ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

TAM VE SANSÜRLÜ ÖRNEKLEM DURUMLARINDA WEIBULL DAĞILIMI İÇİN BAZI İSTATİSTİKİ SONUÇ ÇIKARIMLARI

Dilşen TAMAM

İSTATİSTİK ANABİLİM DALI

ANKARA
2008

Her hakkı saklıdır

ÖZET

Yüksek Lisans Tezi

TAM VE SANSÜRLÜ ÖRNEKLEM DURUMLARINDA WEIBULL DAĞILIMI İÇİN BAZI İSTATİSTİKİ SONUÇ ÇIKARIMLARI

Dilşen TAMAM

Ankara Üniversitesi

Fen Bilimleri Enstitüsü

İstatistik Anabilim Dalı

Danışman: Prof. Dr. Fahrettin ARSLAN

Yaşam analizi, yapılacak çalışmalarda başarısızlık olarak adlandırılan ve genellikle bozulma, ölüm, çürüme v.b. olarak karşımıza çıkan olayların meydana gelmesine kadar geçen süre olarak elde edilen verilerin analizidir. Yaşam analizi, mühendislik, sosyal bilimler, aktüerya ve tıp alanlarında yapılan çalışmalarda kullanılmaktadır ve bu bilimler için önemli bir alandır. Bu çalışmada, yaşam analizinin temel kavramları verilecektir. Yaşam analizini diğer istatistiksel analiz yöntemlerinden ayıran en önemli özellik sansürlü veri ile çalışılmasıdır. Çalışmada, yaşam analizinin temel karakteristiği olan sansürlemenin çeşitleri için olasılık yoğunluk fonksiyonlarının elde edilişi gösterilecektir. Ayrıca, yaşam modellerinde en yaygın kullanılan dağılım olan Weibull dağılımı verilerek, dağılımın parametre tahminleri, tam ve sansürlü veri durumlarının her ikisi için yapılacaktır. Yaşam analizinde faydalı olan yaşam ve Hazard fonksiyonları, Weibull dağılımına uygun olduğu gösterilen bir örnekle ifade edilecektir. Tezdeki amaç, yaşam sürelerine yönelik uygun dağılım için parametre tahminlerini elde etmektir.

Ekim 2008, 55 sayfa

Anahtar Kelimeler: Yaşam Analizi, Sansürleme, Sansürlü Veri, Parametrik Yaşam Modelleri, Weibull Dağılımı, Parametre Tahmini.

ABSTRACT

Master Thesis

SOME INFERENCE ABOUT THE WEIBULL DISTRIBUTION BY USING THE COMPLETE AND CENSORED SAMPLING DATA

Dilşen TAMAM

Ankara University
Graduate School of Natural and Applied Sciences
Department of Statitistics

Supervisor: Prof. Dr. Fahrettin ARSLAN

The survival analysis is the analysis of data that are called as a failure within the studies that will be done, and, the data obtained as elapsed time which could be seen until the formulation of events such as break, dead and putrefaction. The survival analysis is used in the areas of engineering, social science, actuaria and the medicine and it is very important area for these sciences. Within the framework of this study the basic concepts of survival analysis will be detailed. The major characteristic of survival analysis with the comparison of other studies is to be studied with especially censored data. In this study it will be shown the acquisition of probability density function for the censorization that is the basic charecteristic of survival analysis. Also, the Weibull distiribution that is common for survival modeling will be given and the parameter estimation of distribution will be performed for both complete and censored data situations. The survival and hazard functions that are useful for survival analysis will be denoted by an example that shows the convenience of Weibull distribution. The purpose of this thesis is to obtain the estimation of parameters for the convenient distribution for the duration of survival.

October 2008, 55 pages

Key Words: Survival Analysis, Censoring, Censored Data, Parametric Survival Models, Weibull Distribution, Parameter Estimation

TEŞEKKÜR

Çalışmalarımı yönlendiren, bana duyduğu güven ile çalışmamda bilgi, öneri ve yardımlarıyla bana yol gösteren danışman hocam, Sayın Prof. Dr. Fahrettin ARSLAN (Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü)'a ve hiçbir zaman benden desteğini esirgemeyen sevgili aileme ve arkadaşlarıma teşekkürlerimi sunarım.

Dilşen TAMAM
Ankara, Ekim 2008

İÇİNDEKİLER

ÖZET……. i

ABSTRACT…….. ii

TEŞEKKÜR……..... iii

SİMGELER DİZİNİ……... vi

ŞEKİLLER DİZİNİ……. viii

ÇİZELGELER DİZİNİ……... ix

1. GİRİŞ……..... 1

2. YAŞAM ANALİZİ……... 3

2.1 Yaşam Süresi Dağılımları……..... 4

2.1.1 Sürekli modeller……..... 4

2.1.2 Kesikli modeller…….…… 9

2.2 Veri Yapısı…….. 10

3. SANSÜRLEME…….. 12

3.1 Sağdan Sansürleme…… 13

3.1.1 I.Tür sansürleme…… 13

3.1.2 II.Tür sansürleme……... 14

3.1.3 Bağımsız rasgele sansürleme……..... 15

3.1.4 İlerletilmiş II. tür sansürleme…….……..... 15

3.2 Soldan Sansürleme……..... 17

3.3 Aralık Sansürlemesi……... 17

3.4 İkili Sansürleme…….. 18

4. PARAMETRİK YAŞAM MODELLERİ……..... 19

4.1 Weibull Dağılımı……..... 19

5. PARAMETRİK YAŞAM MODELLERİNDE TAHMİN……..... 22

5.1 Tam ve Sansürlü Örneklem Durumlarında En Çok Olabilirlik Yöntemi 22

5.2 Weibull Dağılımı İçin Parametre Tahmini……..... 24

5.2.1 Tam örneklem durumunda parametre tahmini……..…..... 25

5.2.2 Sansürlü (tam olmayan) örneklem durumunda parametre tahmini 26

5.2.2.1 I. Tür sansürleme durumunda parametre tahmini…….. 26

5.2.2.2 II. Tür sansürleme durumunda parametre tahmini……. 28
6. UYGULAMA……... 31
7. TARTIŞMA VE SONUÇ…… 43
KAYNAKLAR……..... 44
EKLER …… 46
EK 1 Akciğer Kanseri Hastalarına Ait Veriler…..……. 47
EK 2 Minitab 14 Sonuçları …… 51
EK 3 I.Tip Sansürlenmiş Veri İçin Weibull Dağılımının Parametrelerini Hesaplayan Matlab Programı…… 53
ÖZGEÇMİŞ……..... 55

SİMGELER DİZİNİ

T	Yaşam süresi rasgele değişkeni
$f(t)$	Olasılık yoğunluk fonksiyonu
$F(t)$	Dağılım fonksiyonu
$S(t)$	Yaşam fonksiyonu
$h(t)$	Hazard fonksiyonu
$\Lambda(t)$	Birikimli Hazard Fonksiyonu
L_i	Sansürleme zamanı
min	Minimum
max	Maksimum
r	Gözlenen başarısızlık sayısı
T_i	Bireyin yaşam süresi
t_r	r-inci başarısızlık sayısı
λ	Lamda
β	Beta
θ	Teta
$\hat{\lambda}$	Lamda parametresinin tahmini
$\hat{\beta}$	Beta parametresinin tahmini
$\hat{\theta}$	Teta parametresinin tahmini
δ , Δ	Delta

$E(T)$	Beklenen değer
$Var(T)$	Varyans
$E(T^r)$	r-inci moment
$s(U)$	Sansürlenmemiş gözlem kümesi eleman sayısı
$\sum$	Toplam
$\prod$	Çarpım
$\int$	İntegral
∞	Sonsuz
$>$	Büyük
$<$	Küçük
!	Faktöriyel

ŞEKİLLER DİZİNİ

Şekil 2.1 Sansürsüz veri……. 11

Şekil 2.2 Sağdan sansürlü veri……. 11

Şekil 3.1 İlerletilmiş II. tür sansürleme…… 17

Şekil 6.1 Dört farklı dağılım için yaşam olasılık grafikleri…….. 34

Şekil 6.2 Weibull dağılımının parametreleriyle ilgili grafikler…… 36

Şekil 6.3 Akciğer kanseri hastalarına ait yaşam fonksiyonu grafiği…… 40

Şekil 6.4 Akciğer kanseri hastalarına ait Hazard fonksiyonu grafiği……... 41

ÇİZELGELER DİZİNİ

Çizelge 6.1 Akciğer kanseri verisinde kullanılan değişkenlerin kodlarına göre yaşam süreleri..... 32

Çizelge 6.2 Durum değişkenleri frekans tablosu…… 33

Çizelge 6.3 Weibull dağılımının bazı karakteristikleri……... 35

1. GİRİŞ

Yaşam modelleri, ölüm, bozulma, çürüme gibi başarısızlık (failure) olarak adlandırılan olayların meydana çıkmasına kadar geçen sürenin modellemesidir. Yaşam analizi de bu olayların meydana gelmesine kadar geçen süre olarak elde edilen verilerin analizidir. Canlılar için genellikle "ölüm" kavramından bahsedilirken, cansızlar için "bozulma" kavramından bahsedilir. Yani yapılacak çalışmaya göre olayın ne olduğu belirlenir. Bu çalışmada da genellikle canlılar söz konusu olacağı için 'başarısızlık' yerine 'ölüm' kavramı kullanılacaktır.

Yaşam analizi, tıp, mühendislik, sosyal bilimler, sigortacılık gibi birçok alanda kullanılmaktadır.

Yaşam analizinde en önemli unsur yaşam süresidir. Bu çalışmada, yaşam analizinin önemli fonksiyonlarından olan Hazard fonksiyonu, birikimli Hazard fonksiyonu, yaşam fonksiyonu gibi temel kavramlar açıklanacak, olasılık yoğunluk fonksiyonu ve dağılım fonksiyonu ile aralarındaki ilişki verilecektir.

Yaşam modellerinin temel kavramlarından biri de sansürlemedir. Sansürlenmiş verilerin varlığı, yaşam analizini diğer istatistik modellerinden ayıran en belirgin özelliktir. Sansürleme kavramından bahsedilirken, sansürleme çeşitleri verilip sansürleme çeşitlerine göre olasılık yoğunluk fonksiyonları verilecektir.

Parametrik olmayan istatistiksel yöntemler belli bir dağılım varsayımı gerektirmediğinden pratikte daha kullanışlıdır. Ancak, parametrik yöntemler daha profesyonel sonuçlar vereceğinden, yaşam analizinde kullanılan önemli dağılımlarla çalışmak daha çok tercih edilen bir yoldur. Bu nedenle, yaşam analizinde önemli bir yer tutan ve en yaygın kullanılan Weibull dağılımı verilecektir.

Parametrik yaşam modellerinde tahmin yapmak için tam örneklem ve sansürlü örneklem durumlarının her ikisi için de en çok olabilirlik yöntemi verilmiştir. Bu yöntem Weibull dağılımının parametrelerinin tahmininde de kullanılmıştır. Her iki durumda da Weibull dağılımının en çok olabilirlik tahmin edicileri araştırılmıştır.

Uygulama bölümünde, sansürlü veri içeren akciğer kanseri 137 hastaya ilişkin veriler için en uygun dağılım belirlenip, bu dağılım için parametre tahminleri bulunacaktır. Ayrıca, yaşam fonksiyonu ve Hazard fonksiyonu gibi yaşam analizinde önemli yer tutan ve veriler hakkında önemli ipuçları veren fonksiyonların grafikleri MINITAB 14 paket programı yardımıyla çizdirilecektir.

2. YAŞAM ANALİZİ

Belirli bir hastalığa yakalanan bireylerin, hastalığın tanısından sonra, birey gözlem altında iken, çeşitli tedavi yöntemlerinin yaşam süresine etkisini araştırmak ya da uygulanan tedavi yöntemiyle ne kadar süre yaşayabileceğini tahmin etmek, tedavi tiplerinin ve diğer faktörlerin yaşam süresine etkilerini incelemek amacıyla geliştirilmiş olan yöntemler ailesi "Yaşam Analizi" olarak adlandırılır.

Yaşam analizi, T zaman süresinde (araştırma periyodu) gözlenen n sayıda hasta birimden elde edilen yaşam sürelerinin dağılımını açıklayarak, yaşam süresini etkileyen ve etkilemesi olası değişkenleri içeren modeller kurarak, bu modellere göre parametre tahminleri yapmayı amaçlamaktadır.

Bir hastalığa yakalanan kişinin aldığı tedavi türünün ve hastanın zamana bağlı olarak durumunda değişmelere neden olan faktörlerin, yaşam süresi üzerine etkilerini ortaya koymak için yaşam analizi yöntemlerinden yararlanılır.

Hastalık olgularında, hastaların herhangi bir medikal ya da cerrahi girişimden (ilaç ile tedavi, ameliyat, ilaç+ameliyat v.b.) sonraki yaşamlarının yaşam süreleri, gün, ay, yıl gibi sürelerdir. Yaşam analizi yöntemleri, yaşam sürelerini ve diğer faktör değişkenleri içeren veri setlerinde yaşam olasılıkları, ölüm olasılıkları, ortalama yaşam süresi, ortanca yaşam süresi tahminleri yapmayı amaçlayan yöntemlerdir (Özdamar 2003).

Yaşam analizi aynı zamanda mekanik parçaların ve elektronik eşyaların farklı şartlar altında ne kadar bozulmadan kalabileceği ile ilgili tahminlerde veya hayvanların kullanıldığı laboratuar deneylerinde de uygulanabilir.

Yaşam analizi, bazı uygulamalarda güvenilirlik analizi veya sağ kalım analizi olarak da isimlendirilmektedir.

Yaşam analizinde kullanılan "başarısızlık" terimi, incelenen konunun denekte görülmesi durumudur. Canlılar için genelde ölüm veya hastalık, mekanik aletler için ise bozulma anlamına gelir (Nelson 1982).

