ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

GREENWOOD VE KAPLAN-MEIER METODU YARDIMI İLE VARYANS TAHMİNİ

Çiğdem TOPÇU

İSTATİSTİK ANABİLİM DALI

ANKARA
2007

ÖZET

Yüksek Lisans Tezi

GREENWOOD VE KAPLAN-MEIER YÖNTEMİYLE VARYANS TAHMİNİ

Çiğdem TOPÇU

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Doç. Dr. Fahrettin ARSLAN

Yaşam analizi, mühendislik, tıp ve biyoloji bilimleriyle uğraşan istatistikçiler ve araştırmacılar için önemli bir çalışma alanıdır. Yaşam analizi ile diğer istatistik analizileri arasındaki ayrım, sansürlü veri ile çalışılmasıdır. Bu çalışmanın amacı, sansürlenmiş gözlemler için elde edilen yaşam fonksiyonunu tahmin etmek için kullanılan ve parametrik olmayan bir yöntem olan Kaplan-Meier tahmin yöntemini ve bu yöntemle elde edilen tahminin varyansını hesaplamak için kullanılan Greenwood formülünü tanıtmaktır. Bu amaçla, ilk olarak yaşam analizi için temel tanım ve kavramlar üzerinde durulmuştur. Daha sonra, sansürleme kavramı ve çeşitli sansürleme türleri için olasılık fonksiyonları verilmiştir. Diğer bölümlerde, Kaplan-Meier tahminlerinin elde edilişi ve bu tahminlerin özellikleri incelenmiştir. Son olarak, Ankara Numune Eğitim ve Araştırma Hastanesi Radyasyon Onkolojisi Servisi'nden elde edilen bir veri seti kullanılarak analiz yapılmıştır.

2007, 48 sayfa

Anahtar Kelimeler: Yaşam Fonksiyonu, Sansürleme, Kaplan- Meier Tahmini, Çarpım-Limit Tahmini

ABSTRACT

Master Thesis

VARIANCE ESTIMATION WITH GREENWOOD AND KAPLAN-MEIER METHODS

Çiğdem TOPÇU

Ankara University
Graduate School of Natural and Applied Science
Department of Statistics

Advisor: Assoc. Prof. Dr. Fahrettin ARSLAN

The statistical analysis of lifetime or response time data has become a topic of considerable interest to statisticians and workers in areas such as engineering, medicine, and biological sciences. The distinction between lifetime analysis and other statistical analysis is that lifetime analysis deals with censored data. The aim of this study is, to introduce Kaplan-Meier estimate which is a nonparametric estimation method for the survivor function for the incomplete samples and the Greenwood formula used to obtain the variance of the estimate obtained from the Kaplan-Meier estimation method. For this purpose, basic definitions and concepts for the survival analysis are considered. Then, concept of censoring and the probability functions corresponding to various censoring mechanisms are given. In the other chapters, the methodology for obtaining Kaplan-Meier estimates and properties of this estimation method are examined. Finally, a data set which is obtained from Radiation Oncology Service, Ankara Numune Education and Research Hospital is analyzed.

2007, 48 pages

Key Words: Survivor Functions, Censoring, Kaplan-Meier Estimation, Product-Limit Estimatio

TEŞEKKÜR

Çalışmalarım süresince benden desteğini esirgemeyen, birikimleriyle ilerlememe yardımcı olan danışman hocam Sayın Doç. Dr. Fahrettin ARSLAN'a ve her zaman yanımda olan aileme sonsuz teşekkürlerimi sunarım.

Çiğdem TOPÇU
Ankara, Ocak 2007

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Yaşam analizi, başarısızlık olarak adlandırılan, genel olarakta ölüm, çürüme veya bozulma olarak belirlenen olayın ortaya çıkmasına kadar geçen süre olarak elde edilen verilerin analizidir. Söz konusu olay, çalışmanın türüne göre belirlenir. Bu olay canlılar için genellikle ölüm, cansız nesneler içinde bozulma olarak alınır.

