HAFTA 4

6.3. Normal Dağılımdan Örneklem

Merkezi limit teoremine göre bazı koşullar altında, kitle dağılımı ne olursa olsun n örneklem hacmi yeterince büyükse, örneklem ortalaması dağılımda normal dağılıma yaklaşmaktadır. Bu kısımda, $X_1, X_2 \ldots, X_n$ $N(\mu,\sigma^2)$ dağılımından alınan $X_1, X_2 \ldots, X_n$ örnekleminin bazı özellikleri incelenecektir. Her bir i için $X_i \sim N(\mu,\sigma^2)$ olup olasılık yoğunluk fonksiyonu,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right),\ x \in \mathbb{R}$$

şeklindedir.

Teorem 6.3.1 $X_1, X_2, \ldots, X_n$ ler $N(\mu,\sigma^2)$ dağılımlı bağımsız rasgele değişkenler olsun. Bu durumda,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{ve} \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

olmak üzere,

a) $\bar{X}_n \sim N(\mu, \sigma^2/n)$ b) $\bar{X}_n$ ile S_n^2 bağımsız c) $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$

dır.

İspat: a) Bağımsız normal dağılıma sahip rasgele değişkenlerin toplamının da normal olduğunu biliyoruz. O halde, $\bar{X}_n$ nin beklenen değeri ile varyansını hesaplamak yeterlidir. $E(\bar{X}_n) = \mu$ ve $Var(\bar{X}_n) = \sigma^2/n$ olduğundan $\bar{X}_n \sim N(\mu, \sigma^2/n)$ dir.

b) Bu ifadenin değişik kaynaklarda değişik ispatları vardır (Öztürk ve digerleri (2006) teoremin ispatını beş farklı yoldan vermiştir). X ve Y iki rasgele değişken olmak üzere, ortak olasılık yoğunluk fonksiyonu marjinal olasılık yoğunluk fonksiyonlarının çarpımı olarak yazılabilirse, X ve Y bağımsızdır. Örneklem ortalamasının tanımından

$$\sum_{i=1}^{n}(X_i - \bar{X}_n) = 0 \Rightarrow X_1 - \bar{X}_n = -\sum_{i=2}^{n}(X_i - \bar{X}_n)$$

yazılır. Ayrıca,

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \left[(X_1 - \bar{X}_n)^2 + \sum_{i=2}^{n}(X_i - \bar{X}_n)^2 \right] = \left[\left(\sum_{i=2}^{n}(X_i - \bar{X}_n) \right)^2 + \sum_{i=2}^{n}(X_i - \bar{X}_n)^2 \right]$$

eşitliğinden, $S_n^2 = h(X_2 - \bar{X}_n,\ X_3 - \bar{X}_n,\ \ldots,\ X_n - \bar{X}_n)$ şeklinde yazılabilir. Yani, S_n^2 sadece $X_2 - \bar{X}_n,\ X_3 - \bar{X}_n,\ X_4 - \bar{X}_n,\ \ldots,\ X_n - \bar{X}_n$ lerin bir fonksiyonudur. Böylece, $\bar{X}_n$ ile S_n^2 nin bağımsız olduğunu göstermek için,

$\bar{X}_n$ ile $\left(X_2 - \bar{X}_n,\ X_3 - \bar{X}_n,\ X_4 - \bar{X}_n,\ \ldots,\ X_n - \bar{X}_n\right)$

nin bağımsız olduğunu göstermek yeterlidir. $X_1, X_2, \ldots, X_n$ rasgele değişkenlerinin ortak olasılık yoğunluk fonksiyonu $\mu = 0$ ve $\sigma^2 = 1$ için,

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right)$$

şeklindedir. Şimdi,

$$Y_1 = \bar{X}_n \ ,\ Y_2 = X_2 - \bar{X}_n,\ Y_3 = X_3 - \bar{X}_n, \ldots,\ Y_n = X_n - \bar{X}_n$$

dönüşümlerini tanımlayalım. Böylece, $\bar{X}_n$ ile S_n^2 nin bağımsız olduğunu göstermek için Y_1 ile $(Y_2, Y_3, ..., Y_n)$ rasgele değişkenlerinin bağımsız olduğunu göstermek yeterli olacaktır. Buradan

$$\begin{aligned} nY_1 &= X_1 + X_2 + \ldots + X_n = X_1 + (Y_1 + Y_2) + (Y_1 + Y_3) + ... + (Y_1 + Y_n) \\ &= X_1 + (n-1)Y_1 + (Y_2 + Y_3 + ... + Y_n) \end{aligned}$$

olduğundan ters dönüşümler

$$X_1 = Y_1 - (Y_2 + Y_3 + ... + Y_n),\ X_2 = Y_1 + Y_2,\ X_3 = Y_1 + Y_3,\ X_4 = Y_1 + Y_4, ..., X_n = Y_1 + Y_n$$

şeklindedir. Bu dönüşümlere ait Jacobien matrisi ve determinantı

$$J = \begin{bmatrix} 1 & -1 & -1 & . & . & . & -1 \\ 1 & 1 & 0 & . & . & . & 0 \\ 1 & 0 & 1 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & 0 & 0 & . & . & . & 1 \end{bmatrix}, \ \det(J) = n$$

olup $(Y_1, Y_2, Y_3, ..., Y_n)$ rasgele değişkenlerinin ortak olasılık yoğunluk fonksiyonu $c = n(2\pi)^{-n/2}$ için,

$$f(y_1, y_2, ..., y_n) = c \exp\left(-\frac{1}{2}\left(y_1 - \sum_{i=2}^{n} y_i \right)^2 \right) \exp\left(-\frac{1}{2} \sum_{i=2}^{n} (y_1 + y_i)^2 \right)$$

$$= \left[\frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left(-\frac{y_1^2}{2} \right) \right] \left\{ \frac{\sqrt{n}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left(\left[\sum_{i=2}^{n} y_i^2 + \left(\sum_{i=2}^{n} y_i \right)^2 \right] \right) \right) \right\}$$

$$= f_{Y_1}(y_1)\, f_{Y_2,\ldots,Y_n}(y_1, \ldots, y_n)$$

şeklindedir. Dolayısı ile, Y_1 ile $(Y_2, Y_3, ..., Y_n)$ rasgele değişkenleri bağımsızdır. Yani,

$\bar{X}_n$ ile $\left(X_2 - \bar{X}_n,\ X_3 - \bar{X}_n,\ X_4 - \bar{X}_n,\ \ldots,\ X_n - \bar{X}_n \right)$

rasgele değişkenleri bağımsızdır. Buradan da,

$\bar{X}_n$ ile $\left(X_1 - \bar{X}_n,\ X_2 - \bar{X}_n,\ X_3 - \bar{X}_n,\ X_4 - \bar{X}_n,\ \ldots,\ X_n - \bar{X}_n \right)$

rasgele değişkenlerinin bağımsızlığı elde edilir. Dolayısı ile,

$\bar{X}_n$ ile $\{(X_1 - \bar{X}_n)^2, (X_2 - \bar{X}_n)^2, (X_3 - \bar{X}_n)^2,\ \ldots,\ (X_n - \bar{X}_n)^2\}$

rasgele değişkenleri bağımsız olup,

$$\bar{X}_n \text{ ile } \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \text{ veya } \bar{X}_n \text{ ile } S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

rasgele değişkenlerinin bağımsız olduğu elde edilmiş olur.