2.1 Yaşam Süresi Dağılımları

Yaşam analizi, başarısızlık olarak adlandırılan bir olayın meydana gelmesine kadar geçen sürede elde edilen verilerin analizidir.

Yaşam analizinde geçen bazı kavram ve gösterimler aşağıda kısaca açıklanmıştır (Özdamar 2003).

Yaşam süresi, bir bireyin belirli girişime ya da etkene maruz kaldıktan sonra iyileşmesine, hastalığın tekrarlamasına ya da ölüme kadar geçen süreye denilmekte ve t_i ile gösterilmektedir.

Yaşam fonksiyonu, yaşam sürelerinin olasılık dağılımına denilmektedir. Fonksiyon, yaşamsal verilerin genel eğilimini matematiksel bir modelle ifade eder. Yaşam fonksiyonu bir olasılıktır ve $S(t)$ ile gösterilmektedir.

Ani ölüm olasılığı (Hazard fonksiyonu), sağ olan bir kişinin, belirli bir zamanda (anda) ölüm olasılığı, taşıdığı ölüm riskidir ve $h(t)$ ile gösterilir.

Birikimli ölüm fonksiyonu (birikimli Hazard fonksiyonu), T zamanı içinde belirli bir t zamanı (anı) için hesaplanmış olan ölüm olasılıklarının birikimli fonksiyonudur ve $\Lambda(t)$ ile gösterilir.

Yaşam süresi dağılımları sürekli ve kesikli modeller olmak üzere ikiye ayrılır.

2.1.1 Sürekli Modeller

Yaşam süresinin negatif olamayacağı kabulü göz önüne alınarak, bir kitledeki canlıların yaşam süresi, negatif olmayan sürekli bir T rasgele değişkeni ile gösterilsin.

Herhangi bir canlının t zamanından önce ölmesi olasılığı,

$$F(t) = P(T \le t) = \int_0^t f(x)dx \quad , \; t > 0 \tag{2.1}$$

olarak tanımlanan "Dağılım Fonksiyonu" yardımıyla bulunur. Burada $f(t)$ ise, T rasgele değişkeninin olasılık yoğunluk fonksiyonudur.

Yukarıda tanımlanan T rasgele değişkeni için, herhangi bir canlının t zamanına kadar yaşadığı biliniyor olsun. Bu canlının t zamanından sonra da yaşaması olasılığı,

$$S(t) = P(T > t) = \int_{t}^{\infty} f(x)dx \tag{2.2}$$

şeklinde tanımlanan fonksiyon yardımı ile bulunur. Bu fonksiyona "Yaşam Fonksiyonu" adı verilir.

Yaşam fonksiyonu ile dağılım fonksiyonu arasında,

$$S(t) = 1 - F(t) \tag{2.3}$$

şeklinde bir ilişki vardır (Lawless 2003, London 1988, Miller 1981).

Dağılım fonksiyonu $F(t)$ azalmayan bir fonksiyon,

$$\lim_{t \to -\infty} F(t) = 0$$

$$\lim_{t \to \infty} F(t) = 1$$

dir. Burada,

$$\lim_{t \to 0} S(t) = S(0) = 1 - F(0) = 1 - 0 = 1$$

$$\lim_{t \to \infty} S(t) = S(\infty) = 1 - F(\infty) = 1 - 1 = 0$$

olması sebebiyle yaşam fonksiyonu $S(t)$'nin yukarıdaki özellikleri taşıyan azalan bir fonksiyon olduğu söylenir.

T rasgele değişkeninin olasılık fonksiyonu, küçük bir zaman aralığında bir bireyin başarısız olma olasılığının limitidir.

Bu fonksiyon,

$$f(t) = \lim_{\Delta t \to 0} \left[\frac{P(t < T \le t + \Delta t)}{\Delta t} \right]$$

şeklinde yazılır.

t zamanından sonra yaşadığı bilinen bir bireyin, t zamanındaki ani başarısızlık ya da ölüm oranı,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \,|\, T > t)}{\Delta t} \quad , \quad t \ge 0 \tag{2.4}$$

biçiminde tanımlanan "Hazard Fonksiyonu" ile belirlenir (Lee 1984).

Hazard fonksiyonu, T yaşam süresinin $t \ge 0$ ve pozitif küçük bir Δt değeri için $[t, t + \Delta t)$ aralığında yaşamın koşullu dağılımıdır. $\lambda(t)$ veya $r(t)$ olarak da gösterilir.

Yaşam fonksiyonu, yaşama olasılığını incelerken; Hazard fonksiyonu başarısızlığı (ölümü) inceler ve zamana bağlı ölüm riskini belirler. Hazard fonksiyonunun grafiği, yapılan çalışma için model oluşturmada önemli ipuçları verir. Bu sebeple, yaşam fonksiyonu gibi, Hazard fonksiyonu da, yaşam modellerinin önemli bir karakteristiğidir.

Ayrıca,

$$\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \,|\, T > t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t, T > t)}{\Delta t} \cdot \frac{1}{P(T > t)} \\
&= \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t)}{\Delta t} \cdot \frac{1}{P(T > t)} \\
&= \frac{1}{P(T > t)} \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t)}{\Delta t} \\
&= \frac{1}{P(T > t)} \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \\
&= \frac{f(t)}{S(t)}
\end{aligned} \tag{2.5}$$

dır. Başka bir ifadeyle $h(t)$ fonksiyonu,

$$h(t)=\frac{f(t)}{S(t)}=\frac{f(t)}{1-F(t)}=-\frac{d}{dt}\ln(S(t)) \tag{2.6}$$

olarak da ifade edilebilir (Gross anda Clark 1975, Cox and Oakes 1984).

(2.6) eşitliğinde her iki tarafın integrali alınırsa,

$$\int_0^t h(x)dx=-\ln(S(t))$$

$$S(t)=e^{-\int_0^t h(x)dx}$$

$$S(t)=e^{-\Lambda(t)} \tag{2.7}$$

sonucuna ulaşılır.

Burada $\Lambda(t)$, "Birikimli (Kümülatif) Hazard Fonksiyonu" olarak adlandırılır ve

$$\Lambda(t)=\int_0^t h(x)dx \tag{2.8}$$

şeklinde ifade edilir (Lawless 2003).

Kümülatif Hazard fonksiyonu aşağıdaki özellikleri taşır:

(i) $\Lambda(0)=0$ 'dır. Çünkü,

$$\begin{aligned}\lim_{t\to 0}\Lambda(t)&=\lim_{t\to 0}\int_0^t h(x)dx\\&=\lim_{t\to 0}\int_0^t \frac{f(x)}{1-F(x)}dx\\&=\lim_{t\to 0}\left(-\ln\left(1-F(t)\right)\right)\end{aligned}$$

$$= \lim_{t\to 0}\left[-\ln\left(S(t)\right)\right]$$
$$= 0 \tag{2.9}$$

Veya, $S(0)=1$ olduğundan $\Lambda(0)=0$ olduğu görülür.

(ii) $\Lambda(\infty)=\infty$ 'dır. Çünkü,

$$\lim_{t\to\infty}\Lambda(t) = \lim_{t\to\infty}\int_0^t h(x)dx$$
$$= \lim_{t\to\infty}\left(\int_0^t \frac{f(x)}{1-F(x)}dx\right)$$
$$= \lim_{t\to\infty}\left(-\ln\left(1-F(t)\right)\right)$$
$$= \lim_{t\to\infty}\left(-\ln\left(S(t)\right)\right)$$
$$= \infty \tag{2.10}$$

Veya, $S(\infty)=0$ olduğundan $\Lambda(\infty)=\infty$ olduğu görülür.

x anından sonra yaşadığı bilinen canlının $(x, x+t)$ yaşam aralığında (t zaman sonra) ölmesi olasılığı ise,

$$ {}_tq_x = q(x, x+t)$$
$$= P(x < X \le x+t \,|\, X > x)$$
$$= \frac{P(x < X \le x+t, X > x)}{P(X > x)}$$
$$= \frac{P(x < X \le x+t)}{P(X > x)}$$
$$= \frac{\int_x^{x+t} f(y)dy}{\int_x^{\infty} f(y)dy}$$
$$= \frac{F(x+t)-F(x)}{1-F(x)}$$
$$= \frac{1-S(x+t)-1+S(x)}{S(x)}$$

$$= \frac{S(x) - S(x+t)}{S(x)}$$
$$= 1 - \frac{S(x+t)}{S(x)} \qquad (2.11)$$

dir.

Sonuç olarak, yaşam fonksiyonu, dağılım fonksiyonu ve Hazard fonksiyonu arasındaki ilişki (2.12) ile ifade edilir (Le 1997).

$$S(t) = 1 - F(t) = \exp\{-\Lambda(t)\} \qquad (2.12)$$

2.1.2 Kesikli modeller

Yaşam süreleri bazen kesikli olarak ifade edilmiş olabilir. T rasgele değişkeni $0 \le t_1 \le t_2 \le \ldots$ değerlerini alıyorsa, T rasgele değişkeninin olasılık fonksiyonu,

$$f(t_j) = P(T = t_j) \quad , \quad j = 1, 2, \ldots \qquad (2.13)$$

ilgili yaşam fonksiyonu,

$$S(t_j) = P(T \ge t_j) = \sum_{j:t_j > t} f(t_j) \qquad (2.14)$$

Hazard fonksiyonu,

$$h(t_j) = P(T = t_j \,|\, T \ge t_j)$$
$$= \frac{P(T = t_j, T \ge t_j)}{P(T \ge t_j)}$$
$$= \frac{f(t_j)}{S(t_j)} \quad , \quad j = 1, 2, \ldots \qquad (2.15)$$

şeklinde tanımlanır.

Ayrıca,

$$f(t_j) = P(T \geq t_j) - P(T \geq t_{j+1}) \tag{2.16}$$

olduğundan,

$$\begin{aligned} h(t_j) &= \frac{f(t_j)}{S(t_j)} \\ &= \frac{P(T \geq t_j) - P(T \geq t_{j+1})}{S(t_j)} \\ &= \frac{S(t_j) - S(t_{j+1})}{S(t_j)} \\ &= 1 - \frac{S(t_{j+1})}{S(t_j)} \end{aligned} \tag{2.17}$$

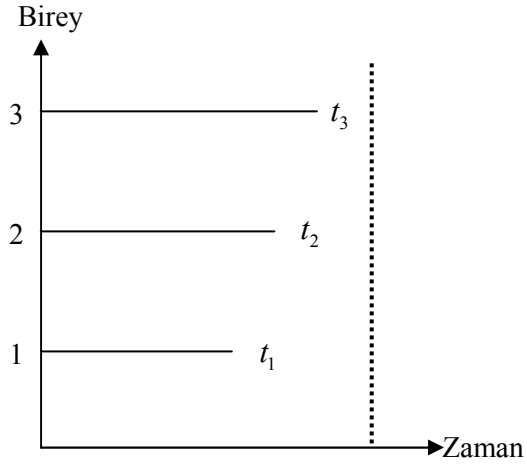
eşitliği yazılır (Lawless 2003).

2.2 Veri Yapısı

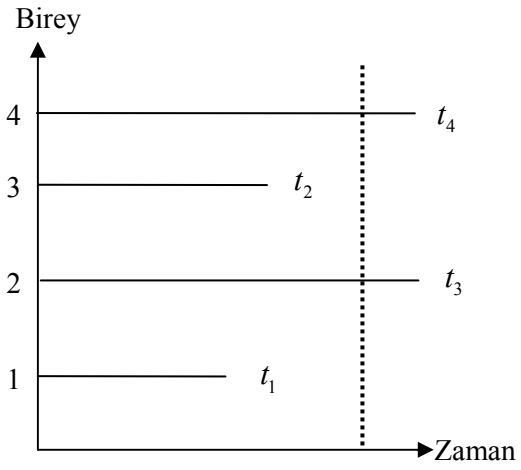
Yaşam analizindeki en önemli değişken yaşam süresidir. Yaşam süresi de çoğu kez sansürlü olmaktadır. Yani, bireyin yaşam süresi hakkında her zaman tam bir bilgiye ulaşılamayabilir. Veri yapısındaki bu farklılık, yaşam analizini diğer istatistiksel analiz yöntemlerinden ayıran en önemli özelliktir.

Tam örneklem durumunda, her bir bireyin yaşamı, çalışmanın periyodu içerisinde son bulmuştur. Veriler her bir bireyin ölüm zamanını kapsar. Ancak sansürlü örneklem durumunda çalışma tamamlandığında bireyler hala yaşıyor veya çalışma süresinin sonuna gelinmiş olmasına rağmen bireylerin yaşam durumu bilinmiyor olabilir. Ya da birey çalışma süresi içerisinde çeşitli nedenlerle kayıp gözlem durumuna düşmüş, herhangi bir nedenden dolayı çalışmadan çıkmış olabilir. Yani, tam örneklem durumunda bireylerin başarısızlık zamanı kesin belli iken, sansürlü örneklem durumunda ise başarısızlık zamanı hakkında kesin bir bilgiye ulaşılamaz.

Tam ve sansürlü veri yapıları aşağıdaki gibi Şekil 2.1 ve Şekil 2.2 ile ifade edilebilir (Nelson 1982).



Şekil 2.1 Sansürsüz Veri



Şekil 2.2 Sağdan Sansürlü Veri

3. SANSÜRLEME

Sansürleme; zaman ve maliyet gibi birtakım sınırlamalar nedeniyle, kesin olarak bilinmeyen, herhangi bir sebeple gözlenemeyen verilerin göz ardı edilmesidir. Bir çalışmada, ilgilenilen olay bir bireyin yaşam süresi olduğunda, her bir bireyin çalışmanın başlangıcından sonuna kadar gözlem altında bulundurulması çeşitli nedenlerden dolayı olanaksızdır. Bu durumda veri "sansürlüdür" denir.