Sansürlenmiş yaşam süresi verileri için, yaşam fonksiyonu tahmini, t zamanından sonra yaşayan bireylerin sayısı tam olarak bilinemediğinden, bilinen istatistiksel yöntemler ile elde edilemez. Bu durumda, bu tahmini elde etmek için bilinen yöntemler üzerinde bir takım uyarlamalar yapmak gerekir. Bu uyarlamalar sonucu elde edilen yaşam fonksiyonu tahmini literatürde Product Limit (çarpım limit, Ç-L) tahmini yada Kaplan-Meier (K-M) tahmini olarak bilinir. İlk olarak, 1958 yılında, Kaplan ve Meier, sansürlenmiş veriler için yaşam fonksiyonunun tahmni ve bu tahminin özellikleri üzerinde çalışmıştır. Bu çalışmadan sonra, Çarpım Limit (Ç-L) tahmini, Kaplan-Meier tahmini olarak anılmaya başlamıştır.

Bu çalışmada amaç; Greenwood ve K-M yöntemi yardımıyla elde edilen varyans tahminini anlatmaktır. K-M tahmini literatürde, Ç-L tahmini olarakta geçmektedir. Bu tahminin, yaşam fonksiyonunun parametrik olmayan tahminidir. Ç-L tahmini, $[0,t)$ olarak ifade edilen zaman aralığını, alt zaman aralıklarına bölerek, her $\left[u_{j-1},u_j\right)$ alt zaman aralığındaki p_j oranının tahmin edilmesine dayanmaktadır. Sözü edilen p_j oranı, u_{j-1}'de yaşadığı bilinen bir bireyin, u_j'den sonrada yaşama olasığı olarak tanımlanan koşullu olasılıktır ve $p_j = \Pr\left(T > u_j \mid T > u_{j-1}\right) = \dfrac{\Pr\left(T > u_j\right)}{\Pr\left(T > u_{j-1}\right)} = \dfrac{P_j}{P_{j-1}}$ olarak gösterilir. Burada, P_j, bireyin u_j'den sonra yaşama olasılığıdır. Sonuç olarak, kitle için yaşam fonksiyonu $P(t) = S(t) = \Pr(T > t)$ ile gösterilir ve p_j oranlarının çarpımı olarak elde edilir. Bu tahmin, K-M yada Ç-L tahmini olarak adlandırılır. Bu tahminin elde edilmesinden sonra, Greenwood formülü olarak bilinen bu tahmine ilişkin varyans elde

edilişi anlatılmıştır. Ayrıca, bu varyansın tahmini ise, delta yöntemi yarmıyla $S(t)$ fonksiyonun logaritmasının, $c>0$ gibi bir sabit etrafınfa Taylor serisine açılmasıyla elde edilmiştir.

Çalışmanın amacı doğrultusunda ilk olarak, yaşam analizine ilişkin temel tanım ve kavramların üzerinde durulmuştur. Bu kavramların, matematiksel ifadeleri ayrıntılarıyla anlatılmıştır. Yaşam analizinde önemli bir kavram olan hazard fonksiyonu kısaca açıklanmış, farklı özellik gösteren populasyonlara ait yoğunluk fonksiyonu ve ilgili hazard fonksiyonu grafikleri incelenmiştir.

Üçüncü bölümde ise, yaşam analizinin temel karakteristiği olan ve yaşam süresi verilerinin analizinde bilinenden farklı istatistiksel yöntemlerin oluşturulmasını zorunlu kılan sansürleme kavramı incelenmiştir. En çok karşılaşılan sansürleme türleri örneklerle anlatılmış, çeşitli türlerde sansürlenen veriler için olasılık yoğunluk fonksiyonunun ve olabilirlik yapılarının nasıl elde edileceği gösterilmiştir.

Dördüncü bölümde, sayma süreçlerinin yaşam analizinde nasıl bir uygulama alanına sahip olduğuna kısaca değinilmiştir. Sayma süreçleri gösterimleri kullanılarak, sansürlü veriler için olabilirlik fonksiyonun elde edilişi anlatılmış ve Kaplan-Meier tahmin edicisinin parametrik olmayan en çok olabilirlik tahmini olduğunun gösterilmesine hazırlık yapılmıştır.

Beşinci bölümde, parametrik olmayan yaşam fonksiyonu tahmininin ilk olarak tam örneklem (sansürlemenin olmadığı durum) için elde edilmiştir. Daha sonra, sağdan sansürlenmiş örneklem üzerinde, Kaplan-Meier tahmini olarak bilinen yaşam fonksiyonu tahmininin elde edilişi bir örnek ile açıklanmıştır. Bu tahminin, aynı zamanda parametrik olmayan en çok olabilirlik tahmin edicisi olduğu incelenmiştir. Kaplan-Meier tahmininin ortalama ve Greenwood formülü olarak bilinen varyansı elde edilerek özellikleri üzerinde durulmuştur. Kaplan-Meier tahminleri kullanılarak elde edilen varyansın tahmininin elde edilmesi, delta yöntemi yardımıyla da gösterilmiştir. Bu bölümde son olarak, geniş örneklem teorisi kullanılarak Kaplan-Meier tahmininin asimptotik varyansı elde edilmiştir.