c) Teoremin bu kısmının da değişik şekilde ispatlanabilir. Örneğin, çok değişkenli normal dağılımın özelliklerinden teoremin ispatı yapılabilir. Ayrıntılı bilgi için herhangi bir lineer model kitabında karesel formların dağılımına bakılabilir. Bunun için tümevarım tekniğini kullanalım. İşlemlerin kolay yürütülebilmesi için, $\sigma^2 = 1$ alalım ve iddianın $n = 2$ için doğru olduğunu gösterelim. $n = 2$ ise $(n-1) S_n^2$,

$$(n-1) S_n^2 = S_2^2 = \left(X_1 - \frac{X_1 + X_2}{2} \right)^2 + \left(X_2 - \frac{X_1 + X_2}{2} \right)^2 = \frac{1}{2} (X_1 - X_2)^2 = \left(\frac{X_1 - X_2}{\sqrt{2}} \right)^2$$

olarak yazılabilir. Buradan da,

$$\frac{X_1 - X_2}{\sqrt{2}} \sim N(0,1) \ \Rightarrow \left(\frac{X_1 - X_2}{\sqrt{2}} \right)^2 \sim \chi_1^2 \Rightarrow (2-1) S_2^2 \sim \chi_{(2-1)}^2 \Rightarrow \frac{(2-1) S_2^2}{\sigma^2} \sim \chi_{(2-1)}^2$$

elde edilir. Yani, iddia $n=2$ için doğrudur. Şimdi, iddia $n=k$ için doğru olsun. Yani, $(k-1)S_k^2/\sigma^2 \sim \chi_{k-1}^2$ olsun ve iddianın $n=k+1$ için doğru olduğunu gösterelim. Bunun için,

$$(n-1)S_n^2 = (n-2)S_{n-1}^2 + \left(\frac{n-1}{n}\right)\left(X_n - \bar{X}_n\right)^2$$

eşitliği (Bkz. Problem (6.5.1)) $n=k+1$ için

$$k\,S_{k+1}^2 = (k-1)S_k^2 + \left(\frac{k}{k+1}\right)\left(X_{k+1} - \bar{X}_k\right)^2$$

olarak yazılır. Varsayımdan $(k-1)S_k^2 \sim \chi_{k-1}^2$ olup X_{k+1} ile $\bar{X}_k$ bağımsızdır. Ayrıca,

$$X_{k+1} - \bar{X}_k \sim N\left(0, \frac{k+1}{k}\right) \Rightarrow \left(\frac{k}{k+1}\right)\left(X_{k+1} - \bar{X}_k\right)^2 \sim \chi_1^2$$

dir. Teoremin (b) kısmından $\bar{X}_k$ ile S_k^2 bağımsız olup

$$(k-1)S_k^2 \quad \text{ile} \quad \left(\frac{k}{k+1}\right)\left(X_{k+1} - \bar{X}_k\right)^2$$

rasgele değişkenleri de bağımsızdır. Bağımsız ki-kare dağılımına sahip rasgele değişkenlerin toplamı yine *ki-kare* dağılımına sahip olduğundan

$$k\,S_{k+1}^2 = (k-1)S_k^2 + \left(\frac{k}{k+1}\right)\left(X_{k+1} - \bar{X}_k\right)^2 \sim \chi_{k-1}^2 + \chi_1^2 = \chi_k^2$$

elde edilir. Böylece, her $n \in \mathbb{N}$ n için $(n-1)S_n^2/\sigma^2 \sim \chi_{n-1}^2$ dir ◊

Bu teoremin bir sonucu olarak, istatistikte ve diğer alanlarda çok kullanılan ve *örneklem dağılımları* olarak bilinen Z, χ_{n-1}^2, t ve F dağılımlarını inceleyelim. Aslında, χ_{n-1}^2, t ve F dağılımları da normal dağılımla ilişkilidir. Hemen hemen bütün istatistiki sonuç çıkarımlar bu dört dağılıma dayanmaktadır. Bu dağılımlar kullanılarak, kitlenin parametreleri hakkında hipotez testleri yapılabilmekte, yine parametreler için güven aralıkları yazılabilmektedir. Bunların her birinde de verilerin normalliği ön plana çıkmaktadır. Normal olmayan veriler için istatistiki sonuç çıkarımlar ancak MLT nin geçerli olduğu durumlarda yapılabilir.

$X_1, X_2, \ldots, X_n$ bağımsız $N(\mu, \sigma^2)$ dağılımlı rasgele değişkenlerin bir dizisi ise, $Z_n = \sqrt{n}\,(\bar{X}_n - \mu)/\sigma \sim N(0,1)$ olduğunu biliyoruz. Ancak σ^2 bilinmediği zaman, σ^2 yerine tahmin edicisi S_n^2 kullanılır. Böylece, istatistiki sonuç çıkarım için

$$t_n = \sqrt{n}\left(\bar{X}_n - \mu\right)/S_n$$

istatistiğinin dağılımına ihtiyaç duyulur. Teorem (6.3.1c) ye göre, $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$ olduğundan $Var(S_n^2) = 2\sigma^4/(n-1)$ olup $n \to \infty$ iken $Var(S_n^2) \to 0$ dır. Dolayısı ile, Örnek (6.2.2) ye göre $n \to \infty$ iken $S_n^2 \xrightarrow{P} \sigma^2$ dir. Buradan da, Slutsky Teoremine göre $n \to \infty$ iken, $t_n = \sqrt{n}\,(\bar{X}_n - \mu)/S_n \xrightarrow{\mathcal{D}} N(0,1)$ olur.

t − dağılımı:

X ile Y bağımsız ve $X \sim N(0,1)$, $Y \sim \chi_p^2$ olsun. $T = X/\sqrt{Y/p}$ rasgele değişkeninin olasılık yoğunluk fonksiyonu (Bölüm 3),

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)(p\pi)^{1/2}} \frac{1}{\left(1 + t^2/p\right)^{(p+1)/2}}, \quad t \in \mathbb{R}$$

dir. Bu olasılık yoğunluk fonksiyonuna sahip T rasgele değişkeni *serbestlik derecesi* p *olan* $t-$*dağılımına sahiptir* denir ve $T \sim t_p$ ile gösterilir. Dağılımın oasılık yoğunluk fonksiyonunun grafiği Şekil (6.3.1) de verilmiştir. Grafikten de görüldüğü gibi, dağılımın şekli normal dağılıma benzer ve standart normal dağılımda olduğu gibi fonksiyon sıfır noktasına göre simetriktir. Dağılımın beklenen değer ve varyansı,

$p > 1$ için $E(T) = 0$ ve $p > 2$ için $Var(T) = p/(p-2)$

dir.

