böyle durumlarda, daha esnek ve uyarlanabilir bir alternatif olan Negatif Binom

regresyon modeli tercih edilmelidir (Hilbe, 2011). Ayrıca, Poisson modeli, sıfır sayısının orantısız şekilde fazla olduğu veri setlerinde de yetersiz kalabilmektedir. Bu tür veri yapılarında, Zero-Inflated (Sıfır-Enflasyonlu) ya da Hurdle (Eşik) modelleri daha etkili sonuçlar sunmaktadır (Lambert, 1992).

Poisson regresyonunun bazı önemli avantajları şunlardır; modelin yapısı basittir, parametreler açık bir şekilde yorumlanabilir ve veriler modele iyi uyum sağladığında yüksek doğruluk sağlamaktadır. Öte yandan, modelin dezavantajları da vardır; eş-dağılım varsayımı çoğu zaman gerçek verilere uymamaktadır, sıfır sayısı fazla olan veri setlerinde performans düşmektedir ve varyans için değişiklikler modelin doğruluğunu olumsuz etkilemektedir.

Negatif Binom (NB) regresyon modeli, Poisson regresyonunun bazı temel eksiklerini gidermek amacıyla geliştirilmiştir. Özellikle varyansın ortalamadan daha büyük olduğu, yani; aşırı saçılmanın (overdispersion) söz konusu olduğu durumlarda tercih edilen bu model, sayım verileri üzerinde daha esnek ve güvenilir sonuçlar sağlamaktadır (Hilbe, 2011; Cameron & Trivedi, 2013). Poisson modelinde, varyansın ortalamaya eşit olduğu varsayımı çoğu gerçek veri setinde geçerli olmadığından, NB modeli bu dezavantajı ortadan kaldırarak daha gerçekçi bir yaklaşım sunmaktadır. Temel olarak, Poisson modeline benzer şekilde çalışmaktadır; olayların belirli bir zaman dilimi ya da mekanda ne sıklıkla gerçekleştiğini tahmin etmeye yöneliktir. Ancak önemli fark, varyansın ortalamadan büyük olabileceğini kabul etmesidir. Bu yönüyle model, daha esnektir. NB modelinin varsayımları arasında şunlar bulunmaktadır; gözlemler birbirinden bağımsızdır, veriler sıfır ve pozitif tam sayılardan oluşmaktadır ve varyans değeri ortalamadan büyüktür (overdispersion).

Teknik açıdan bakıldığında, NB modeli genellikle gamma karışımı Poisson dağılımı olarak da tanımlanmaktadır. Bu yaklaşım, gözlemler arasında doğrudan ölçülemeyen heterojenliklerin varlığını kabul etmektedir. Bu rastlantısal farklılıklar, Poisson oranı (λ) üzerinde bir çeşit rastgelelik yaratır ve bu da dağılımın NB formuna dönüşmesine neden olmaktadır (Zeileis vd., 2008).

NB regresyonunun genel matematiksel yapısı (3) no'lu eşitlikteki gibidir:

$$Y_i \sim NB(\mu, \alpha) \qquad (3)$$

Burada: μ, beklenen olay sayısını (yani ortalamayı) temsil etmektedir. α,varyans için aşırı saçılmayı açıklayan saçılma parametresidir. Varyans, (4) no'lu eşitlikteki gibi ifade edilir:

$$\mathrm{Var}(Y_i) = \mu_i + \alpha\mu_i^2 \qquad (4)$$

NB modelinde logaritmik bağlantı fonksiyonu, Poisson modeline benzer şekildedir ve (5) no'lu eşitlikte ifade edildiği gibidir:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \qquad (5)$$

Bu bağlantı fonksiyonu sayesinde μ_i her zaman pozitif kalmaktadır.

Eğer α değeri sıfıra eşitse, NB modeli Poisson modeline dönüşmektedir. Bu nedenle; NB modeli, Poisson modelinin daha genel bir versiyonu olarak değerlendirilebilmektedir. Bu ilişki sayesinde iki model karşılaştırılabilir ve veri setinde aşırı saçılma olup olmadığı istatistiksel testlerle belirlenebilmektedir. Aşırı saçılma olduğunda, standart hata tahminleri gerçekte olduğundan daha düşük çıkabilmektedir ve bu da yanlış anlamlılık yorumlarına yol açmaktadır. NB modeli, varyansın artmasına izin vererek bu sorunu doğrudan çözmektedir.

NB modeli, özellikle aşırı saçılma içeren ve sayım verilerinin sık kullanıldığı alanlarda yaygın şekilde tercih edilmektedir. Bunlar arasında sağlık araştırmaları, suç analizi (kriminoloji), ekonomi ve sigortacılık gibi uygulama alanları sayılabilir.

Model parametreleri genellikle Maksimum Olabilirlik Tahmini (MLE) yöntemiyle hesaplanmaktadır. Katsayılar, Poisson modelinde olduğu gibi logaritmik olarak yorumlanmaktadır. Daha açık ifadeyle, bağımsız değişkenlerden biri bir birim arttığında, olay oranı e^{β} kadar değişmesi olarak yorumlanmaktadır.

NB modelinin başlıca avantajları arasında şunlar yer almaktadır; aşırı saçılmayı doğrudan modelleme yeteneği, birçok gerçek veri setinde daha doğru sonuçlar sunması ve Poisson modelinin özel bir durumu olarak esnek bir yapıya sahip olmasıdır. Ancak modelin bazı dezavantajları da vardır; çok sayıda sıfır içeren veri setlerinde yetersiz kalabilmesi; verideki heterojenlik kaynakları net bir şekilde açıklanmadığında model çıktıları yorumlanması güç hale gelebilmesidir.

Hurdle regresyon modeli, sıfır değerlerin oldukça yaygın olduğu sayım verilerinin analizinde kullanılan iki aşamalı bir istatistiksel yöntemdir. Geleneksel Poisson veya Negatif Binom modelleri bu tür verilerde genellikle varsayımları ihlal ederken, Hurdle modeli hem aşırı sıfırları hem de aşırı saçılmayı başarılı bir şekilde ele alabilmektedir (Mullahy, 1986; Cameron & Trivedi, 2013).

Bu model, verinin iki farklı sürece göre üretildiği varsayımına dayanmaktadır;

Birinci Aşama (Hurdle kısmı): Bir gözlemin sıfır mı yoksa sıfırdan farklı mı olduğu belirlenmektedir. Bu aşamada, logit veya probit gibi ikili (binary) modeller tercih edilmektedir.

İkinci Aşama (Truncated Count kısmı): Sıfır olmayan değerlerin sayım modeli ile analiz edildiği bölümdür. Bu aşamada, yalnızca sıfır dışı değerleri içeren

(0'dan itibaren kesilmiş) Poisson ya da Negatif Binom modelleri kullanılmaktadır.

Modelin toplam olasılık yapısı, (6) eşitliğindeki gibi özetlenmektedir;

$$P(Y_i = 0) = 1 - \pi_i$$

$$P(Y_i = y_i) = \frac{\pi_i f(y_i)}{1 - f(0)}, y_i > 0 \qquad (6)$$

Burada; π_i, bir gözlemin sıfır dışı gruba ait olma olasılığıdır (birinci aşamada hesaplanır). $f(y_i)$, ikinci aşamada kullanılan sayım modelinden elde edilen olasılığı, f(0) ise, bu modelde sıfır değerinin olasılığını temsil etmektedir.

Hurdle modelinde, her iki aşamanın parametreleri birbirinden bağımsız olarak tahmin edilmektedir. İlk aşamada (ikili modelde) π_i değeri Maksimum Olabilirlik Tahmini (MLE) yöntemiyle hesaplanırken, ikinci aşamada seçilen dağılıma bağlı olarak sıfır dışı gözlemler için de ayrı MLE hesaplamaları yapılmaktadır. Bu noktada, aynı bağımsız değişkenlerin iki aşamada farklı etkiler gösterebileceği unutulmamalıdır. Modelin sunduğu başlıca avantajlar arasında, aşırı sıfır değerleri içeren verilerde esnek ve etkili çözüm sunması, her aşamada farklı açıklayıcı değişkenlerin kullanılabilmesi gibi özellikler yer almaktadır. Ancak bazı dezavantajları da vardır; iki ayrı modelin birlikte kullanılması tahmin sürecini karmaşık hale gelmesi; yorumlama güçlükleridir. Sağlık, sigorta ve ulaşım gibi alanlarda yaygın bir şekilde uygulanmaktadır.

Zero-Inflated Regresyon (ZIR) modelleri, özellikle sayım verilerinde çok sayıda sıfır değerin yer aldığı durumlarda kullanılan güçlü istatistiksel yöntemlerdir. Klasik Poisson veya Negatif Binom modelleri, bu tür aşırı sıfır içeren verileri yeterince açıklayamadığında, Zero-Inflated Poisson (ZIP) ve Zero-Inflated Negative Binomial (ZINB) gibi modeller devreye girmektedir (Lambert, 1992; Cameron & Trivedi, 2013).

Bu modellerin temel mantığı, verilerin iki farklı süreçten üretildiği varsayımına dayanmaktadır. İlki, sıfır üretici süreç (yapısal sıfırlar) olarak tanımlanır. Gözlemin kesinlikle sıfır olma ihtimali, genellikle logit veya probit gibi ikili regresyon modelleriyle tahmin edilmektedir. İkincisi, sayım süreci (rasgele sıfırlar dahil) olara görülür. Hem sıfır hem de pozitif değerlerin klasik sayım dağılımlarıyla (Poisson veya NB) modellendiği bölümdür. Bu iki süreç ayrı ayrı ele alınır ve her birinin kendine özgü regresyon denklemleri bulunmaktadır.

ZIP modelinde bir gözlemin dağılımı, (7) no'lu eşitlikteki gibi tanımlanır:

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) \times f(0 \mid \mu_i)$$

$$P(Y_i = y) = (1 - \pi_i) \times f(y \mid \mu_i),\ y > 0 \qquad (7)$$

Burada: π_i, gözlemin yapısal sıfır grubuna ait olma olasılığıdır. $f(y \mid \mu_i)$, klasik sayım modeli (Poisson veya NB) tarafından verilen olasılıktır.

Modelde genellikle şu iki ayrı regresyon yapısı tanımlanmaktadır. İlki, sıfır süreci (logit modeli)dir. İkincisi, sayım süreci (Poisson veya NB modeli)dir. Bu iki süreçte yer alan açıklayıcı değişkenler ise aynı yada farklı olarak seçilebilmektedir.

ZIP modeli, varyansın ortalamaya eşit olduğu varsayımına uygun Poisson dağılımına dayanmaktadır ve saçılmanın düşük olduğu durumlarda kullanılmaktadır. Öte yandan ZINB modeli, aşırı saçılma bulunan veri setlerinde tercih edilmektedir. Eğer bir veri setinde hem aşırı sıfır hem de aşırı saçılma söz konusuysa, ZINB modeli en uygun seçenek olmaktadır (Hilbe, 2011).

Zero-inflated modellerin temel avantajı, sıfır değerlerin yoğun olduğu durumları anlamlı şekilde modellemesidir. Bununla birlikte, modellerin iki aşamalı yapısı analiz sürecini karmaşıklaştırabilir ve bazı durumlarda modelin aşırı uyum sağlaması (overfitting) gibi sorunlar da görülebilmektedir. Bu modeller, eğitim, sağlık, sigorta ve ulaşım gibi alanlarda yaygın olarak uygulanmaktadır.

Conway-Maxwell-Poisson (COM-Poisson) regresyon modeli, klasik Poisson modelinin sabit varyans ile ortalama varsayımını esnetmek için geliştirilmiş, esnek bir sayım verisi modelidir. Bu model, hem aşırı saçılma (overdispersion) hem de az saçılma (underdispersion) durumlarında etkili çalışarak, geleneksel Poisson ve Negatif Binom modellerine göre önemli avantajlar sağlamaktadır (Shmueli et al., 2005).

COM-Poisson dağılımı, ilk kez 1962 yılında Conway ve Maxwell (1962) tarafından, kuyruklarda bekleyen müşterilerin davranışlarını incelemek amacıyla önerilmiştir. Ancak bu dağılımın istatistiksel modelleme ve özellikle regresyon analizlerinde kullanımı 2000'li yıllarda yaygınlaşmaya başlamıştır (Shmueli et al., 2005; Sellers & Shmueli, 2010). Klasik Poisson dağılımından daha esnek bir yapıya sahip olması sayesinde, COM-Poisson modeli, özellikle varyansın ortalamadan önemli ölçüde farklılık gösterdiği sayım verilerinde oldukça uygundur.

Regresyon analizinde, bu model genellikle λ üzerine kurulu bir dönüşüm ile parametrik hale (8) no'lu eşitlik yardımıyla getirilmektedir.

$$\log(\lambda_i) = x_i'\beta \qquad (8)$$

Alternatif olarak, beklenen değer üzerinden yaklaşık bir dönüşüm de yapılabilmektedir ve bu eşitlik (9) no'lu eşitlikteki gibidir (Huang, 2017):

$$\log(E[Y_i]) \approx \log(\lambda_i) - (1 / 2\nu) \times \log(\lambda_i) \qquad (9)$$

Bazı uygulamalarda ν parametresi sabit olarak kabul edilirken, daha gelişmiş modellerde ν de ayrı bir regresyon denklemiyle açıklanabilmektedir. Modelin parametreleri genellikle Maksimum Olabilirlik Tahmini (MLE) veya sayısal (iteratif) yöntemler kullanılarak hesaplanmaktadır (Sellers & Shmueli, 2010).