Çalışmanın uygulama bölümünde, 2001-2006 yılları arasında, Ankara Numune Eğitim ve Araştırma Hastanesi Radyasyon Onkolojisi servisinde yürütülen bir çalışma için oluşturulan sağdan sansürlü bir veri seti kullanılmıştır. Sözü edilen serviste, RT (Radyoterapi) alan ileri evre meme CA tanılı 63 kadın hastanın, tanı konulmasından Eylül 2006 tarihine kadar olan yaşam süresi verileri kullanılarak, temel istatistikler, Kaplan-Meier tahminleri, bu tahminlere ilişkin standard hatalar ve güven aralıkları elde edilmiştir. Ayrıca, yaşam ve hazard fonksiyonu tahmininin grafikleri çizdirilmiş ve yorumlanmıştır.

2. TANIMLAR VE GENEL KAVRAMLAR

Genel olarak yaşam analizi; çalışmada başarısızlık kabul edilen bir olayın (failure) ortaya çıkmasına kadar geçen süre olarak elde edilen verilerin analizidir.

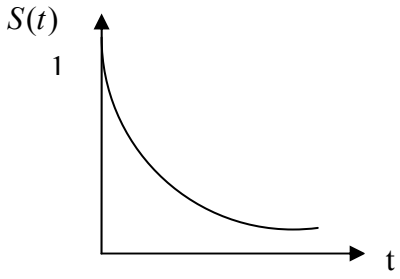
Negatif olmayan sürekli bir T rasgele değişkeni, bir kitledeki bireylerin yaşam süresini göstersin. Herhangi bir bireyin *t* zamanından (*t* zamanına kadar yaşadığı biliniyor) sonra da yaşama olasılığı;

$$S(t) = P(T > t) = \int_{t}^{\infty} f(t)dt \qquad (2.1)$$

şeklinde tanımlanır ve 'Yaşam Fonksiyonu' olarak adlandırılır.

Yaşam fonksiyonu $S(t)$ aşağıdaki özellikleri taşıyan, monoton azalan bir fonksiyondur (Şekil 2.1).

$$\lim_{t\to\infty} S(t) = 0 \quad ve \quad \lim_{t\to 0} S(t) = 1.$$

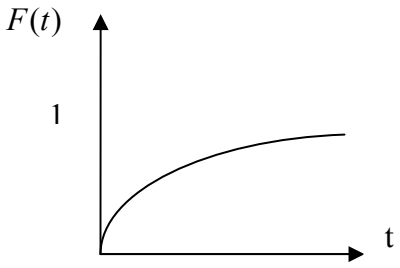


Şekil 2.1 Yaşam fonksiyonu grafiği

T rasgele değişkeninin olasılık yoğunluk fonksiyonu, $f(t)$ olmak üzere,

$$F(t) = P(T \le t) = \int_0^t f(t)dt, \ t > 0 \qquad (2.2)$$

şeklinde tanımlanan fonksiyona, 'Birikimli Başarısızlık (Failure) Fonksiyonu' denir ve bireyin, t zamanından önce ölmesi olasılığını verir (Şekil 2.2).



Şekil 2.2 Dağılım Fonksiyonu Grafiği

Yaşam fonksiyonu ile Birikimli Başarısızlık Fonksiyonu arasında;
$S(t) = 1 - F(t)$ ile ifade edilen bir ilişki vardır.

t zamanına kadar yaşadığı bilinen bir bireyin t zamanındaki ani başarısızlık (failure) ya da ölüm oranı (fonksiyonu),

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, T \ge t)}{P(T \ge t)} = \frac{f(t)}{S(t)} \qquad (2.3)$$

biçiminde tanımlanır ve $h(t)$ fonksiyonuna 'hazard fonksiyonu' denir.

