İSTATİSTİK KURAMINA MATEMATİKSEL YAKLAŞIM

Mustafa AYTAÇ *

1- GİRİŞ

İstatistik bilimi, insanlığı ilgilendiren birbirinden farklı iki sahada aynı zamanlarda yayılmaya başlamıştır diyebiliriz. Bunlar şans oyunları ve politik bilimlerdir. Özellikle politikacıların istatistik bilgileri kendilerini haklı gösterecek şekilde — istatistik biliminin yöntemlerine bağlı kalmaksızın — yorumlamaları, onun bir tür yalan söyleme yöntemi olarak ün yapmasını sağlamıştır. Öyleyse istatistik nedir? [1]

"İstatistik, yığınların özellikleri üzerinde sayma veya ölçme işlemleri ile elde edilen verilerle ilgilenen bilimsel bir yöntemdir."

Şans oyunları üzerine yapılan çalışmalar, ölçme hataları üzerinde bir takım kuramlar geliştirerek matematiksel istatistiğin temelini oluşturmuştur. Politik bilimlerle ilgilenen istatistik ise, verilerin tablo ve diyagramlarla gösterilmesini sağlamış ve daha ilerdeki yıllarda değişik ortalamalarla verilerin özetlenmesini de içererek gittikçe gelişmiş ve betimleyici istatistiği (descriptive statistics) oluşturmuştur.

Son yıllarda istatistiğin gelişmesi betimleyici istatistikten, istatistik tahmine (Statistical Inference veya inductive statistics) kaydığı oranda hızlı olmuştur. İstatistik tahminleri, örnek verilerine dayanarak genellemelerde bulunmaktır.

Biz bu çalışmamızda istatistikte yer alan bazı konu ve kavramlar üzerinde durmaya çalıştık. Şüphesiz bunların dışında daha birçok konular vardır. Bunların hepsini böyle bir makalede ele almak olanaksızdır. Bizim amacımız herhangi bir matematiksel istatistik kitabı veya makalesi okuyacak kişilere bunların anlaşılmasını sağlamak üzere istatistik konularından önemli saydıklarımızı matematiksel bir temele dayanarak sunmaktır. Son yıllarda büyük gelişme göstererek hemen hemen bütün bilim dallarında yaygın bir kullanma alanı bulan parametrik olmayan istatistik yöntemlerini de bu çalışmamızın kapsamı dışında bıraktık. Çünkü parametrik olmayan istatistik yöntemlerinde, hem ana kütle dağılımının belirlenmesine gerek duyulmaz hem de bunları uygulamak için ileri düzeyde matematik ve istatistik bilgisi istenmez. Bu nedenden dolayı sadece parametrik istatistik yöntemleri üzerinde durduk.

Bunun yanında çalışmamızda sürekli tesadüfi değişken durumunu ele aldık. Süreksiz tesadüfi değişkenleri gözönüne almak istersek, makalemizdeki integral işa-

---

* *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Yardımcı Doçenti.*

1 Kendall, M.G. — Stuart, A., The Advanced Theory of Statistics Griffin Co., London, 1963, Vol. 1, s. 1.

retlerinin yerine toplama işaretini koymak yeterlidir. Bunun yanında f(x) olarak gösterdiğimiz olasılık yoğunluk fonksiyonlarını da p(x) ile yer değiştirmemiz gerekir.

2- TESADÜFİ DEĞİŞKENLERLE İLGİLİ KAVRAMLAR ve BAZI DAĞILIMLAR

Tesadüfi değişkenler olasılık uzayında belirlenmiş gerçek fonksiyonlardır. Bunlar genellikle X, Y ve Z gibi büyük harflerle gösterilirler. Tesadüfi değişkenlerin almış oldukları değerler ise aynı harflerin küçükleri ile belirtilir. Bununla beraber, amacımıza uygun olarak tesadüfi değişkeni, tesadüfi bir deney sonucunun bir fonksiyonu olarak da düşünebiliriz. X belli bir A olayı içindedir cümlesi ile olasılık kuramında sık sık ilgileniriz ve bunu $Pr(X \epsilon A)$ şeklinde gösteririz.