Avantajları; hem az hem de çok saçılmalı verileri modelleyebilmesi, Poisson ve Bernoulli gibi klasik dağılımları özel durumlar olarak kapsaması, ortalama ve varyansın birbirinden bağımsız şekilde kontrol edilebilmesidir. Dezavantajları; karmaşık yapısı nedeniyle hesaplamaların zor olması, parametrelerin yorumu açısından teknik bilgi gerektirmesidir.

COM-Poisson modeli, doğrudan sıfır enflasyonunu modellemez; ancak bazı çalışmalar bu eksikliği gidermek için Zero-Inflated COM-Poisson (ZICMP) gibi modeller geliştirmiştir (Sellers & Raim, 2016).

Model, sağlık bilimlerinden ulaşıma, kriminolojiden ekonomiye kadar birçok alanda başarıyla uygulanmaktadır. Özellikle varyansın belirsiz olduğu sayım verilerinde yüksek doğrulukla sonuç vermektedir.

3. UYGULAMA

Bu çalışma, sıfır şişkinliği ve aykırı değerler içeren sayısal verilerin modellenmesinde farklı regresyon yaklaşımlarının performansını karşılaştırmak amacıyla tasarlanmış bir Monte Carlo simülasyon çalışmasıdır. Simülasyon, beş farklı sayısal veri modelinin tahmin performanslarını Ortalama Kare Hatası (MSE) metriği üzerinden değerlendirmektedir.

Her bir simülasyon iterasyonunda n=100,200,300 ve 400 gözlemlik bir veri seti oluşturulmaktadır. Normal dağılım ve binom dağılımından üretilen 2 adet bağımsız değişken ve negatif binom dağılımdan üretilen bağımlı değişken kullanılmıştır. Bağımlı değişken, iki adet sıfır şişkinlik oranını (%30 ve %40 oranında sıfır şişkinliği) ve %5, %10, %20, %30 ve %35 oranında aykırı değere sahip gözlemlerden oluşmaktadır. 5 adet model (Poisson Regresyon, Negatif Binom Regresyon (NB), Sıfır Şişkin Negatif Binom Regresyon (ZINB), Hurdle Negatif Binom Regresyon, COM-Poisson Regresyon) karşılaştırılmıştır. Değerlendirme kriteri olarak, gerçek y değerleri ile tahmini y değerleri arasındaki Ortalama Kare Hatası (MSE) ölçülmüştür. 1000 bağımsız tekrardan oluşan simülasyon çalışması yapılmıştır.

Bu çalışmada, farklı veri özelliklerine sahip sayısal veri setlerinde hangi modelin daha iyi performans gösterdiğini objektif olarak değerlendirmeyi amaçlamaktadır.

İlgili simülasyon sonuçları, Tablo 1-4'te özetlendiği gibidir.

Tablo 1 incelendiğinde, en iyi performans (En Düşük MSE), ZINB (Sıfır Şişkin Negatif Binom) ve Hurdle modelleri, her iki sıfır şişkinlik senaryosunda da en düşük MSE değerlerine sahiptir. Bu, bu modellerin sıfır şişkinliği olan verilere uyum sağlamada daha başarılı olduğunu gösterir. Özellikle ZINB, %30 ve %40 sıfır şişkinlik durumlarında diğer yöntemlere göre belirgin şekilde daha iyi performans göstermektedir. En kötü performans (En Yüksek MSE), Poisson regresyonu, hem %30 hem de %40 sıfır şişkinlik durumlarında en yüksek MSE değerlerine sahiptir. Bu, Poisson modelinin aşırı dağılım (overdispersion) ve sıfır şişkinliği olan verilerde yetersiz kaldığını doğrulamaktadır. Com-Poisson ve NB modelleri Poisson'a göre daha iyi olsa da, ZINB ve Hurdle kadar iyi değildir. %40 sıfır şişkinlik durumunda, tüm modellerin MSE değerleri genel olarak %30'a göre daha düşüktür. Bu durum, sıfır şişkinliğin artmasının modellerin tahmin performansını iyileştirebileceğini göstermektedir. Aykırı değer oranı artmasıyla, tüm modellerin MSE değerleri artmaktadır. Bu, model karmaşıklığı arttıkça tahmin hatasının da arttığını göstermektedir. Ancak ZINB ve Hurdle, diğerlerine göre daha yavaş bir artış göstermektedir, bu da bu modellerin daha kararlı olduğunu düşündürmektedir.

Tablo 2 incelendiğinde, en iyi performans (En Düşük MSE), ZINB (Sıfır Şişkin Negatif Binom) ve Hurdle modelleri, her iki sıfır şişkinlik senaryosunda da en düşük MSE değerlerine sahiptir. Bu, bu modellerin sıfır şişkinliği olan verilere uyum sağlamada daha başarılı olduğunu gösterir. Özellikle ZINB, %30 ve %40 sıfır şişkinlik durumlarında diğer yöntemlere göre belirgin şekilde daha iyi performans göstermektedir. En kötü performans (En Yüksek MSE), Poisson regresyonu, hem %30 hem de %40 sıfır şişkinlik durumlarında en yüksek MSE değerlerine sahiptir. Bu, Poisson modelinin aşırı dağılım (overdispersion) ve sıfır şişkinliği olan verilerde yetersiz kaldığını doğrulamaktadır. Com-Poisson ve NB modelleri Poisson'a göre daha iyi olsa da, ZINB ve Hurdle kadar iyi değildir. %40 sıfır şişkinlik durumunda, tüm modellerin MSE değerleri genel olarak %30'a göre daha düşüktür. Bu durum, sıfır şişkinliğin artmasının modellerin tahmin performansını iyileştirebileceğini göstermektedir. Aykırı değer oranı artmasıyla, tüm modellerin MSE değerleri artmaktadır. Bu, model karmaşıklığı arttıkça tahmin hatasının da arttığını göstermektedir. Ancak ZINB ve Hurdle, diğerlerine göre daha yavaş bir artış göstermektedir, bu da bu modellerin daha kararlı olduğunu düşündürmektedir.

Örneklem büyüklüğü (n) arttıkça MSE değerleri düşmektedir, ancak model sıralaması değişmemektedir. Aykırı değer oranı arttıkça, tüm modellerin performansı düşmektedir, ancak ZINB/Hurdle en dirençli olan modellerdir.

Tablo 3'e göre, MSE değerleri belirgin şekilde düştüğü görülmektedir. Hurdle modeli, bazı durumlarda ZINB'den daha iyi performans göstermektedir. Sıfır şişkinlik arttıkça ZINB/Hurdle'nin diğer modellere göre daha iyi olduğu

belirlenmiştir. NB, Poisson'a göre çok daha iyi olduğu tespit edilmiştir. Com-Poisson, NB'ye yakın MSE değeri vermiştir, ancak sıfır şişkinlikte daha kötü olduğu belirlenmiştir. NB ve Com-Poisson arasındaki MSE farkının arttığı belirlenmiştir, bu da Com-Poisson'un sıfır şişkinlikte daha az etkili olduğunu göstermektedir. Poisson MSE değeri, diğer modellere daha kötüdür. Örneklem hacmi arttıkça, Poisson modelinin performansı da kötüleşmektedir. Örneklem hacmi arttıkça, modeller daha kararlı bir yapıya ulaşmaktadır. Aykırı değer oranı arttıkça, MSE değerleri de artmaktadır. %40 sıfır şişkinlikte bazı modellerin MSE'si artmıştır. Bu durum, sıfır şişkinliğin çok yüksek olması, tahminleri zorlaştırabileceğinin göstergesidir.

Tablo 4'e göre, MSE değerleri belirgin şekilde düştüğü görülmektedir. Hurdle modeli, bazı durumlarda ZINB'den daha iyi performans göstermektedir. Sıfır şişkinlik arttıkça ZINB/Hurdle'nin diğer modellere göre daha iyi olduğu belirlenmiştir. NB, Poisson'a göre çok daha iyi olduğu tespit edilmiştir. Com-Poisson, NB'ye yakın MSE değeri vermiştir, ancak sıfır şişkinlikte daha kötü olduğu belirlenmiştir. NB ve Com-Poisson arasındaki MSE farkının arttığı belirlenmiştir, bu da Com-Poisson'un sıfır şişkinlikte daha az etkili olduğunu göstermektedir. Poisson MSE değeri, diğer modellere daha kötüdür. Örneklem hacmi arttıkça, Poisson modelinin performansı daha da kötüleşmektedir. Örneklem hacmi arttıkça, modeller daha kararlı bir yapıya ulaşmaktadır. Aykırı değer oranı arttıkça, MSE değerleri de artmaktadır. %40 sıfır şişkinlikte bazı modellerin MSE'si artmıştır. Bu durum, sıfır şişkinliğin çok yüksek olması, tahminleri zorlaştırabileceğinin göstergesidir.

Tablo 1. n=100, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	99.7205	145.7247	205.4721	268.4135	332.1827
NB	73.8596	89.4601	125.8354	153.8239	172.6347
ZINB	58.1488	65.2812	78.3489	92.5743	102.2862
Hurdle	60.8246	68.9373	82.4758	97.1828	109.5269
Com-Poisson	76.4227	94.1841	132.5709	161.7833	181.4233
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	88.4114	132.7238	187.5361	250.3573	328.7273
NB	65.4359	78.5422	104.2746	142.6739	168.6528
ZINB	45.2862	51.8254	71.1623	80.1444	95.4315
Hurdle	48.1776	54.6341	77.8218	85.7257	99.9244
Com-Poisson	71.2935	85.4148	112.4536	148.4362	175.3851

Tablo 2. n=200, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	82.4164	127.8547	187.5323	245.7349	312.4724
NB	59.7319	84.1423	112.6741	142.1632	168.2553
ZINB	54.2833	62.0804	72.8511	85.2727	98.7217
Hurdle	56.1586	65.3786	76.9358	89.6388	103.8573
Com-Poisson	61.4273	89.2572	121.5466	148.9287	175.3862
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	71.4124	122.4281	182.9196	243.7226	308.4729
NB	52.7336	72.1763	101.6782	138.1573	165.3233
ZINB	42.1578	48.6344	67.2422	78.6311	89.7448
Hurdle	45.6202	51.4545	71.0508	82.2539	94.8211
Com-Poisson	59.2851	81.2437	109.8313	145.8284	172.8571

Tablo 3. n=300, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	13.9253	18.6428	28.4177	42.8648	58.3474
NB	8.8803	11.4216	16.5553	25.8317	33.4519
ZINB	5.8211	7.2548	11.2466	18.4536	24.6726
Hurdle	6.7589	7.8936	9.8709	14.7269	18.9263
Com-Poisson	10.2146	13.0703	19.8263	32.1627	42.1623
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	17.6434	24.5642	41.5616	63.4553	72.8968
NB	11.4228	15.6749	26.8333	39.8279	47.5622
ZINB	6.1509	8.7224	17.4569	28.6333	34.1207
Hurdle	7.0811	9.4579	14.2812	22.4748	28.6387
Com-Poisson	10.9578	17.2357	32.1703	48.9132	58.2344

Tablo 4. n=400, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	9.7844	14.6731	29.1257	52.6779	64.1241
NB	6.4532	9.4544	19.7865	34.1227	41.2538
ZINB	4.2321	6.1212	14.1233	24.5684	29.8722
Hurdle	4.8758	6.7866	11.4572	18.9281	23.1504
Com-Poisson	7.1269	10.2354	23.4548	41.2393	49.8349
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Poisson	12.7829	18.6754	38.9172	58.6736	67.2313
NB	8.4513	12.3476	24.5651	37.1294	43.1238
ZINB	5.2336	7.4523	17.1244	26.7855	31.4575
Hurdle	6.1258	8.2345	13.7867	20.4582	24.7822
Com-Poisson	9.1264	13.1277	29.3423	45.2341	51.8954

4. SONUÇ VE ÖNERİLER

Bu çalışmada, sıfır şişkinlik (zero-inflation) ve aşırı dağılım (overdispersion) içeren verilerde beş farklı istatistiksel modelin (Poisson, Negatif Binom (NB), Sıfır Şişkin Negatif Binom (ZINB), Hurdle ve Conway-Maxwell-Poisson (Com-Poisson)) performansları MSE (Ortalama Kare Hatası) üzerinden karşılaştırılmıştır. Analizler, n=100, 200, 300 ve 400 örneklem büyüklükleri, p=2 bağımsız değişken ve %30/%40 sıfır şişkinlik senaryoları altında 1000 simülasyon tekrarı ile gerçekleştirilmiştir.

Tüm tablolarda, ZINB ve Hurdle modelleri en düşük MSE değerleriyle açık ara diğer modellere göre en iyi performansı göstermiştir. Bu sonuç, sıfır şişkinlik durumunda bu modellerin sıfırları ayrı bir süreçle modellemesinin avantajını doğrulamaktadır. NB ve Com-Poisson, Poisson'a göre daha iyi performans sergilemiş, ancak ZINB/Hurdle kadar etkili olamamıştır. Poisson regresyonu ise tüm senaryolarda en yüksek MSE değerleriyle başarısız olmuştur. Bu, Poisson modelinin aşırı dağılım ve sıfır şişkinliği elde edememesiyle açıklanmaktadır.