Daha önce tanımları verilen $S(t)$ ve $f(t)$ fonksiyonları (2.3) ifadesi kullanılarak aşağıdaki biçimlerde ifade edilebilir.

$$h(t) = -\frac{d}{dx} \ln S(t)$$

$$\ln S(x) \Big|_0^t = -\int_0^t h(x)dx$$

$S(0)=1$ olduğundan

$$S(t)=\exp\left(-\int_0^t h(x)dx\right) \tag{2.4}$$

Bu noktada , birikimli hazard fonksiyonunun tanımını yapmak mümkün olacaktır.

$$H(t)=\int_0^t h(x)dx \tag{2.5}$$

Birikimli hazard fonksiyonunun tanımından sonra (2.4) ile gösterilen yaşam fonksiyonu

$$S(t)=\exp\left[-H(t)\right] \tag{2.6}$$

olarakta yazılabilir.

Son olarak, olasılık yoğunluk fonksiyonu;

$h(t)=f(t)S(t)$ olduğundan,

$$f(t)=h(t)\exp\left(-\int_0^t h(x)dx\right) \tag{2.7}$$

şeklinde gösterilir.

Bazen yaşam süreleri, gruplanmış olarak ya da bir sıralamanın döngü sayısı olarak elde edildiğinde, T rasgele değişkeni kesikli bir değişken olarak davranabilir. Bu durumda, $T_1,T_2,...$ rasgele değişkenlerinin $t_1,t_2,...\ \left(0<t_1<t_2<...\right)$ değerlerini aldığı düşünülürse, bu değişkenlerin olasılık fonksiyonu;

$$f(t_j)=P(T=t_j),\quad j=1,2,... \tag{2.8}$$

Yaşam fonksiyonu ise;

$$S(t) = P(T \geq t) \sum_{j:t_j \geq t} P\left(T = t_j\right) = \sum_{j:t_j \geq t} f(t_j) \tag{2.9}$$

gibi olacaktır.

Yaşam fonksiyonu, tüm $t \geq 0$ için $S\left(0\right) = 1$ ve $S\left(\infty\right) = 0$ özelliklerini taşıyan soldan sürekli artmayan bir basamak fonksiyonudur.

Kesikli zaman hazard fonksiyonu ise;

$$h(t_j) = P(T = t_j \mid T \geq t_j) \tag{2.10}$$

$$h(t_j) = \frac{f(t_j)}{S(t_j)}, \quad j = 1, 2, \ldots \tag{2.11}$$

şeklinde tanımlanır.

Ayrıca;

$$f(t_j) = S(t_j) - S(t_{j+1}) \tag{2.12}$$

olduğundan,

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)}, \quad j = 1, 2, \ldots \tag{2.13}$$

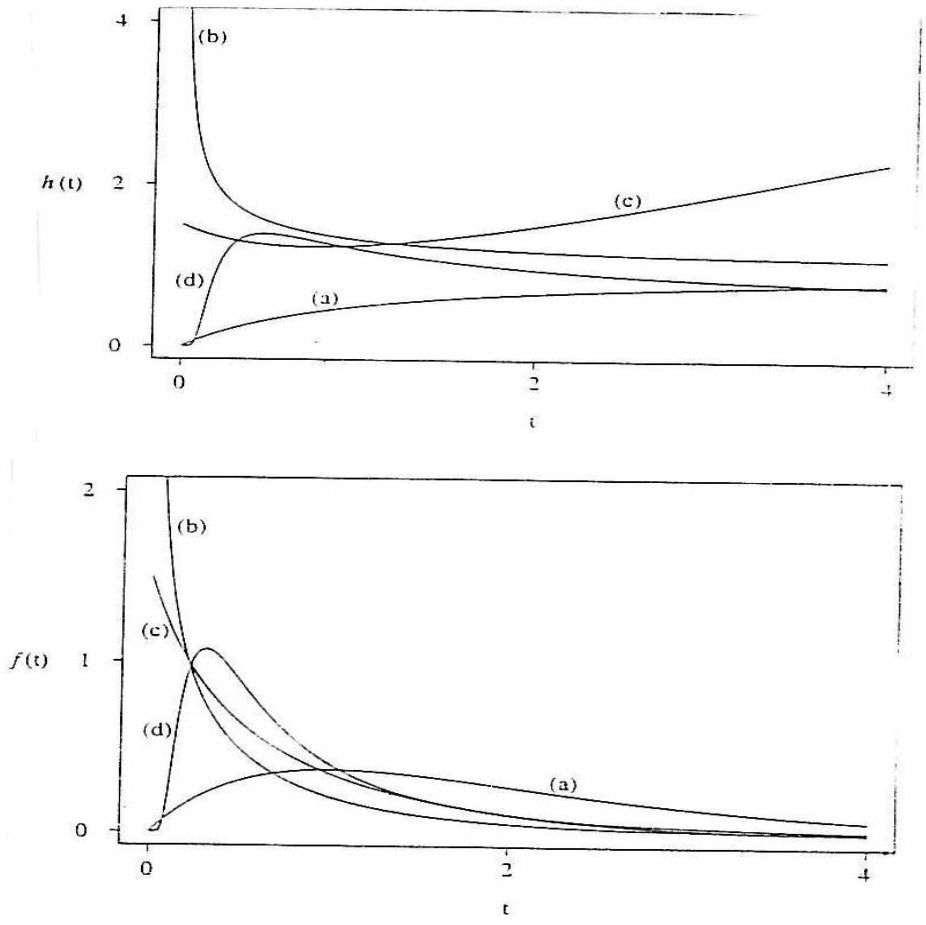
eşitliği yazılabilir ve aşağıdaki eşitlik elde edilir.