$F(x) = Pr(X \leq x)$, x'e kadar birikmiş olasılığı gösterir ve birikimli olasılık fonksiyonu olarak tanımlanır. Bu durumda

$$P(A) = Pr(X \epsilon A) = \int_A dF(x) \quad \text{dir.}$$

Eğer X sürekli bir tesadüfi değişkense $f(x) = F'(x)$ olasılık yoğunluk fonksiyonu olduğu yerlerde ,

$$P(A) = Pr(X \epsilon A) = \int_A f(x)dx \quad \text{dir.}$$

Yani bu olasılık $y = f(x)$ eğrisi ve x-ekseni üzerinde bulunan A seti ile arasında kalan alandır. Bu aynı zamanda A setindeki x-değerlerinin altında bulunan yığının yüzdesi olarak da düşünülebilir.

X tesadüfi bir değişken ve U(X)'de X'in bir fonksiyonu olsun. U(X)'in beklenen değeri (veya matematiksel ümidi)

$$E[u(X)] = \int_{-\infty}^{+\infty} u(x)dF(x)$$

$$= \int_{-\infty}^{+\infty} u(x)f(x)dx \quad \text{dir.}$$

X tesadüfi değişkeni dağılımının iki önemli karakteristiği vardır. Bunlar ortalama $\mu = E(X)$ ve varsayans $\sigma^2 = E[(X - \mu)^2]$ dir. Varsayans Var (X) veya V(X) ile de gösterilir ve şöyle hesaplanır.

$$V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 . f(x)\, dx$$

Varsayansın kare kökü σ, X'in standart sapması olarak bilinir. X'in karakteristik fonksiyonu olan $\varphi(t)$, f(x)'in bir Fourier dönüşümüdür ve aşağıdaki gibi tanımlanır.

$$\varphi(T) = E(e^{itX}) = \int_{-\infty}^{+\infty} e^{itX} f(x)\, dx$$

Dönüşüm teorisinde $\varphi(t)$'nin tanımı f(x) veya F(x) ile eşdeğerdir. Bu ise dönüşüm kuramına dağılım kuramı içinde büyük bir kullanma alanı ve geniş yararlar sağlar.

Normal ve Ki-Kare dağılımları istatistikte önemli olan iki sürekli dağılımdır. Normal dağılımın şekli çan eğrisi gibi olup olasılık yoğunluk fonksiyonu şöyledir.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Burada μ ve σ parametreleri sırasıyla ortalama ve standart sapmayı gösterir. Eğer bir tesadüfi değişkenin sahip olduğu yoğunluk fonksiyonu normalse, bunu $x \approx N(\mu, \sigma^2)$ şeklinde de gösterebiliriz. $\mu = 0$ ve $\sigma = 1$ ise bu dağılım standart normal dağılım olarak bilinir ve N(0, 1) olarak gösterilir.

Ki-Kare dağılımının sahip olduğu olasılık yoğunluk fonksiyonu ise şöyledir[2].

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\frac{r}{2})\, 2^{r/2}}\, x^{\frac{r}{2}-1}\, e^{-\frac{x}{2}} & 0 < x < \infty \\ 0 & \text{Diğer yerlerde} \end{cases}$$

Bu dağılımın kısaca gösterilişi ise $X^2_{(r)}$ şeklindedir. r parametresi artı değerli bir tamsayı olup serbestlik derecesi olarak da bilinir. Ki-Kare dağılımında $\mu = r$ ve $\sigma^2 = 2r$'dir.

İki veya daha fazla tesadüfi değişkene sahip olduğumuzu düşünelim. Aynı olasılık uzayında tanımlanmış (X, Y) tesadüfi değişkenleri çifti, eğer f(x, y) gibi bir fonksiyon kartezyen düzlemi üzerinde tanımlanmış bir A olayı için

$$\Pr[(X, Y) \in A] = \int_A \int f(x, y)\, dx\, dy$$

eşitliğini sağlıyorsa, bileşik sürekli tesadüfi değişken sayılır ve f(x, y)'de bileşik sıklık fonksiyonu adını alır. Bu olasılık, kartezyen düzlemi üzerinde olup, Z = f(x, y)'nin altında kalan toplam hacim bire eşittir. U(X, Y)'nin beklenen değeri ise şöyle tanımlanır,

$$E[U(X, Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} u(x, y)\, f(x, y)\, dx\, dy$$

