SIRA İSTATİSTİKLERİ VE UYGULAMA ALANLARINDAN BİR ÖRNEĞİN DEĞERLENDİRMESİ

Esin Cumhur PİRİNÇCİLER
Araş. Gör. Dr., Çanakkale Onsekiz Mart Üniversitesi
Biga İİBF, Ekonometri Bölümü
esin.pirincciler@gmail.com

ÖZET

Sıra istatistikleri dağılımdan bağımsız olma özelliği, yoğunluk fonksiyonu ve momentlerinin elde edilebilmesi özellikleri, örneklem hakkındaki tüm bilgiyi içermesi sayesinde istatistiksel çalışmalarda çok önemli bir yere sahiptir. Sıra istatistikleri pek çok mühendislik tasarımında ve dağılış teorisi çalışmalarında şans değişkenin olasılık yoğunluk fonksiyonunun kantil değerlerinin elde edilmesinde kullanılan önemli araçlardır. Bu çalışmada sıra istatistiklerinin temel kavramları ve sıra istatistiğinin dağılım özellikleri incelenerek uygulama alanlarına ilişkin örnek verilmektedir.

Anahtar Kelimeler: Sıra İstatistikleri, Maksimumun Dağılımı, Minimumun Dağılımı, Sıra İstatistiklerinin Dağılımı

ABSTRACT

Because order statistics contains all the information about the sample, characteristics of independent distribution, having density functions and moments has a very important place in statistical studies. Order statistics are very important tools for obtaining quantiles of random variable probability density function in distributional theory and at some engineering design. In this study is examined fundamental concepts of order statistics and distributional property of order statistic and gave example about application areas.

Keywords: Order Statistics, Distribution of Maximum, Distribution of Minimum, Distribution of Order Statistics

GİRİŞ

Sıra istatistikleri, istatistik teorisinde önemli bir yer tutmakta ve istatistiksel tahminleme yöntemlerinde oldukça yaygın kullanılan bir kavramdır. Aynı zamanda sigortacılık, deprem ölçümleri ve analizi, doğal afetlerin tahminlenmesi, en zayıf halka prensibi gibi çalışmalarda da sıra istatistiklerinden yararlanılmaktadır.

Matematiksel istatistiğin temel problemlerinden birisi, her zaman şans değişkenin dağılım fonksiyonunun elde edilememesi ve bunun yerine deneysel değerlerin kullanılarak dağılımın tahminlenmesidir. Sıra istatistikleri yeterli istatistikler oldukları için örneklem hakkında tüm bilgiyi içerirler. Sıra istatistiklerine dayalı birçok istatistik, dağılımdan bağımsız özelliği taşıdığı için parametrik olmayan istatistiksel yöntemlerde kullanılan en temel yaklaşımdır.

Bu çalışmada sıra istatistiklerine ait tanımlar, maksimum ve minimumun dağılımı, bazı sıra istatistiklerinin ortak dağılımı, sıra istatistiğinin koşullu dağılımı, sıra istatistiğinin farkının dağılımı ve yağış verileri ile yapılan uygulama incelenecektir.

1. SIRA İSTATİSTİKLERİ İLE İLGİLİ TEMEL KAVRAMLAR

$(X_1, X_2, \dots, X_n)$ verilen bir populasyondan alınan bir örnek ve ardışık $X_1, X_2, \dots, X_n$ değerleri büyüklüklerine göre sıraya konulduğunda $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ oluşan yeni sıralamanın r-inci üyesi örneğin r-inci sıra istatistiğidir. Notasyon olarak sıra istatistiği $X_{r:n}$ yerine $X_{(r)}$ veya X_r şeklinde de gösterilebilir. İki önemli terim olan $X_{1:n} = Min(X_1, X_2, \dots, X_n)$ ve $X_{n:n} = Max(X_1, X_2, \dots, X_n)$ uç değerler olarak adlandırılır ve uygulamalarda önemli rol oynarlar.