$$S(t) = \prod_{j:t_j < t} \left[1 - h(t_j)\right] \tag{2.14}$$

2.1 Yaşam Analizinde Hazard Fonksiyonu

Hazard fonksiyonu, yaşam fonksiyonunun aksine başarısızlığa (failure) odaklanır. Bu fonksiyon, bir yaşam dağılımının dikkate değer bir karakteristiğidir. Hazard fonksiyonu, bir çok uygulamanın merkezini oluşturan, zamanla değişen başarısızlık riskini belirler. Hazard fonksiyonunun grafiğinin şeklinden elde edilen ön bilgiler, veri için model seçiminde iyi bir yol göstericidir.

Şekil 2.3'deki, dört olasılık yoğunluk fonksiyonuna ve ilgili hazard fonksiyonlarına aittir.



Şekil 2.3 Hazard ve yoğunluk fonksiyonları

→ ***monoton artan hazard fonksiyonu***

Bir tür yaşlanma yada tükenmenin var olduğu populasyondaki bireylere ait yaşam sürelerinin dağılımının hazard fonksiyonu artan bir fonksiyondur.

→ ***monoton azalan hazard fonksiyonu***

Belirgin tipteki elektronik aletlerin yaşam sürelerinin dağılımı azalan hazard fonksiyonuna sahiptir.

U-şeklindeki hazard fonksiyonu

Doğumundan ölümüne kadar takip edilen bir populasyondaki bireylere ait yaşam sürelerinin dağılımı, genellikle U-şeklindeki bir hazard fonksiyonuna sahiptir. Aynı zamanda, artan hazard fonksiyonuna sahip yada zayıf bireylerden oluşan alt populasyonlardan kurtulan populasyonlardaki bireylere ait yaşam süreleri dağılımının hazard fonksiyonuda U- şeklindedir.

Ters U- şeklindeki hazard fonksiyonu

Örneğin, bir kanser tedavisi sonrası yaşam sürelerinin dağılımının sahip olduğu hazard fonksiyonu ters U- şeklindedir.

Ek olarak, sabit hazard fonksiyonu, bozulmanın yada ölümün ani şoklar yada kazalar gibi rasgele olduğu dengeli populasyonlarda görülür.

Bireyin yaşam süresi iç ve dış bir takım faktörlerden etkilenir. Bu faktörlere, açıklayıcı değişkenler (covariates) adı verilir. Açıklayıcı değişkenler, iç (internal) açıklayıcı değişkenler ve dış (external) açıklayıcı değişkenler olmak üzere ikiye ayrılır. Örneğin, kan basıncı, ölçüm alınan yaşa göre değişkenlik gösterir. Bu örnekte, yaş bir iç açıklayıcı değişkendir. Örnekten anlaşılacağı gibi, iç değişkenlerdeki değişim, bireyin kendine ait özelliklerinden kaynaklanmaktadır. Yaşam sürecinden bağımsız olarak

gelişen, hava kirliliği, ortam sıcaklığı gibi değişkenler ise dış açıklayıcı değişken olarak adlandırılır. Zamandan bağımsız olarak değişen açıklayıcı değişkenler, dış açıklayıcı değişken olarak kabul edilir.