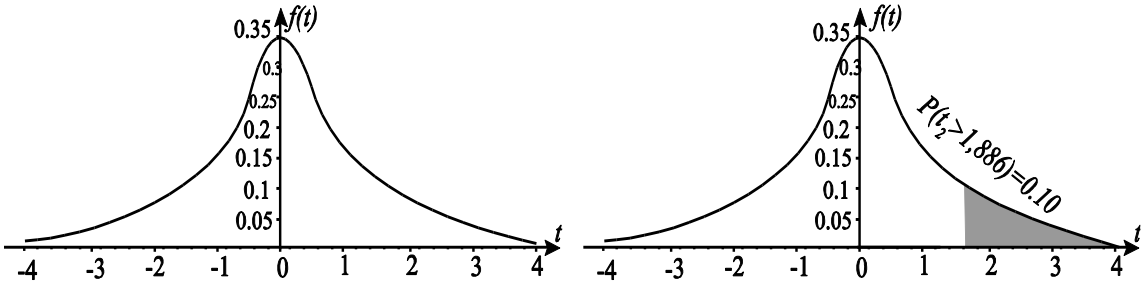
$X_1, X_2, \ldots, X_n$ beklenen değeri μ, varyansı σ^2 olan normal dağılımdan bir örneklem olmak üzere, $t_n = \sqrt{n}\,(\bar{X}_n - \mu)/S_n$ istatistiğinin dağılımını bulalım. Kolayca görüleceği gibi t_n istatistiği,

$$t_n = \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{S_n} = \frac{\dfrac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma}}{\dfrac{S_n}{\sigma}} = \frac{\dfrac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma}}{\sqrt{\dfrac{(n-1)S_n^2}{\sigma^2}/(n-1)}}$$

şeklinde yazılabilir. Buradan, $\bar{X}_n$ ile S_n^2 bağımsız ve dağılımları da

$$\sqrt{n}\,(\bar{X}_n - \mu)/\sigma \sim N(0,1) \quad \text{ve} \quad (n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$$

dir (Teorem (6.3.1)). Buradan, $p = n-1$ olmak üzere $t_n = \sqrt{n}\left(\bar{X}_n - \mu\right)/S_n \sim t_{n-1}$ olduğu görülür.



Şekil: 6.3.1 $p = 2$ için $t-$ dağılımının olasılık yoğunluk fonksiyonu

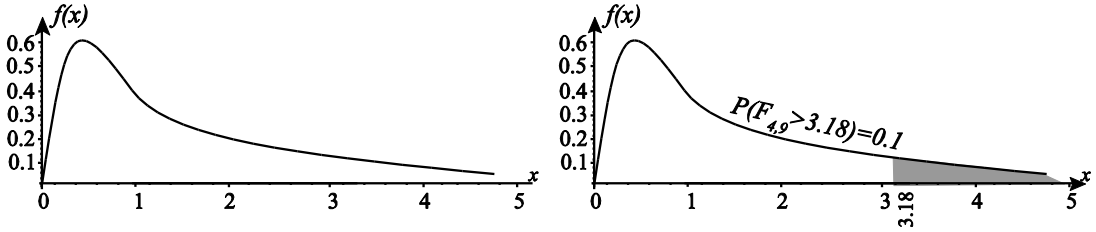
Bu dağılımın bazı kullanım alanlarından (güven aralıkları, hipotez testleri gibi) dokuz ve onuncu bölümlerde bahsedilecektir.

F – dağılımı:

Uygulamada çok kullanılan dağılımlardan biri de F dağılımıdır. Bağımsız X ve Y rasgele değişkenleri $X \sim \chi^2_p$ ve $Y \sim \chi^2_q$ şeklinde ki-kare dağılımına sahip olsunlar. Buradan, üçüncü bölümde bahsedilen bilgiler kullanılarak $F = \dfrac{X/p}{Y/q}$ rasgele değişkeninin olasılık yoğunluk fonksiyonunun,

$$f(x) = \frac{\Gamma\left(\dfrac{p+q}{2}\right)}{\Gamma\left(\dfrac{p}{2}\right)\Gamma\left(\dfrac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2} \frac{x^{(p-2)/2}}{\left(1+\dfrac{p}{q}x\right)^{(p+q)/2}}\ ,\ x \in \mathbb{R}^+$$

olduğu gösterilebilir. Olasılık yoğunluk fonksiyonu bu şekilde olan F rasgele değişkenine *serbestlik dereceleri p ve q olan F dağılımına sahiptir* denir ve $F \sim F(p,q)$ ile gösterilir. Fonksiyonun grafiği şekil (6.3.2) de verilmiştir.



Şekil 6.3.2 F dağılımının $p=4$ ve $q=6$ için olasılık yoğunluk fonksiyonu

$F \sim F(p,q)$ olsun. F dağılımının beklenen değeri, ki-kare dağılımının beklenen değerinden bulunabilir. $F \sim F(p,q)$ ise bağımsız $X \sim \chi_p^2$ ve $Y \sim \chi_q^2$ rasgele değişkenleri için $F = \dfrac{X/p}{Y/q}$ dir. Buradan, X ve Y bağımsız olduğundan, F nin beklenen değeri,

$$E(F) = E\left(\frac{X/p}{Y/q}\right) = E\left(\frac{X}{p}\right)E\left(\frac{q}{Y}\right) = \frac{q}{q-2}$$

dir(Bkz. Problem (5.5.7)).

$X_1, X_2 \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ ve $Y_1, Y_2 \ldots, Y_m \sim N(\mu_y, \sigma_y^2)$ şeklinde birbirinden bağımsız iki farklı örneklem olsun. σ_x^2 / σ_y^2 şeklinde varyansların oranını tahmin etmek için $S_{n,X}^2 / S_{m,Y}^2$ örneklem varyanslarının oranını gözönüne alalım. Buradan,

$$F = \frac{S_{n,X}^2 / S_{m,Y}^2}{\sigma_x^2 / \sigma_y^2} = \frac{S_{n,X}^2 / \sigma_x^2}{S_{m,Y}^2 / \sigma_y^2} = \frac{\dfrac{(n-1)S_{n,X}^2}{\sigma_x^2} / (n-1)}{\dfrac{(m-1)S_{m,Y}^2}{\sigma_y^2} / (m-1)}$$

nin dağılımı için,

$$\frac{(n-1)S_{n,X}^2}{\sigma_x^2} \sim \chi_{n-1}^2 \quad \text{ve} \quad \frac{(m-1)S_{m,Y}^2}{\sigma_y^2} \sim \chi_{m-1}^2$$

olduğunu biliyoruz. Ayrıca, $S_{n,X}^2$ ile $S_{m,Y}^2$ bağımsız olduğundan

$$F = \frac{S_{n,X}^2 / S_{m,Y}^2}{\sigma_x^2 / \sigma_y^2} \sim F(n-1, m-1)$$

elde edilir. Buna göre $\sigma_x^2 = \sigma_y^2$ olduğu varsayımı altında, $F = S_{n,X}^2 / S_{m,Y}^2 \sim F(n-1, m-1)$ olup bu istatistiğin değeri varyansların aynı olduğunu sınamak için kullanılır. Bu dağılımın da bazı kullanım alanlarından dokuz ve onuncu bölümlerde biraz bahsedilecektir.