Örneklem büyüklüğü (n) arttıkça tüm modellerin MSE değerleri düşmüştür. Ancak, model sıralaması (ZINB/Hurdle > NB > Com-Poisson > Poisson) değişmemiştir. Bu, örneklem arttıkça tahminlerin daha kararlı hale geldiğini, ancak sıfır şişkinlikte özel modellerin avantajının korunduğunu göstermektedir.

Aykırı değer oranı arttıkça tüm modellerin MSE değerleri artmıştır. Ancak, ZINB ve Hurdle modelleri bu artışa en dirençli olanları olarak karşımıza çıkmıştır. Bu durum, ZINB/Hurdle'nin karmaşık modellerde bile kararlı kaldığını göstermektedir.

%40 sıfır şişkinlikte, ZINB ve Hurdle'nin göreceli performansı daha da iyileşmiş olduğu görülmüştür. Buna karşılık, Poisson'un performansı daha da kötüleşmiştir.

Sıfır şişkin verilerde ZINB veya Hurdle modelleri tercih edilmelidir. Poisson kesinlikle kullanılmamalıdır; NB veya Com-Poisson daha iyi alternatiflerdir, ancak ZINB kadar etkili değildir. Örneklem büyüdükçe modellerin performansı artar, ancak sıfır şişkinlik durumunda model seçimi değişmez. Aykırı değer oranı artıkça tüm modellerin performansı düşmektedir, bu nedenle model karmaşıklığı dengelenmelidir.

Sonuç olarak, bu çalışma, sıfır şişkinlik ve aşırı dağılım içeren sayma verilerinde, ZINB ve Hurdle modellerinin üstünlüğünü net bir şekilde ortaya koymaktadır. Poisson modelinin yetersizliği ve NB/Com-Poisson'un kısıtlı avantajları da tutarlı bir şekilde gözlemlenmiştir. Gelecek çalışmalarda, daha yüksek sıfır şişkinlik oranları ve daha büyük örneklemler ile test edilerek bu bulguların sağlamlığı doğrulanabilir.

KAYNAKLAR

Cameron, A. C. & Trivedi, P. K. (2013). *Regression Analysis of Count Data (2nd ed.)*. Cambridge University Press.

Hilbe, J. M. (2011). *Negative Binomial Regression (2nd ed.).* Cambridge University Press.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34* (1), 1-14.

Zeileis, A., Kleiber, C. & Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software, 27* (8), 1-25.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33* (3), 341-365.

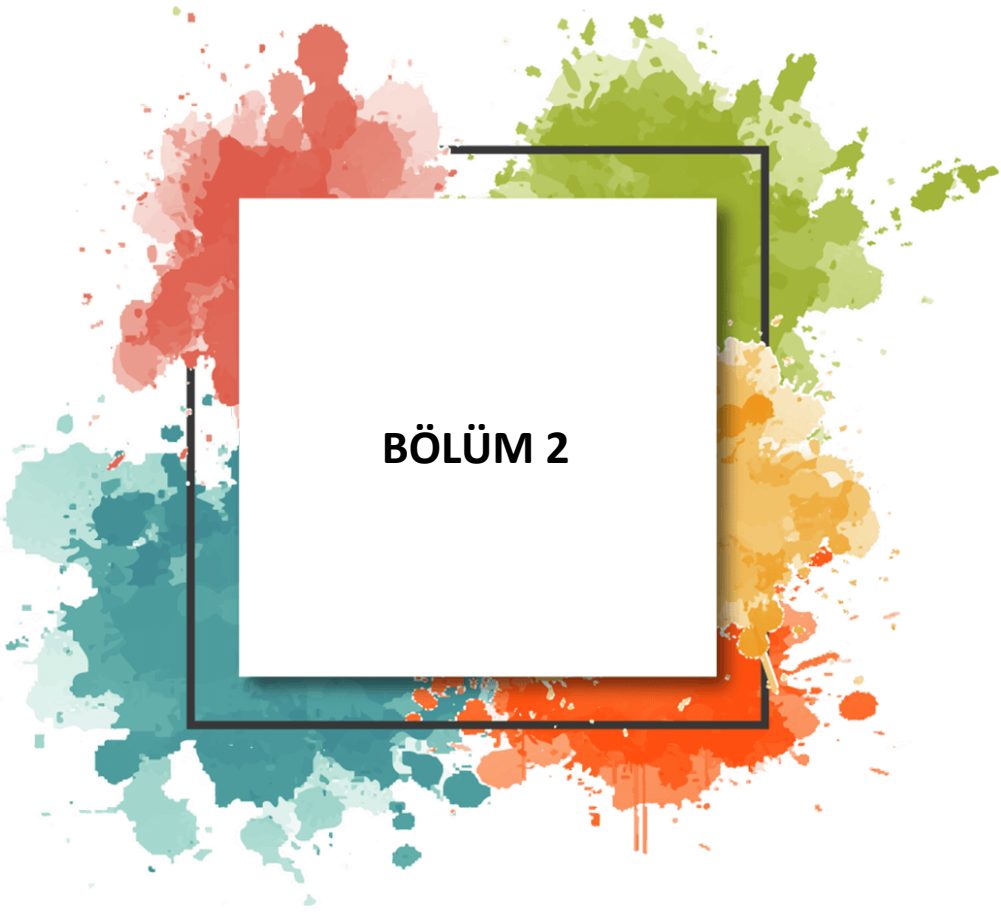
Conway, R. W. & Maxwell, W. L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering, 12,* 132–136.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. & Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54* (1), 127–142.

Sellers, K. F. & Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics, 4* (2), 943–961.

Sellers, K. F., & Raim, A. M. (2016). Zero-inflated Conway-Maxwell-Poisson regression model for overdispersed count data with excess zeros. *Journal of Statistical Computation and Simulation, 86 (*15), 3108–3120.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling, 17* (2), 109–130.


BÖLÜM 2

Sıfır Şişkin ve Aykırı Değer İçeren Sayma Verilerinde Model Performanslarının Karşılaştırılması

Barış Ergül[1] & Arzu Altın Yavuz[2]

1. GİRİŞ

Sayma verileri (count data), yalnızca sıfır ve pozitif tam sayı değerleri alabilen ve istatistiksel analizde özel bir inceleme alanı oluşturan veri türüdür (Cameron & Trivedi, 2013). Bu veri türü; sağlık bilimlerinden sosyal bilimlere kadar çok çeşitli disiplinlerde geniş bir uygulama alanına sahiptir. Sayma verilerinin analitik yaklaşımları, sürekli verilerden yapısal olarak farklılık arz etmektedir ve bu farklılıklar, uygun modelleme tekniklerinin seçiminde belirleyici bir rol oynamaktadır (Hilbe, 2011).

Sayma verilerinin temel karakteristikleri aşağıdaki gibidir; sürekli değişkenlerden farklı olarak, sayma verileri yalnızca ayrık ve pozitif tam sayı değerleri almaktadır. Bu özellik, normal dağılım varsayımının geçerliliğini ortadan kaldırmaktadır (McCullagh & Nelder, 1989). Sayma verileri çoğunlukla sağa çarpık bir dağılım sergilemektedir. Bu çarpıklık, özellikle ortalama değeri düşük olan veri setlerinde daha belirgin hale gelmektedir (Agresti, 2002). Sayma verilerinde; varyans, sıklıkla ortalamaya bağlıdır. Poisson dağılımında olduğu gibi, varyansın ortalama ile eşit olması beklenmektedir ($E[X] = Var[X] = \lambda$) (Winkelmann, 2008). Sayma verilerinde sıfır değerlerin yüksek oranda gözlemlenmesi yaygındır. Bu durum, özellikle düşük gerçekleşme olasılığına sahip olayların modellenmesinde önemli bir analitik güçlük yaratmaktadır (Lambert, 1992).

Sayma verilerinin analizinde en yaygın kullanılan yöntemlerin başında Poisson regresyon modeli gelmektedir. Bu model, üç temel varsayıma dayanmaktadır (Hilbe, 2011); koşullu ortalama ve varyans eşittir ve $E[Y|X] = Var[Y|X] = \exp(X\beta)$ şeklinde tanımlanmaktadır. Tüm gözlemlerin birbirinden bağımsız olduğu kabul edilmektedir. Bağımsız değişkenlerin, bağımlı değişken üzerindeki etkisinin logaritmik ölçekte lineer olduğu varsayılmaktadır.

Ancak uygulamada, bu varsayımların çoğu zaman sağlanmadığı gözlemlenmektedir (Cameron & Trivedi, 2013). Bu bağlamda, Poisson modelinin yetersiz kaldığı bazı tipik durumlar bulunmaktadır. İlki, aşırı yayılım

[1] Dr. Öğr. Üy., Eskişehir Osmangazi Üniversitesi Fen Fakültesi, İstatistik Bölümü
0000-0002-1811-5143
[2] Prof. Dr., Eskişehir Osmangazi Üniversitesi Fen Fakültesi, İstatistik Bölümü
0000-0002-3277-740X

olarak bilinmektedir ve modelin varsaydığı eşitlikten saparak varyansın ortalamadan daha büyük olması durumudur. Bu sapma, Poisson modelinde standart hataların küçültülmesine ve dolayısıyla güven aralıklarının yanıltıcı biçimde dar olmasına yol açmaktadır (Dean & Lawless, 1989). İkincisi, sıfır şişkinliği (Zero-Inflation) olarak karşımıza çıkmaktadır. Sıfır şişkinliği, veri setinde beklenenden çok daha fazla sıfır değerinin gözlenmesiyle ortaya çıkmaktadır (Lambert, 1992). Bu gibi durumlarda, sıfır-şişkin Poisson (ZIP) ya da sıfır-şişkin negatif binom (ZINB) modelleri tercih edilmektedir. Üçüncüsü, aykırı değerlerin etkisidir. Poisson regresyon modeli, veri setinde yer alan aykırı biçimde yüksek sayım değerlerine karşı oldukça duyarlıdır. Bu tür aykırı gözlemler, modelin parametre tahminlerini önemli ölçüde çarpıtmaktadır (Cantoni & Ronchetti, 2001). Hatta tek bir aykırı gözlem dahi, modelin genel geçerliliğini ve yorumlanabilirliğini ciddi şekilde zedeleyebilmektedir (Huber, 1981).

Bu çalışma, sıfır şişkinliği (zero-inflation) ve aykırı değerler (outliers) içeren sayma verileri üzerinde dayanıklı (robust) regresyon modellerinin performansını değerlendirmek amacıyla tasarlanmış bir simülasyon ile 5 farklı modelin performansları değerlendirilmiştir.

2. MATERYAL VE METOT

Poisson regresyon modeli, özellikle sayısal biçimde ifade edilen bağımlı değişkenlerin modellenmesinde yaygın olarak kullanılan temel bir istatistiksel tekniktir (Cameron & Trivedi, 2013). Nadir olayların sayım verilerinin analizi için oldukça elverişli olması, bu modeli araştırmalar açısından cazip kılmaktadır. Ancak, modelin temel varsayımlarının çoğu zaman gerçek veri setlerinde karşılanmaması, araştırmacıları daha esnek ve güvenilir alternatif modeller geliştirmeye itmiştir (Hilbe, 2011).

Klasik Poisson modeli, üç temel varsayıma dayanır (McCullagh & Nelder, 1989); ilki, olayların ortalaması ile varyansının birbirine eşit olması varsayımıdır. İkincisi, gözlemler arasında bağımsızlık olması varsayımı olarak bilinir. Sonuncusu, bağımlı değişken ile bağımsız değişkenler arasında log-lineer bir ilişki bulunması varsayımı olarak bilinmektedir.

Ancak gerçek dünyadaki veri uygulamalarında bu varsayımların sıkça ihlal edildiği görülmektedir (Agresti, 2015). Bu ihlaller genellikle şu sorunlarla kendini gösterir: aşırı dağılım (overdispersion), sıfır şişkinliği ve aykırı değerler sorunlarıdır.

Bu sorunları aşmak için sağlam yöntemler geliştirilmiştir. Bunlardan ilki, Quasi-Poisson yaklaşımıdır. Poisson modelinin aşırı dağılıma duyarlılığına karşı geliştirilmiştir. Bu modelde varyans, ortalama ile doğrusal bir ilişki içinde olacak şekilde yeniden tanımlanır (Ver Hoef & Boveng, 2007):

Diğer yöntem, sağlam standart hatalar yaklaşımı olarak bilinir. Sandviç kovaryans tahmincisi olan bu yöntem, modelin varyans-covaryans yapısının daha doğru tahmin edilmesini sağlamaktadır. Bu yaklaşım, klasik varsayımlar bozulsa dahi güvenilir standart hata tahminleri elde etmeyi mümkün kılmaktadır. (Zeileis, 2006). 1 no'lu eşitlikteki gibi ifade edilmektedir.

$$\text{Var}\,(\hat{\beta}) = (X'WX)^{-1}(X'\Omega X)(X'WX)^{-1} \qquad (1)$$

Sağlam Poisson modelleri, çeşitli uygulama alanlarında kullanılmaktadır. Özellikle, epidemiyolojik ve finansal risk modellemesi çalışmalarında sıkça kullanılmaktadır (Hardin & Hilbe, 2018; Agresti, 2015).