$m_n(x)$, $X_j \leq x$ değerleri ile örnekteki eleman sayısıdır ve binomial tipin rassal bir değişkenidir, $B(n, F(x))$, çünkü n bağımsız Bernolli denemesi tekrarlandığında gerçekleşme (başarı) sayısı ile aynı zamanda meydana gelir. Bir Bernolli deneyi üst populasyondan rastgele bir değerin çekilmesiyle oluşur. Bu değer x' e (başarı) eşit ya da küçük olabilir, olasılığı $p = F(x)$ ya da x(hata)' dan daha büyük olduğunda olasılığı $1 - p$' dir. Sonuç olarak $m_n(x)$' in kümülatif dağılım fonksiyonu binomial dağılıma sahiptir.

$$F_{m_n(x)}(r) = Prob[m_n(x) \leq r] = \sum_{k=0}^{r} \binom{n}{k} F^k(x)[1 - F(x)]^{n-k} \qquad (1)$$

Fakat $\{X_{r:n} \leq x\}$ olayı yani r-inci sıra istatistiği x' e eşit ya da küçük bir değer olduğunda, $\{m_n(x) \geq r\}$ olur. (1) denkleminden;

$$F_{X_{r:n}}(x) = P[X_{r:n} \leq x] = 1 - F_{m_n(x)}(r-1)$$

$$= \sum_{k=0}^{r} \binom{n}{k} F^k(x)[1-F(x)]^{n-k}$$

$$= r\binom{n}{r}\int_0^{F(x)} u^{r-1}(1-u)^{n-r}du$$

$$= I_{F(x)}(r, n-r+1) \qquad (2)$$

elde edilir. $F_{X_{r:n}}(x)$, $X_{r:n}$' nin kümülatif dağılım fonksiyonudur ve $I_p(a,b)$ tamamlanmamış beta fonksiyonudur.

Eğer üst populasyon tamamen sürekli ise (2) denkleminin x' e göre birinci türevi alınarak $X_{r:n}$' nin olasılık yoğunluk fonksiyonu

$$f_{X_{r:n}}(x) = r\binom{n}{r}F^{r-1}(x)[1-F(x)]^{n-r}f(x)$$

$$= F^{r\;\;1}(x)[1 - F(x)]^{n\;\;r}\frac{f(x)}{B(r,n-r+1)} \qquad (3)$$

belirlenir. B(a, b) beta fonksiyonudur (David, 1970).

1.1 Maksimum'un Dağılımı

Maksimum son sıra istatistiği olduğu için denklem (2) ve (3)'deki r=n durumunu sağlamaktadır. Maksimum kümülatif dağılım fonksiyonu ve olasılık yoğunluk fonksiyonu

$$F_{X_{n:n}}(x) = F^n(x)$$

$$f_{X_{n:n}}(x) = nF^{n-1}(x)f(x)$$

olarak elde edilir.

1.2 Minimum' un Dağılımı

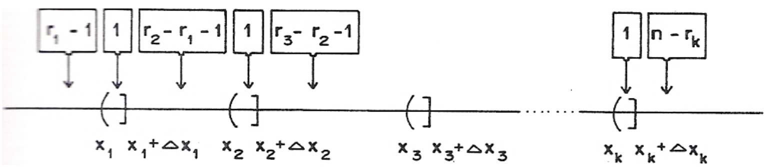
Bir rassal örneğin minimumunun dağılımı denklem (2) ve (3)'de r=1 yerine konularak elde edilir.

$$F_{X_{1:n}}(x) = 1 - [1-F(x)]^n$$

$$f_{X_{1:n}}(x) = n[1-F(x)]^{n-1}f(x)$$

1.3 Bazı Sıra İstatistiklerinin Ortak Dağılımı

Olasılık yoğunluk fonksiyonu f(x) ve kümülatif dağılım fonksiyonu F(x) olan, bir populasyondan alınan n hacimlik bir şans örneğinin $X_{r_{1:n}}, X_{r_{2:n}}, \ldots, X_{r_{k:n}}$ k sıra istatistikleri $r_1 < r_2 < \ldots < r_k$'dir. Sıra istatistiklerinin bu kümesinin ortak olasılık yoğunluk fonksiyonunu elde etmek için $\Delta x_j, 1 \le j \le k$ için $\{x_j \le X_{rj:n} < x_j + \Delta x_j; 1 \le j \le k\}$ örnekteki k eleman $(x_j, x_j + \Delta x_j)$ aralığında ve $r_j - r_{j-1} - 1$ ise $(x_{j-1} + \Delta x_{j-1}, x_j)$ aralığında yayılmaktadır. $\Delta x_0 = 0,\ r_0 = 0,\ r_{k+n} = n+1,\ x_0 = -\infty$ ve $x_{k+1} = \infty$'dur.