Z, t, χ^2 ve F dağılımları literatürde örneklem dağılımları olarak bilinir. Aynı normal dağılımda olduğu gibi, bu dağılımlar için de tablolar düzenlenmiştir. Verilerin normallik varsayımı bazen sağlanmayabilir. Böyle durumlarda, bazen merkezi limit teoremi bazen de dönüşümler ile normallik sağlanır.

Örnek 6.3.1 a) $X \sim F(p,q)$ ise $Y = 1/X$ rasgele değişkeninin olasılık yoğunluk fonksiyonunu bulalım. X rasgele değişkeninin olasılık yoğunluk fonksiyonu,

$$f_X(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2} \frac{x^{(p-2)/2}}{\left(1+\frac{p}{q}x\right)^{(p+q)/2}} \ , \ x \in \mathbb{R}^+$$

olup Y nin olasılık yoğunluk fonksiyonu $f_Y(y) = f_X(x(y))|dx/dy|$ dir. Burada, $x = 1/y$ denirse $dx/dy = -1/y^2$ olup Y nin olasılık yoğunluk fonksiyonu,

$$f_Y(y) = \frac{1}{y^2}\frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2} \frac{(1/y)^{(p-2)/2}}{\left(1+\frac{p}{q}\frac{1}{y}\right)^{(p+q)/2}} \ , \ y \in \mathbb{R}^+$$

şeklinde bulunur. Bu eşitlik biraz düzenlendiğinde,

$$c = \Gamma\left(\frac{p+q}{2}\right)\left[\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)\right]^{-1}\left(\frac{p}{q}\right)^{p/2}$$

olmak üzere Y nin olasılık yoğunluk fonksiyonu,

$$f_Y(y) = c\frac{1}{y^2}\frac{(1/y)^{(p-2)/2}}{\left(1+\frac{p}{q}\frac{1}{y}\right)^{(p+q)/2}} = c\frac{y^{-2}y^{1-p/2}}{\left(\frac{qy+p}{qy}\right)^{(p+q)/2}} = c\frac{y^{-2}y^{1-p/2}(qy)^{(p+q)/2}}{(qy+p)^{(p+q)/2}}$$

$$= cq^{(p+q)/2}p^{-(p+q)/2}\frac{y^{(q-2)/2}}{(qy+p)^{(p+q)/2}} = cq^{(p+q)/2}\frac{y^{(q-2)/2}}{\left(p\left[1+\frac{q}{p}y\right]\right)^{(p+q)/2}}$$

şeklinde yazılabilir. Ayrıca c değeri yerine konulduğunda,

$$c\,q^{(p+q)/2}p^{-(p+q)/2} = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}q^{(p+q)/2}p^{-(p+q)/2} = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{q}{p}\right)^{q/2}$$

olup Y rasgele değişkeninin olasılık yoğunluk fonksiyonu,

$$f_Y(y) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{q}{p}\right)^{q/2}\frac{y^{(q-2)/2}}{\left(1+\frac{q}{p}y\right)^{(p+q)/2}}\ ,\ y\in\mathbb{R}^+$$

olarak elde edilir. Yani, $X \sim F(p,q)$ ise $Y = 1/X \sim F(q,p)$ dir.

b) $X \sim t_q$ ise $Y = X^2 \sim F(1,q)$ dir. $X \sim t_q$ ise dağılımın olasılık yoğunluk fonksiyonu,

$$f_X(x) = \frac{\Gamma((q+1)/2)}{\Gamma(q/2)(q\pi)^{1/2}}\frac{1}{(1+x^2/q)^{(q+1)/2}}\ ,\ x\in\mathbb{R}$$

dir. $Y = X^2$ denirse $X = \sqrt{Y}$ ve $dX/dY = 1/(2\sqrt{Y})$ olup Y nin olasılık yoğunluk fonksiyonunun $f_Y(y) = 2\,|1/(2\sqrt{y})|\,f_X(\sqrt{y})$ şeklinde olduğunu biliyoruz (X in olasılık yoğunluk fonksiyonu simetrik olduğundan 2 ile çarpılır). $\Gamma(1/2) = \sqrt{\pi}$ olduğundan Y nin olasılık yoğunluk fonksiyonu $y\in\mathbb{R}^+$ için,

$$f_Y(y) = \frac{1}{\sqrt{y}}\frac{\Gamma((q+1)/2)}{\Gamma(q/2)(q\pi)^{1/2}}\frac{1}{(1+y/q)^{(q+1)/2}} = \frac{\Gamma((q+1)/2)}{\Gamma(q/2)(q\pi)^{1/2}}\frac{y^{(1-2)/2}}{(1+y/q)^{(q+1)/2}}$$

$$= \frac{\Gamma((q+1)/2)}{\Gamma(q/2)\Gamma(1/2)}\left(\frac{1}{q}\right)^{1/2}\frac{y^{(1-2)/2}}{(1+y/q)^{(q+1)/2}}$$

şeklinde bulunur. Bu da serbestlik dereceleri 1 ve q olan F dağılımının olasılık yoğunluk fonksiyonudur. Yani, $X \sim t_q$ ise $Y = X^2 \sim F(1,q)$ dir.

c) $X \sim F(p,q)$ ise $Y = \dfrac{pX/q}{1+(pX/q)} \sim Beta\left(\dfrac{p}{2},\dfrac{q}{2}\right)$ olduğunu gösterelim. Y nin değer kümesi $D_Y = (0,1)$ dir. Ters dönüşüm ve türevi

$$X = \frac{q}{p}\frac{Y}{1-Y} \qquad \text{ve} \qquad \frac{dX}{dY} = \frac{q}{p}\left(\frac{1}{(1-Y)^2}\right)$$

olup Y nin olasılık yoğunluk fonksiyonu $y \in (0,1)$ için,

$$f_Y(y) = \frac{q}{p}\left(\frac{1}{(1-y)^2}\right) f_X\left(\frac{q}{p}\frac{y}{1-y}\right)$$

$$= \frac{q}{p}\left(\frac{1}{(1-y)^2}\right)\frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{p}{q}\right)^{p/2}\frac{\left(\frac{q}{p}\frac{y}{1-y}\right)^{(p-2)/2}}{\left[1+\frac{p}{q}\left(\frac{q}{p}\frac{y}{1-y}\right)\right]^{(p+q)/2}}$$

$$= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\frac{1}{(1-y)^2}\frac{\left(\frac{y}{1-y}\right)^{(p-2)/2}}{\left(\frac{1}{1-y}\right)^{(p+q)/2}}$$

$$= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\frac{1}{(1-y)^2}\frac{y^{(p/2)-1}(1-y)^{(p+q)/2}}{(1-y)^{(p/2)-1}} = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} y^{(p/2)-1}(1-y)^{(q/2)-1}$$

şeklinde yazılır. Bu da $Beta(p/2,\ q/2)$ dağılımının olasılık yoğunluk fonksiyonudur $\oplus$

6.4. Sıra İstatistikleri

İstatistiki veriler incelenirken genellikle dağılım hakkında herhangi bir bilgi yoktur. Veri analizine başlamadan önce dağılım hakkında görsel bazı bilgilere başvurulur. Bunlar genellikle *histogram, Box-Cox çiziti* ve *normal olasılık grafiği* dir. Bu grafiklerin oluşturulmasında verilerin sıralanmış halinden yararlanılır. Ayrıca, *mode, medyan, yüzdelikler, çeyreklikler* gibi değerler hesaplanırken de verilerin sıralanmış hali kullanılır.