Bileşik dağılımının önemli karakteristikleri ise şunlardır:

ortalamaları ; $\mu_X = E(X)$ ve $\mu_Y = E(Y)$

varyansları ; $\sigma^2_X = V(X) = E[(X - \mu_X)^2]$, $\mu_Y = V(Y) = E[(Y - \mu_Y)^2]$

ve kovaryansı; $\sigma_{XY} = Kov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

X ve Y'nin korelasyon katsayısı ise,

$$P_{XY} = \frac{\sigma_{XY}}{\sigma_X\, \sigma_Y} \quad \text{dir.}$$

---

2 Formüldeki Γ (r/2) gamma fonksiyonu olup, olasılık yoğunluk fonksiyonu şöyledir:

$$f(x; \alpha, \beta) = \frac{1}{\alpha' . \beta^{\alpha+1}}\, x^{\alpha}\, e^{-\frac{x}{\beta}}, \quad 0 < x < \infty$$

$\alpha > 1$ ve $\beta > 0$ olup bu fonksiyonun parametreleridir.

X ve Y'nin varyans-kovaryans matrisi olarak tanımlanan matris ise [3]

$$\begin{bmatrix} \sigma^2_X & \sigma_{XY} \\ \sigma_{XY} & \sigma^2_Y \end{bmatrix} \quad \text{dir.}$$

$X' = (X_1, X_2, ..., X_n)$ olduğu yerlerde, X'in önemli bir p-değişkenli dağılımı çok değişkenli normal dağılım olarak bilinir ve onun sahip olduğu olasılık yoğunluk fonksiyonu şöyle tanımlanır:

$$\frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}} \exp\left[-\frac{(X-\mu)\Sigma^{-1}(X-\mu)}{2}\right]$$

μ ve Σ; ortalama ve varyans-kovaryans matrisleridir. Bu dağılım $N_p(\mu, \Sigma)$ şeklinde de gösterilir.

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)dy \quad \text{ve} \quad f_2(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

fonksiyonları sırasıyla X ve Y'nin marjinal dağılımlarıdır. Y = y verildiği zaman, X'in koşullu olasılık yoğunluk fonksiyonu

$$f_1(x/y) = \frac{f(x,y)}{f_2(y)}$$

ve X = x verildiği zaman, Y'nin koşullu olasılık yoğunluk fonksiyonu ise

$$f_2(y/x) = \frac{f(x,y)}{f_1(x)} \quad \text{dir}^{4}.$$

---

3 Eğer n tane X, Y, Z ..... gibi tesadüfi değişken varsa bunların varyans-kovaryans matrisi şöyledir:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

4 Genel olarak $X_1, X_2, ..., X_n$ tesadüfi değişkenler ve bu tesadüfi değişkenlerin bileşik olasılık yoğunluk fonksiyonu da $f(x_1, x_2, ..., x_n)$ olsun. $X_{k+1} = x_{k+1}, ..., X_n = x_n$'nin bileşik koşullu beklenen ise,

$$f(x_{k+1}, ... x_n | x_1, ... x_k) = \frac{f(x_1, x_2, ..., x_k, x_{k+1}, ..., x_n}{f(x_1, x_2, ..., x_k)} \quad \text{dır.}$$

işaretlerinin sayısı (n − k) tane ve $f(x_1, x_2, ..., x_k) > 0$ olmak üzere şöyledir:

P-değişkenli normal durumunda, her marjinal olasılık yoğunluk fonksiyonu uygun ortalama ve varyansa sahip olarak normaldir. Bunun yanında verilen $X_j = x_j$, $j \neq i$ için X_i'nin koşullu olasılık yoğunluk fonksiyonuda normaldir.