2k+1 mümkün sonuç ile çok terimli deneyi düşünüldüğünde: verilen populasyondan rassal olarak n eleman ve 2k+1 aralıkları belirtilmiştir. Bağımsızlık ve tekrarlamalı varsayımı altında her bir aralığın eleman sayısı multinomial şans değişkenidir.

$$M\{n; f(x_1)\Delta x_1, f(x_2)\Delta x_2, \ldots, f(x_k)\Delta x_k, [F(x_1) - F(x_0)], [F(x_2) - F(x_1)], \ldots, [F(x_{k+1}) - F(x_k)]\}$$

Parametreler n (örnek hacmi) ile ve olasılıklar 2k+1 aralıkları ile ilişkilidir. Sonuç olarak multinomial şans değişkenlerinin özellikleri kullanılabilir ve istatistik kümesinin ortak olasılık yoğunluk fonksiyonu;

$$f_{r_1,r_2,\ldots,r_{k:n}}(x_1,x_2,\ldots,x_k) = n! \prod_{i=1}^{k} f(x_i) \prod_{j=1}^{k+1} \left(\frac{[F(x_j)-F(x_{j-1})]^{r_j-r_{j-1}-1}}{(r_j-r_{j-1}-1)!} \right) \qquad (4)$$

$$x_1 < x_2 \le \ldots \le x_k$$

şeklinde elde edilir. (4) denklemindeki k=2, ve olarak verildiğinde n hacimlik örneğin maksimumu ve minimumunun ortak olasılık yoğunluk fonksiyonu;

$$f_{1,n:n}(x_1,x_2) = n(n-1)f(x_1)f(x_2)[F(x_1)-F(x_2)]^{n-2}; \qquad x_1 \leq x_2 \qquad (5)$$

olarak belirlenir ve (4) denklemdeki k=2, r ve r $_2 = i+1$ olarak verildiğinde iki ardışık sıra istatistiğinin ortak olasılık yoğunluk fonksiyonu;

$$f_{i,i+1:n}(x_1,x_2) = \frac{n!\,f(x_1)f(x_2)F^{i-1}(x_1)[1-F(x_2)]^{n-i-1}}{(i-1)!\,(n-i-1)!}; \qquad x_1 \leq x_2 \qquad (6)$$

elde edilir ve herhangi iki sıra istatistiği $X_{r:n}$ ve $X_{s:n}$ (r<s)'nin ortak olasılık yoğunluk fonksiyonu; denklem (4)'e göre

$$f_{r,s:n}(x_1,x_2) = n!\,f(x_1)f(x_2)\frac{F^{r-1}(x_1)\left[F\left(x_2-F(x_1)\right)\right]^{s-r-1}[1-F(x_2)]^{n-s}}{(r-1)!\,(s-r-1)!\,(n-s)!};$$

$$x_1 \leq x_2 \qquad (7)$$

olarak elde edilir.

1.4 İki Sıra İstatistiğin Koşullu Dağılımı

İki sıra istatistiğin koşullu dağılımında ilgilenilen ana konularından biri simulasyon çalışmalarıdır. Verilen $X_{s:n} = x_2(r<s)$ ile $x_{r:n}$'nin olasılık yoğunluk fonksiyonu denklem (7) ve (3)'de yerine konulduğunda;

$$f_{X_{r:n}/X_{s:n}}(x_1/x_2) = \frac{f_{r,s:n}(x_1,x_2)}{f_{s:n}(x_2)} = \frac{(s-1)!\,f(x_1)F^{r-1}(x_1)\left[F\left(x_2-F(x_1)\right)\right]^{s-r-1}F^{1-s}(x_2)}{(r-1)!\,(s-r-1)!}$$

$$x_1 \leq x_2 \qquad (8)$$

elde edilir. Bu ifade (3) denklemi ile karşılaştırıldığında, $x_{s:n} = x_2$ parent populasyonundan s-1 hacimlik örneğindeki r'inci sıra istatistiğinin dağılımıdır ve $x_{r:n}$'nin dağılımı x_2'nin sağ tarafında kesilmiştir.

$x_{r:n} = x_1$ iken $x_{s:n}$'nin olasılık yoğunluk fonksiyonu;

$$f_{X_{s:n}/X_{r:n}}(x_2/x_1) = \frac{f_{r,s:n}(x_1, x_2)}{f_{s:n}(x_1)}$$

$$= \frac{(n-1)!\, f(x_2)[1-F(x_2)]^{n-s}\left[F(x_2 - F(x_1))\right]^{s-r-1}[1-F(x_1)]^{r-n}}{(n-s)!\,(s-r-1)!} \qquad (9)$$

şeklindedir. Benzer olarak bu ifade (3) denklemi ile karşılaştırıldığında parent populasyonundan alınan (n-r) hacimlik örnekte (s-r)'inci sıra istatistiğinin dağılımı $x_{s:n}$'nin dağılımıdır ve x_1'in solunda kesilmiştir.