Gözlemlenen birey;

- Tedavi gördüğü süre içerisinde trafik kazası gibi farklı bir sebepten ölmüş,
- Kalp yetmezliği, kan değerlerinin artması v.b. gibi sebeplerle tedaviye ara vermek zorunda kalmış,
- Başka bir hastanede veya başka bir şehirde tedaviye devam etmek zorunda kalmış,
- Tedaviye cevap vermemiş,
- Tedavi süresi içerisinde başka bir hastalığa yakalanmış,
- Tedaviden vazgeçmiş

olabilir. Bu gibi durumlarda yaşam süresi kesin olarak bilinemeyeceğinden sansürlüdür.

Yaşam modelinde meydana gelebilecek 3 durumdan söz edilebilir (Kleinbaum 1996):

(a) Birey gözlem esnasında ölebilir.
(b) Birey gözlemden geri çekilebilir. İlgilenilen olay dışında bir başka nedenden dolayı ölebilir veya uygulanan yöntemlerden beklenmeyen bir sonuç alınabilir.
(c) Birey gözlemin sonunda hala yaşıyor olabilir.

(a) durumunda bireyin yaşam süresi bilindiğinden sansürlü değildir. (b) durumunda bireyin yaşam süresi, gözlemden çekilme zamanından itibaren sansürlüdür. (c) durumunda ise, bireyin yaşam süresi çalışmanın sonlandırılma zamanına kadar bilinmesine rağmen, gözlem sonrası hakkında bir bilgi olmadığından bu bireyin yaşam süresi de sansürlüdür (Kleinbaum 1996).

Sansürleme, sağdan sansürleme ve soldan sansürleme olarak iki ana gruba ayrılır. Ayrıca, sağdan ve soldan sansürlemeler kullanılarak elde edilen aralık sansürlemesi ve ikili sansürleme de genelleştirilmiş sansürleme çeşitleri olarak incelenebilir.

Bu çalışmada, soldan sansürleme sık karşılaşılan bir sansürleme çeşidi olmadığından, en çok karşılaşılan sansürleme çeşidi olması sebebiyle, sağdan sansürleme üzerinde durulacaktır. Bölüm 6'daki uygulamalar da sağdan sansürleme üzerine yapılacaktır.

3.1 Sağdan Sansürleme

Başarısızlık olarak adlandırılan (ölüm, bozulma, çürüme v.b.) olay, çalışma için belirlenen bir durma zamanına kadar gerçekleşmezse, bireyin yaşam süresinin uzunluğu çalışmanın durma zamanının sağ tarafına geçer. Böyle bir durumda, bu bireyin yaşam süresi kesin olarak bilinmeyecek ve birey gözleme alınmayacaktır. Yani, bireyin yaşam süresi sansürlenecektir. Bu tip sansürlemeye "sağdan sansürleme" denir.

L_i sansürleme zamanı, T_i bireyin yaşam süresi olmak üzere; $T_i > L_i$ olduğunda bu bireyin yaşam süresinin sağdan sansürlenmiş olduğu söylenir.

$$\delta_i = \begin{cases} 0 & , \ T_i > L_i \\ 1 & , \ T_i \le L_i \end{cases} \qquad , \qquad i = 1, 2, ..., n \tag{3.1}$$

Eğer $\delta_i = 0 \ \left(T_i > L_i\right)$ ise birey sansürlenmiş, $\delta_i = 1 \ \left(T_i \le L_i\right)$ ise gözlenmiştir (Nelson 1982, Lawless 2003).

Sağdan sansürleme, kendi içinde bazı alt gruplara ayrılır:

1. I. Tür Sansürleme (Type I Censoring)
2. II. Tür Sansürleme (Type II Censoring)
3. Bağımsız Rasgele Sansürleme (Independent Random Censoring)
4. İlerletilmiş II. Tür Sansürleme (Progressive Type II Censoring)

3.1.1 I. Tür sansürleme

I. tür sansürlemede, her bireyin bir sansürleme zamanının olduğu düşünülür ($L_i > 0$). Bireyler sürece herhangi bir zamanda dahil olurlar ve belirlenmiş durma zamanına kadar gözlenirler. T_i bireyin çalışma süresince gözlenebildiği süre, L_i sansürleme

zamanı ise çalışmanın başlama zamanı ile bitiş zamanı arasında bir zamandır. Bu durumda, L_i sansürleme zamanı sabit bir sayıdır.

I. tür sansürleme için genel gösterim, $t_i = \min(T_i, L_i)$ ve $\delta_i = I(T_i \le L_i)$ olmak üzere, (t_i, δ_i) için olasılık yoğunluk fonksiyonu,

$$\begin{aligned} P(t_i, \delta_i) &= f\left(t_i\right)^{\delta_i} P\left(T_i > L_i\right)^{1-\delta_i} \\ &= f\left(t_i\right)^{\delta_i} S\left(t_i\right)^{1-\delta_i} \end{aligned} \tag{3.2}$$

olarak tanımlanır.

(3.1) ve (3.2) eşitliğinden,

$$P(t_i = L_i, \delta_i = 0) = P(T_i > L_i) = S(L_i) \quad , \; i = 1, 2, ..., n \qquad \text{(sansürlü)}$$

$$P(t_i, \delta_i = 1) = f(t_i) \quad , \quad t_i \le L_i \qquad \text{(sansürsüz)}$$

olarak elde edilir (Lawless 2003).

3.1.2 II. Tür sansürleme

II. tür sansürlemede, başlangıçta belirlenen bir başarısızlık sayısı vardır. n birey aynı anda gözlenmeye başlanır ve çalışmanın başında belirlenen sabit bir r tane başarısızlık gözlendiği anda çalışmaya son verilir. Çalışmanın toplam süresi, r-inci başarısızlık zamanı olan $t_{(r)}$' ye eşittir. Bu zaman, çalışmanın başında bilinmemektedir.

$T_{(1)} \le T_{(2)} \le ... \le T_{(r)}$ rasgele örneklem olmak üzere,

$T_{(1)}, T_{(2)}, ..., T_{(r)}$' nin ortak olasılık yoğunluk fonksiyonu;

$$\frac{n!}{(n-r)!}\left\{\prod_{i=1}^{r} f(t_{(i)})\right\} S(t_{(r)})^{n-r} \tag{3.3}$$

biçiminde tanımlanır. Bu ifade sıralı istatistiklerin genel biçimidir (Lawless 2003).

3.1.3 Bağımsız rasgele sansürleme

Bağımsız rasgele sansür modelinde birçok durumda sansürleme süreci, başarısızlık zamanı (failure time) ile ilişkilidir. Sonlandırma zamanı rasgele olup çalışmadan önce belli değildir, daha sonradan seçilir. Fakat bu seçim, sonlandırma sürecine kadar çalışmanın sonuçlarından etkilenir.

T, her bireyin yaşam süresi iken L de sansürleme zamanıdır. T ile L rasgele değişkenleri bağımsız sürekli rasgele değişkenlerdir.

Ayrıca; $S(t)$, T rasgele değişkeninin, $G(t)$ de L rasgele değişkeninin yaşam fonksiyonu iken $f(t_i)$, T rasgele değişkeninin ve $g(t_i)$ de L rasgele değişkeninin olasılık yoğunluk fonksiyonudur.

$t_i = \min(T_i, L_i)$ ve $\delta_i = I(T_i \le L_i)$ olmak üzere, (3.1) eşitliğinden, (t_i, δ_i) çifti için olasılık yoğunluk fonksiyonu,

$$P(t_i = t, \delta_i = 0) = P(L_i = t, T_i > L_i) = g(t)S(t)$$

$$P(t_i = t, \delta_i = 1) = P(T_i = t, T_i \le L_i) = f(t)G(t)$$

$$P(t_i, \delta_i) = [f(t_i)G(t_i)]^{\delta_i} [g(t_i)S(t_i)]^{1-\delta_i} \tag{3.4}$$

olarak elde edilir (Lawless 2003).

3.1.4 İlerletilmiş II. tür sansürleme

İlerletilmiş II. tür sansürleme, II. tür sansürlemenin genelleştirilmiş halidir. Bu sansürleme çeşidinde, söz konusu olayda yer alan n tane bireyden, başarısız olan r_1 tane birey gözlenir. Geriye kalan $n - r_1$ tane bireyden n_1 tanesi çalışmadan uzaklaştırılır. Böylece, $n - r_1 - n_1$ birey çalışmada kalmış olur. Daha sonra, başarısız olan r_2 tane birey gözlenir. Geriye $n - r_1 - n_1 - r_2$ birey kalır. Kalan bireylerden n_2

tanesi çalışmadan alınarak geriye kalan bireyler ile çalışmaya devam edilir. İşleyiş bu şekilde devam ettirilir (Şekil 3.1).

$T_{(1)} \le T_{(2)} \le ... \le T_{(r_1)}$: İlk gözlenen r_1 başarısızlık sayısı

$T^*_{(1)} \le T^*_{(2)} \le ... \le T^*_{(r_2)}$: Daha sonraki r_2 başarısızlık sayısı

olmak üzere, verinin dağılımı;

$$g_1(t_{(1)}, t_{(2)}, ..., t_{(r_1)}) g_2(t^*_{(1)}, t^*_{(2)}, ..., t^*_{(r_2)} \mid t_{(1)}, t_{(2)}, ..., t_{(r_1)}) \tag{3.5}$$

şeklindedir.

(3.3) eşitliğinden, (3.5) ifadesinin ilk terimi aşağıdaki gibi yazılabilir:

$$\frac{n!}{(n-r_1)!} \left\{ \prod_{i=1}^{r_1} f(t_{(i)}) \right\} S(t_{(r_1)})^{n-r_1} \tag{3.6}$$

(3.5) ifadesinin ikinci terimi, $t_{(1)}, t_{(2)}, ..., t_{(r_1)}$ başarısızlık zamanları gözlendiğinde geriye kalan bireylerin yaşam sürelerinin dağılımının, olasılık yoğunluk fonksiyonu ve yaşam fonksiyonu aşağıdaki gibi olan dağılıma sahip olduğunu göstermektedir.

$$f_1(t) = \frac{f(t)}{S\left(t_{(r_1)}\right)} \tag{3.7}$$

$$S_1(t) = \frac{S(t)}{S\left(t_{(r_1)}\right)} \quad , \quad t \ge t_{(r_1)} \tag{3.8}$$

$T^*_{(1)} \le T^*_{(2)} \le ... \le T^*_{(r_2)}$, $(n - r_1 - n_1)$ genişlikli rasgele örneklemden alınan gözlemlerdir.

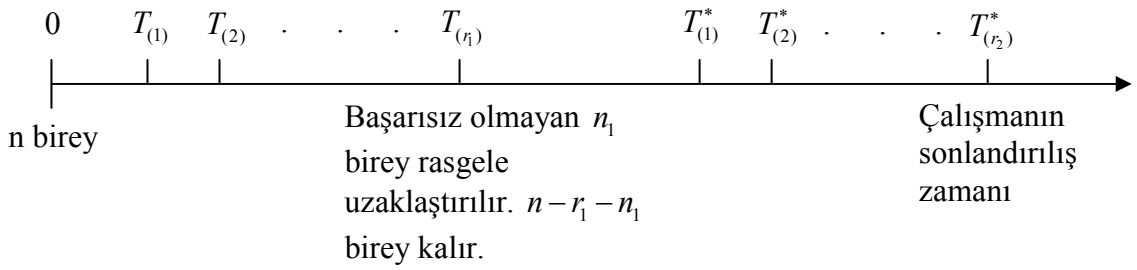
(3.5) ifadesinin ikinci terimi ise aşağıdaki gibi olacaktır:

$$\frac{(n-r_1-n_1)!}{(n-r_1-n_1-r_2)!} f_1(t^*_{(1)}) ... f_1(t^*_{(r_2)}) \left[S_1(t^*_{(r_2)})\right]^{n-r_1-n_1-r_2} \tag{3.9}$$

(3.6) ve (3.9)' daki ifadeler birleştirilerek;

$$c\, f_1(t_{(1)})...f_1(t_{(r_1)})[S_1(t^*_{(r_1)})]^{n-r_1}\, f_1(t^*_{(1)})...f_1(t^*_{(r_2)})[S_1(t^*_{(r_2)})]^{n-r_1-n_1-r_2} \qquad (3.10)$$

ifadesine ulaşılır. Burada $c = n! \cdot (n-r_1-n_1)! \,/\, [(n-r_1)!\,(n-r_1-n_1-r_2)!]$' dır (Lawless 2003).



Şekil 3.1 İlerletilmiş II. Tür Sansürleme (Lawless 2003)

3.2 Soldan Sansürleme

L_i sansürleme zamanı, T_i bireyin yaşam süresi olmak üzere; $T_i < L_i$ olduğunda bu bireyin yaşam süresinin soldan sansürlenmiş olduğu söylenir.

$$\delta_i = \begin{cases} 0 & , \quad T_i \le L_i \\ 1 & , \quad T_i > L_i \end{cases}$$

Eğer $\delta_i = 0$ $(T_i \le L_i)$ ise birey sansürlenmiş, $\delta_i = 1$ $(T_i > L_i)$ ise gözlenmiştir (Nelson 1982, Lawless 2003).

$$t_i = \max(T_i, L_i)$$

3.3 Aralık Sansürlemesi

Aralık sansürlemesi, genelleştirilmiş bir sansürleme çeşididir. Genellikle takip gerektiren olaylarda kullanılır. Çalışmaya konu olan olayın meydana gelme süresi, bir aralıkta ifade edilir. Yaşam süresi $(L_i, R_i]$ aralığında yer alır.

Eğer aralık sansürlemesi, sağdan sansürlemenin genelleştirilmiş biçimi olarak ifade ediliyorsa, sol sınır noktasının 0, sağ sınır noktasının ise L_i olarak alındığı söylenir.

Soldan sansürlemenin genelleştirilmiş biçimi olarak ifade ediliyorsa da, sol sınır noktasının L_i , sağ sınır noktasının ise (∞) olarak alındığı söylenir. Yani, aralık sansürlemesi, sağdan sansürlemenin ve soldan sansürlemenin genelleştirilmiş şeklidir (Nelson 1982).