Bireyin yaşam süresinin etkileyen faktörler zamana göre değişim gösteriyorsa bu faktörlere, 'zamana bağımlı açıklayıcı değişkenler' (time-dependent covariate) yada 'zamanla değişen açıklayıcı değişkenler' denir. Zamana bağımlı açıklayıcı değişkenlerin varlığında, modelleme için en uygun yaklaşım, önceki açıklayıcı değişkenler geçmişine (history) bağımlı hazard fonksiyonunun belirlenmesidir.

t zamanına kadar olan, açıklayıcı değişkenler geçmişi $X(\infty) = X$ özelliğine sahip, $X(t) = \{X(s),\ 0 \le s \le t\}$ olarak ifade edilirse, T rasgele değişkenine ait $X(t)$ etki değişkeni geçmişine bağımlı hazard fonksiyonu; $h(t \mid X(t))$ olarak gösterilir ve aşağıdaki şekilde tanımlanır.

$$h(t \mid X(t)) = h_0(t)\exp(\quad \boldsymbol{w}(t)) \qquad (2.1.1.)$$

Bu ifade, orantısal hazard fonksiyonunun bir uzantısıdır.

Bu ifadede, $\boldsymbol{\beta}$, $p \times 1$ boyutlu, regresyon katsayılarının vektörüdür. Aynı zamanda, $\boldsymbol{w}(t)$ ifadesi ise, $X(t)$'nin özelliklerini belirleyen bir vektördür.

3. SANSÜRLEME VE SANSÜRLEME ÇEŞİTLERİ

En genel tanımıyla sansürleme, zaman ve maliyet gibi bir takım sınırlamalar nedeniyle, örneklemdeki birimlerden elde edilen gözlemlerin analize dahil edilememesi veya elde edilemeyen bilgilerin göz ardı edilmesidir.

Sansürlemenin ortaya çıkış nedenleri şöyle sıralanabilir:

- Çalışma süresince birim takip edilemeyebilir.
- Birim, beklenmeyen bir etki sonucu çalışmadan ayrılmak zorunda kalabilir.
- Birim için gözlenmek istenen olay çalışma için planlanan zaman aralığı içinde meydana gelmeyebilir.
- Birim için gözlenmek istenen olay, çalışmanın amacı dogrultusunda beklenen nedenden farklı bir nedenle meydana gelebilir.

Güvenilirlik analizinde, sansürleme, olabileceğinden daha kısa zamanda bilgiyi elde etmek için kullanılır. Örneğin; r tane birimi çalışmaya alıp her birimin bozulmasını beklemek yerine $(n \geq r)$ olacak şekilde n tane birim çalışmaya alınır ve C gibi belirlenmiş bir zamana kadar gözlenir. (I. Tip sansürleme) C zamanından önce bozulan birimlerden gözlem alınır ve diğerleri sağdan sansürlenir. Bu biçimdeki bir düzen ile, r genişliğinde sansürlenmemiş bir örneklemden elde edilebilecek kadar bilgi elde edilebilir. Ayrıca, n-genişliği yeterince büyük seçilerek, deney daha kısa zamanda tamamlanabilir.

Yaşam analizinde, birey için gerçekleşmesi beklenen olay (failure) gözlenene kadar geçen geçen süre ilgilenilen değişkendir ve çoğu durumda yaşam süresi olarak adlandırılır. Yaşam analizinde elde edilen veriler çoğunlukla sağdan sansürlü verilerdir.

Sansürleme;

- Sağdan sansürleme (Right censoring),
- Soldan sansürleme (Left censoring),
- İkili sansürleme (Double censoring),
- Aralık sansürlemesi (Interval censoring) gibi alt gruplara ayrılır.

Sansürlenmiş veri için, olabilirlik fonksiyonunu veya isatistiksel sonuç çıkarımlarını elde etmek için sansürleme mekanizmasına ilişkin olasılık modeline ihtiyaç vardır. Daha sonra görülecektir ki, çeşitli sansürleme mekanizmalarında, yaşam süresinin dağılımına ait parametre için, olabilirlik fonksiyonu $L = \prod_{i=1}^{n} f\left(t_i\right)^{\delta_i} S\left(C_i\right)^{1-\delta_i}$ biçimde olacaktır.

Öncelikle sansürlenmiş veri için bir takım gösterimleri tanımlamak gerekir. n tane bireyin yaşam süresinin aynı dağılımlı bağımsız $T_1,...,T_n$ rasgele değişkenleri olduğu varsayılsın. Bu değişkenlerin gözlenmiş değerleri t_i ile gösterilir ve t_i yaşam süresi yada sansürleme süresi olarak adlandırılır. $\delta_i = I\left(T_i = t_i\right)$ olarak tanımlanan durum değişkeni, t_i'nin gözlenmiş bir yaşam süresi yada sansürleme zamanı olduğunu gösterir. Eğer, $\delta_i = 1$ ise, t_i gözlenmiş bir yaşam süresi $\delta_i = 0$ ise, t_i sansürleme zamanıdır. Bu durumda, gözlenmiş veri, $\left(t_i, \delta_i\right)$ $i = 1,...,n$ ikililerinden oluşacaktır.