1.5 İki Ardışık Sıra İstatistiğinin Koşullu Dağılımı

İki ardışık sıra istatistiği için $x_{i+1:n} = x_2$ koşulu altında $x_{i:n}$'nin olasılık yoğunluk fonksiyonu denklem (8)'da yerine konularak;

$$f_{X_{i:n}/X_{i+1:n}}(x_1/x_2) = \frac{f_{i,i+1:n}(x_1, x_2)}{f_{i+1:n}(x_2)}$$

$$= \frac{i f(x_1) F^{i-1}(x_1)}{F^{i}(x_2)} \qquad (10)$$

Ve kümülatif dağılım fonksiyonu;

$$F_{X_{i:n}/X_{i+1:n}}(x_1/x_2)$$

$$= \left(\frac{F(x_1)}{F(x_2)}\right)^{i} \qquad (11)$$

$x_{i:n} = x_1$ koşulu ile $x_{i+1:n}$'nin olasılık yoğunluk fonksiyonu ve kümülatif dağılım fonksiyonu;

$$f_{X_{i+1:n}/X_{i:n}}(x_2/x_1) = \frac{f_{i,i+1:n}(x_1, x_2)}{f_{i:n}(x_1)}$$

$$= \frac{(n-i)!\, f(x_2)[1-F(x_2)]^{n-i-1}}{[1-F(x_1)]^{n-i}} \qquad (12)$$

ve

$$F_{X_{i+1:n}/X_{i:n}}(x_2/x_1) = 1 - \left(\frac{1-F(x_2)}{1-F(x_2)}\right)^{n-1} ;x_1 \leq x_2 \qquad (13)$$

1.6 İki Sıra İstatistiğinin Farkının Dağılımı

İki sıra istatistiği $x_{s:n}$ ve $x_{r:n}$ $(r \prec s)$ arasındaki farkın dağılımını türetmek için (6)'de verilen ortak olasılık yoğunluk fonksiyonu dönüşümü uygulamalıyız.

$$\left.\begin{array}{r} U = X_{r:n} \\ V = X_{s:n} - X_{r:n} \end{array}\right\} \qquad (14)$$

Tersi alınırsa;

$$\left.\begin{array}{r} X_{r:n} = U \\ X_{s:n} = U + V \end{array}\right\} \qquad (15)$$

Jakobien ile;

$$J = \begin{vmatrix} 1 & 0 \\ 1 & 1 \end{vmatrix} = 1 \qquad (16)$$

Böylece (U,V)'nin ortak olasılık yoğunluk fonksiyonu;

$$g(u,v) = \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} f(u)f(u+v)F^{r-1}(u).[F(u+v)-F(u)]^{s-r-1}[1-F(u+v)]^{n-s}$$

$$0 \leq v \qquad (17)$$

olarak elde edilir ve aşağıda gösterilen V marjinal olasılık yoğunluk fonksiyonuna sahiptir.

$$h(v) = \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!} \int_{-\infty}^{\infty} f(u)f(u+v)F^{r-1}(u).[F(u+v)-F(u)]^{s-r-1}[1-F(u+v)]^{n-s}du$$

$$0 \leq v \qquad (18)$$

2. BİR UYGULAMA ÖRNEĞİ

Eğer $X_1, X_2, \dots, X_n$ birbirini izleyen günlerde bir nehrin su seviyeleri ise alt sıra istatistiğinin özellikleri $X_{1:n}, X_{2:n}, \dots, X_{k:n}$ kıtlığın modellenmesi için ve $X_{n-k:n}, X_{n-k+1:n}, \dots, X_{n:n}$ üst sıra istatistiğinin özellikleri ya da maksimum $X_{n:n}$ su baskını riski ve selden korunma tasarımları için kullanılır.

Bir nehirde son 60 yıl boyunca metre küp başına yıllık maksimum su baskınları Tablo 1' de gösterilmiştir. Gelecek 20 yılda aşanların ortalama değeri 4 olması için bir su baskını tasarım değerinin seçileceği varsayılsın (Castillo, 1988:334).