3.4 İkili Sansürleme

Bazı çalışmalarda soldan sansürlemenin meydana geldiği durumlarda, sağdan sansürleme de aynı zamanda ortaya çıkabilir. Böyle durumlarda, yaşam sürelerinin ikili sansürlendiği ifade edilir.

Burada, L_i ele alınan olayın birey için gerçekleşmesinden önceki zaman iken, L_j ele alınan olayın birey için gerçekleşmesinden sonraki zamandır.

$$\delta = \left\{ \begin{array}{ll} 1\ , & \textit{eğer } T \textit{ başarısızlık zamanı ise} \\ 0\ , & \textit{eğer } T \textit{ sağdan sansürlenmiş ise} \\ -1\ , & \textit{eğer } T \textit{ soldan sansürlenmiş ise} \end{array} \right\}$$

Eğer, $X \le L_j$ ya da $X \ge L_i$ ise bireyin yaşam süresi kesin olarak biliniyor demektir (Nelson 1982).

4. PARAMETRİK YAŞAM MODELLERİ

Üstel, log-normal, extreme değer dağılımları gibi negatif olmayan bazı olasılık dağılımlarına uygun yaşam modelinin yorumunu yapmak mümkündür. Ancak bu bölümde sadece Weibull dağılımı ele alınmıştır.

4.1 Weibull Dağılımı

Weibull dağılımı yaşam modellerinde en yaygın kullanılan dağılımdır. Risk ve sigortacılık gibi alanlarda sıkça kullanılmaktadır. Başarısızlığın oluşmasına kadar geçen süreyi ya da başarısızlıktan sonra ikinci bir başarısızlığın oluşmasına kadar geçen süreyi modellemede kullanılan iki parametreli bir dağılımdır.

T, Weibull dağılımından alınan bir yaşam süresi rasgele değişkeni olsun. T rasgele değişkeni sürekli bir dağılıma sahiptir ve aşağıdaki olasılık yoğunluk fonksiyonuna sahiptir:

$$f(t) = (\lambda\beta)(\lambda t)^{\beta-1} e^{-(\lambda t)^{\beta}} \quad , \; t > 0, \; \beta > 0, \; \lambda > 0 \tag{4.1}$$

Burada β biçim, λ ölçek parametresidir. $\beta = 1$ olduğunda, T rasgele değişkeni üstel dağılıma sahip olur. Dolayısıyla, Weibull dağılımı üstel dağılımın bir genellemesidir denilebilir.

(2.5) ve (2.6) eşitliklerinden Hazard fonksiyonu,

$$h(t) = (\lambda\beta)(\lambda t)^{\beta-1} \tag{4.2}$$

olduğunda,

$$\begin{aligned} \int_0^t h(x)dx &= \int_0^t (\lambda\beta)(\lambda x)^{\beta-1} dx \\ &= \lambda\beta(\lambda)^{\beta-1} \frac{t^{\beta}}{\beta} \\ &= (\lambda t)^{\beta} \end{aligned} \tag{4.3}$$

olur. Dolayısıyla,

$$S(t) = e^{-\int_0^t h(x)dx}$$
$$= e^{-(\lambda t)^{\beta}} \tag{4.4}$$

olarak bulunur. Böylece (2.5) eşitliği de sağlanmış olur.

Dağılım fonksiyonu ise (2.3) eşitliğinden,

$$F(t) = 1 - e^{-(\lambda t)^{\beta}} \tag{4.5}$$

olarak yazılır (Lee 1984).

$\beta \geq 1$ ise Weibull dağılımının Hazard fonksiyonu monoton artan, $\beta < 1$ ise monoton azalandır (Barlow and Proschan 1975, Kleinbaum 1996).

Gamma fonksiyonu,

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\, dx \tag{4.6}$$

olarak tanımlandığından, yaşam süresi rasgele değişkeninin r-inci dereceden beklenen değeri (r-inci momenti),

$$E\left(T^r\right) = \int_0^{\infty} t^r f(t) dt$$
$$= \int_0^{\infty} t^{r+\beta-1} \lambda^{\beta} \beta \left(e^{-(\lambda t)^{\beta}}\right) dt \tag{4.7}$$

elde edilir. (4.6) eşitliği ve $u = (\lambda t)^{\beta}$ ifadesi (4.7) eşitliğinde yerine konulduğunda,

$$E\left(T^r\right) = \frac{1}{\lambda^r} \int_0^{\infty} u^{r/\beta} e^{-u} du$$
$$= \frac{\Gamma\left(1 + \frac{r}{\beta}\right)}{\lambda^r} \tag{4.8}$$

elde edilir.

Böylece yaşam süresi rasgele değişkeninin beklenen değeri,

$$E(T) = \frac{\Gamma\left(1+\frac{1}{\beta}\right)}{\lambda} \tag{4.9}$$

ve varyansı,

$$\begin{aligned} Var(T) &= E\left(T^2\right) - \left[E\left(T\right)\right]^2 \\ &= \frac{\Gamma\left(1+\frac{2}{\beta}\right)}{\lambda^2} - \left[\frac{\Gamma\left(1+\frac{1}{\beta}\right)}{\lambda}\right]^2 \\ &= \lambda^{-2}\left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma\left(1+\frac{1}{\beta}\right)^2\right] \end{aligned} \tag{4.10}$$

olur (Gross and Clark 1975, Lee 1984, London 1988).

Pth quantile ise;

$$\begin{aligned} F(t_p) = p \quad \Rightarrow \quad 1 - e^{-(\lambda t_p)^\beta} &= p \\ 1 - p &= e^{-(\lambda t_p)^\beta} \\ \ln(1-p) &= -(\lambda t_p)^\beta \\ \left(-\ln(1-p)\right)^{1/\beta} &= \lambda t_p \\ t_p &= \frac{\left(-\ln(1-p)\right)^{1/\beta}}{\lambda} \end{aligned} \tag{4.11}$$

dir (Lee 1984).

5. PARAMETRİK YAŞAM MODELLERİNDE TAHMİN

Bu bölümde, T yaşam süresi rasgele değişkeni iken, $S(t)$ yaşam modeli kullanılarak, Weibull dağılımının parametreleri için tahmin ediciler ve güven aralıkları bulunmuştur. Bu tahminler, tam ve sansürlü örneklem durumlarının her ikisi için de, en çok olabilirlik tahmin yöntemi ile elde edilmiştir.

Parametrelerin en çok olabilirlik tahmin edicilerini elde edebilmek için olabilirlik fonksiyonlarından yararlanılır. Böylece olasılık dağılımlarının parametrelerinin en çok olabilirlik tahmin edicileri bulunabilir. Bunun için Weibull dağılımının tek değişkenli parametre tahminleri verilecektir.

5.1 Tam ve Sansürlü Örneklem Durumlarında En Çok Olabilirlik Yöntemi

θ parametresine bağlı olarak, olasılık yoğunluk fonksiyonu,

$$f(t;\theta) = \prod_{i=1}^{n} f(t_i;\theta) \tag{5.1}$$

ile gösterilsin. $(t_1, t_2, ..., t_n)$, n birimlik bir rasgele örneklemi göstermek üzere, olabilirlik fonksiyonu,

$$L(\theta;t) = f(t;\theta) \tag{5.2}$$

ile verilir. θ parametre kümesi üzerinde maksimum yapan $\hat{\theta}(t_1, t_2, ..., t_n)$ değerine, var olması halinde θ 'nın "en çok olabilirlik tahmini" ve $\hat{\theta} = (T_1, T_2, ..., T_n)$ istatistiğine de "en çok olabilirlik tahmin edicisi" denir. Buna göre,

$$L(\theta;t) = \max \prod_{i=1}^{n} f(t_i;\theta) \tag{5.3}$$

dır. Logaritmik fonksiyonu,

$$\ln L(\theta;t) = \max \prod_{i=1}^{n} \ln f(t_i;\theta) \tag{5.4}$$

yazılabilir (Haris and Albert 1991, Miller and Miller 2001).

Eğer belirlenen bir t_i zamanında bireylerin tamamı gözlenmiş ise, yani tam veri örneklemesi söz konusu ise, olabilirlik fonksiyonu,

$$L(\theta;t) = \prod_{i=1}^{n} f(t_i;\theta) = \prod_{i=1}^{n} h(t_i)S(t_i) \tag{5.5}$$

olur. Ancak, sansürlü veri örneklemesi söz konusu olduğunda, olabilirlik fonksiyonu,

$$L(\theta;t) = \prod_{i=1}^{n} S(t_i) \tag{5.6}$$

olacaktır.

Tamamı gözlemlenmiş (sansürsüz) bireyler ile sansürlü (sağdan) gözlemlerin birlikte ifade edildiği, dağılımın olabilirlik fonksiyonu,

$$\begin{aligned} L(\theta;t) &= \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i)]^{\delta_i} [S(t_i)] \end{aligned} \tag{5.7}$$

şeklinde yazılabilir.

Burada $\delta_i = 1$ sansürsüz bireyi, $\delta_i = 0$ ise sansürlü bireyi göstermektedir. Yukarıdaki ifadenin e tabanında logaritması alınırsa, logaritmik olabilirlik fonksiyonu,

$$\begin{aligned} \ln L(\theta;t) &= \sum_{i=1}^{n} \{\delta_i \ln h(t_i) + \ln S(t_i)\} \\ &= \sum_{i=1}^{n} \{\delta_i \ln h(t_i) - \Lambda(t_i)\} \end{aligned} \tag{5.8}$$

olur. Dağılımın Hazard ve yaşam fonksiyonları bu ifadede yerine yazılarak, parametrelerin en çok olabilirlik tahmin edicileri elde edilir (Haris and Albert 1991).

5.2 Weibull Dağılımı İçin Parametre Tahmini

Weibull dağılımından alınan bir T rasgele değişkeninin olasılık yoğunluk fonksiyonu, yaşam fonksiyonu ve Hazard fonksiyonu,

$$f(t) = (\lambda\beta)(\lambda t)^{\beta-1} e^{-(\lambda t)^{\beta}} \quad , \; t > 0, \; \beta > 0, \; \lambda > 0$$

$$S(t) = \exp[-(\lambda.t)^{\beta}]$$

$$h(t) = (\lambda\beta)(\lambda t)^{\beta-1}$$

olarak 4. Bölüm'de verilmişti.

$\theta = \lambda^{-\beta}$ olarak alındığında Weibull dağılımının olasılık yoğunluk fonksiyonu,

$$f(t) = \frac{\beta}{\theta}(t)^{\beta-1}\left(e^{-\frac{1}{\theta}(t)^{\beta}}\right) \quad , t > 0 \; , \beta > 0 \; , \theta > 0 \tag{5.9}$$

yaşam fonksiyonu,

$$S(t) = \exp\left\{-\frac{1}{\theta}(t)^{\beta}\right\} \tag{5.10}$$

ve Hazard fonksiyonu,

$$h(t) = \frac{\beta}{\theta}(t)^{\beta-1} \tag{5.11}$$

olarak elde edilir.

Bu bölümde, tam gözlenmiş (sansürsüz) veri ile I. ve II. tür sağdan sansürlü veri durumlarında Weibull dağılımı için en çok olabilirlik yöntemiyle parametre tahmini yapılmıştır.

5.2.1 Tam örneklem durumunda parametre tahmini

$t_1, t_2, ..., t_n$ yaşam süreleri, λ ve β parametreli Weibull dağılımından alınan n birimlik bir örneklem olsun. (5.3) ve (5.9) eşitliklerinden olabilirlik fonksiyonu,

$$L(t_1, t_2, ..., t_n) = \theta^{-n} \beta^n \prod_{i=1}^{n} (t_i)^{\beta-1} e^{-\frac{1}{\theta}(t_i)^{\beta}} \tag{5.12}$$

olarak elde edilir.

(5.12) eşitliği, (5.4) eşitliğinde yerine yazıldığında, logaritmik olabilirlik fonksiyonu,

$$\ln L = n \ln \beta - n \ln \theta + (\beta - 1) \sum_{i=1}^{n} \ln t_i - \frac{1}{\theta} \sum_{i=1}^{n} (t_i)^{\beta} \tag{5.13}$$

olur. (5.13) eşitliğinde, θ parametresine göre türev alındığında,

$$\frac{\partial \ln L}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} (t_i)^{\beta} \tag{5.14}$$

eşitliği elde edilip, bu türev sıfıra eşitlendiğinde θ parametresinin tahmin edicisi,

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (t_i)^{\hat{\beta}} \tag{5.15}$$

olarak elde edilir.

Aynı şekilde, (5.13) eşitliğinde, β parametresine göre türev alındığında,

$$\frac{\partial \ln L}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^{n} \ln(t_i) - \frac{1}{\theta} \sum_{i=1}^{n} \left(t_i\right)^{\beta} \ln(t_i) \tag{5.16}$$

eşitliği elde edilip, bu türev sıfıra eşitlenip (5.15) eşitliği yerine yazıldığında,

$$\hat{\beta} = \left(\frac{\sum_{i=1}^{n} (t_i)^{\hat{\beta}} \ln(t_i)}{\sum_{i=1}^{n} (t_i)^{\hat{\beta}}} - \frac{1}{n} \sum_{i=1}^{n} \ln(t_i) \right)^{-1} \tag{5.17}$$

elde edilir. β parametresinin tahmin edicisi (5.17) eşitliğini sağlayan değerdir (Cohen 1965).

β ve θ parametreleri için $100(1-\alpha)\%$' lık asimptotik güven aralıkları,

$$P\left\{\hat{\beta}-Z_{\alpha/2}\,sh\left(\hat{\beta}\right)<\beta<\hat{\beta}+Z_{\alpha/2}\,sh\left(\hat{\beta}\right)\right\}=1-\alpha \tag{5.18}$$

$$P\left\{\hat{\theta}-Z_{\alpha/2}\,sh\left(\hat{\theta}\right)<\theta<\hat{\theta}+Z_{\alpha/2}\,sh\left(\hat{\theta}\right)\right\}=1-\alpha \tag{5.19}$$

olarak verilir (Lee 1984).