U(X), X'in bir fonksiyonu olsun. Y = y verildiği zaman U(X)'in koşullu beklenen değeri,

$$E[u(X) \mid y] = \int_{-\infty}^{+\infty} u(x).f(x \mid y)\, dx \quad {}^{5}.$$

ve koşullu varyansı

$$\begin{aligned} \text{Var}[u(X) \mid y] &= E[X - E(X \mid y)]^2 \\ &= E(X^2 \mid y) - [E(X \mid y)]^2 \\ &= \int_X [y - E(X \mid y)]^2 . f(y \mid x)\, dy \quad \text{dir.} \end{aligned}$$

Benzer şekilde g(Y), Y'nin bir fonksiyonu olsun. X = x verildiği zaman g(Y)'nin koşullu beklenen değeri ve varyansı şöyledir.

$$E[g(Y) \mid x] = \int_{-\infty}^{+\infty} g(y) . f(y \mid x)\, dy$$

$$\text{Var}[g(Y) \mid x] = \int_Y [y - E(Y \mid x)]^2 . f(y \mid x)\, dy$$

Eğer f(x,y,z), X, Y ve Z tesadüfi değişkenlerinin bileşik olasılık yoğunluk fonksiyonu ise bu üç değişkeni bağımsız olması için gerekli ve yeterli koşul

$$f(x,y,z) = f_1(x) . f_2(y) . f_3(z) \quad \text{olmasıdır.}$$

Buradaki $f_1(x)$, $f_2(y)$ ve $f_3(z)$ sırasıyla X, Y ve Z'nin marjinal olasılık fonksiyonlarıdır. Gerçekte bu üç marjinal dağılım aynı ise, o zaman bu üç tesadüfi değişken birbirleri ile ikişer ikişer bağımsızdır denir.

İstatistikte zor gibi görünen konulardan birisi de bir, iki veya daha fazla tesadüfi değişkenin dağılım fonksiyonunu bulmaktır. Gerçekte kuramsal olarak bunu

---

$$f(x_1, x_2, \ldots, x_k) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f(x_1, x_2, \ldots, x_k, x_{k+1}, \ldots x_n)\, dx_{k+1} \ldots dx_n \quad k = 1$$

için bu formül $f_2(x_2, x_3, \ldots, x_n \mid x_1)$ ve $f_1(x_1)$'i verir.

5 Genel durumda incelersek; $X_1 = x_1$, $X_2 = x_2$, ..., $X_k = x_k$ verildiği zaman $u(X_{k+1}, \ldots, X_n)$'nin koşullu beklenen değeri integral işaretlerinin sayısı (n-k) tane ve $f(x_1, \ldots, x_k) > 0$ olmak üzere şöyledir:

$$E[U(X_{k+1}, \ldots X_n) \mid x_1, \ldots, x_k] = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} u(x_{k+1}, \ldots, x_n)$$

$$[f(x_{k+1}, \ldots, x_n \mid x_1, \ldots, x_k)\, dx_{k+1}, \ldots, dx_n]$$

$k = 1$ için bu formül, $E[U(X_2, \ldots, X_k, \ldots, X_n) \mid x_1]$'i verir.

elde etmek çok kolaydır. Z = u(X,Y)'nin X ve Y tesadüfi değişkenlerinin bir fonksiyonu olduğunu ve aynı zamanda da f(x,y) olasılık yoğunluk fonksiyonuna sahip olduğunu varsayalım. O zaman Z'nin dağılım fonksiyonu

$$G(Z) = Pr(Z \leq z) = \int_A \int f(x,y)\, dx\, dy \quad \text{dir.}$$

Tanımlamadaki A-seti $u(x,y) \leq Z$ tarafından tanımlanmış olup kartezyen düzlemindedir. Özel durumlarda G(Z)'yi veya g(z) ile bağlantılı olarak g(z) = G'(Z)'yi bulmak matematiksel istatistikte daha fazla bir çaba gerektirir. Buna üç tane önemli örnek gösterilebilir.

a) Eğer X ve Y sırasıyla N(0,1) ve $\chi^2_{(r)}$ dağılımlarına sahip olup bunlar bağımsız iseler o zaman,

$$T = \frac{X}{\sqrt{Y/r}}$$

bir t-dağılımına sahiptir ve serbestlik derecesi de r'dir.