Eğer geçmiş *n* denemedeki *m*.inci en büyük gözlemi aşma olasılığı p_m ise gelecek *N* denemedeki *r* aşanların olasılığı;

$$w\left(n, m, N, \frac{r}{p_m}\right) = \binom{N}{r} p_m^r [1 - p_m]^{N-r}$$

şeklindedir. Aşanların ortalaması, $\bar{r}(n, m, N)$, binomial değişkenin ortalaması Np_m' dir, toplam olasılık kuralı ve $U_{m:n}$ m.inci sıra istatistiğinin ortalaması m/(n+1)' dir.

$$\bar{r}(n, m, N) = \int_0^1 Np_m \, f(p_m) dp_m = N\mu_{U_{m:n}} = \frac{Nm}{n+1}$$

$$\bar{r}(60, m, 20) = \frac{20m}{61} \approx 4 \quad \rightarrow \quad m \approx 12$$

Tablo 1:
Yağış Verileri

24.21	26.46	29.48	30.32	31.60
32.88	33.03	33.63	35.14	35.23
35.59	35.89	35.95	36.7	36.49
36.50	37.13	37.48	38.01	38.21
38.53	38.91	39.26	39.45	40.32
40.36	40.49	40.69	41.03	41.05
41.54	42.62	42.82	42.91	43.05
43.31	43.34	43.42	43.65	43.87
44.71	45.04	45.58	46.00	48.29
48.76	49.28	49.43	50.17	50.45
50.73	51.90	52.54	52.94	54.01
57.84	60.10	61.95	67.76	75.70

Seçilmesi gereken değer serideki 12.inci en büyük sıra istatistiğidir. Metreküp başına 50.17 olarak bulunur.

Yıllık olarak maksimum akışın kümülatif dağılım fonksiyonu metre küp başına;

$$F(x) = \exp\left[-\exp\left(-\frac{x-38.5}{7.8}\right)\right]$$

Yıllık maksimum 60 ve 70 metreküp akışın geri dönüş periodu;

$$\tau_{60} = \frac{1}{1-F(60)} = 16.25 \text{ yıl}$$

$$\tau_{70} = \frac{1}{1-F(70)} = 57.24 \text{ yıl}$$

Bunun anlamı 60 ve 70 metreküp akış gerçekleşmesi her 16.25 ve 57.24 yılda bir kez olmaktadır.

Periyottaki karakteristik en büyük değeri aşma olasılığı;

$$1 - F^n(u_n) = 1 - \left(1 - \frac{1}{n}\right)^n$$

Büyük örnek için 1-exp(-1)=0.6321’ dir.

Yıllık olarak bir nehrin maksimum su akışının kümülatif dağılım fonksiyonu metreküp başına aşağıdaki şekilde verildiğinde;

$$F(x) = \exp\left[-\exp\left(-\frac{x-38.5}{7.8}\right)\right]$$

Eğer birim periyotta belirli bir değeri aşanların sayısının beklenen değeri 1 ise *X* şans değişkeninin belirli bir değeri *x*, *n* birim period zamanı için karakteristik en büyük değerdir. Değer n’ e bağlı olduğu için karakteristik en büyük değer u_n olarak gösterilir.

Aynı bakış açısıyla karakteristik en küçük değer de v_n olarak gösterilir.

Karakteristik en büyük değer aşağıdaki eşitliği sağlar:

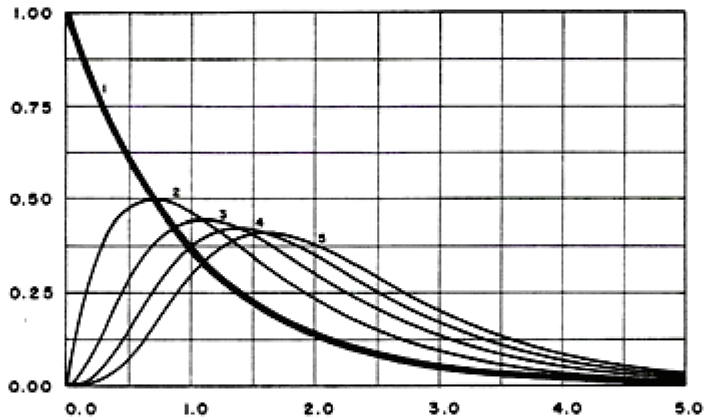
$$n[1 - F(u_n)] = 1 \quad \rightarrow \quad F(u_n) = 1 - \frac{1}{n}$$

beş yıllık dönem için su akışının karakteristik en büyük değeri;