5.2.2 Sansürlü (tam olmayan) örneklem durumunda parametre tahmini

5.2.2.1 I. Tür sansürleme durumunda parametre tahmini

En çok olabilirlik yöntemi ile uygun bir şekilde çalışmak için gözlemlerin kümesini U ve C gibi iki alt kümeye ayıralım. U gözlenmiş bozulma zamanlarını ve C sansürlenmiş gözlemlerin indislerinin kümesini göstersin.

$$L(\delta_i,t_i)=\begin{cases}f(t_i) & ,\ \delta_i=1\\ S(t_i) & ,\ \delta_i=0\end{cases} \tag{5.20}$$

$$L(\delta_i,t_i)=f(t_i)^{\delta_i}\,S(L_i)^{1-\delta_i}$$

$$\begin{aligned}L(t_1,...,t_n)&=\prod_{i=1}^{n}L(\delta_i,t_i)\\&=\prod_{i=1}^{n}f(t_i)^{\delta_i}\,S(L_i)^{1-\delta_i}\\&=\prod_{U}f(t_i)\prod_{C}S(L_i)\end{aligned} \tag{5.21}$$

(5.9) eşitliğiyle verilen olasılık yoğunluk fonksiyonu ile (5.10) ile verilen yaşam fonksiyonu yerine konulduğunda olabilirlik fonksiyonu,

$$L(t_1,...,t_n)=\prod_{i\in U}\left(\frac{\beta}{\theta}(t_i)^{\beta-1}e^{-\frac{1}{\theta}(t_i)^{\beta}}\right)\prod_{i\in C}\left(e^{-\frac{1}{\theta}(t_i)^{\beta}}\right) \tag{5.22}$$

olarak elde edilir. (5.22) eşitliği, (5.4) eşitliğinde yerine yazıldığında, logaritmik olabilirlik fonksiyonu,

$$\ln L=\sum_{i\in U}\ln\left(\frac{\beta}{\theta}(t_i)^{\beta-1}e^{-\frac{1}{\theta}(t_i)^{\beta}}\right)+\sum_{i\in C}\ln\left(e^{-\frac{1}{\theta}(t_i)^{\beta}}\right)$$

$$\ln L=s(U)\ln\beta-s(U)\ln\theta+(\beta-1)\sum_{i\in U}\ln t_i-\frac{1}{\theta}\sum_{i=1}^{n}t_i^{\beta} \tag{5.23}$$

olur. (5.23) eşitliğinin θ'ya göre türevi alınırsa,

$$\frac{\partial \ln L}{\partial\theta}=-\frac{s(U)}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}t_i^{\beta}$$

eşitliği elde edilip, bu türev sıfıra eşitliğinde θ parametresinin tahmin edicisi,

$$\hat{\theta}=\frac{\sum_{i=1}^{n}t_i^{\hat{\beta}}}{s(U)} \tag{5.24}$$

olarak elde edilir.

Aynı şekilde, (5.23) eşitliğinde, β parametresine göre türev alındığında,

$$\frac{\partial \ln L}{\partial\beta}=\frac{1}{\beta}+\frac{\sum_{i=1}^{n}\ln t_i}{s(U)}-\frac{\sum_{i=1}^{n}t_i^{\beta}\ln t_i}{\sum_{i=1}^{n}t_i^{\beta}} \tag{5.25}$$

elde edilir. Bu türev sıfıra eşitlendiğinde,

$$\hat{\beta} = \left(\frac{\sum_{i=1}^{n} t_i^{\hat{\beta}} \ln t_i}{\sum_{i=1}^{n} t_i^{\hat{\beta}}} - \frac{\sum_{i=1}^{n} \ln t_i}{s(U)} \right)^{-1} \tag{5.26}$$

bulunur. β parametresinin tahmin edicisi (5.26) eşitliğini sağlayan değerdir (Cohen 1965, Gross and Clark 1975).

β ve θ parametreleri için $100(1-\alpha)\%$' lık asimptotik güven aralıkları,

$$P\left\{ \hat{\beta} - Z_{\alpha/2}\, sh\left(\hat{\beta}\right) < \beta < \hat{\beta} + Z_{\alpha/2}\, sh\left(\hat{\beta}\right) \right\} = 1-\alpha \tag{5.27}$$

$$P\left\{ \hat{\theta} - Z_{\alpha/2}\, sh\left(\hat{\theta}\right) < \theta < \hat{\theta} + Z_{\alpha/2}\, sh\left(\hat{\theta}\right) \right\} = 1-\alpha \tag{5.28}$$

olarak verilir (Lee 1984).

5.2.2.2 II. Tür sansürleme durumunda parametre tahmini

λ ve β parametreli Weibull dağılımından yaşam zamanları $t_{(1)}, t_{(2)}, ..., t_{(n)}$ olan n birimlik bir örneklem alınsın. Herhangi r tane bireyin yaşam sürelerinin bilindiği kabul edildiğinde yaşam zamanları,

$$t_{(1)} \le t_{(2)} \le ... \le t_{(r)} = t_{(1)}^{+} = t_{(2)}^{+} = ... = t_{(n-r)}^{+}$$

olarak sıralansın.

(5.9) eşitliğinde verilen Weibull dağılımının olasılık yoğunluk fonksiyonundan, (5.3) eşitliğindeki olabilirlik fonksiyonu,

$$L(t_{(1)}, ..., t_{(n)}) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} \left[\frac{\beta}{\theta} (t_i)^{\beta-1} \left(e^{-\frac{1}{\theta}(t_i)^{\beta}} \right) \right] \left(e^{-\frac{1}{\theta}(t_r)^{\beta}} \right)^{n-r} \tag{5.29}$$

olarak elde edilir. olarak elde edilir. (5.29) eşitliği, (5.4) eşitliğinde yerine yazıldığında, logaritmik olabilirlik fonksiyonu,

$$\ln L = \ln\left(\frac{n!}{(n-r)!}\right) + r\ln(\beta) - r\ln(\theta) + (\beta-1)\sum_{i=1}^{r}\ln(t_{(i)}) - \frac{1}{\theta}\sum_{i=1}^{r}\left(t_{(i)}\right)^{\beta}$$

$$-\frac{(n-r)}{\theta}t_{(r)}^{\beta} \qquad (5.30)$$

olarak bulunur.

(5.30) eşitliğinin θ 'ya göre türevi alındığında,

$$\frac{\partial \ln L}{\partial \theta} = -\frac{r}{\theta} + \frac{1}{\theta^2}\left[\sum_{i=1}^{n}\left(t_{(i)}\right)^{\beta} + (n-r)\left(t_{(r)}\right)^{\beta}\right] \qquad (5.31)$$

eşitliği elde edilip, bu türev sıfıra eşitliğinde θ parametresinin tahmin edicisi,

$$\hat{\theta} = \frac{1}{r}\left(\sum_{i=1}^{r}t_{(i)} + (n-r)\left(t_{(r)}\right)^{\hat{\beta}}\right) \qquad (5.32)$$

olarak bulunur.

Aynı şekilde, (5.32) eşitliğinde, β parametresine göre türev alındığında,

$$\frac{\partial \ln L}{\partial \beta} = \frac{r}{\beta} + \sum_{i=1}^{r}\ln(t_{(i)}) - \frac{1}{\theta}\left(\sum_{i=1}^{r}t_{(i)}^{\beta}\ln t_{(i)} + (n-r)t_{(r)}^{\beta}\ln\left(t_{(r)}\right)\right) \qquad (5.33)$$

olarak bulunur. Bu türev sıfıra eşitlenip, (5.32) eşitliği yerine konduğunda,

$$\hat{\beta} = \left(\frac{\sum_{i=1}^{r}t_{(i)}^{\hat{\beta}}\ln t_{(i)} + (n-r)t_{(r)}^{\hat{\beta}}\ln\left(t_{(r)}\right)}{\sum_{i=1}^{r}t_{(i)}^{\hat{\beta}} + (n-r)t_{(r)}^{\hat{\beta}}} - \frac{1}{r}\sum_{i=1}^{r}\ln t_{(i)}\right)^{-1} \qquad (5.34)$$

bulunur. β parametresinin tahmin edicisi (5.34) eşitliğini sağlayan değerdir (Cohen 1965, Gross and Clark 1975).

β ve θ parametreleri için $100(1-\alpha)\%$' lık asimptotik güven aralıkları,

$$P\left\{\hat{\beta}-Z_{\alpha/2}\,sh\left(\hat{\beta}\right)<\beta<\hat{\beta}+Z_{\alpha/2}\,sh\left(\hat{\beta}\right)\right\}=1-\alpha \tag{5.35}$$

$$P\left\{\hat{\theta}-Z_{\alpha/2}\,sh\left(\hat{\theta}\right)<\theta<\hat{\theta}+Z_{\alpha/2}\,sh\left(\hat{\theta}\right)\right\}=1-\alpha \tag{5.36}$$

olarak verilir (Lee 1984).

6. UYGULAMA

Bu çalışmada kullanılacak olan uygulama verisi Kalbfleisch and Prentice (page 223-224) kitabından alınmıştır. Veriler, akciğer kanseri hastalarına aittir.

Akciğer kanseri, dünyada en sık görülen kanser çeşidi olup, akciğer dokusunu oluşturan hücrelerin kontrol dışı çoğalmasıyla meydana gelir. Akciğer kanseri riski; ailede akciğer kanseri hastası bireylerin olması, kimyasal maddelerle iç içe çalışılması, sigara kullanılması, daha önceden akciğerle ilgili bir hastalık geçirilmesi, bazı çevresel faktörler (hava kirliliği v.b.), beslenme özellikleri gibi nedenlerden ötürü artmaktadır.

Çalışmada, akciğer kanserinin meydana geldiği dört tip hücre gözlemlenmiştir. Squamous tip, küçük hücre tipi, adeno tip ve geniş hücre tipi için yaşam zamanları incelenmiştir.

Veriler, akciğer kanseri tanısı konulan hastaların, 1000 günlük gözlem sürelerine ilişkin olup, teşhis zamanı ortalama 9 ay olan, ortalama 58 yaşındaki 137 hastadan oluşmaktadır. Hastalara ait yaşam süreleri gün olarak elde edilmiştir. 137 akciğer hastasının 128 tanesi sansürsüz (% 93,43), 9 tanesi ise sansürlüdür (% 6,57). Burada 128 hastaya ait veriler tam gözlemlenmiş olup, 9 hastaya ait veriler ise sağdan sansürlü veri olarak alınmıştır. Sağdan sansürleme çeşidinin I. tür sansürlemeye uygun olduğu görülmektedir.

Bu bölümde yapılan yaşam analizlerinde Minitab 14, parametre tahminlerinde Matlab 5.2 programlarından yararlanılmış olup, hücre tipleri dikkate alınmadan, sadece yaşam süresi ve durum değişkeni göz önüne alınarak analiz yürütülmüştür.

Uygulamada öncelikle verilerin bir istatistiksel dağılıma uygun olup olmadığı kontrol edilmiş, Weibull dağılımına uygunluğu görülmüştür. Weibull dağılımı için yaşam grafikleri çizdirilmiş ve dağılımın parametre tahminleri yapılmıştır.

137 akciğer kanseri hastalarına ait veriler EK 1'de verilmiştir. 4 hücre tipine göre akciğer kanseri hastalarının sansürlü/sansürsüz durumları ile en yüksek ve en düşük yaşam süreleri Çizelge 6.1 ile verilmiştir.

Çizelge 6.1 Akciğer kanseri verisinde kullanılan değişkenlerin kodlarına göre yaşam süreleri

Hücre Tipi	Durumu	Birey Sayısı	En Düşük Yaşam Süresi (gün)	En Yüksek Yaşam Süresi (gün)
Squamous tip (1)	Sansürsüz ($\delta_i = 1$)	31	1	999
	Sansürlü ($\delta_i = 0$)	4	25	231
Toplam		35	1	999
Küçük hücre tipi (2)	Sansürsüz ($\delta_i = 1$)	45	2	392
	Sansürlü ($\delta_i = 0$)	3	97	123
Toplam		48	2	392
Adeno tip (3)	Sansürsüz ($\delta_i = 1$)	26	3	186
	Sansürlü ($\delta_i = 0$)	1	83	83
Toplam		27	3	186
Geniş hücre tipi (4)	Sansürsüz ($\delta_i = 1$)	26	12	553
	Sansürlü ($\delta_i = 0$)	1	182	182
Toplam		27	12	553
GENEL TOPLAM	Sansürsüz ($\delta_i = 1$)	128	1	999
	Sansürlü ($\delta_i = 0$)	9	25	231

Çizelge 6.1'de uygulamaya konu olan dört tip akciğer kanseri çeşidinin, kullanılan veri setindeki dağılımı anlatılmıştır. Her hücre tipi için, sansürsüz (tam gözlenen) birey sayısı $\delta_i = 1$ ile ifade edilirken, sansürlenen (yaşamı devam eden, bilgi kaybı olan, takip edilemeyen v.s.) birey sayısı ise $\delta_i = 0$ ile ifade edilmiştir.

$$\delta_i = \begin{cases} 1\,, & \textit{eğer } i-\textit{inci bireyin yaşam süresi belli ise } (\textit{sansürsüz}) \\ 0\,, & \textit{eğer } i-\textit{inci bireyin yaşam süresi belli değil ise } (\textit{sansürlü}) \end{cases}$$

Çizelgede, sansürsüz durumdaki birey sayısı ve sansürlü durumdaki birey sayısı görülebilmektedir. Bu anlatım her tip hücre için yapıldığından, her hücre tipi kanser hastalığından toplam kaç birey gözlendiği de görülmektedir. Örneğin, squamous hücre tipi kanser hastalığı için, 35 hasta birey gözlenmiş olup, bu bireylerin 31'i sansürsüz durumdayken 4'ü sansürlüdür. Aynı şekilde, diğer üç tip hücre için de sansürsüz ve sansürlü durumlarına bakıldığında, genel toplamda, 137 hastanın 9 tanesinin sansürlü, 128 tanesinin ise tam gözlenmiş durumda olduğu görülebilir.