b) Eğer X ve Y bağımsız olup, dağılımları $\chi^2_{(r_1)}$ ve $\chi^2_{(r_2)}$ ise, o zaman

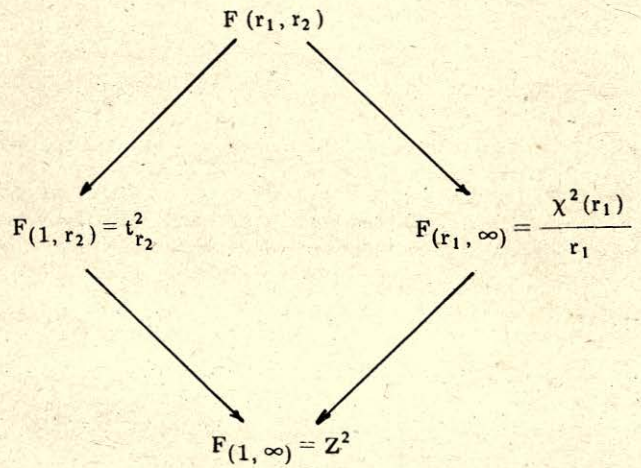
$$F = \frac{(X / r_1)}{(Y / r_2)}$$

r_1 ve r_2 serbestlik derecelerinde bir F dağılımıdır.

c) Eğer X bir P-tesadüfi değişkenli normal dağılıma sahipse $N_p(\mu, \Sigma)$, c sabit sayıların bir vektörü olmak üzere; c'X,ortalaması c'μ ve varyansı c'Σc olacak şekilde bir normal dağılım gösterir. Yani c'X, N(c'μ, c'Σc) dir. Aynı zamanda $(X - \mu)'\ \Sigma(X - \mu)$ P-serbestlik derecesinde bir Ki-Kare dağılımıdır. Başka bir anlatımla

$$(X - \mu)'\ \Sigma^{-1}\ (X - \mu) \approx \chi^2_{(p)} \quad \text{dir.}$$

F, χ^2, t ve Z tesadüfi değişkenleri arasında daha önce de değindiğimiz gibi kuvvetli bir ilişki vardır. r_1 ve r_2 serbestlik dereceleri olmak üzere bunu bir şema yardımı ile göstermeye çalışalım.



Başka bir ifade ile belli bir α-önem seviyesinde F-dağılımı $r_1 = 1$ ve r_2 serbestlik derecesine sahip olan t-dağılımının tablodan bulunan değerinin karesine eşittir. Bu ilişki şemadaki koşullar içerisinde χ^2 ve Z dağılımı içinde geçerlidir.

3. İSTATİSTİK TAHMİNLERİ

Örnek değerlerinden giderek ana kütle hakkında bir takım önerilerde bulunma yöntemi, bazı istatistik karar kuramlarına önderlik eder. Örnek ile ana kütle arasındaki bu tür bir ilişkiye kısaca istatistik tümevarım denir. İstatistik bilimi son senelerdeki hızlı gelişmesini, istatistik tümevarım yöntemlerinin hemen hemen her alanda büyük bir etkinlikle kullanma olanağı bulmasına borçludur, diyebiliriz.

Eğer $X_1, X_2, \ldots\ldots, X_n$ tam bağımsız iseler, o zaman bütün n değişkenin ortak marjinal dağılımı olan f(x) dağılımının, tesadüfi bir örneğinin elemanları olduğunu söyleriz. Tesadüfi bir örnek olan $X_1, X_2, \ldots\ldots, X_n$ elemanlarının gözlenen değerleri $x_1, x_2, \ldots\ldots, x_n$ olsun. Her x_i'ye (1/n) ağırlığı koyarak bir olasılık dağılımı elde ederiz. Bunlarla bağlantılı olan $F_n(x)$, gözlenen dağılım fonksiyonu olarak bilinir ve süreksiz bir dağılım özelliği gösterir. Bu süreksiz dağılımın ortalama ve varyansı sırasıyla şöyledir:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad , \qquad S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Bunlar daha çok gözlenen bir örneğin ortalama ve varyansı olarak bilinirler. Çünkü bağlantılı oldukları tesadüfi değişkenler X ve S^2 olup

$$E(X) = \mu \qquad \text{ve} \qquad E(S^2) = \frac{n-1}{n} \sigma^2 \quad \text{dir.}$$