Yaşam analizinde, önemli olan ve ayrı bir dikkat gerektiren diğer bir varsayım ise, bireyin yaşam süresinin, sansürleme zamanından bağımsız olmasıdır.

3.1 Sağdan Sansürleme

Yaşam analizinde, başarısızlık (failure) olarak adlandırılan ve genel olarak; ölüm, bozulma, çürüme, v.b., olarak kabul edilen olay, çalışma için belirlenen bir bitiş

zamanına kadar gerçekleşmezse, bireyin yaşam süresinin uzunluğu çalışmanın bitiş zamanının "sağ" tarafına geçer. Böyle bir durumda, bu birey için gözlem alınamayacak ve yaşam süresi kesin olarak bilinmeyecektir. Bu nedenle, bireyin yaşam süresi sansürlenecektir. Bu tip sansürlemeye "sağdan sansürleme" denir.

Sağdan sansürleme, kendi içinde;

- I. tür sansürleme (Type I censoring)
- II. tür sansürleme (Type II censoring)
- İlerleyen II. tür sansürleme (Progressive type II censoring)
- Rasgele sansürleme (Random censoring) gibi alt gruplara ayrılır.

Bu sansürleme tipinde, başarısızlık gözlenilmemiştir. Başarısızlık zamanının sansürleme zamanından daha büyük olduğu bilinir.

3.1.1 I. tür sansürleme

I. tür sansürleme mekanizmasında, her bireyin sabit bir sansürleme zamanının olduğu düşünülür. ($C_i > 0$)

Bireyler sürece herhangi bir zamanda dahil olurlar vedaha önceden belirlenmiş olan çalışmanın bitiş zamanına kadar gözlenirler. Burada C_i sansürleme zamanı, çalışmanın başlama zamanı ile bitiş zamanı arasında bir zamandır. Bu sansürleme türünde, C_i sansürleme zamanının sabit bir sayı olduğu açıktır. T_i ise, bireyin çalışma süresince gözlenebildiği süredir ve rasgele değişkendir.

Çalışmanın sonunda elde edilen örneklem n tane (t_i, δ_i) çiftinden oluşacaktır.

Burada;

$\left.\begin{array}{l}\delta_i = 0, \quad T_i > C_i \rightarrow i.birey\ sansürlenmiştir. \\ \delta_i = 1, \quad T_i \leq C_i \rightarrow i.birey\ ölmüştür.\end{array}\right\}$ şeklinde tanımlanmıştır.

I. Tür sansürleme için genel gösterim;

$t_i = \min\left(T_i, C_i\right)$ ve $\delta_i = I(T_i \leq C_i)$ olmak üzere, $\left(t_i, \delta_i\right)$ için birleşik olasılık yoğunluk fonksiyonu,

$$P(t_i, \delta_i) = f(t_i)^{\delta_i} S(C_i)^{1-\delta_i} \qquad (3.1.1.1)$$

Şeklindedir.

Çünkü;

$t_i = \min(T_i, C_i) \quad \delta_i = I(T_i \leq C_i)$

$\delta_i = 0 \rightarrow T_i > C_i \quad t_i = C_i \quad olur.$ (*Sansürlenmiş gözlem*)

$\delta_i = 1 \rightarrow T_i \leq C_i \quad t_i = T_i \quad olur.$ (*Sansürlenmemiş gözlem*)

(t_i, δ_i) rasgele örneklemin oyf:

$P(t_i, \delta_i = 0) = P(t_i, T_i > C_i) = P(T_i = C_i, T_i > C_i) = P(T_i \geq C_i) = S(t_i+) = S\left(C_i\right)$

$P(t_i, \delta_i = 1) = P(t_i, T_i \leq C_i) = P(t_i = T_i, T_i \leq C_i) = P(0 < T_i < t_i) = f(t_i)$

Burada, $S(t_i+) = P(T > t_i) = S\left(C_i\right)$ 'dir. Eğer, $S(t)$ t_i zamanında sürekli ise $S(t_i+) = S(t_i)$ olacaktır.