Ayrıca çizelgede, en düşük ve en yüksek yaşam süreleri de verilmiş olup, yine bu da her tip hücre ve durum için tekrarlanmıştır. Örneğin, küçük hücre tipi kanser hastalığının, sansürsüz durumda en düşük yaşam süresi 2 gün iken en yüksek yaşam süresi ise 392 gündür. Sansürlü durumunda ise, en düşük yaşam süresi 97 gün iken, en yüksek yaşam süresi 123 gündür. Aynı şekilde, diğer üç tip hücre için de yaşam sürelerine bakıldığında, genel toplamda, sansürsüz durumdaki 128 hastanın en düşük yaşam süresi 1 gün iken, en yüksek yaşam süresi 99 gün ve sansürlü durumdaki 9 hastanın en düşük yaşam süresi 25 gün iken, en yüksek yaşam süresi 231 gündür.

Çizelge 6.2 Durum Değişkeni Frekans Tablosu

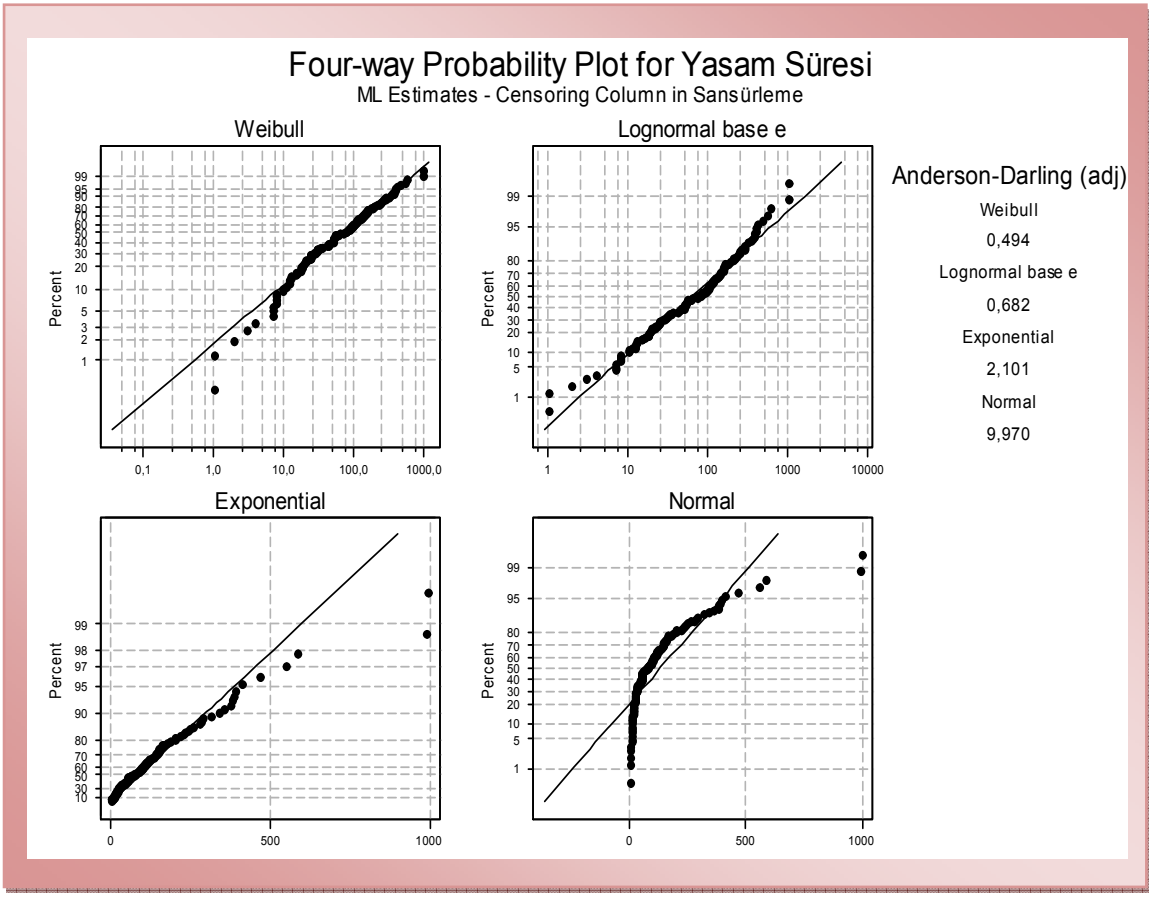
DURUM	FREKANS	YÜZDE
0	9	6.57
1	128	93.43
TOPLAM	137	100.00

Başka bir deyişle, uygulama verisinde, 137 olgudan elde edilen yaşam süresi verilerinin % 6.57'si tamamlanmamış (sansürlü), % 93.43'ü ise tamamlanmış (sansürsüz) verileri göstermektedir (Çizelge 6.2).

Bu şekilde verilen akciğer kanseri verilerinin, öncelikle Weibull dağılımına uygun olup olmadığı araştırılmıştır.

Yaşam verilerinin analizinde kullanılan önemli olasılık dağılımları ile ileri analizleri yapabilmek daha uygundur. Yaşam analizinde kullanılan en önemli dağılımlardan Weibull, log-normal, üstel ve normal dağılımlar karşılaştırılmıştır.

Karşılaştırma için, küçük ve büyük uç değerlere karşı duyarsızlığı giderilmiş, Anderson-Darling uyum iyiliği test istatistiği kullanılmıştır. Bu test, Normal, Weibull, log-normal ve fonksiyonları bilinen dağılımlara uygulanabilir. Hesaplanan olasılık fonksiyonları ile n birimden elde edilen verilerin deneysel dağılım fonksiyonları arasındaki farkların beklenen uyum içinde olup olmadıklarını test eder (Özdamar 2004).



Şekil 6.1 Dört farklı dağılım için yaşam olasılık grafikleri

Şekil 6.1'de akciğer kanseri verilerine en uygun dağılımın Weibull dağılımı olduğu görülmektedir. Ayrıca Anderson-Darling uyum iyiliği test istatistiği değeri en küçük olan dağılım veri için en uygun dağılım olduğundan, buradan da Weibull dağılımının uygunluğu tespit edilebilir.

Veri setlerinde kolay bilgi edinme yollarından birisi de belirtici istatistiklerden yararlanmaktır. Belirtici istatistikler, sayısal verileri özet olarak tanıtan, özetleyen, birimlerin yığıldıkları tipik değerleri ve bu değerler etrafında değerlerin yayılması ve dağılımları hakkında bilgi veren değerlerdir. Adından da anlaşılacağı gibi belirtici istatistikler, veri setinde yer alan değişkenleri belirten, tanıtan, açıklayan, özet olarak tanıtan değerlerdir. Belirtici istatistiklerden, n sayıda birime ait verinin genel eğilimini, yayılımını, belirli değerler etrafında toplanma özellikleri hakkında bilgi edinmek için ve topluma ait veriler için hesaplanabilecek parametrelerin tahmininde yararlanılır. Bu sebeple Çizelge 6.3'te bazı belirtici istatistikler verilmiştir (Özdamar 2004).

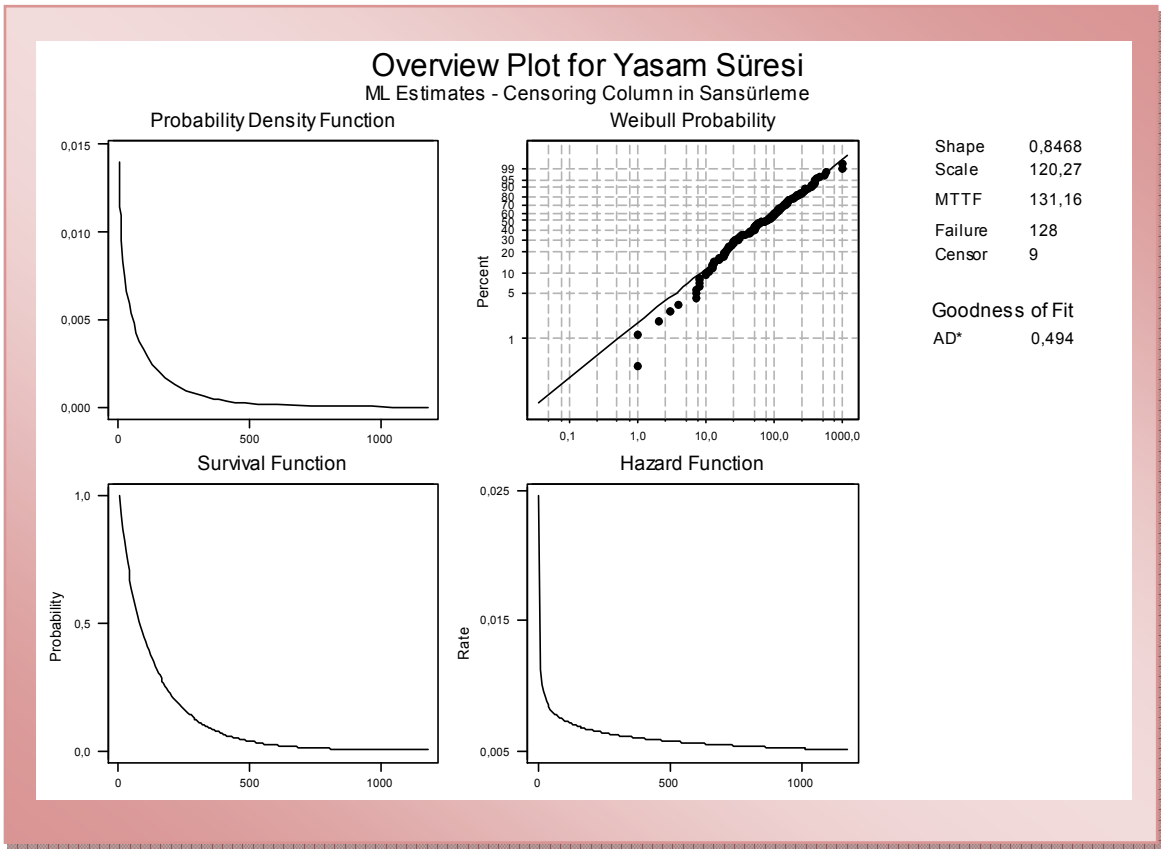
Çizelge 6.3 Weibull Dağılımının Bazı Karakteristikleri

	Tahmin	Standart Hata	95,0 % Güven Aralığı	
			Alt Sınır	Üst Sınır
Ortalama (Mean)	131,1614	13,7918	106,7336	161,1800
Standart Sapma	155,5844	20,8052	119,7129	202,2046
Ortanca (Medyan)	78,0188	9,3361	61,7079	98,6412
İlk Çeyreklik (Q1)	27,6182	4,5648	19,9760	38,1842
Üçüncü Çeyreklik (Q3)	176,8838	18,4741	144,1408	217,0647

Akciğer kanseri hastalarının yaşam sürelerine bakıldığında, ortalama yaşam süresi 131 gün olarak elde edilirken, standart sapma 155,5 gün ve ortanca yaşam süresi (medyan) 78 gün olarak elde edilmiştir.

137 akciğer kanseri hastasının ortanca yaşam süresi 78 gün ve % 95 güven aralığında bu hastaların en az yaşama süresi yaklaşık 62 gün, en çok yaşama süresi ise yaklaşık 99 gün olarak elde edilmiştir. Burada yaşam süresi incelenirken ortalama yaşam süresi yerine, ortanca yaşam süresi göz önüne alınmıştır. Çünkü yaşam analizinde, yaşam sürelerini içeren veri setlerinde ana eğilimden aşırı derecede küçük ya da aşırı derecede büyük değerler olabilir. Değişim genişliği büyük olacağından ortalamalar bu değerlerden etkilenerek gerçek eğilimi yansıtmaktan uzak kalırlar. Bu nedenle, aşırı uç değerlerden etkilenmeyen ortanca yaşam süresi, veriyi temsil eden istatistiktir. Ortanca yaşam süresini dikkate almak daha sağlıklı sonuç verir (Çizelge 6.3).

Weibull dağılımının parametreleri olan β-biçim (shape) ve λ-ölçek (scale) parametreleri akciğer kanseri hastaları için tahmin edilmiştir. Ayrıca dağılımın olasılık yoğunluk fonksiyonu, yaşam fonksiyonu ve Hazard fonksiyonuna ait grafikler de elde edilmiştir.



Şekil 6.2 Weibull dağılımının parametreleriyle ilgili grafikler

Şekil 6.2'de Weibull dağılımı için şekil ve ölçek parametreleri tahminleri bulunmuştur. Ancak, Minitab programında Weibull dağılımının ölçek parametresi $\frac{1}{\lambda}$ şeklinde ifade edildiği için, $\hat{\beta} = 0,8468$ ve $\hat{\lambda} = \frac{1}{120,27} = 0,0083$ olarak bulunmuş olur. Bu tahminler en çok olabilirlik yöntemi kullanılarak hesaplanmıştır.

Bu parametre tahminleri teorik olarak Bölüm 5'te elde edilmişti. I. tür sağdan sansürlemeye uygun akciğer kanseri verilerinin parametrelerinin tahminleri Newton-Raphson yönteminden yararlanılarak da elde edilebilir.

Öncelikle Newton-Raphson yönteminin işleyişinden kısaca bahsedilecek olunursa;

$i = 1, 2, ..., r$ için θ^0 , θ_i parametresinin gerçek değeri,

$$U(\theta_1, \theta_2, ..., \theta_r) = \frac{\partial \log L(\theta_1, ..., \theta_r)}{\partial \theta_i} \quad , \quad i = 1, 2, ..., r$$

ve

$$V(\theta_1, ..., \theta_r) = \begin{pmatrix} \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_1^2} & \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_1 \partial \theta_r} \\ \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_2 \partial \theta_r} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_r \partial \theta_1} & \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_r \partial \theta_2} & \cdots & \frac{\partial^2 \log L(\theta_1, ..., \theta_r)}{\partial \theta_r^2} \end{pmatrix}$$

olsun.