$X_1, X_2, \ldots, X_n$ olasılık veya olasılık yoğunluk fonksiyonu $f(x;\theta)$ olan kitleden bir örneklem olsun. Bu rasgele değişkenler aynı $(\Omega, \mathcal{U}, P)$ olasılık uzayı üzerinde tanımlıdır. Her $w \in \Omega$ için $X_1(w) \le X_2(w)$ oluyorsa X_1 rasgele değişkeni X_2 rasgele değişkeninden küçük ya da eşittir denir ve $X_1 \le X_2$ gösterimi kullanılır. Buna göre,

$X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$, $X_{(2)} =$ ikinci en küçük ,..., $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$

olmak üzere, rasgele değişkenlerin $X_{(1)} \le X_{(2)} \le X_{(3)} \le \ldots \le X_{(n)}$ şeklinde bir sıralanmasından söz edilebilir. Buradaki, $X_{(i)}$ lerin her biri örneklemin bir fonksiyonu olup birer tahmin edicidir. Bu istatistiklere *sıra istatistikleri* denir. Bu istatistikler için bazen $X_{1:n} \le X_{2:n} \le X_{3:n} \le \ldots \le X_{n:n}$ gösterimi de kullanılır.

$X_1, X_2, \ldots, X_n$ örneklemi için, $X_{(1)} \le X_{(2)} \le X_{(3)} \le \ldots \le X_{(n)}$ sıra istatistikleri göz önüne alındığında,

Örneklem genişliği : $\mathcal{R} = X_{(n)} - X_{(1)}$

Örneklem medyanı (Ortanca): $M = \begin{cases} X_{((n+1)/2)} & , \text{ n tek olduğunda} \\ \frac{1}{2}[X_{(n/2)} + X_{((n/2)+1)}] & , \text{ n } \textit{çift} \text{ olduğunda} \end{cases}$

Uzunluk ortası (Midrange): $V = (X_{(1)} + X_{(n)})/2$

gibi bazı istatistikler de sıra istatistikleri ile tanımlanır.

Örnek 6.4.1 Bir istatistik dersinden sınava giren öğrencilerin aldığı notlar aşağıdadır. Görüldüğü gibi dağılım hakkında herhangi bir bilgi yoktur.

66	71	67	69	75	66	64	70	62	83
70	79	74	74	79	94	76	69	88	72
84	76	63	70	77	80	77	72	78	73
75	78	90	76	62	78	78	72	77	72
72	59	73	75	76	80	56	67	69	80

Örneklem ortalaması ile örneklem varyansının değerleri

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = 73.36 \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \cong 61.79$$

olarak hesaplanmıştır. Ayrıca, veriler küçükten büyüğe sıralandığında,

$x_{(1)} = 56$, $x_{(50)} = 94$, $x_{(25)} = 74$, $x_{(26)} = 74$, $x_{(48)} = 88$

değerleri gözlenmiştir. Bunlarla birlikte, bu verilere ait bazı özet bilgiler de

$m = 0.5[x_{(25)} + x_{(26)}] = 0.5(74+74) = 74$, $r = x_{(50)} - x_{(1)} = 94 - 56 = 38$

olarak elde edilmiştir. Verilerin küçükten büyüğe sıralanmış hali de aşağıdadır.

56	58	60	62	62	62	64	66	66	66
68	68	68	70	70	70	70	72	72	72
72	72	72	72	74	74	74	74	76	76
76	76	76	76	76	78	78	78	78	78
78	80	80	80	82	84	84	88	90	94

Tablo: Verilerin küçükten büyüğe doğru sıralanmış hali

Veriler küçükten büyüğe sıralandığında, %50 si medyan değerinden küçük ya da eşittir. %25'i bir Q_L sayısından küçük ya da eşit kalıyorsa, Q_L sayısına birinci çeyreklik, %75'i bir Q_U sayısından küçük ya da eşit kalıyorsa Q_U sayısına da üçüncü çeyreklik denir. İkinci çeyreklik medyandır. Benzer şekilde, verilerin %95'i bir $c(95\%)$ sayısından küçük ya da eşit kalıyorsa $c(95)$ sayısı dağılımın %95 lik kritik değeridir. Bu kritik değerlerden bazıları,

$c(99\%) = 94$, $c(95\%) = 88$, $c(90\%) = 83$
$c(10\%) = 62$, $c(5\%) = 60$, $c(1\%) = 56$

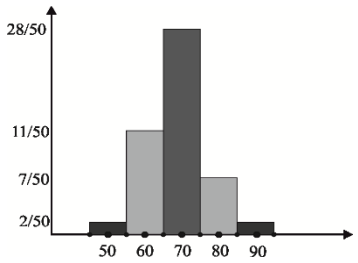
olarak gözlenmiştir. Bu değerlere göre, verilerin %95'i 88 sayısından küçük ya da eşittir. Benzer şekilde, %90'ı 83 sayısından küçük ya da eşittir. Verilere ait diğer yüzdelikler de hesaplanabilir. Örneğin, verilerin %60'ı bir $c(60\%)$ değerinden küçük ya da eşit oluyorsa $c(60\%)$ sayısı %60 lık kritik değer olarak alınır. Buradaki $c(60\%)$ sayısı örneklem içinde bir sayı olmak zorunda değildir.

Veriler belli kategorilere ayrılarak da grafikler oluşturulur. Örneğin 90 ve üzerinde not alan öğrenciler birinci grup, 80 ile 90 arasında not alan öğrenciler ikinci grup olarak ele alınarak bir gruplandırma yapılabilir. Her gruptaki gözlem sayısına o grubun *sıklığı* (*frekansı*) denir. Her bir aralıktaki gözlem sayısı (frequency, sıklık) f_i, aralıktaki gözlem sayısının toplam gözlem sayısına oranı (relative frequency) f_i / n dir. Buna göre $n = 50$ olmak üzere aşağıdaki tablo oluşturulmuştur.