S^2, σ^2'nin taraflı bir tahmini olurken; $\bar{X}$, μ'nun tarafsız bir tahminidir. σ^2'nin tarafsız tahminini elde etmek amacı ile birçok yazar örneğin varyansı olarak aşağıdaki formülü tanımlamışlardır.

$$S^2 = \frac{nS^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Şüphesiz iyi bir tahmincinin tarafsızlık ilkesi yanında etkinlik, tutarlılık ve yeterlilik koşullarına da uyması gerekir.

Daha genel olarak $Z = u(X_1, X_2, \ldots\ldots, X_n)$ gibi örnek birimlerinin bir fonksiyonu olan bir istatistiğe sahipsek ve $Z = u(x_1, x_2, \ldots\ldots, x_n)$ olarak gözlenen Z değerleri, θ gibi bir parametreye yakınlaşmayı sağlıyorsa; o zaman Z, θ'nın iyi bir tahminidir. Bu yakınlığı elde etmek için Z'nin küçük taraflı (veya tam tarafsız) ve küçük varyansa sahip olmasını amaçlarız. Bunu sağlamak için iki yol vardır.

i) Z, θ'nın bütün tarafsız tahminleri arasında minimum varyanslı tarafsız bir tahminidir veya

ii) Z, E $[(Z - \theta)^2]$'yı minimum yapan bir istatistiktir. Başka bir ifade ile Z, θ nın minimum varyanslı sapmasız bir tahminidir. Bir tahmincinin daha küçük varyansa sahip olup olmadığını ise belirli varsayımlar altında Rao-Blackwell teoremi yardımı ile sağlayabiliriz.

Nokta tahmininde iyi yöntemlerden birisi de maximum benzerlik yöntemidir. Örneğin, θ'nın bilinmeyen $f(x/\theta)$ gibi bir dağılımdan çıktığını varsayalım. θ'nın bir fonksiyonu olarak düşünülen $X_1, X_2, \ldots\ldots, X_n$'nin bileşik olasılık yoğunluk fonksiyonu, $L(\theta)$ benzerlik fonksiyonu olarak bilinir ve aşağıdaki gibi tanımlanır.

$$L(\theta) = f(x_1/\theta) \cdot f(x_2/\theta) \ldots\ldots f(x_n/\theta)$$

Bu ilişkiden bulunan θ, $L(\theta)$'yı maximum değere sahip kılar ve bileşik olasılık yoğunluk fonksiyonu θ'nın maximum tahmincisi olarak bilinir ve genellikle θ olarak gösterilir. Normal dağılım için,

$$\hat{\mu} = \bar{X} \qquad \text{ve} \qquad \sigma^2 = S^2 \quad \text{Dir.}$$

İstatistikçiler bir θ parametresinin nokta tahmininden giderek daha fazla bilgi sağlama yolunu bulmaya çalışırlar. Başka bir anlatımla, nokta tahmini ile birlikte onlar bu tahminin hata yapısının bir ölçüsünü de birlikte vermeye çalışacaklardır. Daha iyi açıklayabilmek amacı ile normal dağılım durumunda $\hat{\mu} = \bar{X}$' dır. Fakat $T = \sqrt{n}(\bar{x} - \mu)/S_1$'in, $n - 1$ serbestlik derecesinde bir t-dağılımına sahip olduğu açıklıkla gösterilebilir. Eğer t_{tab} (t-tablodeğeri);

$$\Pr(|T| \leq t_{tab}) \leq 0.95$$

şeklinde seçilirse, o zaman $X \mp t_{tab} S_1/\sqrt{n}$ alt ve üst değerleri ile tanımlanan aralık bilinmeyen ortalama μ'ı 0.95 olasılıkla içerir. Çünkü

$$|T| \leq t_{tab}, \ \bar{X} - t_{tab} S_1/\sqrt{n} \leq \mu \leq \bar{X} + t_{tab} S_1/\sqrt{n}$$

ile eşdeğerdedir. Böylece gözlenen aralık $\bar{x} \mp t_{tab} s_1/\sqrt{n}$ için 0.95 güven aralığını ihtiva eder. Yani $(t_{tab} s_1)/\sqrt{n}$ (veya yalnızca $S_1/\sqrt{n}$), X tahmininin hata yapısının bir ölçüsünü içerir. Güven aralıklarını ana kütlenin değerleri olan oranlar ve varyanslar için oluşturabilir ve benzer değerlendirmeler yapabiliriz.