$U(\theta_1, \theta_2, ..., \theta_r) = \begin{pmatrix} U_1(\theta_1, ..., \theta_r) \\ \vdots \\ U_r(\theta_1, ..., \theta_r) \end{pmatrix}$ diyelim. $i = 1, 2, ..., r$ için, θ_i^* , θ_i ile θ_i^0 arasında olmak üzere, $U(\theta_1, ..., \theta_r)$'nin $(\theta_1^0, ..., \theta_r^0)$ etrafında Taylor serisine açılması ile,

$$U(\theta_1,...,\theta_r) = U(\theta_1^0,...,\theta_r^0) + V(\theta_1^0,...,\theta_r^0)\begin{bmatrix}(\theta_1^* - \theta_r^0)\\ \vdots \\ (\theta_1^* - \theta_r^0)\end{bmatrix} \tag{6.1}$$

yazılabilir (Gertsbakh 1989). (6.1) eşitliğinde, $\theta_i = \hat{\theta}_i \quad (i = 1,2,...,r)$ alınsın. $\hat{\theta}_i$, θ_i 'nin en çok olabilirlik tahmin edicisidir. Bu durumda, $U(\hat{\theta}_1,...,\hat{\theta}_r) = 0$ olup, (6.1) eşitliğinden,

$$\begin{bmatrix}\theta_1^*\\ \vdots \\ \theta_r^*\end{bmatrix} = \begin{bmatrix}\theta_1^0\\ \vdots \\ \theta_r^0\end{bmatrix} - V^{-1}\left(\theta_1^0,...,\theta_r^0\right) U\left(\theta_1^0,...,\theta_r^0\right) \tag{6.2}$$

bulunur. (6.2) ifadesiyle, $i = 1,2,...,r$ için $\theta_i^{(1)}$, $\hat{\theta}_i$ 'nin başlangıç değeri olmak üzere,

$$\begin{bmatrix}\theta_1^{(m+1)}\\ \vdots \\ \theta_r^{(m+1)}\end{bmatrix} = \begin{bmatrix}\theta_1^{(m)}\\ \vdots \\ \theta_r^{(m)}\end{bmatrix} - V^{-1}\left(\theta_1^{(m)},...,\theta_r^{(m)}\right) U\left(\theta_1^{(m)},...,\theta_r^{(m)}\right), \qquad m = 1,2,... \tag{6.3}$$

iterasyonları izlenir. $i = 1,2,...,r$ için $\theta_i^{(m)}$, $\theta_i^{(m+1)}$ sayısına yeterince yakınsa iterasyon durdurulur. $m \to \infty$ iken $i = 1,2,...,r$ için $\theta_i^{(m)} \to \hat{\theta}_i$ 'dir.

Bu bilgiler ışığında, I. tür sağdan sansürleme örneği için parametre tahmininde bulunulduğunda,

$$U(\beta) = \frac{1}{\beta} + \frac{\sum_{i=1}^{n} \ln t_i}{s(U)} - \frac{\sum_{i=1}^{n} t_i^{\beta} \ln t_i}{\sum_{i=1}^{n} t_i^{\beta}} \tag{6.4}$$

$$V(\beta) = -\frac{1}{\beta^2} + \frac{\sum_{i=1}^{n} t_i^{\beta} \left(\ln t_i\right)^2}{\sum_{i=1}^{n} t_i^{\beta}} + \left(\frac{\sum_{i=1}^{n} t_i^{\beta} \ln t_i}{\sum_{i=1}^{n} t_i^{\beta}}\right)^2 \tag{6.5}$$

olur. $\beta(1)$ başlangıç değeri ile

$$\beta(m+1)=\beta(m)-\frac{U\left(\beta(m)\right)}{V\left(\beta(m)\right)}$$

$$\beta(m+1)=\beta(m)-\frac{\frac{1}{\hat{\beta}}+\frac{\sum_{i=1}^{n}\ln t_i}{s(U)}-\frac{\sum_{i=1}^{n}t_i^{\hat{\beta}}\ln t_i}{\sum_{i=1}^{n}t_i^{\hat{\beta}}}}{-\frac{1}{\hat{\beta}^2}+\frac{\sum_{i=1}^{n}t_i^{\hat{\beta}}\left(\ln t_i\right)^2}{\sum_{i=1}^{n}t_i^{\hat{\beta}}}+\left(\frac{\sum_{i=1}^{n}t_i^{\hat{\beta}}\ln t_i}{\sum_{i=1}^{n}t_i^{\hat{\beta}}}\right)^2} \qquad (6,6)$$

iterasyon işlemine ulaşılır. $\beta(m+1)$ yeterince $\beta(m)$ sayısına yeterince yakınsa iterasyon durdurularak β 'nın en çok olabilirlik tahmin değeri olan $\hat{\beta}=\beta(m+1)$ olarak alınır.

(5.24) eşitliğinde, $\hat{\theta}=\frac{\sum_{i=1}^{n}t_i^{\hat{\beta}}}{s\left(U\right)}$ olarak bulunmuştu.

Ayrıca, $\theta=\lambda^{-\beta}$ alındığından, her iki tarafın e tabanında logaritması alınırsa, λ parametresi,

$$\ln\theta=\ln\lambda^{-\beta}$$

$$\ln\theta=-\beta\ln\lambda$$

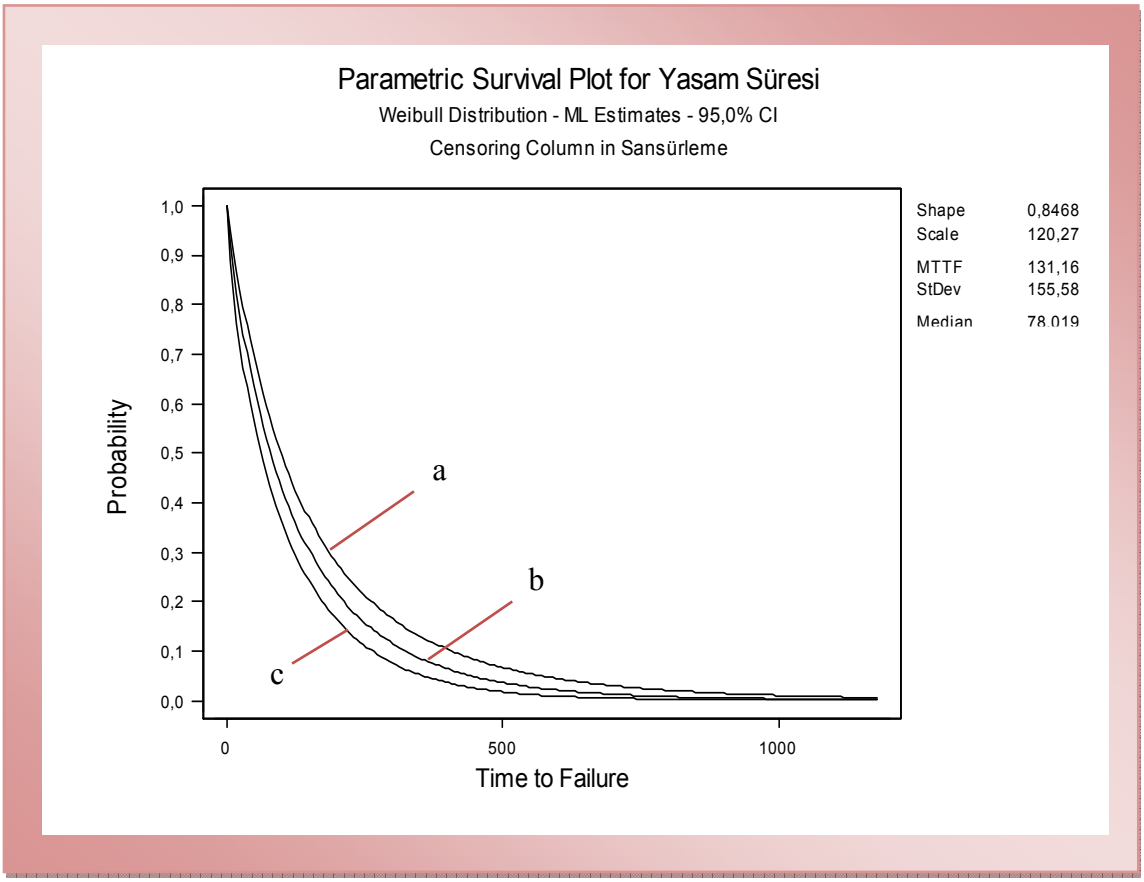
$$\ln\lambda=-\frac{\ln\theta}{\beta}$$

$$\hat{\lambda}=\mathrm{e}^{\left\{-\frac{\ln\theta}{\beta}\right\}} \qquad (6.7)$$

olarak elde edilir.

(5.24), (6.6) ve (6.7) eşitlikleri göz önünde bulundurularak yazılan Matlab bilgisayar programı kodları (EK 3) sonucunda, λ ve β parametrelerinin değeri, Newton-Rapson yöntemine göre de, $\hat{\beta} = 0,8468$ ve $\hat{\lambda} = 0,0083$ olarak bulunmuştur.

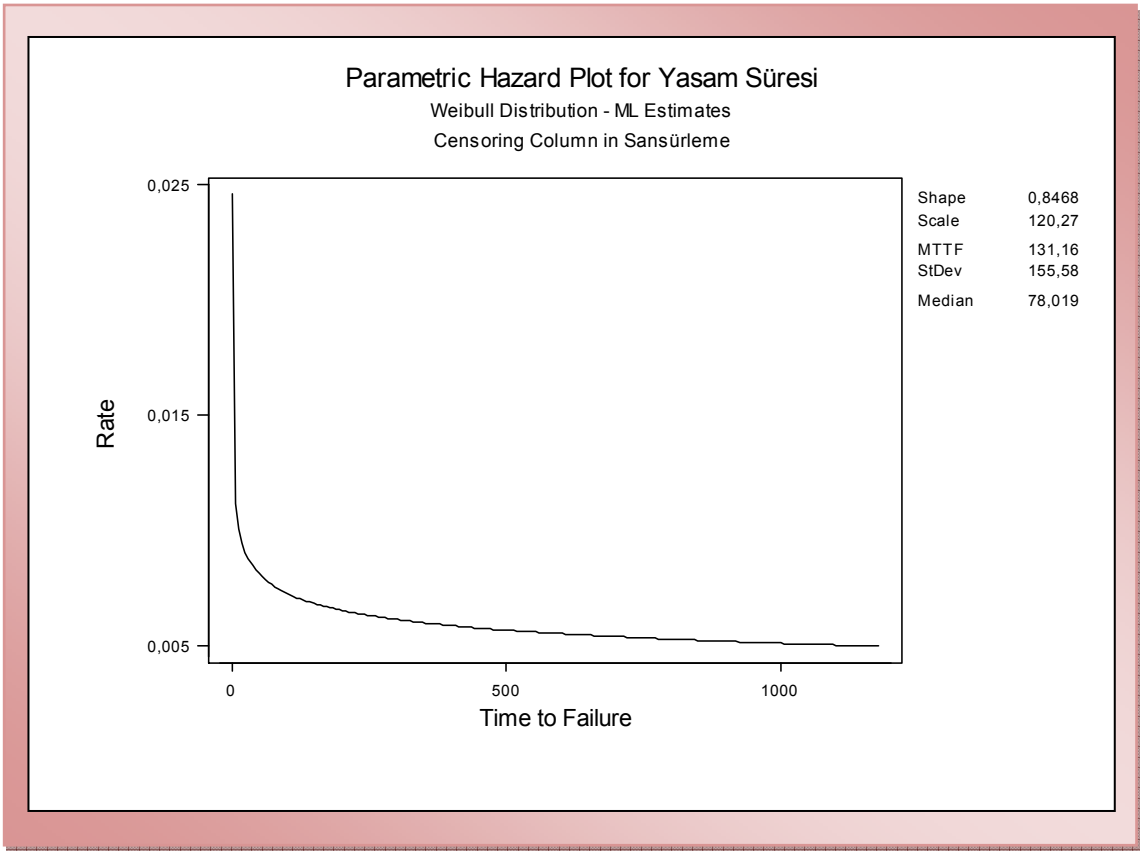
Akciğer kanseri hastalarına ait yaşam fonksiyonu Şekil 6.3 ile verilmiştir.



Şekil 6.3 Akciğer kanseri hastalarına ait yaşam fonksiyonu grafiği

Weibull dağılımı için % 95 güven aralığındaki yaşam fonksiyonu Şekil 6.3 ile gösterilmiştir. Burada, a ile ifade edilen fonksiyon yaşam fonksiyonunun alt sınırı iken, c ile ifade edilen fonksiyon ise yaşam fonksiyonunun üst sınırıdır. b ile ifade edilen fonksiyon da akciğer kanseri hastaların yaşam sürelerine ilişkin yaşam fonksiyonudur. Burada, b yaşam fonksiyonunun % 95 güven aralığı içerisinde olduğu görülmektedir.

Akciğer kanseri hastalarına ait Hazard fonksiyonu Şekil 6.4 ile verilmiştir.



Şekil 6.4 Akciğer kanseri hastalarına ait Hazard fonksiyonu grafiği

Elde edilen parametre tahminlerinden faydalanarak, Hazard ve yaşam fonksiyonu değerlerini hesaplayabiliriz. Bu sayede, her t anı için yaşam ve ani ölüm olasılıklarını elde etme şansımız olur.

$$h(t) = (\lambda\beta)(\lambda t)^{\beta-1} = (0.0083)(0.8468)(0.0083t)^{-0.1532}$$
$$= \frac{(0,00702844)}{(0.0083t)^{0.1532}}$$

$$S(t) = e^{-(\lambda t)^{\beta}} = \mathrm{e}^{-(0.0083t)^{0.8468}}$$

olarak bulunur.