Negatif binom regresyonu, sayım verilerinin analizinde sıklıkla karşılaşılan aşırı dağılım sorununu aşmak amacıyla geliştirilen, Poisson modeline kıyasla daha esnek bir genelleştirilmiş lineer modeldir (Hilbe, 2011). Özellikle varyansın ortalamadan büyük olduğu durumlarda, klasik Poisson modeline göre daha güvenilir sonuçlar sunmaktadır (Cameron & Trivedi, 2013). Negatif binom dağılımı, Poisson dağılımının genişletilmiş bir biçimi olarak, aşırı dağılımı modelleyebilmek için bir ek parametre (α veya k) içermektedir (McCullagh & Nelder, 1989). Bu modelin varyans fonksiyonu (Lawless, 1987) ise (2) no'lu eşitlikteki gibi tanımlanmaktadır.

$$\text{Var}(Y) = \mu + \alpha\mu^2 \qquad (2)$$

Burada, μ: Beklenen değer, α: Dağılım (dispersiyon) parametresidir. Bu ifade, varyansın ortalamanın bir fonksiyonu olarak büyüdüğünü gösterir ve aşırı dağılımı açıkça yansıtır.

Aşırı dağılımı modellemek için, Gourieroux ve arkadaşları (1984) tarafından geliştirilen bu yaklaşım, klasik varyans tahminlerinin yetersiz kaldığı durumlarda daha sağlam sonuçlar elde etmek için sandviç kovaryans tahmincileri kullanır. Bu sayede, aykırı gözlemlere karşı dirençli bir yapı elde edilmektedir.

Breslow (1984) tarafından önerilen diğer teknik, verideki etkili gözlemlere daha fazla ağırlık vererek sonuçların güvenilirliğini artırmaktadır. M-tahmin temelli olan bu yöntem, özellikle veri setinde bozulma olduğunda sağlamlık özelliği göstermesiyle dikkat çekmektedir.

Negatif binom regresyonu epidemiyoloji, trafik güvenliği ve kriminoloji gibi alanlarda yaygın olarak kullanılmaktadır (Gardner et al., 1995; Miaou, 1994; Osgood, 2000).

Sağlam Negatif Binom, aşırı sıfır şişkinliği içeren veri setlerinde model yetersiz kalabilmektedir. Ek olarak, hesaplama açısından karmaşıktır, özellikle büyük veri setlerinde işlem yükünü artırabilmektedir.

Sıfır şişkin negatif binom (ZINB) regresyonu, sayım verilerinde hem aşırı sıfır gözlem hem de aşırı dağılım durumlarıyla başa çıkmak için geliştirilmiş esnek ve güçlü bir karma modeldir (Lambert, 1992). Bu model, gözlenen sıfırların yalnızca rastlantısal değil, aynı zamanda yapısal nedenlerle de oluşabileceğini varsaymaktadır (Mullahy, 1986). Bu nedenle ZINB modeli, sıfır değerleri iki farklı süreçle açıklar; yapısal sıfırlar ve dağılımsal sıfırlar olmak üzere açıklanan modelin bu özelliği, onu özellikle epidemiyoloji, ekoloji ve sosyal bilimlerdeki sayım verilerine yönelik analizlerde değerli bir araç haline getirmiştir (Zeileis vd., 2008).

ZINB modeli iki ayrı bileşenden oluşmaktadır (Greene, 1994); ilki, sıfır üreten bileşen (logit model), ikincisi, sayım bileşeni (negatif binom dağılımı)dir. Bu yapı, gözlemlerin bir kısmının hiçbir koşulda pozitif değer alamayacağını, geri kalanların ise negatif binom dağılımına göre sıfır ya da pozitif değer alabileceğini kabul etmektedir.

Cameron ve Trivedi (1998) tarafından geliştirilen sağlam ZINB modelleri, klasik modelin kırılganlığını azaltmak amacıyla bazı istatistiksel iyileştirmeler sunmaktadır. Bu model, sandviç kovaryans tahmincileri aracılığıyla standart hataları daha güvenilir hale getirmektedir. Ek olarak değişen varyans durumuna karşı dayanıklıdır. Modelin hatalı olarak elde edilmesine karşı daha esnek bir yapıdadır.

Bu konudaki diğer yaklaşım, Heilbron (1994) tarafından önerilmiştir. Bu yöntem, özellikle veri setinde aykırı ya da dengesiz gözlemlerin bulunduğu durumlarda tercih edilmektedir. Bu yöntemde, gözlemlere ağırlık vererek aykırı değerlerin etkisini sınırlamaktadır.

Hurdle modelleri, sayım verilerinde sıkça karşılaşılan aşırı sıfır gözlem problemiyle başa çıkmak için geliştirilmiş iki aşamalı modellerdir (Mullahy, 1986). Bu modeller, sıfır ve sıfır dışı gözlemleri birbirinden tamamen farklı mekanizmalarla açıklayarak, ZINB (sıfır şişkin negatif binom) modellerine güçlü bir alternatif sunmaktadır (Cameron & Trivedi, 2013). Sağlık ekonomisi, tıbbi araştırmalar ve tüketici davranışlarının incelendiği iktisadi çalışmalarda yaygın olarak kullanılmaktadır (Winkelmann, 2008).

Hurdle modelleri, iki temel süreçten oluşur (Cragg, 1971); ilki, sıfır eşiği bileşeni olarak adlandırılan bileşendir. Bu aşamada, gözlemin sıfır mı yoksa sıfırdan büyük mü olduğunu belirlemek üzere ikili bir model (logit ya da probit) uygulanmaktadır. Diğeri, kesilmiş sayım bileşenidir. Sıfır olmayan gözlemler, sıfır değerlerin dahil edilmediği bir dağılım (örneğin Poisson ya da negatif binom) kullanılarak modellenmektedir.

Sağlam Hurdle için ilk yaklaşımlardan birisi, Wooldridge (2010) tarafından geliştirilen bu yaklaşımdır. Dağılım varsayımlarındaki sapmalara karşı dayanıklı analizler yapma olanağı sağlamaktadır. QMLE (Quasi-Maximum Likelihood) yöntemini kullanarak doğru dağılım varsayımı yapılmasa bile tutarlı tahminler sunmaktadır. Sandviç kovaryans matrisi sayesinde güvenilir standart hatalar elde edilmektedir. Modelin yapısal hatalarına karşı istatistiksel esneklik sağlamaktadır.

Diğer yaklaşım, ağırlıklı Hurdle Modelleridir. Terza (1998) tarafından önerilen bu yöntem, örneklem yapısına duyarlılığı artırmak için gözlemlere ağırlık atanmasına yönelik bir yaklaşımdır. Gözlemlere ağırlık verilerek örneklem farklılıklarının etkisi azaltılmaktadır. Eşik (sıfır/pozitif karar) ve sayım bileşenleri ayrı ayrı ağırlıklandırılabilmektedir. Özellikle karmaşık örnekleme tasarımlarında daha tutarlı ve sağlam sonuçlar üretmektedir.

Conway-Maxwell-Poisson (COM-Poisson) regresyonu, sayım verilerinin modellenmesinde Poisson ve negatif binom modellerinin sınırlılıklarını aşmak amacıyla geliştirilmiş esnek bir genelleştirilmiş lineer modeldir (Shmueli vd., 2005). Özellikle hem under-dispersed (varyans < ortalama) hem de over-dispersed (varyans > ortalama) verileri tek bir çatı altında modelleyebilme yeteneğiyle dikkat çekmektedir (Sellers & Shmueli, 2010).

Sağlam COM-Poisson Modelleri için Sellers ve Shmueli (2013) tarafından geliştirilen yaklaşım, sandviç kovaryans tahmincileri kullanmaktadır. Değişen varyanslılığa karşı dayanıklı bir yaklaşımdır.

3. BULGULAR

Bu çalışma, sıfır şişkinliği (zero-inflation) ve aykırı değerler (outliers) içeren sayma verileri üzerinde dayanıklı (robust) regresyon modellerinin performansını değerlendirmek amacıyla tasarlanmış bir simülasyondur. Farklı sıfır şişkinlik oranları altında beş farklı modelin (Sağlam Poisson, Sağlam Negatif Binom, Sağlam ZINB, Sağlam Hurdle, Sağlam Com-Poisson) tahmin doğruluğu (Ortalama Kare Hatası, MSE) karşılaştırılmaktadır.

Örneklem büyüklüğü (n) olarak, 100, 200, 300, 400 gözlem; aykırı değer oranı (τ) olarak %5, %10, %20, %30, %35 ve sıfır şişkinlik oranları (ς) olarak %30, %40 alınmıştır. Bağımsız değişkenler olarak, Normal dağılımlı sürekli değişken ve Binom dağılımlı kategorik değişken kullanılmıştır. Bağımlı değişken olarak, sayma verisi özelliğindeki değişken kullanılmıştır. Simülasyonda, Bootstrap tabanlı katsayı tahmini (100 bootstrap) ve tekrarlı Monte Carlo Simülasyonu (1000 iterasyon) kullanılmıştır.

Bu çalışma, aykırı değerlere ve sıfır şişkinliğe karşı dirençli modellerin hangi koşullarda daha iyi performans gösterdiğini ortaya koymayı amaçlamaktadır.

Elde edilen bulgular, pratik uygulamalarda hangi modelin tercih edilmesi gerektiğine dair bir rehber sunacaktır.

Simülasyon sonuçları Tablo 1-4'te özetlendiği gibidir.

Tablo 1'e göre, n=100 gözlem ve p=2 bağımsız değişken içeren bir simülasyon çalışmasında, %30 ve %40 sıfır şişkinlik (zero-inflation) oranlarına sahip veriler üzerinde dayanıklı (robust) regresyon tekniklerinin performansını karşılaştırmaktadır. Modellerin performansı, y değişkeninin tahmin indeki ortalama kare hatası (MSE) ile ölçülmüştür. En Düşük MSE, Robust ZINB ve Robust Hurdle Modelleri vermiştir. %30 sıfır şişkinlik için, Robust ZINB ve Robust Hurdle en iyi performansı gösteriyor. Benzer bir durum %40 şişkinlik için de bulunmuştur. Sonuç olarak, sıfır şişkinliği yüksek verilerde, ZINB ve Hurdle modelleri diğerlerine göre belirgin şekilde daha başarılıdır. Sağlam Poisson modeli, hem aşırı yayılıma hem de sıfır şişkinliğine karşı hassas olduğu için en yüksek MSE değerlerine sahiptir. Sağlam Negatif Binom modeli, aşırı yayılım (overdispersion) olduğunda iyi sonuç vermektedir, ancak sıfırları modellemede ZINB ve Hurdle kadar başarılı olmadığı tespit edilmiştir. Tüm modeller dayanıklı tekniklerle güçlendirildiği için aykırı değerlere karşı dirençli, ancak ZINB ve Hurdle en iyi dengeyi sağlayan modeller olarak görülmektedir. Aykırı değer oranı arttıkça, MSE değerlerinin de arttığı tespit edilmiştir. Bu sonuçlar, sıfır şişkinliği yüksek ve aykırı değerler içeren veri setlerinde, ZINB veya Hurdle modellerinin tercih edilmesi gerektiğini desteklemektedir.

Tablo 2'ye göre, Robust ZINB, en iyi performansı göstermektedir. %30 ve %40 sıfır şişkin olması durumlarında, en düşük MSE değerlerini Robust ZINB modeli vermiştir. Robust ZINB, hem sıfırları hem de aşırı yayılımı modelleyebildiği için örneklem hacmi arttığında daha da güçlü model olduğu görülmektedir. Benzer bir durum Robust Hurdle modeli için de geçerlidir. Robust ZINB modeline en iyi alternatifi olarak Robust Hurdle modeli olduğu tespit edilmiştir. En kötü performansı Robust Poisson modeli göstermiştir.

Tablo 3'e göre, MSE değerleri, örneklem arttıkça düşmeye devam etmektedir. %30 sıfır şişkinlikte; Robust ZINB (2.87) < Robust Hurdle (3.18) < Robust NB (3.49) < Robust COMP (3.72) < Robust Poisson (5.24) ve %40 sıfır şişkinlikte; Robust ZINB (5.21) < Robust Hurdle (5.47) < Robust NB (6.32) < Robust COMP (6.55) < Robust Poisson (8.97) olarak gerçekleşmiştir. Örneklem hacmi arttıkça, tüm modellerin MSE değerleri düşmektedir, ancak sıfır şişkinliği yüksekse ZINB ve Hurdle dışındakilerde ciddi bir iyileşme olmamaktadır. Aykırı değer oranı arttıkça, aykırı değerlerin etkisini azaltmada başarılı bulunmuştur.

Tablo 4 incelendiğinde, benzer sonuçlar elde edilmiştir. Aykırı değer oranı %10 ve daha yüksek olduğunda Robust Hurdle MSE en düşük değerdedir. Bu da, Robust Hurdle modelinin, diğer modellere göre daha iyi olduğunun göstergesidir.

Örneklem hacmi arttıkça, MSE değerlerinde belirgin bir düşüş gözlenmemiştir. %30 ve %40 şişkinlikte en kötü performansı Robust Poisson modeli vermiştir. MSE değerleri, diğer modellere göre daha yüksektir. Robust ZINB/Hurdle'ın sıfır şişkinliğindeki üstünlüğü, Robust Poisson'un yetersizliği tespit edilmiştir. Robust NB, sıfır şişkinlik oranı düşük olan veri setlerinde daha iyi performans sergilemektedir.

Örneklem büyüklüğü (n) artarsa, tüm modellerin MSE değerleri düşmektedir, ancak sıfır şişkinliği yüksekse hata artmaya devam etmektedir. Aykırı değerlerin oranı attıkça, Robust teknikler sayesinde modeller aykırı değerlere karşı dirençli kalmaktadır.