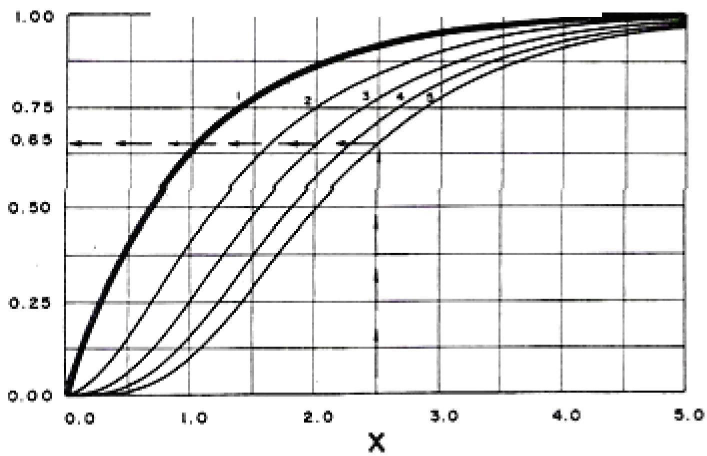
$$F(u_5) = 1 - \frac{1}{5} = 0.8$$

$u_5 = 50.2$ metreküptür.

Şekil 1.
E(1) Parent Populasyonundan 1'den 5'e Kadar Alınan Örneklerin Maksimumunun Olasılık Yoğunluk Fonksiyonu Ve Kümülatif Dağılım Fonksiyonu



Olasılık Yoğunluk Fonksiyonu



Kümülatif Dağılım Fonksiyonu

Bir kuraklık I_0'dan daha fazla yağış olmayan t_0 süreci için tanımlanmıştır. Eğer yağış yoğunluğu I_0'dan büyük olduğu varsayılırsa, ortalama μ yağmur/yıl 'nın poisson süreci olarak davrandığını varsayılmaktadır. Üstel bir parent populasyonu $E(\mu)$,'den alınan n hacimlik rasaal örneğin maksimumunun olasılığı ile aynı zamanda kuraklılığın olmama olasılığı t_0'dan daha küçük olacaktır. Çünkü ardışık yağışla arasındaki tüm süre t_0'dan daha küçük olursa kuraklık olmaz. (Birbirini izleyen yağışlar arası periyotlar üstel şans değişkenleridir. $E(\mu)$ Şekil 1' de E(1) parent populasyonundan 1'den 5'e kadar alınan örneklerin maksimumunun olasılık yoğunluk fonksiyonu ve kümülatif dağılım fonksiyonu gösterilmiştir. Eğer $t_0 = 3$ ay ve $\mu = 10$ olursa I_0'dan daha fazla 5 yağış gerçekleşmesinden sonra kuraklığın olmama olasılığı 0.65'dir.

SONUÇ

Sıra istatistikleri sigortacılık, deprem ölçümleri ve analizi, doğal afetlerin tahminlenmesi, en zayıf halka prensibi, mühendislik tasarımı çalışmaları gibi çalışmalarda da kullanılmakta ve bu alanlarda da pek çok uygulamalar bulunmaktadır. Bu çalışmada sıra istatistiklerinin temel kavramları ve sıra istatistiğinin dağılım özellikleri, maksimum ve minimumun dağılımı, bazı sıra istatistiklerinin ortak dağılımı, sıra istatistiğinin koşullu dağılımı, sıra istatistiğinin farkının dağılımı incelenerek uygulama alanlarına ilişkin bir yağış verisi örneği değerlendirilmiştir.

KAYNAKÇA

BALAKRISHNAN, N. and CHEN, W. W. S.(1999), *Handbook of Tables for Order Statistics from Lognormal Distributions with Applications.* Amsterdam, Netherlands: Kluwer.

BALAKRISHNAN, N. and COHEN, A. C.(1991), *Order Statistics and Inference.* New York: Academic Press.

BALAKRISHNAN, N. and RAO, C. R. (Eds.)(1998), *Handbook of Statistics, Vol. 16: Order Statistics: Theory and Methods.* Amsterdam, Netherlands: Elsevier.

BALAKRISHNAN, N. and RAO, C. R. (Eds.).(1998) *Order Statistics: Applications.* Amsterdam, Netherlands: Elsevier.

CASTILLO E., (1988), Extreme Value Theory in Engineering, London: Academic Press Inc.

DAVID, H. A.(1981), *Order Statistics, 2nd ed.*, New York: Wiley.