$t = 50,100,250,500,750,1000$ zamanları için Hazard ve yaşam olasılıkları,

$$h(1) = 0.01464388056 \qquad S(1) = 0.9828554706$$
$$h(50) = 0.008042183408 \qquad S(50) = 0.6219737993$$
$$h(100) = 0.007231962707 \qquad S(100) = 0.4256940726$$
$$h(250) = 0.006284804441 \qquad S(250) = 0.1563814374$$
$$h(500) = 0.005651633273 \qquad S(500) = 0.03554153791$$
$$h(750) = 0.005311250591 \qquad S(750) = 0.00905799795$$
$$h(1000) = 0.005082251796 \qquad S(1000) = 0.002474506123$$

olarak elde edilir. Buradan da görüldüğü gibi, yaşam olasılığı zaman ilerledikçe azalmaktadır. Ayrıca, Şekil 6.3 ve Şekil 6.4'te de yaşam olasılığı ile ani ölüm riskinin giderek azaldığı görülmektedir. Bu azalış, Hazard fonksiyonunda daha hızlıdır.

İlk 50 günde Hazard fonksiyonu ani düşüş gösterirken, 50'inci günden itibaren kademeli olarak azalmaya devam etmektedir. Hazard fonksiyonunun ani ölüm riskini ifade etmesinden dolayı, grafiği, yaşam fonksiyonu grafiğine göre daha keskindir.

7. TARTIŞMA VE SONUÇ

Tam ve sansürlü veri örneklemesi durumunda Weibull dağılımı için parametre tahminleri yapmak, bu çalışmanın esas amacıdır. Bu nedenle önce, yaşam sürelerinin analizinde, yaşam fonksiyonu, dağılım fonksiyonu, Hazard fonksiyonu gibi bazı temel kavramlar verilmiş ve bu fonksiyonlar arasındaki ilişkiler irdelenmiştir. Daha sonra, yaşam modellerini diğer istatistik modellerinden ayıran en önemli fark olan "sansürleme" kavramından ve sansürleme çeşitleri konularına değinilmiştir.

Belli bir dağılım varsayımı gerektirmediği için pratikte daha kullanışlı olan parametrik olmayan yöntemlere göre, parametrik istatistiksel yöntemleri kullanmak daha profesyonel sonuçlar verir. Bu sebeple, yaşam analizinde kullanılan önemli dağılımlarla çalışmak daha çok tercih edilen bir yoldur. Bu çalışmada, yaşam analizinde önemli bir yer tutan ve en yaygın kullanılan Weibull dağılımı ele alınmıştır. Weibull dağılımının olasılık yoğunluk fonksiyonu, yaşam fonksiyonu, Hazard fonksiyonu verilmiş, beklenen değer ve varyansı elde edilmiştir.

Bu çalışmada, parametrik yaşam modellerinde tahmin yapmak için en çok olabilirlik yöntemi kullanılmıştır. Bu yöntem, tam örneklem ve sansürlü örneklem durumlarının her ikisi için de verilmiştir. Ayrıca en çok olabilirlik yöntemi, Weibull dağılımının parametrelerinin tahmini için de kullanılmış olup, her iki durumda da Weibull dağılımının en çok olabilirlik tahmin edicileri araştırılmıştır. Uygulamada en sık rastlanan sansürleme yöntemi olması nedeniyle, sansürlü örneklem durumunda Weibull dağılımının parametre tahminleri, sağdan sansürleme üzerine yapılmıştır.

Uygulama bölümünde, sansürlü veri içeren akciğer kanser hastalarına ilişkin veriler ele alınmıştır. Bu hastalara ait verilerin sağdan sansürlü veriler olduğu görülmüştür. Yaşam analizi yapabilmek için Minitab 14 paket programından faydalanılmıştır. Önce, veriler için en uygun dağılım araştırılmış ve bu dağılımın Weibull dağılımı olduğu görülmüştür. Bu dağılım için önemli karakteristikler elde edilmiştir. Daha sonra, Weibull dağılımı için parametre tahminleri elde edilmiş olup, aynı tahminler Newton-Raphson yöntemiyle de yapılmıştır. Ayrıca, yaşam fonksiyonu ve Hazard fonksiyonu gibi yaşam analizinde önemli yer tutan fonksiyonların grafikleri elde edilmiştir.

KAYNAKLAR

Barlow, R.E. and Proschan. 1975. Statistical Theory of Reliability and Life Testing. Holt, Rinehart and Winston, New York.

Cohen, A.C.Jr. 1965. Maximum Likelihood Estimation in the Weibull Distribution Based on Complete Censored Samples. Techometrics.

Cox, D.R. and Oakes, D. 1984. Analysis of Survival Data, Chapman and Hall, London.

Gertsbakh, I. B. 1989. Statistical Reliability Theory. Marcel Dekker, New York.

Gross, A.J. and Clark, V.A. 1975. Survival Distributions: Reliability Applications in the Biomedical Science. John Wiley, New York.

Haris, E. and Albert, A. 1991. Survivorship Analysis for Clinical Studies. Marcei Dekker, New York, USA.

Kalbfleisch, J. D. and Prentice, R. I. 2002. The Statistical Analysis of Failure Time Data.

Kleinbaum, D.G. 1996. Survival Analysis a Self Learning Text. Springer, New York.

Lawlees, J. F. 2003. Statistical Models and Methods for Lifetime Data. John Wiley, New York, USA.

Le, C.T. 1997.Applied Survival Analysis. John Wiley, New York.

Lee, E.T. 1984.Statistical Methods for Survival Data Analysis. Lifetime Learning Publications, Belmont.

London, D. 1998. Survival Models and Their Estimation. Actex Publications Winsted and Avon, Connecticut.

Miller, I. and Miller, M. 2001. John E. Freund'dan Matematiksel İstatistik. 6. Baskı (Çeviren: Ümit Şenesen). Literatür Yayıncılık, İstanbul.

Miller, R.G. 1981. Survival Analysis. John Wiley&Sons, New York.

Nelson, W. 1982. Applied Life Data Analysis, John Wiley & Sons, Inc., Canada.

Özdamar, K. 2003. SPSS ile Biyoistatistik. Kaan Kitabevi, Eskişehir.

Özdamar, K. 2004. Paket Programlar İle İstatistiksel Veri Analizi. Kaan Kitabevi. Eskişehir.

EKLER

EK 1 Akciğer Kanseri Hastalarına Ait Veriler ……... 47
EK 2 Minitab 14 Sonuçları ……. 51
EK 3 I.Tip Sansürlenmiş Veri İçin Weibull Dağılımının Parametrelerini Hesaplayan Matlab Programı …… 53

EK 1 Akciğer Kanseri Hastalarına Ait Veriler

Hasta No	Hücre Tipi (Kod)	Yaşam Süresi (gün)	Sansürleme (Kod)
1	1	72	1
2	1	411	1
3	1	228	1
4	1	126	1
5	1	118	1
6	1	10	1
7	1	82	1
8	1	110	1
9	1	314	1
10	1	100	0
11	1	42	1
12	1	8	1
13	1	144	1
14	1	25	0
15	1	11	1
16	1	999	1
17	1	12	1
18	1	87	0
19	1	231	0
20	1	242	1
21	1	991	1
22	1	111	1
23	1	1	1
24	1	587	1
25	1	389	1
26	1	33	1
27	1	25	1
28	1	357	1
29	1	467	1
30	1	201	1
31	1	1	1
32	1	30	1
33	1	44	1
34	1	283	1
35	1	15	1
36	2	30	1

Hasta No	Hücre Tipi (Kod)	Yaşam Süresi (gün)	Sansürleme (Kod)
37	2	384	1
38	2	4	1
39	2	54	1
40	2	13	1
41	2	123	0
42	2	97	0
43	2	153	1
44	2	59	1
45	2	117	1
46	2	16	1
47	2	151	1
48	2	22	1
49	2	56	1
50	2	21	1
51	2	18	1
52	2	139	1
53	2	20	1
54	2	31	1
55	2	52	1
56	2	287	1
57	2	18	1
58	2	51	1
59	2	122	1
60	2	27	1
61	2	54	1
62	2	7	1
63	2	63	1
64	2	392	1
65	2	10	1
66	2	25	1
67	2	103	0
68	2	21	1
69	2	13	1
70	2	87	1
71	2	2	1
72	2	20	1
73	2	7	1
74	2	24	1
75	2	99	1

Hasta No	Hücre Tipi (Kod)	Yaşam Süresi (gün)	Sansürleme (Kod)
76	2	8	1
77	2	99	1
78	2	61	1
79	2	25	1
80	2	95	1
81	2	80	1
82	2	51	1
83	2	29	1
84	3	8	1
85	3	92	1
86	3	35	1
87	3	117	1
88	3	132	1
89	3	12	1
90	3	162	1
91	3	3	1
92	3	95	1
93	3	24	1
94	3	18	1
95	3	83	0
96	3	31	1
97	3	51	1
98	3	90	1
99	3	52	1
100	3	73	1
101	3	8	1
102	3	36	1
103	3	48	1
104	3	7	1
105	3	140	1
106	3	186	1
107	3	84	1
108	3	19	1
109	3	45	1
110	3	80	1
111	4	177	1
112	4	162	1
113	4	216	1
114	4	553	1

Hasta No	Hücre Tipi (Kod)	Yaşam Süresi (gün)	Sansürleme (Kod)
115	4	278	1
116	4	12	1
117	4	260	1
118	4	200	1
119	4	156	1
120	4	182	0
121	4	143	1
122	4	105	1
123	4	103	1
124	4	250	1
125	4	100	1
126	4	152	1
127	4	164	1
128	4	19	1
129	4	53	1
130	4	15	1
131	4	43	1
132	4	340	1
133	4	133	1
134	4	111	1
135	4	231	1
136	4	378	1
137	4	49	1

EK 2 Minitab 14 Sonuçları

Distribution ID Plot

Variable: Yasam Süresi

Goodness of Fit

Distribution	Anderson-Darling (adj)
Weibull	0,494
Lognormal base e	0,682
Exponential	2,101
Normal	9,970

Table of MTTF

		Standard	95% Normal CI	
Distribution	Mean	Error	Lower	Upper
Weibull	131,161	13,7918	106,734	161,180
Lognormal base e	165,926	28,7663	118,125	233,070
Exponential	130,180	11,5064	109,473	154,803
Normal	130,678	14,0676	103,106	158,250

Distribution Overview Plot

Distribution: Weibull

Variable: Yasam Süresi

Anderson-Darling (adj)
0,494

Distribution Analysis: Yasam Süresi

Variable: Yasam Süresi

Censoring Information	Count
Uncensored value	128
Right censored value	9

Censoring value: Sansürleme = 0

Estimation Method: Maximum Likelihood

Distribution: Weibull

Parameter Estimates

		Standard	95,0% Normal CI	
Parameter	Estimate	Error	Lower	Upper
Shape	0,84681	0,05678	0,74252	0,96574
Scale	120,27	13,05	97,24	148,76

Goodness-of-Fit

Anderson-Darling (adjusted) = 0,4940

Characteristics of Distribution

		Standard	95,0% Normal CI	
	Estimate	Error	Lower	Upper
Mean(MTTF)	131,1614	13,7918	106,7336	161,1800
Standard Deviation	155,5844	20,8052	119,7129	202,2046
Median	78,0188	9,3361	61,7079	98,6412
First Quartile(Q1)	27,6182	4,5648	19,9760	38,1842
Third Quartile(Q3)	176,8838	18,4741	144,1408	217,0647
Interquartile Range(IQR)	149,2656	15,8470	121,2246	183,7928

EK 3 I.Tip Sansürlenmiş Veri İçin Weibull Dağılımının Parametrelerini Hesaplayan Matlab Programı

```
clear
%Birinci tip sansürlenmiş veri için Weibull dağılımının %parametrelerinin en çok
%olabilirlik tahmin edicilerinin Newton-Raphson yöntemi yardımıyla hesabı.
%c değeri sanürlenmemiş gözlemlerin sayısı ve d değeri bu gözlemlerin ortalamasıdır.
t=[72 411 228 126 118 10 82 110 314 42 8 144 11 999 12 242 991 111 1 587 389 33 25
357 467 201 1 30 44 283 15 30 384 4 54 13 153 59 117 16 151 22 56 21 18 139 20 31
52 287 18 51 122 27 54 7 63 392 10 25 21 13 87 2 20 7 24 99 8 99 61 25 95 80 51 29 8
92 35 117 132 12 162 3 95 24 18 31 51 90 52 73 8 36 48 7 140 186 84 19 45 80 177
162 216 553 278 12 260 200 156 143 105 103 250 100 152 164 19 53 15 43 340 133
111 231 378 49 100* 25* 87* 231* 123* 97* 103* 83* 182*];
c=128;
T=0;
for i=1:c
T=T+log(t(i));
end
d=T/c;
b(1)=.5;
for j=1:5000
T1=sum(t.^b(j));T2=sum((t.^b(j)).*log(t));
T3=sum((t.^b(j)).*(log(t).^2));
R(j)=(1/b(j))+d-(T2/T1);
S(j)=(-1/(b(j)^2))-(T3/T1)+((T2/T1)^2);
b(j+1)=b(j)-(R(j)/S(j));
%İterasyonların durdurulmasında yeterince yakınlığın ölçüsü .00001 olarak alındığında,
if abs(b(j+1)-b(j))<.00001
BETA=b(j+1);
break
end
end
```

```
TETA=(T1/c);
LAMDA=exp(-log(TETA)/BETA);
BETA
TETA
LAMDA
```

ÖZGEÇMİŞ

Adı Soyadı : Dilşen TAMAM

Doğum Yeri : Ankara

Doğum Tarihi : 03.06.1983

Medeni Hali : Bekar

Yabancı Dili : İngilizce

Eğitim Durumu

Lise : Yahya Kemal Beyatlı Lisesi (1997-2000)

Lisans : Ankara Üniversitesi Fen Fakültesi İstatistik Bölümü (2001-2005)

Yüksek Lisans: Ankara Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı (Eylül 2005-Kasım 2008)

Çalıştığı Kurum/ Kurumlar ve Yıl

Jandarma Genel Komutanlığı (2007-)