Tablo 1. n=100, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Robust Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	15.7241	19.9488	24.5102	32.1861	34.8258
Robust NB	10.4526	13.8764	16.2384	21.4507	21.9344
Robust ZINB	9.2118	12.8573	13.4556	16.9247	19.6737
Robust Hurdle	9.8564	13.4251	14.1243	18.3722	20.8525
Robust Com-Poisson	11.0357	17.8633	17.8426	23.1783	26.1576
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	16.4573	18.6766	23.9121	29.8226	33.1577
Robust NB	10.0352	12.8541	15.4779	18.9466	20.9104
Robust ZINB	8.7236	11.3524	12.8251	14.1586	17.8344
Robust Hurdle	9.1511	12.5802	13.6489	15.8755	19.4526
Robust Com-Poisson	10.8793	16.8386	16.8321	22.4548	25.7234

Tablo 2. n=200, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Robust Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	5.6767	7.8541	18.7214	21.0344	40.4769
Robust NB	3.7873	5.1217	11.8982	14.5736	26.8872
Robust ZINB	3.1216	4.2548	9.8763	11.7655	24.0764
Robust Hurdle	3.4547	4.6321	10.5214	12.6392	23.2101
Robust Com-Poisson	4.0218	5.3432	12.6437	14.3377	28.3433
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	10.4768	13.5225	27.4549	33.6753	42.8321
Robust NB	7.0549	10.5778	16.2874	20.1431	28.9547
Robust ZINB	5.9103	8.8336	12.4766	17.7219	26.8322
Robust Hurdle	6.2326	9.3031	13.1543	17.8544	25.1764
Robust Com-Poisson	7.3431	11.8334	16.0437	22.9102	30.1435

Tablo 3. n=300, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Robust Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	5.2438	7.1877	14.8326	16.0547	38.9104
Robust NB	3.4983	4.6762	9.6767	10.1433	24.1213
Robust ZINB	2.8762	3.8989	7.9211	8.4219	21.4769
Robust Hurdle	3.1876	4.2213	8.4102	8.9106	20.8542
Robust Com-Poisson	3.7217	4.9106	10.3222	10.8768	26.8327
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	8.9766	13.6211	25.8336	32.4981	41.1439
Robust NB	6.3211	9.4764	15.2863	19.4219	26.5338
Robust ZINB	5.2107	7.8545	14.0541	17.1569	24.6268
Robust Hurdle	5.4764	8.1217	12.4218	16.2872	23.4671
Robust Com-Poisson	6.5554	9.8329	15.9177	21.0744	28.7653

Tablo 4. n=400, p=2, %30 ve %40 (sıfır-şişkinlik) olduğunda Robust Tekniklerin y tahmini için MSE değerleri (1000 tekrar)

%30 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	4.6753	7.0436	13.2674	15.8924	32.8577
Robust NB	3.1247	4.3751	8.4531	9.7738	19.4653
Robust ZINB	2.5436	3.6548	7.9087	8.3762	17.0227
Robust Hurdle	2.6871	3.9288	7.2817	8.1211	15.6763
Robust Com-Poisson	3.2898	4.7117	9.9646	10.2321	21.3329
%40 ς					
Teknik	% τ				
	5	10	20	30	35
Robust Poisson	7.2567	10.3577	21.5592	28.1686	38.3444
Robust NB	4.6572	7.1673	12.3344	17.7674	24.8369
Robust ZINB	3.8235	4.6981	10.2126	15.2785	21.5101
Robust Hurdle	3.9896	4.2438	9.8971	14.3873	20.7582
Robust Com-Poisson	4.8788	8.2453	12.7887	19.8166	25.2236

4. SONUÇ VE ÖNERİLER

Bu çalışmada, sıfır şişkinliği (zero-inflation) ve aykırı değerler (outliers) içeren sayma verilerinin modellenmesinde beş farklı dayanıklı (robust) regresyon tekniğinin performansı, örneklem büyüklüğüne (n=100, 200, 300, 400) ve sıfır şişkinlik oranları (%30, %40) ile farklı aykırı değer oranlarına (%5, %10, %20, %30, %35) bağlı olarak karşılaştırılmıştır.

Örneklem hacmi arttıkça, MSE değerlerinin azaldığı tespit edilmiştir. %30 ve %40 sıfır şişkinlikte Robust ZINB ve Robust Hurdle modelleri en düşük MSE değerlerini vermiştir. Robust NB ve Robust COMP modelleri, sıfırları modelleyemediğinden Robust ZINB ve Robust Hurdle modellerine göre yetersiz kalmıştır. Robust Poisson modeli, diğer modellere göre daha yüksek MSE değerleri verdiğinden en kötü model olmuştur.

Bu kapsamlı simülasyon çalışması, sıfır şişkinliği ve aykırı değerler içeren verilerde hangi modelin ne zaman kullanılması gerektiğine dair net bir rehber sunmaktadır. Uygulamacılar, özellikle, Robust ZINB/Hurdle'ın sıfır

şişkinliğindeki üstünlüğünü, Robust NB'nin sınırlı ancak basit çözüm olduğunu, Robust Poisson ve Robust COMP'un dezavantajlarını dikkate almalıdır.

Gelecek çalışmalar için, farklı sıfır şişkinlik mekanizmaları (Hurdle'a daha uygun veri yapıları) test edilebilir. COM-Poisson'un nu parametresi optimize edilerek performansı iyileştirilebilir. Gerçek veri setlerinde bu modellerin karşılaştırılması yapılabilir.

Sıfır şişkinlik değeri yüksekse, Robust ZINB veya Robust Hurdle kullanılması önerilmektedir; örneklem ne kadar büyük olursa olsun bu modeller vazgeçilmezdir.

KAYNAKLAR

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* Wiley.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data (2nd ed.)*. Cambridge University Press

Hardin, J. W., & Hilbe, J. M. (2018). *Generalized Linear Models and Extensions (4th ed.).* Stata Press.

Hilbe, J. M. (2011). *Negative Binomial Regression (2nd ed.).* Cambridge University Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.).* Chapman & Hall.

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. Negative Binomial Regression. *Ecology, 88* (11), 2766-2772.

Zeileis, A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software, 16* (9), 1-16.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics, 33* (1), 38-44.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics, 5 (*3), 236-244.

Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118* (3), 392.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica, 52* (3), 681-700.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics, 15 (*3), 209-225.

Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention, 26* (4), 471-482.

Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology, 16* (1), 21-43.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data.* Cambridge University Press.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *NYU Working Paper* No. EC-94-10.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal, 36* (5), 531-547.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34* (1), 1-14.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics, 33* (3), 341-365.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica, 39* (5), 829-844.

Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics, 84* (1), 129-154.

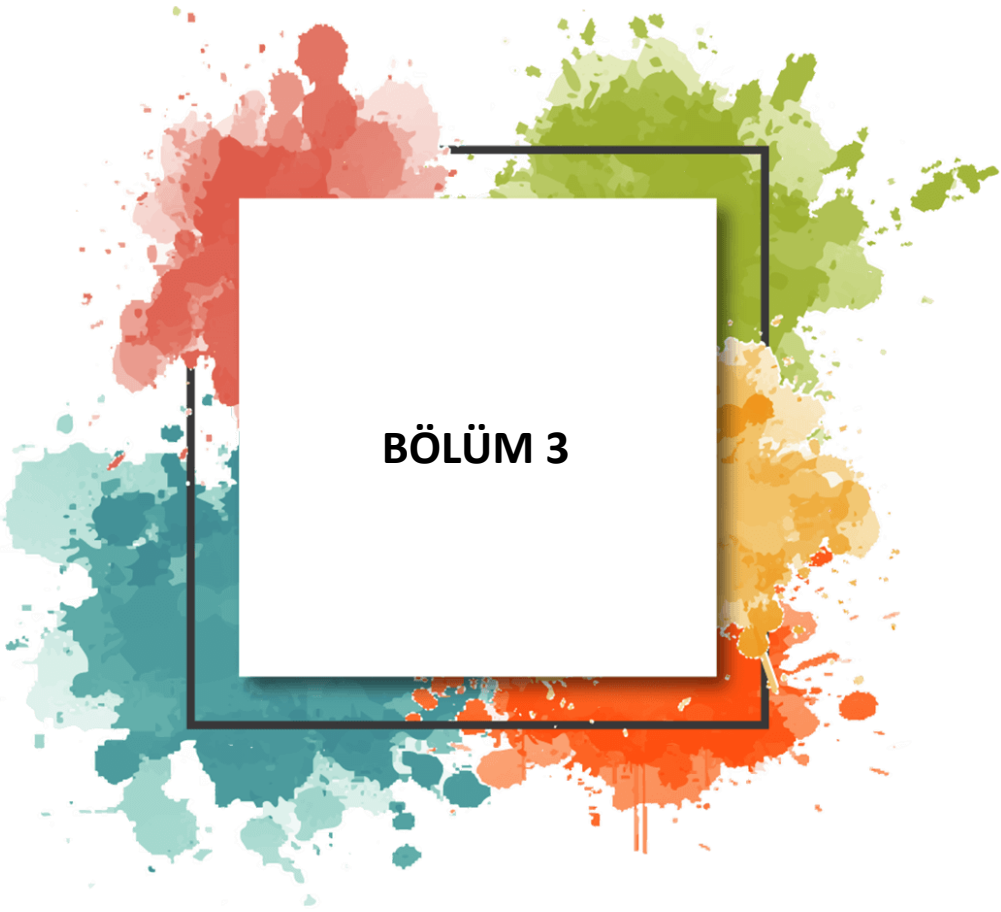
Winkelmann, R. (2008). *Econometric analysis of count data (5th ed.).* Springer-Verlag.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data (2nd ed.).* MIT Press.

Sellers, K. F., & Shmueli, G. (2010). A flexible regression model for count data. *Annals of Applied Statistics, 4* (2), 943-961.

Sellers, K. F., & Shmueli, G. (2013). Data dispersion: Now you see it... now you don't. Communications in Statistics - Theory and Methods, 42 (17), 3134-3147.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Series C, 54* (1), 127-142.


BÖLÜM 3

İstatistiksel Düşünme ve Belirsizlik Kavramı

Mine Doğan[1] & Mehmet Gürcan[2]

1. Giriş

İnsanlık tarihi boyunca karşılaşılan en temel sorunlardan biri, belirsizlikle baş etme ihtiyacıdır. Geleceği öngörmek, kararları sağlam temellere dayandırmak ve karmaşık olguların içindeki düzeni açığa çıkarmak, bilimin olduğu kadar gündelik yaşamın da değişmez uğraşları arasındadır. Bu bağlamda istatistiksel düşünme, modern çağın en güçlü zihinsel araçlarından biri olarak ortaya çıkmıştır. İstatistiksel düşünme, yalnızca sayıların ya da tabloların işlenmesi değildir; esasen veri, belirsizlik ve çıkarım arasındaki döngüsel ilişkiyi kavrama biçimidir. Bu düşünme tarzı, üç temel sütun üzerine oturur: verinin gözlenmesi, modelin kurulması ve çıkarımın yapılması. Her yeni veri, mevcut modeli sınar; her yeni model, elde edilen çıkarımı yeniden şekillendirir. Bu dinamik döngü, istatistiksel yöntemin evrensel gücünü belirler.

Belirsizlik, doğası gereği ölçülebilir ve yönetilebilir bir olgudur. Burada olasılık kuramı, belirsizliği nicel bir çerçeveye oturtarak istatistiksel analizin temel zeminini sağlar. İstatistiksel düşünmenin özgünlüğü, belirsizliği bir engel olarak değil, karar vermede hesaba katılması gereken bir gerçeklik olarak görmesidir. Böylece, örneğin bir klinik çalışmadan elde edilen sınırlı verilerle toplumsal bir sağlık politikası inşa etmek ya da bir üretim sürecinde kaliteyi güvence altına almak mümkün olur. Bu bağlamda, istatistiksel yöntemler iki ana boyutta işlev görür: tanımlayıcı istatistik ve çıkarımsal istatistik. Tanımlayıcı istatistik, verinin sunduğu bilgiyi özetler, düzenler ve görselleştirir; yani mevcut tabloyu anlamamıza aracılık eder. Çıkarımsal istatistik ise, eldeki örnek veriden hareketle daha geniş bir kitle hakkında genellemeler yapmamızı sağlar. İkisi bir arada, istatistiksel düşünmeyi yalnızca geçmişi açıklamakla sınırlamaz, aynı zamanda geleceğe yönelik tahminler ve kararlar için bir pusula haline getirir. Sonuç olarak, istatistiksel düşünme, belirsizliği yok etmeye değil, onu anlamlandırmaya ve yönetilebilir kılmaya yöneliktir. Bu yaklaşım, veriyi salt bir bilgi kaynağı değil, anlam üreten bir süreç olarak ele alır. Böylece bilimden mühendisliğe, sağlıktan ekonomiye kadar pek çok alanda karar vericilerin ve araştırmacıların temel referans noktası haline gelir.