Grup i	Aralık	f_i	f_i / n	$\sum_{k=1}^{i} f_k / n$
1	$x \geq 90$	2	2 / 50	2 / 50
2	$80 \leq x < 90$	7	7 / 50	9 / 50
3	$70 \leq x < 80$	28	28 / 50	37 / 50
4	$60 \leq x < 70$	11	11 / 50	48 / 50
5	$50 \leq x < 60$	2	2 / 50	50 / 50

Bu tablo başka bir şekilde aşağıdaki gibi özetlenebilir. Bu özet şekline verilerin *Dal-Yaprak Grafiği* denir.

9	04
8	0002448
7	000022222224444666666888888
6	02224666888
5	68



Şekil 6.4.1 Dal Yaprak Grafiği(solda) ve Histogram(sağda)

Bu oranlardan verilerin histogramı oluşturulur. Histogramdan yararlanarak dağılımın şekli hakkında görsel bilgiler elde edilir ⊕

Sıra istatistiklerinin olasılık fonksiyonları aşağıdaki teoremde özetlenmiştir. Teoremin ispatı burada verilmemiştir. Ayrıntılı bilgi için Casella ve Berger (2002) ye bakılabilir.

Teorem 6.4.1 Dağılım fonksiyonu $F(x)$, olasılık veya olasılık yoğunluk fonksiyonu da $f(x)$ olan kitleden bir örneklem $X_1, X_2, \ldots, X_n$ olsun. Bu örnekleme ilişkin sıra istatistikleri $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ olmak üzere,

a) $X_{(j)}$ rasgele değişkeninin olasılık yoğunluk fonksiyonu $x \in D_X$ için

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)^{j-1}[1-F(x)]^{n-j},$$

b) $X_{(i)}$ ve $X_{(j)}$ rasgele değişkenlerinin ortak olasılık yoğunluk fonksiyonu $x < y$ için

$$f_{X_{(i)},X_{(j)}}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(x) f(y)[F(x)^{i-1} * [F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j},$$

c) $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ rasgele değişkenlerinin ortak olasılık yoğunluk fonksiyonu da,

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(x_1, x_2, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^{n} f(x_i) & , \quad x_1 < x_2 < \ldots < x_n \\ 0 & , \quad d.y. \end{cases}$$

dir (Casella ve Berger, 2002, sayfa 229-230) ◊

Örnek 6.4.2 $X_1, X_2, \ldots, X_n$, $U(0,\theta)$ dağılımından bir örneklem olsun. Yani X lerin olasılık yoğunluk fonksiyonu ve dağılım fonksiyonu sırası ile

$$f(x;\theta)=\begin{cases}1/\theta & , \quad 0<x<\theta \\ 0 & , \quad d.y.\end{cases} \quad , \quad F(x;\theta)=\begin{cases}0 & , \quad x<0 \\ x/\theta & , \quad 0\le x\le\theta \\ 1 & , \quad x>\theta\end{cases}$$

olarak verilmiş olsun.

a) $X_{(n)}$ sıra istatistiğinin olasılık yoğunluk fonksiyonu Teorem (6.4.1a) dan, $0<x<\theta$ için

$$f_{X_{(n)}}(x;\theta)=\frac{n!}{(n-1)!(n-n)!}\frac{1}{\theta}\left(\frac{x}{\theta}\right)^{n-1}\left(1-\frac{x}{\theta}\right)^{n-n}=\frac{n}{\theta^n}x^{n-1}$$

eşitliğinden

$$f_{X_{(n)}}(x;\theta)=\begin{cases}\dfrac{n}{\theta^n}x^{n-1} & , \quad 0<x<\theta \\ 0 & , \quad d.y.\end{cases}$$

olarak yazılır. $X_{(n)}$ sıra istatistiğinin olasılık yoğunluk fonksiyonu dağılım fonksiyonu yardımı ile de bulunabilir. $F_{X_{(n)}}(x)$, $X_{(n)}$ rasgele değişkeninin dağılım fonksiyonunu göstermek üzere, $x<0$ için $F_{X_{(n)}}(x)=0$ ve $x\ge\theta$ için $F_{X_{(n)}}(x)=1$ olup $0<x<\theta$ için,

$$F_{X_{(n)}}(x)=P(X_{(n)}\le x)=P(\max\{X_1,\ldots,X_n\}\le x)=P(X_1\le x, X_2\le x,\ldots,X_n\le x)$$
$$=P(X_1\le x)P(X_2\le x)...P(X_n\le x)=[P(X_1\le x)]^n=[x/\theta]^n=\theta^{-n}x^n$$

dir. Buradan $X_{(n)}$ nin olasılık yoğunluk fonksiyonu dağılım fonksiyonunun türevinden (dağılım fonksiyonunun türevlenebildiği yerlerde),

$$f_{X_{(n)}}(x)=\begin{cases}\dfrac{n}{\theta^n}x^{n-1} & , \quad 0<x<\theta \\ 0 & , \quad d.y.\end{cases}$$

şeklinde elde edilir. $X_{(n)}$ nin ilk iki momenti ve varyansı,

$$E(X_{(n)})=\frac{n}{\theta^n}\int_0^\theta x^n\,dx=\frac{n}{n+1}\theta, \quad E(X_{(n)}^2)=\frac{n}{\theta^n}\int_0^\theta x^{n+1}\,dx=\frac{n}{n+2}\theta^2$$

ve

$$Var(X_{(n)})=E(X_{(n)}^2)-[E(X_{(n)})]^2=\frac{n\ \theta^2}{(n+1)^2(n+2)}$$

şeklinde hesaplanmıştır. Buradan da, $T=(n+1)n^{-1}X_{(n)}$ rasgele değişkeninin beklenen değeri ile varyansı da

$$E(T)=E\left(\frac{n+1}{n}X_{(n)}\right)=\theta \quad \text{ve} \quad Var(T)=Var\left(\frac{n+1}{n}X_{(n)}\right)=\frac{\theta^2}{n(n+2)}$$

olur.

b) $\mathcal{R}=X_{(n)}-X_{(1)}$ örneklem genişliği istatistiğinin olasılık yoğunluk fonksiyonunu bulalım. Önce, $X_{(1)}$ ve $X_{(n)}$ nin ortak olasılık yoğunluk fonksiyonu Teorem (6.4.1b) den

$$f_{X_{(1)},X_{(n)}}(x,y)=\frac{n(n-1)}{\theta^n}(y-x)^{n-2} \quad , \quad 0<x<y<\theta$$

şeklindedir. $\mathcal{R}=X_{(n)}-X_{(1)}$ nin olasılık yoğunluk fonksiyonu için $V=(X_{(n)}+X_{(1)})/2$ yardımcı dönüşümünü tanımlayalım. Ters dönüşümler $X_{(1)}=V-\mathcal{R}/2$, $X_{(n)}=V+\mathcal{R}/2$ şeklinde olup Jacobien matrisi ve determinantı

$$J=\begin{bmatrix}1 & -1/2\\ 1 & 1/2\end{bmatrix} \quad \text{ve} \quad \det(J)=1$$