İstatistik tümevarımda en önemli aşamalardan birisi tahmin ise bir diğeri de istatistik hipotez testleridir. İstatistiksel hipotez olasılığın dağılımı üzerine kurulmuş bir cümledir. Örneğin $H_0 : \mu = \mu_0$. H_0'ın testi öyle bir yöntemdir ki, H_0', red veya kabul etmemiz ana kütleden alınan tesadüfi örnek üzerine kurulmuştur. H_0'ı reddetme düşüncemiz, aklımızda $H_1 : \mu > \mu_0$ gibi bazı almaşık hipotezlere sahip olduğumuzu gösterir. Normal örneğimizde $H_0 : \mu = \mu_0$ veya $H_1 : \mu = \mu_1$'ı kabul edebilmemiz gözlenen örnek ortalaması $\bar{x}$'ın μ_0 veya μ_1'e yakın olup olmamasına bağlıdır. Böyle bir yöntemde açıkca iki tip hatadan birisini yapabiliriz. H_0 doğru olduğu zaman H_0'ı reddeden I. tip hata ve H_0 yanlış olduğu zaman H_0'ı kabul eden II. tip hata. Bu iki tip hatanın olasılıklarını sırasıyla α ve β ile gösteririz. $1 - \beta$, deki testin gücü olarak bilinir. α ise testin önem (anlamlılık) seviyesi olup, araştırıcının ne ölçüde iddialı olduğunun bir göstergesidir.

Normal örneğimize dönersek $H_0 : \mu = \mu_0$ istatistik hipotezimizi iki yanlı almaşık hipotez olan $H_1 : \mu \neq \mu_0$'a ($\mu > \mu_0$ veya $\mu < \mu_0$ olabilir) karşı test etmek istediğimizi düşünelim. H_0 doğruluğunu içeren $n - 1$ serbestlik derecesi ile t-dağılımına sahip $|T| = \sqrt{n}(\bar{X} - \mu_0)/S_1$ istatistiğini düşünelim.

$$\Pr(|T| \leq t_{tab} \mid H_0) \leq \alpha/2$$

olacak şekilde bir t_{tab} -kritik değeri buluruz. Eğer gözlenen $|T| \geq |t_{tab}|$ ise, H_0'ı reddedip H_1 hipotezini α-önem seviyesinde kabul ederiz. Daha doğru bir deyimle, H_0 hipotezini kabul etmemiz için yeterli nedenimiz yoktur deriz.

Büyük ölçüde hipotez testlerinden yararlanan ve günümüzde uygulamada en çok kullanılan yöntemlerden birisi de varyans analizleridir. Birkaç cümle ile de buna değinmeye çalışalım.

Bir bağımlı değişken üzerinde etkide bulunan bağımsız değişkenlerin etkilerini karşılaştırmak amacı ile kurduğumuz modellerde varyans analizi tekniğinden yararlanırız. Varyans analizinin uygulandığı modellerde bir veya birkaç faktör vardır. Ve bu analiz faktörlere maledilecek etkinin önemli olup olmadığını ortaya çıkarır. Bunun için faktörlerin uygulanacağı yığından tesadüfi olarak seçilmiş bulunan birimler gruplara ayrılır ve grupların herbirisine faktörlerin değişik düzeylerini olası kombinasyonlarından biri uygulanarak, grup ortalamalarının birbirine eşit olduğu şeklinde oluşturulan sıfır hipotezi (H_0) test edilir.