[1] Arş. Gör. Dr., Fırat Üniversitesi Fen Fakültesi İstatistik Bölümü
ORCID: 0000-0002-2745-9909
[2] Prof. Dr., Fırat Üniversitesi Fen Fakültesi İstatistik Bölümü
ORCID: 0000-0002-3641-8113

Belirsizlik tek bir şey değildir; farklı kaynakları, doğası ve giderilebilirlik derecesi farklıdır. Yaygın sınıflandırma genelde şu şekildedir:

- Risk ölçülebilir bir belirsizliktir. Olasılık dağılımı bilinir veya güvenilir biçimde tahmin edilebilir. Örnek olarak, adil bir zar atışında 6 gelme olasılığı = 1/6 olur, fakat oyuncunun aleyhine olan zar gelişinin yönetilme imkanı yoktur.
- Olasılıkların bilinmediği ya da tanımlanamadığı durumlar. Yeni teknolojiyle ilgili bilinmezlikler bu kategoriye girebilir [1].
- Aleatory (stochastic) belirsizlik, doğrudan "rastgelelik" ile ilgili, doğal varyasyondur. Örnek olarak, bireyler arasındaki biyolojik farklılıklar gösterilebilir. Genellikle veriyle karakterize edilir ve tekrarlı ölçümlerle istatistiksel olarak nicelendirilebilir.
- Epistemic (bilgi kaynaklı) belirsizlik, bilgisizlikten kaynaklanır. Ek veriyle veya daha iyi modellerle azaltılabilme imkanı vardır. Örnek olarak, bir parametrenin bilinmeyen değeri veya eksik model bilgisi verilebilir.
- Model belirsizliği (structural/model-form uncertainty), hangi modelin doğru olduğu kesin değilse ortaya çıkar. Farklı model formları farklı sonuç verebilmektedir.
- Gözlem / ölçüm belirsizliği (measurement error), alet, protokol veya insan kaynaklı hatalardan oluşur. Sistematik (bias) veya rasgele olabilmektedir.
- Derin belirsizlik (deep uncertainty), olasılıkların veya sonuçların tanımlanmasının çok zor olduğu durumlarda karşımıza çıkar. Politika, iklim projeksiyonları gibi olgularda karşımıza çıkar.

Peki neden sınıflandırmak önemlidir? Çünkü her biri anlamlandırılmak için farklı stratejilere gerek duymaktadır. Olasılık, belirsizliği nicelleştirmenin en yaygın ve matematiksel çerçevesidir. Ancak yorum ve kullanım biçimi farklılık gösterir. Örnek olarak bir deney sonsuz kez tekrar edildiğinde deney durağan hale gelerek mümkün olayların oranları, olasılıkları, öğrenilebilir [2]. Bayesçi, öznel denebilecek yorumlara göre olasılık, derecelendirmeyi ifade eder. Kişisel ama mantıklı kurallara göre güncellenmelidir [3-4]. Cox yaklaşımı ise tutarlı koşullar altında olasılık derecelerinin modellenebileceğini savunur [5].

Bu niceleştirmenin temel araçlarının başında olasılık dağılımları, beklenen değer, varyans, kuantil gibi özet istatistikler gelmektedir. Güven aralıkları, tekrarlanan deneylerde belirli bir yüzdeye, güvene, bağlı olarak kısmen belirsizlik ortamını açıklamaya çalışır. Bu sayede tahminlerin belirsizlikleri anlamlı bir

ifadeye oturur. Belirsizlikle çalışırken kullanılan pratik yöntemler, hem rastgele varyasyonu hem de bilgi eksikliğini yönetmeye yönelik sistematik yaklaşımları içerir. Bu bağlamda, deneysel tasarım ile varyasyon kontrol altına alınırken, daha fazla veri toplama, yeniden örnekleme teknikleri, Bayesci modeller veya duyarlılık analizleri epistemik belirsizliği azaltmaya ya da görünür kılmaya yardımcı olur. Ayrıca model ortalaması, ensemble yöntemler ve senaryo analizi gibi stratejiler, model formu belirsizliğine karşı daha dengeli sonuçlar üretir. Kalibrasyon, doğrulama ve bağımsız veriyle test ise tahminlerin güvenilirliğini artırır. Bu yöntemlerin ortak amacı, belirsizliği ortadan kaldırmak değil, onu saydamlaştırarak karar vericiler için yönetilebilir hale getirmektir [6].

Tüm bu çerçeve bize gösteriyor ki, belirsizlik yalnızca kaçınılmaz değil, aynı zamanda bilginin gelişmesi için verimli bir zemindir. İstatistiksel düşünme, bu zeminde hem bilimsel ilerlemenin hem de gündelik kararların en güvenilir yol göstericisidir. Veriyi anlamlandırmak, riskleri yönetilebilir hale getirmek ve bilinmeyene karşı daha sağlam adımlar atmak için istatistiksel araçlar vazgeçilmezdir. Bu bölümde tartışılan kavramlar, okuyucuya yalnızca teknik bir yöntem seti değil, aynı zamanda belirsizliği yapıcı bir fırsata dönüştürebilecek zihinsel bir bakış açısı kazandırmayı amaçlamaktadır.

2. İstatistiksel Akıl Yürütme

İstatistiksel düşünmenin merkezinde yer alan yapı, veri, model ve çıkarım arasındaki döngüsel ilişkidir. Bu döngü, belirsizliği yönetmenin ve bilgiyi sistematik olarak üretmenin en güçlü aracıdır. İstatistiksel süreçte veri, yalnızca gözlemlerden ibaret değildir; uygun bir çerçeveye oturtulduğunda, modeller aracılığıyla daha geniş anlamlara ve genellemelere ulaşmamıza imkân verir. Her istatistiksel analiz, veriye dayanır. Veri toplama biçimi, elde edilecek çıkarımların kalitesini belirler. Örnekleme, kitlenin tamamına erişilemediği zaman uygun yöntemlerle kitleden temsili bir alt grup seçme sürecidir. Kontrollü deneyler, neden-sonuç ilişkilerini araştırmak için güçlü bir zemin sunar ve gözlemsel çalışmalar doğal süreçlerdeki örüntüleri anlamamıza yardımcı olur. Eksik değerler, ölçüm hataları veya önyargılı kayıtlar, elde edilen sonuçların güvenilirliğini doğrudan etkiler. Verinin niteliği ve toplama biçimi, sonraki modelleme adımının sağlamlığı için kritik önemdedir [7].

Model, gerçekliğin basitleştirilmiş temsilidir. Tıpkı örneklemin, kitlenin bir temsili olduğu gibi. Veri tek başına anlamsızdır. Anlamı, istatistiksel çıkarımı, veriyi bir model içinde yorumladığımızda ortaya çıkar. Model, gerçeğin bire bir kopyası değil, araştırma sorusu için yeterli bir soyutlamadır [8]. Belirsizlik altında sonuç üretmek istatistiksel bir sanattır. Veriden hareketle bilinmeyen hakkında sonuç üretme sürecidir. Bu süreç tahmin, hipotezlerin test edilmesi, öngörü ve çıkarım aşamalarıyla başarılı bir şekilde tamamlanır. Bu döngünün

dinamiği bilimin gelişmesiyle her geçen gün daha da sağlam bir zemine oturmaktadır.

İstatistiksel akıl yürütme hangi adımla başlar? Sorusunun en önemli cevabı değişken seçiminin düzgün yapılmasında saklıdır. Yapılan deneylerin tamamı tek türlü düşünülemez. Deneyler farklı kategorilerde değerlendirilir. Bir deney deterministik olabilir. Bu bir sürecin veya sistemin tamamen belirli kurallarla işlediğini ve aynı başlangıç koşullarında her zaman aynı sonucu ürettiğini ifade eder. Rastlantısallık veya belirsizlik içermez. Deneyin elde edilen sonuçlarının ölçüm hataları hiçbir zaman deneyi raslantısal hale getirmez. Örneğin suyun kaynama noktasının deneyin yapıldığı rakıma göre farklılık göstermesi onun rastlantısal olduğunu belirtmez. Rastlantısal deneylerin en basit örneği herkesin bildiği zar atma deneyidir. Ancak bu deneyin rastlantısallığında bile şüphe vardır. Şayet zar atan kişi hileye başvurmadığında deney rastlantısal, aksi halde tesadüfi veya stokastik hale dönüşür. Kısacası bu üç kavramı amiyane tabirle birbirinin yerine kullanmak hiç de doğru değildir. Zar atan kişi hileye başvurduğunda ve hilenin gerçekleşme olasılığı da işin işine girdiğinde deney tamamen tesadüfi, hileye başvurulduğu takdirde kişinin olayı gerçekleştirme başarısına etki eden faktörler göz önüne alındığında ise deney tamamen stokastik duruma gelir. Bu şekilde deneyin sınıflandırılması dolayısıyla deneyin değişkeninin de sınıflandırılmasını sağlamaktadır.

İstatistiksel akıl yürütme, bilgi felsefesinin en köklü sorunlarından biri olan belirsizlik ve tümevarım meselesiyle doğrudan ilişkilidir. Mantık geleneği Aristoteles'ten itibaren, doğru öncüllerden zorunlu sonuçlara ulaşmayı amaçlamış; yani akıl yürütmeyi kesinlik arayışı üzerine kurmuştur. Ancak gerçek dünya, kesinlikten ziyade olasılıklarla biçimlenir. David Hume, tümevarım probleminde bu gerilimi ortaya koymuştur [9]. Geçmişte gözlenen düzenliliklerin gelecekte de geçerli olacağına dair hiçbir mantıksal zorunluluk yoktur. İşte istatistiksel akıl yürütme, bu sorunu çözmeye yönelik pratik bir yol sunar: geçmiş verilerden elde edilen örüntüler aracılığıyla geleceğe dair tahmin yapar, fakat bunu kesinlik iddiasıyla değil, olasılıksal güven düzeyinde gerçekleştirir. Yirminci yüzyılda Karl Popper, bilimin özünü doğrulama değil yanlışlanabilirlik ilkesini de görerek istatistiksel düşünceyle kesişmiştir. Bir hipotez asla kesin olarak doğrulanmaz, sadece belirli bir güven düzeyinde desteklenebilir ya da verilerle çeliştiği ölçüde reddedilebilir. Bu yaklaşım, istatistiksel testlerin "sıfır hipotezi" etrafında örgütlenmesine paralel bir düşünce yapısıdır [10]. Rudolf Carnap ve olasılık mantığı üzerine çalışan diğer mantıkçılar ise, istatistiğin aslında belirsizlik karşısında mantığın doğal bir genişlemesi olduğunu ileri sürmüşlerdir [11]. Bruno de Finetti'nin öznel olasılık yorumu da bu çizgiyi destekler. Olasılık, doğanın gizli bir özelliği değil, rasyonel bir bireyin inanç derecelerini düzenleyen mantıksal bir araçtır [12]. Böylece istatistiksel akıl

yürütme, yalnızca teknik bir yöntemler bütünü değil, aynı zamanda epistemolojik bir duruş halini alır. Bizim için önemli olan, mutlak hakikati keşfetmekten çok, elimizdeki veriler ışığında en tutarlı ve en makul çıkarımları yapabilmektir. Bu açıdan istatistik, felsefi olarak kesin bilginin sınırlarını kabul eder, fakat aynı zamanda rasyonel karar vermeyi mümkün kılacak bir olasılıksal mantık geliştirir. Günlük yaşamda, ekonomide, tıpta veya mühendislikte yapılan her tahmin ya da karar, aslında bu felsefi ve mantıksal zeminin pratik bir yansımasıdır, belirsizliği mutlak surette ortadan kaldırmadan, onunla akıllıca yaşamayı öğrenmek.

3. Nicelemeden Sonraki İlk Adım: Matematiksel Yöntemler

Yapılan bir deneyin gözlem değerleri, veriler, toplandıktan sonra araştırmacının hiç şüphe yok ki çoğu kez merak ettiği değer, ortalamadır. Ortalama, istatistiğin en temel ve en sık kullanılan araçlarından biri olmasına rağmen, basit bir sayıdan çok daha fazlasını temsil eder. Felsefi açıdan bakıldığında, ortalama bireyin tekliği karşısında topluluğun özünü arayan bir kavramdır. Her bir veri noktası, kendi içinde eşsiz bir gözlemi, bir hikâyeyi veya bir gerçeği temsil eder. Ancak ortalama, bu tekil gerçeklikleri bir araya getirerek, dağınıklık içinden ortak bir merkez, bir temsil noktası çıkarma çabasıdır. Bu, tıpkı Platon'un idealar kuramında olduğu gibi, görünen fenomenlerin, bireysel verilerin, arkasındaki temel formu, ortalamayı, keşfetme arayışıdır. Aynı zamanda, ortalama bir model olarak gerçekliğin basitleştirilmiş bir temsilidir, çünkü G. E. Box'un dediği gibi, "tüm modeller yanlıştır, ama bazıları faydalıdır". Ortalamanın faydası da, verinin karmaşıklığını feda ederek, bize genel bir eğilim ve bir referans noktası sunmasında yatar. Bu sayede, tekil sapmaları değil, tüm kümenin yönünü anlamamızı sağlar [13].

Ortalamanın veriyi temsil etmesi onun en önemli özelliğidir. Bu noktada matematiksel olarak tanımlanan ortalamalar her veri grubu için geçerli olmayabilir. Bazı veri grubunda aritmetik ortalama temsil kabiliyetine sahipken, bazı veri grubunda ise medyan veya mod temsil kabiliyetine sahiptir. Hesaplama yöntemi matematiksel olsa da bu yönteme anlam yükleyen istatistiksel gereksinimlerdir.