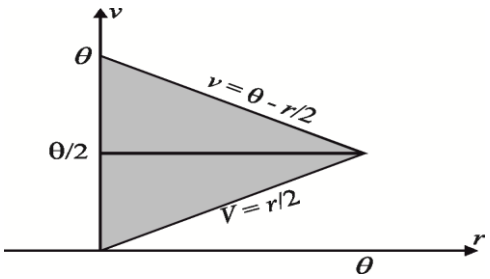
olarak bulunmuştur. Burada, rasgele değişkenlerin sınırlarının belirlenmesi gerekir. $X_{(1)}$ ve $X_{(n)}$ rasgele değişkenlerinin her ikisinin değer kümesi $(0,\theta)$ ve $X_{(1)}\le X_{(n)}$ olduğundan $0<r<\theta$ olduğu açıktır. R nin üst sınırı için $X_{(n)}$ nin alacağı en büyük değerden $X_{(1)}$ rasgele değişkeninin alacağı en küçük değer çıkartıldığında $\mathcal{R}$ nin üst sınırının θ olduğu görülür. Diğer taraftan, $0<x_{(1)}=v-r/2<x_{(n)}=v+r/2$ olduğundan, $0<x_{(1)}=v-r/2$ ise $v>r/2$ dir. Ayrıca, $x_{(n)}=v+r/2<\theta$ olduğundan $v<\theta-r/2$ olup $\mathcal{R}$ ve V nin ortak olasılık yoğunluk fonksiyonu $r/2<v<\theta-r/2$ ve $0<r<\theta$ için

$$f_{\mathcal{R},V}(r,v)=\frac{n(n-1)}{\theta^n}\big((v+r/2)-(v-r/2)\big)^{n-2}=\frac{n(n-1)r^{n-2}}{\theta^n}$$

şeklinde olur. Yani, R ve V nin ortak olasılık yoğunluk fonksiyonu,

$$f_{\mathcal{R},V}(r,v)=\begin{cases}\dfrac{n(n-1)r^{n-2}}{\theta^n} & , \quad 0<r<\theta,\ r/2<v<\theta-r/2\\ 0 & , \quad d.y.\end{cases}$$

dir. Ortak olasılık yoğunluk fonksiyonunun değer kümesi Şekil (6.4.2) de gösterilmiştir.



Şekil 6.4.2 Örnek (5.2.4b) Olasılık yoğunluk fonksiyonunun değer kümesi

Buna göre, $\mathcal{R}$ nin olasılık yoğunluk fonksiyonu için $f_{\mathcal{R},V}(r,v)$ ortak olasılık yoğunluk fonksiyonunun V nin değer kümesi üzerinden integrali alınır. Bu integral,

$$f_{\mathcal{R}}(r)=\int_{v=r/2}^{\theta-r/2} f_{R,V}(r,v)\,dv=\int_{v=r/2}^{\theta-r/2}\frac{n(n-1)\,r^{n-2}}{\theta^n}\,dv=\frac{n(n-1)}{\theta^n}\,r^{n-2}(\theta-r)$$

olup $\mathcal{R}$ nin olasılık yoğunluk fonksiyonu,

$$f_{\mathcal{R}}(r)=\begin{cases}\dfrac{n(n-1)}{\theta^n}\,r^{n-2}(\theta-r) & ,\quad 0<r<\theta\\ 0 & ,\quad d.y.\end{cases}$$

olarak bulunmuştur. V nin olasılık yoğunluk fonksiyonu için grafikten de görüldüğü gibi integralin iki ayrı şekilde hesaplanması gerekir. İntegral değerleri $0<r<\theta/2$ için,

$$f_V(v)=\int_{r=0}^{2v}\frac{n(n-1)\,r^{n-2}}{\theta^n}\,dr=\frac{n\,(2v)^{n-1}}{\theta^n},$$

$\theta/2<r<\theta$ için,

$$f_V(v)=\int_{r=0}^{2(\theta-v)}\frac{n(n-1)\,r^{n-2}}{\theta^n}\,dr=\frac{n\,\left(2(\theta-v)\right)^{n-1}}{\theta^n}$$

şeklindedir. Buna göre V nin olasılık yoğunluk fonksiyonu,

$$f_V(v)=\begin{cases} n\,(2v)^{n-1}/\theta^n & ,\quad 0<v<\theta/2\\ n\,\left(2(\theta-v)\right)^{n-1}/\theta^n & ,\quad \theta/2<v<\theta\\ 0 & ,\quad d.y.\end{cases}$$

olarak elde edilmiştir. $\theta=1$ için $\mathcal{R}\sim Beta(n-1,\,2)$ olduğu açıktır.

c) $X_1, X_2, \ldots, X_n$ beklenen değeri θ olan üstel dağılımdan bir örneklem olsun. $X_{(1)}$ ve $X_{(n)}$ sıra istatistiklerinin ortak olasılık yoğunluk fonksiyonu Teorem (6.4.1b) den ($i=1$ ve $j=n$) $0<x<y<\infty$ için,

$$f_{X_{(1)},X_{(n)}}(x,y)=\frac{n(n-1)}{\theta^2}e^{-(x+y)/\theta}[e^{-x/\theta}-e^{-y/\theta}]^{n-2}$$

şeklindedir. Şimdi, $\mathcal{R}=X_{(n)}-X_{(1)}$ örneklem uzunluğu istatistiğinin olasılık yoğunluk fonksiyonunu bulalım. Bunun için $V=X_{(1)}$ yardımcı dönüşümü ile ters dönüşümler $X_{(1)}=V$ ve $X_{(n)}=\mathcal{R}+V$ olup Jacobien matrisinin determinantı 1 dir. Buna göre, $\mathcal{R}$ ve V istatistiklerinin ortak olasılık yoğunluk fonksiyonu $r,v>0$ için

$$\begin{aligned}
f_{\mathcal{R},V}(r,v)&=f_{X_{(1)},X_{(n)}}\left(x(r,v),y(r,v)\right)=\frac{n(n-1)}{\theta^2}e^{-(2v+r)/\theta}\left[e^{-v/\theta}-e^{-(r+v)/\theta}\right]^{n-2}\\
&=\frac{n(n-1)}{\theta^2}e^{-(2v+r)/\theta}\left[e^{-v/\theta}\left(1-e^{-r/\theta}\right)\right]^{n-2}\\
&=\frac{n(n-1)}{\theta^2}e^{-(2v+r)/\theta}e^{-(n-2)v/\theta}\left[\left(1-e^{-r/\theta}\right)\right]^{n-2}\\
&=\frac{n(n-1)}{\theta^2}e^{-r/\theta}e^{-nv/\theta}\left[\left(1-e^{-r/\theta}\right)\right]^{n-2}
\end{aligned}$$

olarak yazılabilir. Bu fonksiyonun V nin değer kümesi ($\mathbb{R}^+$) üzerinden integrali

$$\begin{aligned}
f_{\mathcal{R}}(r)&=\int_{v=0}^{\infty}f_{R,V}(r,v)\,dv=\frac{n(n-1)}{\theta^2}e^{-r/\theta}\left[\left(1-e^{-r/\theta}\right)\right]^{n-2}\int_{v=0}^{\infty}e^{-nv/\theta}dv\\
&=\frac{n(n-1)}{\theta^2}e^{-r/\theta}\left[\left(1-e^{-r/\theta}\right)\right]^{n-2}\left[-\frac{\theta}{n}e^{-nv/\theta}\right]_{v=0}^{\infty}=\frac{(n-1)}{\theta}e^{-r/\theta}\left[1-e^{-r/\theta}\right]^{n-2}
\end{aligned}$$

olup $\mathcal{R}$ nin olasılık yoğunluk fonksiyonu,

$$f_{\mathcal{R}}(r)=\begin{cases}\dfrac{(n-1)}{\theta}e^{-r/\theta}[1-e^{-r/\theta}]^{n-2} & , \quad r>0\\ 0 & , \quad d.y.\end{cases}$$