İstatistikte tahmin yöntemlerinden birisi de regresyon yönetimidir. Bir bağımlı değişkenle bir veya birkaç açıklayıcı değişken arasında kurulan istatiksel modeldeki bilinmeyen katsayıları tahmin ederek, açıklayıcı değişkenlerin belirlenen değerleri için bağımlı değişkenin alacağı değeri tahmin etmeye regresyon problemi adı verilir. Çok değişkenli doğrusal modelimiz ise $Y = X\beta + \epsilon$'dur. Formülde Y elemanları, Y tesadüfi değişkenine ait gözlem değerlerinden oluşan (n x 1), X ilk sütunu bir rakamlarından ve diğer sütunları da açıklayıcı değişkenlere ait gözlem değerlerinden oluşan (n x p), β parametrelerden oluşan (p x 1) ve ϵ'da hata vektörlerinden oluşan (n x 1)'inci dereceden matrisleri gösterir. Bir açıklayıcı ve bir bağımlı değişken oluşan regresyon modelimizde ise, X'lere ait gözlemlerden oluşan (n x 2)'inci dereceden bir matrisdir. β'nın elemanlarının herbirisine regresyon katsayısı adı verilir. Bu modeldeki varsayımlarımız ise şunlardır;

I) $E(\epsilon) = 0$
II) $Var(\epsilon) = I\sigma^2$
III) $E(\epsilon_i \epsilon_j) = 0 \quad i \neq j$ için
IV) X'in rankı n'dir ve
V) X sabit sayılar kümesidir.

β'nın tahmini olan b aşağıdaki eşitlikten bulunur.

$$b = [(X'X)^{-1} XY]$$

Bu eşitlikten bulunan b aynı zamanda şu önemli özelliklere de sahiptir. 1) Hataların kareler toplamını minimum yapar. 2) b'nin elemanları, $Y_1, Y_2, ..., Y_n$ gözlemlerinin doğrusal fonksiyonlarıdır ve minimum varyanslı β'nın elemanlarının tarafsız tahminleridir. 3) Eğer hatalar bağımsız ve $\epsilon_i \approx N(O, \sigma^2)$ ise o zaman b, β'nın maximum benzerlik tahminidir. Bunun yanında $E(b) = \beta$ ve $Var(b) = \sigma^2_{\epsilon} (X'X)$'dür.

Regresyon doğrusunun gözlemlere ne ölçüde uyduğunu ortaya koyan göstergelerden birisi de determinasyon katsayısı R^2'dir ve şöyle bulunur.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \frac{(..... - n\bar{Y})^2}{(..... - n\bar{Y})^2}$$

Regresyonda kullanılan modeller, parametrelerin durumuna göre doğrusal ve doğrusal olmayan modeller olmak üzere iki gruba ayrılabilir. Ayrıca doğrusal kılınabilenler ve kılınamayanlar olarak da iki gruba ayrılabilir. Doğrusal hale dönüştürülebilen modeller için yukarıdaki ilkeler geçerli olur ki, bunlar da üstel fonksiyon, hiperbol, polinom modelleri ve harmonik fonksiyondur.

KAYNAKLAR

Brunk, H.D.; An Introduction to Mathematical Statistics Blaisdell Publ. Comp. 2nd. Ed., Toronto, 1965.

Draper, N.R. and Smith H.; Applied Regression Analysis, Wiley, New York, 1966.

Freund, E. John and Walpole, Ronald E.; Mathematical Statistics, Prentice-Hall, 3rd. Ed., New Jersey, 1980.

Hogg, R.V. and Craig A.T.; Introduction to Mathematical Statistics, Mac Millan, 3rd Ed., Hong-Kong, 1972.

Işıkara, Bâki; Regresyon Yöntemleri ve Sorunları, İktisat Fakültesi Yayınları, İstanbul, 1975.

Kendall, M.G. and Stuart, A.; The Advanced Theory of Statistics, Griffin Co., London, 1963, Vol I. s. 1.

Korum, Uğur; Matematiksel İstatistiğe Giriş, T.O.D.A.İ.E., Ankara, 1977.

Mendanhall, William; Introduction to Linear Models and The Design and Analysis of Experiments, Wadsworth Publishing Comp., Belmont, 1968.

Taylor, Lester D.; Probability and Mathematical Statistics, Harper and Row, New York, 1974.

Studies in Mathematics; The Mathematical Association of America, U.S.A., 1978.