Bir araştırmacının elinde iki tane değişken olduğunda ilgilenmesi gereken ilk yapı, bu iki değişkenin arasında var olan ilişkidir. Bu ilişki nasıl tanımlanmalı ve ölçü olarak nasıl ifade edilmelidir? Bunun için iki değişkeni iki ayrı vektörle göstermek mantıklı olur. Sonuçta vektörler arasında var olan açıklık veya açı küçük olduğunda bu durum iki değişkenin ilişkili olduğunu, büyük olduğunda ise ilişkisiz olduğunu gösterebilir. Basit olarak bu iki vektörü düzlemde düşünelim,

$$A = (a_1, a_2),\ B = (b_1, b_2),\ a_j, b_j \in R,\ j = 1{,}2$$

(1)

Bu iki vektörün arasındaki vektörlerin yatay eksenle yaptığı açıların farkına eşit olacaktır,

$$cosu = \cos(\theta - \alpha) = cos\theta cos\alpha + sin\theta sin\alpha$$

(2)

Burada θ ve α açıları sırasıyla A ve B vektörlerinin X ekseni ile yaptıkları açılar olsun. Buna göre,

$$\cos\alpha = \frac{a_1}{\|A\|},\ cos\theta = \frac{b_1}{\|B\|},\ \sin\alpha = \frac{a_2}{\|A\|},\ \cos\theta = \frac{b_2}{\|B\|}$$

(3)

İfadelerini kullanarak rahatlıkla yazabiliriz,

$$cosu = \frac{a_1 b_1 + a_2 b_2}{\|A\|\|B\|},\ a_1 b_1 + a_2 b_2 = \|A\|\|B\| cosu = \langle A, B\rangle$$

(4)

Bu açıklamalara dikkat ettiğimizde $\langle A, B\rangle$ yapısının gerçek bir anlamı rahatlıkla anlaşılabilir. Ancak bu yapıyı sadece aşağıdaki formda sunduğumuzda ne anlam ifade edebilir?

$$\langle A, B\rangle = a_1 b_1 + a_2 b_2$$

(5)

Kısacası bulgular veya tanımlamalar ihtiyaçtan doğar. Bu ihtiyacın gerekliliğini sunmadığımız takdirde bu eşitliklerin hiçbir anlamı kalmaz. Bundan dolayı kullanılan işlemlere, tanımlamalara, formüllere istatistiksel anlam yüklemek matematiksel öğrenimin birincil kuralı olmalıdır.

4. Merkezi Limit Teoremi İstatistiğe Verilmiş Bir Lütuf mu?

Elbette! Merkezi Limit Teoremi (MLT), istatistiğin ve matematiğin temel taşlarından biri olarak sadece sayısal bir sonuç değil, aynı zamanda olasılık ve belirsizlikle ilgili derin bir felsefi bakış açısı sunar. Felsefi olarak bakıldığında, MLT bize karmaşık ve rastgele görünen olayların bile belli bir düzene ve öngörülebilirliğe sahip olabileceğini hatırlatır. Her gün karşılaştığımız rastgele süreçler, örneğin insanların boyları, test sonuçları veya finansal dalgalanmalar, tek tek oldukça öngörülemez olabilir. Ancak merkezi limit teoremi sayesinde, bu

bağımsız ve aynı dağılımdan gelen rastgele değişkenlerin toplamı veya ortalaması, dağılımların doğasına bakmaksızın yaklaşık olarak normal dağılıma yaklaşır. Bu, kaosun içinde bir düzen bulma yeteneğimizdir. Rastgeleliğin kendi kendini dengede tutması, doğadaki ve sosyal dünyadaki birçok sistemin analiz edilebilir olmasını sağlar. İstatistiğe getirdiği katkılar çok yönlüdür. MLT sayesinde örnekleme yöntemleri güvenilir hale gelir ve örnek ortalamaları üzerinden tüm kitle hakkında tahminlerde bulunabiliriz. Örneğin, nüfusun ortalama geliri veya bir hastalığın yaygınlığı hakkında, tüm nüfusu ölçmeden anlamlı sonuçlar çıkarabiliriz. Bu, istatistiğin temel felsefesi olan "az ile çoğu anlama" yaklaşımının matematiksel temeli gibidir. Teorinin gücü, tek tek ölçümlerin rastgele ve belirsiz olmasına rağmen, topluca büyük ölçüde öngörülebilir ve simetrik bir davranış sergileyeceklerini garanti etmesidir. Böylece istatistikçiler, kaotik görünen verileri yorumlayabilir ve belirsizlik altında mantıklı kararlar alabilir.

Matematik açısından MLT, olasılık teorisinin köprü taşıdır. Tek değişkenli veya çok değişkenli dağılımların toplam davranışlarını anlamamıza imkân verir ve limitler, integral hesapları, momentler gibi daha ileri matematiksel araçları kullanmamıza yol açar. Ayrıca MLT, normal dağılımın evrensel rolünü açıklayarak matematiği doğayla ve deneyimle bağlar. Rastgelelik ve deterministik yapı arasındaki köprüyü kurar. Bu bakış açısıyla MLT, yalnızca istatistiksel bir araç değil, aynı zamanda evrendeki rastgelelik ve düzen arasındaki ilişkileri kavramamızı sağlayan bir felsefi ilke haline gelir: karmaşık sistemlerin bile belirli koşullar altında öngörülebilir bir düzen gösterdiğini anlamamızı sağlar.

Günlük hayatta MLT'nin etkisini görmek aslında oldukça kolaydır. Örneğin, bir şehirdeki insanların boylarını düşünelim. Her yaş gurubuna ait bireylerin boyları farklı ve rastgele dağılım gösteriyor olabilir; kısa veya uzun insanlar olabilir, dağılım asimetrik veya tek tip olabilir. Ancak MLT sayesinde, şehirden rastgele seçilen belli sayıda insanın ortalama boyu alındığında, bu ortalama neredeyse her zaman normal dağılıma yakınsar. Yani örnek ortalamalar, bireysel varyasyonların ötesinde öngörülebilir bir merkez ve yayılma gösterir. Bu sayede şehir nüfusunun ortalama boyu hakkında tahminler yapabiliriz, tek tek herkesin boyunu ölçmeye gerek kalmaz. Benzer şekilde finans dünyasında da MLT kritik bir rol oynar. Bir yatırımcının portföyündeki günlük getiri, tek başına oldukça değişken ve rastgele olabilir. Ancak, uzun bir dönem boyunca günlük getirilerin ortalaması alındığında, MLT'ye göre bu ortalama yaklaşık olarak normal dağılım gösterir. Bu sayede yatırımcılar risk tahminleri yapabilir, güven aralıkları oluşturabilir ve portföy stratejilerini daha bilimsel temellere dayandırabilir. Modern veri biliminde MLT, özellikle makine öğrenmesi ve istatistiksel modelleme süreçlerinde karşımıza çıkar. Büyük veri kümelerinde, örnekleme

yaparak veri analizine başlamak yaygındır. Örneğin bir sosyal medya platformunda kullanıcıların günlük beğeni sayılarının ortalamasını tahmin etmek istediğimizi düşünelim. Tüm kullanıcı verisini işlemek pratik olmayabilir; bu yüzden rastgele bir örnek seçilir. MLT sayesinde bu örnek ortalama, tüm kullanıcı kitlesinin ortalamasını oldukça doğru bir şekilde temsil eder. Bu, veri biliminde "az veri ile güvenilir tahmin" yapabilmemizi sağlar ve model performansını artırır. Ayrıca, hipotez testleri ve güven aralıklarının oluşturulması da MLT'ye dayanır. Örnek ortalamaları yaklaşık olarak normal dağıldığı için, örneklem büyüklüğü yeterli olduğunda popülasyon parametreleri hakkında güvenilir çıkarımlar yapılabilir. Bu durum, tıbbi araştırmalarda ilaç etkinliğini test etmekten, e-ticaret sitelerinin satış analizine kadar geniş bir yelpazede uygulanır.

Özetle, MLT bize şunu söyler: bireysel veriler rastgele ve kaotik olsa da, ortalamalar ve toplamlar çoğunlukla öngörülebilir bir düzen gösterir. Bu felsefi ve matematiksel bakış açısı, günlük yaşamda ve modern veri biliminde karmaşık sistemleri anlamamızı ve kararlar almamızı mümkün kılar.

Jacob Benoulli, 1705 yılında öldüğünde yazdıklarının bu denli bir çığır açabileceğinin farkında mıydı acaba? Hayattayken kaleme aldığı çalışmalar ölümünden sonra 1713 yılında basıldığında Avrupa'da, o zamanlar pek anlaşılmasa da, yeni bir bilimsel çağı başlatmıştı. Özellikle ilk basım " Ars Conjectandi", Olasılık Sanatı, Siméon Denis Poisson tarafından geniş şekilde araştırılmış ve n tekrarlı bağımsız Bernoulli denemelerinde başarı sayısının olasılığı, deneme sayısı çok çok büyük olduğunda Poisson'un bulduğu yaklaşık yöntemle hesaplanabilmiştir. Kısacası binom olasılığı beraberinde Poisson olasılığını da doğurmuştur. Zamana bağlı Poisson olasılıklarında ardışık başarı anları arasında geçen süreler ise üstel tipli olasılıkları doğurmuştur. Yine aynı periyotta binom olasılıkları büyük sayıdaki tekrarlar göz önüne alındığında Gauss olasılıklarını da beraberinde getirmiştir. Kısacası bugün kullandığımız birçok olasılık dağılımının uzaktan veya yakından Bernoulli ile organik bir bağı mutlaka vardır.

Kısacası istatistik, Jacob ve Daniel Bernoulli'den başlayarak Laplace, Gauss ve Quetelet gibi öncülerin katkılarıyla devasa bir bilim haline gelmiş, yirminci yüzyılda Fisher, Neyman ve Pearson sayesinde deney tasarımı, hipotez testleri ve regresyon gibi yöntemlerle sistematikleşmiştir. Günümüzde ise büyük veri, yapay zeka ve derin öğrenme ile birleşerek aleatory ve episdemic belirsizliklerin nicelenmesi, entropi ve bilgi kuramı uygulamaları ile modern problemlere çözüm sunmaktadır. İstatistik hem matematik bilimine hem de fizik, biyoloji, ekonomi ve sosyal bilimler gibi disiplinlere doğrudan katkıda bulunmuş, aynı zamanda makine öğrenmesi ve derin öğrenmede Gaussien Naive Bayes, Bayesien Networks, Cross entropy loss, RNN ve Stochastic Gradient Descent gibi

algoritmaların temelini oluşturmuştur. İleri düzey veri görselleştirme teknikleri ile yüksek boyutlu ve karmaşık veri yapıları görselleştirilebilmekte, böylece istatistik tarihsel köklerinden modern uygulamalara kadar bütünleşik bir bilimsel çerçeve sunmaktadır.

Tüm bunlar sayesinde istatistik, tarihsel olarak Bernoulli'den Fisher'e uzanan kökleriyle tüm bilimlere katkı sağlamış, günümüzde yapay zeka, derin öğrenme ve ileri veri görselleştirme teknikleriyle karmaşık belirsizlikleri modelleyen ve anlam üreten bir disiplin haline gelmiştir.

5. Entropi ve Bilgi Teorisi

Artık on dokuzuncu yüzyıl! Tüm gelişmelerin harmanlandığı çağ. Ne kadar çok şey bilirsen o kadar güçlü ve etkili olursun; bilgi, hem doğayı anlamanın hem de insanlığın sınırlarını genişletmenin en güçlü aracıdır. İyide bilgiyi ölçmenin bir aracı var mı? Bu yüzyılda bu cevabı arayan araştırmacılar entropi kavramı, modern termodinamiğin temel taşlarından biri olarak Rudolf Clausius tarafından 1865 yılında tanıtılmıştır [14-15]. Clausius, ısı ve iş arasındaki ilişkileri matematiksel bir çerçevede inceleyerek, enerji dönüşümlerinin sınırlarını ve doğadaki süreçlerin yönünü açıklamaya çalışmıştır. Bu bağlamda entropi, bir sistemdeki enerji dağılımının ölçüsü ve düzensizlik derecesi olarak tanımlanmıştır. Clausius'e göre, enerji her zaman kullanılabilirlik açısından azalır ve süreçler belirli bir yön izler; entropi ise bu yönlülüğü ve enerji kaybını nicel olarak ifade eder. Termodinamikte entropi, sistemin hangi durumlara ulaşabileceğini ve enerji dönüşümlerinin verimliliğini anlamamıza yardımcı olurken, aynı zamanda evrenin doğal olarak düzensizliğe doğru ilerlediğini gösteren temel bir ölçüt olarak da öne çıkmaktadır [16]. Ludwig Boltzmann, entropi kavramını istatistiksel bir perspektife taşıyan bilim insanıdır. Clausius'un termodinamik yaklaşımının aksine, Boltzmann sistemi mikroskobik parçacıklar düzeyinde inceledi ve bir gazdaki moleküllerin olası düzenlenişlerini olasılık temelli olarak tanımladı. Ona göre, bir sistemin entropisi, parçacıkların belirli bir makroskopik durumu oluşturabilecek farklı mikroskopik dizilimlerinin sayısıyla doğru orantılıdır. Bu düşünceyi matematiksel olarak ifade etmek için ünlü formülü $S = klnW$ eşitliğini geliştirdi; burada S, k Boltzmann sabitini, W ise sistemin olası mikroskobik durumlarının sayısını temsil eder. Bu formül, entropiyi yalnızca enerji dağılımının bir ölçüsü olarak değil, aynı zamanda olasılıksal düzensizlik ve belirsizlik ölçüsü olarak da yorumlamamıza imkân tanır ve modern istatistiksel mekanik ile termodinamik arasındaki köprüyü kurar [15].