şeklinde bulunmuştur⊕

Sıra istastitikleri ile rasgele değişkenlerin yüzdelikleri de tahmin edilir. X rasgele değişkeninin dağılım fonksiyonu $-\infty\le a<b\le\infty$ aralığında sürekli olsun. X in dağılım

fonksiyonu $F(x)$ olmak üzere X in $p.$ yüzdeliği (quantile) $0 < p < 1$ için $\xi_p = F^{-1}(p)$ olarak verilir. Buna göre, X rasgele değişkeninin medyanı, $\xi_{0.5}$ dir.

$X_1, X_2, \ldots, X_n$ bir örneklem, $Y_1, Y_2, \ldots, Y_n$ de bu örnekleme karşılık gelen sıra istatistikleri olsun. $k = p(n+1)$ diyelim. Burada, k sayısı $p(n+1)$ sayısının tam kısmını göstermektedir. X in olasılık yoğunluk fonksiyonu $f(x)$ ise $f(x)$ in y_k ye kadar integrali $F(y_k)$ dır. Buna göre, $Z = F(Y_k)$ denirse, Z nin beklenen değeri,

$$E\left[F(Y_k)\right] = \int_a^b F(y)\, f_{X_{(k)}}(y)\, dy = \int_{z=0}^{1} \frac{n!}{(k-1)!(n-k)!} z^k (1-z)^{n-k}\, dz$$

$$= \frac{n!k!(n-k)!}{(k-1)!(n-k)!(n+1)!} = \frac{k}{n+1}$$

olur (Hogg, Mckean ve Craig, 2005). Buradan, $p \cong k/(n+1)$ olduğunda Y_k sıra istatistiği ξ_p nin bir tahmin edicisi olarak alınabilir.

Herhangi bir istatistiksel analiz için 5 istatistik (five number summary) değeri ön plana çıkmaktadır. Bunlar

$$Y_1 = \min\{X_1, X_2, \ldots, X_n\}, \qquad Y_n = \max\{X_1, X_2, \ldots, X_n\}$$

birinci çeyreklik $Q_1 = \hat{\xi}_{0.25} = Y_{0.25(n+1)}$, üçüncü çeyreklik $Q_3 = \hat{\xi}_{0.75} = Y_{0.75(n+1)}$ ve

$$\text{medyan (ortanca) } M = \hat{\xi}_{0.5} = \begin{cases} Y_{(n+1)/2} & , \quad n \text{ tek olduğunda} \\ \frac{1}{2}\left[Y_{n/2} + Y_{(n/2)+1}\right] & , \quad n \text{ çift olduğunda} \end{cases}$$

dır.

Örnek 6.4.3 15 birimlik örneklem değerleri sıralanmış olarak aşağıda verilmiştir.

56 70 89 94 96 101 102 102 102 105 106 108 110 113 116

Bu verilere göre, $n = 15$ olup 5 özet istatistiğin değeri

$y_1 = 56$, $Q_1 = y_4 = 94$, $Q_2 = M = y_8 = 102$, $Q_3 = y_{12} = 108$, $y_{15} = 116$

olarak elde edilir $\oplus$

Veri analizinde, aykırı değerlerin (outliers) belirlenmesi önemlidir. Bu aykırı değerlerin belirlenmesinde de sıra istatistiklerinden yararlanılır. Bunun için alt ve üst çitler (lower fence, upper fence) belirlenir. $h = 1.5(Q_3 - Q_1)$ olmak üzere, $LF = Q_1 - h$ ve $UF = Q_3 + h$ şeklinde

alt ve üst çitler tanımlanır. (LF, UF) aralığının dışında kalanlar aykırı değerlerdir. Örnek (6.4.3) de $h = 1.5(Q_3 - Q_1) = 1.5(108 - 94) = 21$ olup alt ve üst çitler

$$LF = Q_1 - h = 94 - 21 = 73 \text{ ve } UF = Q_3 + h = 108 + 21 = 129$$

olduğundan $(73, 129)$ aralığının dışında kalan gözlemler aykırı gözlemlerdir. Buradan, 56 ve 70 bu aralığın dışında olduğundan, aykırı gözlemlerdir.

X sürekli rasgele değişkeninin dağılım fonksiyonu $F((x-a)/b)$ şeklinde olsun. Burada, F dağılım fonksiyonu biliniyor olmasına rağmen, a ve b bilinmiyor olabilir. $Z = (X - a)/b$ ise Z nin dağılım fonksiyonu $F(z)$ olur. $0 < p < 1$ için $\xi_{X,p}$ değeri X rasgele değişkeninin p. yüzdeliğini göstersin. $\xi_{Z,p}$ de Z nin p. yüzdeliğini göstersin. $F(z)$ biliniyorsa $\xi_{Z,p}$ de biliniyor demektir. $p = P(X \le \xi_{X,p}) = P(Z \le (\xi_{X,p} - a)/b)$ olduğundan, $\xi_{X,p} = b\,\xi_{Z,p} + a$ şeklinde bir ilişki elde edilir. Pratikte, $\xi_{X,p}$ yüzdelikleri bilinmeyen parametrelerdir. Bu parametrelerin verilen $X_1, X_2, \ldots, X_n$ örneklemine göre tahmin edilmesi gerekir. $X_1, X_2, \ldots, X_n$ örneklemi için $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ sıra istatistikleri olmak üzere $k = 1, 2, 3, ..., n$ için $p_k = k/(n+1)$ diyelim. Böylece, $X_{(k)}$ sıra istatistiği ξ_{X,p_k} nin tahmin edicisidir. $\xi_{Z,p_k} = F^{-1}(p_k)$ olmak üzere, $X_{(k)}$ lerin ξ_{Z,p_k} lere karşı grafikleri çizilir. Bu grafikler de $q-q$ *çizitleridir*. Bu grafik doğrusal ise, X rasgele değişkeninin dağılımının $F((x-a)/b)$ yapısına uygun olduğu söylenebilir. Uygulamada normallik çok kullanıldığından bu grafiğe bazen normal olasılık grafiği de denir.