Claude Shannon, 1948 yılında yayımladığı "A Mathematical Theory of Communication" adlı makalesinde, bilgi teorisinin temellerini atmıştır. Shannon, iletişim sürecini matematiksel bir çerçeveye oturtarak, bir bilgi kaynağının ürettiği sembollerin belirsizliğini ölçmenin bir yolunu geliştirdi. Bu bağlamda entropi kavramını, olasılık teorisinden esinlenerek bilgi miktarının bir ölçüsü

olarak tanımladı ve artık "Shannon entropisi" olarak bilinen formülü ortaya koydu. Shannon entropisi, bir mesajın ne kadar beklenmedik veya şaşırtıcı olduğunu nicelendirir, yani bir haberin bilgi değerini ölçer. Bu yaklaşım, dijital iletişim, veri sıkıştırma ve hata düzeltme kodları gibi modern bilgi ve iletişim teknolojilerinin temelini oluşturmuş ve bilgi kavramını matematiksel olarak incelemeyi mümkün kılmıştır [17]. Shannon entropisi ve Kolmogorov karmaşıklığı, her ikisi de "bilgi miktarını ölçme" amacı taşır, ancak farklı yaklaşımlara dayanır. Shannon entropisi, olasılık dağılımlarına bağlı olarak bir kaynağın ürettiği sembollerin ortalama belirsizliğini ölçerken, Kolmogorov karmaşıklığı, tek bir diziyi üretmek için gereken en kısa bilgisayar programının uzunluğunu hesaplayarak bireysel bir nesnenin bilgi içeriğini belirler. Yani Shannon, ortalama bilgi ölçümü ve istatistiksel yaklaşım sağlarken, Kolmogorov deterministik ve algoritmik bir bakış sunar. Her iki kavram da özellikle rastgelelik ve düzensizlikle ilgili düşüncelerimizi derinleştirir; uzun ve rastgele bir bit dizisi hem yüksek Shannon entropisine hem de yüksek Kolmogorov karmaşıklığına sahiptir, böylece iki yaklaşımın birbirini tamamladığı görülür.

Şimdi entropiyi basit temellerde biraz inceleyelim. Bir deney yapıldığında iki farklı sonuç aynı sayıda karşımıza çıkarsa bu deneyin sonuçları bize neyi anlatır? Tabi ki seçme aşamasında hiçbir bilgimizin olmadığını. Bundan dolayı bilgi doğal olarak, rahatlıkla, olasılıkla ilişkilendirilebilir. Olasılık sıfırsa, olay gerçekleşmemiş ve bizim olay hakkında hiçbir bilgimiz yoktur. Olasılık birse olay mutlaktır. Deney ve olay aynı gerçekliğe işaret eder. Bu durumda bilgi miktarımız tamdır. Öyleyse olasılığı (p) öyle bir fonksiyonda kullanalım ki bize sıfır olduğunda bomba etkisi yapsın, bir olduğunda ise hiç sesi çıkmasın,

$$S = ln\frac{1}{p} \tag{6}$$

Burada ∞ durumu bomba etkisi ile sıfır durumu ise sesin hiç çıkmaması ile özdeşleştirilmiştir. Tek bir olay için bu rahatlıkla hesaplanabilir. Deneyin tüm sonuçları birlikte düşünüldüğünde ise yine devreye ortalama kavramı girer. Tüm deney için entropi, olayların entropisinin ortalaması dır, $E(S)$.

6. İstatistikten Yapay Zekâya: Belirsizlikten Öğrenmeye

İstatistiksel düşünme, insanlığın belirsizlikle baş etme çabalarının sistematik hale gelmiş biçimidir. Olasılık teorisi, çıkarımsal yöntemler ve veri temsili teknikleri, tarihsel süreçte yalnızca bilimsel araştırmaların değil, günümüzde yapay zekâ ve makine öğrenmesi gibi devrimsel teknolojilerin de temelini oluşturmuştur. Bugün derin öğrenme, büyük veri analitiği ve yapay zekâ sistemleri, aslında istatistiksel akıl yürütmenin modern bir uzantısıdır. Bu bölümde istatistik, yapay zekâ ve derin öğrenme arasındaki yapısal ilişki ele

alınacak; belirsizlik, entropi ve bilgi kuramı üzerinden bu teknolojilerin dayandığı epistemolojik ve matematiksel temeller tartışılacaktır.

Makine öğrenmesi, özünde, veriden örüntü çıkarmaya ve geleceğe yönelik tahminler üretmeye dayalıdır. Bu yönüyle istatistikle doğrudan akrabadır. Naive Bayes sınıflandırıcısı, lojistik regresyon, Gaussian karışım modelleri, Markov zincirleri ve Bayes ağları gibi yöntemler, yapay zekânın erken dönem başarılarının istatistiksel modellerle gerçekleştiğini göstermektedir [18]. Yirminci yüzyılın ortalarında yapay zekâ araştırmaları deterministik mantıksal sistemlere ağırlık vermiş olsa da, belirsizliğin yönetilemezliği nedeniyle istatistiksel yaklaşım yeniden ön plana çıkmıştır. Günümüzde yapay zekâ uygulamalarının neredeyse tamamı olasılık temelli modelleme, parametrik, parametrik olmayan tahminler ve optimizasyon yöntemleri üzerine kuruludur [19].

İstatistikte belirsizlik, aleatory (rastgelelik) ve epistemik (bilgi eksikliği) olarak ayrılır. Bu ayrım aynı zamanda makine öğrenmesinde model hataları ve veri belirsizlikleri şeklinde karşımıza çıkar. Örneğin bir sensörün ölçtüğü sıcaklık değerlerindeki doğal varyasyonlar rastgelelikten kaynaklıdır. Bu tür belirsizlik, veri hacmi arttıkça daha iyi modellenebilir. Eksik veri, hatalı etiketleme veya yanlış model seçimi gibi bilgi yetersizliğinden kaynaklanır. Derin öğrenme modellerinde epistemik belirsizlik özellikle "black-box" yapı nedeniyle kritik önemdedir. Makine öğrenmesinde belirsizlik, Bayesçi yöntemlerle nicel hale getirilebilir. Örneğin, bir sınıflandırıcının olasılıksal tahmin üretmesi, yalnızca sonuca değil, sonucun güven düzeyine dair bilgi sunar. Bu durum tıp, finans ve mühendislikte kritik karar destek mekanizmaları için vazgeçilmezdir. Claude Shannon'ın (1948) bilgi kuramı, günümüzde yapay zekâ için hâlâ temel bir yapı taşıdır. Bilgi kazancı entropi azalması ile ölçülür. Entropi, bilgi miktarının ölçüsü olarak karar ağaçlarında, özellik seçme yöntemlerinde ve derin öğrenmede kayıp fonksiyonlarının tanımında kritik rol oynar. Derin öğrenme yöntemlerinde "cross-entropy loss" sınıflandırma problemlerinde yaygın olarak kullanılan kayıp fonksiyonudur. Modellerin genelleme yeteneğinin artırılmasında entropi ve KL-divergence önemli araçlardandır. Entropi kavramı, yalnızca veri sıkıştırma ve iletişim kuramı için değil, aynı zamanda yapay zekânın "öğrenme" sürecinde de merkezi bir rol oynamaktadır [17].

Derin öğrenme, yüksek boyutlu ve karmaşık veri kümelerinden anlamlı temsil öğrenme sürecidir. Sinir ağlarının katmanlı yapısı, istatistikteki çok değişkenli dağılım analizinin ileri bir biçimi olarak yorumlanabilir [20].

- Merkezi Limit Teoremi: Büyük veri örneklerinde, dağılımların normal dağılıma yakınsaması, derin öğrenme optimizasyonunda gradyan yöntemlerinin istikrarını açıklar.

- Stokastik Gradyan İnişi (SGD): Rastgele seçilen küçük veri kümeleri üzerinden yapılan optimizasyon, olasılık teorisi ve istatistiksel örnekleme ilkelerine dayanır.
- Regülerizasyon: Aşırı öğrenmenin (overfitting) önlenmesi için kullanılan dropout, weight decay gibi yöntemler, istatistiksel model seçiminin karşılığıdır.

Derin öğrenme modellerinin en çok eleştirilen yönlerinden biri "kara kutu" olmalarıdır. Açıklanabilir yapay zekâ araştırmaları, istatistiksel belirsizlik kavramını yeniden öne çıkarır:

- Belirsizlik görselleştirmeleri, karar ağaçları, SHAP ve LIME gibi yöntemler, tahminlerin hangi değişkenlerden etkilendiğini gösterir.
- Bu, epistemik belirsizliğin azaltılması ve karar vericilerin modele güvenmesi için kritik bir adımdır.

Son yıllarda öne çıkan iki eğilim:

- Bayesçi Derin Öğrenme: Parametrelerin olasılık dağılımlarıyla modellenmesi, epistemik belirsizliğin doğrudan hesaplanmasını sağlar.
- Robust AI: Veri hataları, saldırılar (adversarial examples) ve beklenmeyen durumlara karşı dayanıklı yapay zekâ geliştirme çabası, klasik istatistiğin robust yöntemleriyle paralellik taşır.

Bu iki yaklaşım, istatistiksel belirsizliği doğrudan yönetmeye yönelik güncel araştırma alanlarıdır [21]. İstatistiksel düşünce, tarihsel olarak Bernoulli, Laplace, Gauss ve Fisher ile başlamış, bugün derin öğrenme ve yapay zekâ çağında yeniden merkezî bir rol kazanmıştır [22]. Olasılık, belirsizlik ve entropi, yapay zekânın en temel yapıtaşlarını oluşturmaktadır. Modern yapay zekâ araştırmaları, aslında istatistiksel düşüncenin genişlemiş hâlidir: belirsizliği yok etmeye değil, yönetilebilir hale getirmeye çalışır. Böylece yapay zekâ, yalnızca tahmin gücü yüksek algoritmalar değil, aynı zamanda belirsizlik karşısında rasyonel karar mekanizmaları sunan bütünleşik sistemler üretir [23].

7. Gelecek Perspektifleri

İstatistiksel düşünce ile yapay zekânın kesişiminde geleceğe dair ufuk açıcı pek çok yönelim ortaya çıkmaktadır. Öncelikle, nedensel yapay zekâ (causal AI) çalışmaları yalnızca korelasyonları değil, istatistiksel nedensellik modellerini [24], merkeze alarak, belirsizlik altında daha güvenilir karar mekanizmaları sunmayı amaçlamaktadır. Bunun yanında, Bayesçi derin öğrenme yaklaşımları [19], epistemik belirsizliği doğrudan modelleyerek, sağlık gibi yüksek riskli

alanlarda yapay zekâ uygulamalarının güvenilirliğini artırmaktadır. Diğer taraftan, kuantum makine öğrenmesi çalışmaları [25], olasılık kuramı ile kuantum hesaplamanın birleşimini kullanarak belirsizliği daha temel seviyede temsil etme potansiyeline sahiptir. Ayrıca, açıklanabilir yapay zekâ (XAI) yaklaşımları [22], modellerin karar süreçlerini şeffaflaştırarak epistemik belirsizlikleri azaltmaya odaklanmaktadır. Tüm bu gelişmeler, istatistiksel düşüncenin yalnızca tarihsel bir miras değil, aynı zamanda yapay zekânın evriminde yön verici bir ilke olduğunu ortaya koymaktadır. Geleceğin biliminde belirsizlik, klasik istatistikte olduğu gibi, yok edilmesi gereken bir sorun değil; aksine bilgi üretiminin ve öğrenmenin en verimli kaynağı olmaya devam edecektir.

Kaynakça

1. Knight, F. H. (1965). *Risk, uncertainty and profit*. Рипол Классик.
2. Kolmogorov, A. N. (2018). *Foundations of the theory of probability: Second English edition*. Courier Dover Publications.
3. Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.
4. De Finetti, B. (1974). Bayesianism: Its unifying role for both the foundations and applications of statistics. *International Statistical Review/Revue Internationale de Statistique*, 42(2), 117–130.
5. Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1), 1–13.
6. Efron, B., Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
7. Moore, D. S. (2010). *The basic practice of statistics*. Palgrave Macmillan.
8. Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.
9. Hume, D. (2016). An enquiry concerning human understanding. In Seven masterpieces of philosophy (pp. 183–276). Routledge.
10. Popper, K. (1979). *Objective knowledge*. Oxford University Press.
11. Carnap, R., Schilpp, P. A. (1963). *The philosophy of Rudolf Carnap* (pp. 3–84). Cambridge University Press.
12. De Finetti, B. (1936). La logique de la probabilité. In *Actes du congrès international de philosophie scientifique* (Fasc. IV, pp. 31–39). Paris: Hermann.
13. Box, G. E. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.
14. Clausius, R. (1865). Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie \[On different convenient forms of the fundamental equations of the mechanical theory of heat]. *Annalen der Physik*, 201(7), 353–400.
15. Boltzmann, L. (2006). *Ludwig Boltzmann 1844–1906: Eine Ausstellung der Österreichischen Zentralbibliothek für Physik*. Wien.
16. Wolff, S. L. (2013). Rudolph Clausius–A pioneer of the modern theory of heat. *Vacuum*, 90, 102–108.
17. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.

18. Bishop, C. M., Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, No. 4). Springer.

19. Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.

20. Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep learning*. MIT Press.

21. MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

22. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

23. Efron, B., Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

24. Pearl, J. (2009). *Causality*. Cambridge University Press.

25. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, 549(7671), 195–202.