$\left(t_i, \delta_i\right)$ için birleşik olasılık fonksiyonu,

$P(t_i, \delta_i) = f(t_i)^{\delta_i} S(C_i)^{1-\delta_i}$ olarak bulunur.

olabilirlik fonksiyonu ise,

$$L = \prod_{i=1}^{n} f\left(t_i\right)^{\delta_i} S\left(C_i\right)^{1-\delta_i} \qquad (3.1.1.2)$$

olarak elde edilir.

3.1.2 II. tür sansürleme

Bu sansürleme türünde, n tane birey aynı anda gözlenmeye başlanır ve ilk r tane başarısızlık gözlendiği anda çalışma kesilir. (Çalışmanın başında, r sayısı belirlenen bir sabittir.) Burada, çalışmanın toplam süresi r. başarısızlık zamanına $\left(t_{(r)}\right)$ eşittir ve çalışmanın başında bilinmemektedir.

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

$t_{(1)} \le t_{(2)} \le ... \le t_{(r)}$ 'nin bileşik olasılık yoğunluk fonksiyonu;

$$\frac{n!}{(n-r)!}\left\{\prod_{i=1}^{r} f(t_{(i)})\right\} S(t_{(r)})^{n-r} \qquad (3.1.2.1)$$

olarak elde edilir.

Bu ifadenin sıralı istatistiklerin genel biçimi olduğu açıktır.

3.1.3 İlerleyen II. tür sansürleme

II. Tür sansürlemenin genelleştirilmiş halidir ve mekanizması şu şekildedir:
Çalışmada yer alan n tane bireyden başarısız ilk r_1 birey gözlenir, daha sonra kalan $n-r_1$ bireyden n_1 tanesi çalışmadan alınır. Geriye, $n - r_1 - n_1$ birey çalışmada kalır. Daha

sonra, başarısız r_2 birey gözlenir ve geriye $n - r_1 - n_1 - r_2$ birey kalır. Kalan bireylerden n_2 tanesi çalışmadan alınarak kalan bireyler ile çalışma devam ettirilir. Bu mekanizma bu şekilde devam eder.

İlk gözlenen r_1 başarısızlık zamanı; $T_{(1)} \le T_{(2)} \le ... \le T_{(r_1)}$ ve daha sonraki r_2 başarısızlık zamanı; $T^*_{(1)} \le T^*_{(2)} \le ... \le T^*_{(r_2)}$ olmak üzere sıralı istatistiklerin gösterimi ile ifade edilir ve verinin dağılımı (3.1.3.1) ile gösterilen şekilde olur.

$$g_1(t_{(1)}, t_{(2)}, ...t_{(r_1)}) g_2(t^*_{(1)}, ..., t^*_{(r_2)} \mid t_{(1)}, ..., t_{(r_1)}) \qquad (3.1.3.1)$$

Burada, g_1 fonksiyonu $T_{(1)} < ... < T_{(r_1)}$ sıralı istatistiklerinin dağılımıdır.

$$g_1(t_{(1)}, t_{(2)}, ...t_{(r_1)}) = \frac{n!}{(n-r_1)!} f(t_{(1)}) ... f(t_{r_1}) \left[S(t_{r_1}) \right]^{n-r_1} \qquad (3.1.3.2)$$

g_2 fonksiyonu ise, aşağıdaki oyf ve yaşam fonksiyonuna sahip rasgele değişkenlerin ($T^*_{(1)} < ... < T^*_{(r_2)}$) oluşturduğu soldan budanmış dağılımdır.

$$f_1(t) = \frac{f(t)}{S(t_{(r_1)})} \qquad S_1(t) = \frac{S(t)}{S(t_{(r_1)})}$$

$T^*_{(1)}, ..., T^*_{(r_2)}$, soldan budanmış dağılımdan alınan $n - n_1 - r_1$ genişlikli rasgele örneklemdeki en küçük gözlemlerdir.

Bu durumda, $g_2(t^*_{(1)}, ..., t^*_{(r_2)} \mid t_{(1)}, ..., t_{(r_1)})$ aşağıdaki şekilde olacaktır.

$$\frac{(n-r_1-n_1)!}{(n-r_1-n_1-r_2)!}\frac{f(t^*_{(1)})}{S(t_{(r_1)})}\cdots\frac{f(t^*_{(r_2)})}{S(t_{(r_1)})}\frac{\left[S(t_{r_2})\right]^{n-r_1-n_1-r_2}}{\left[S(t_{r_1})\right]^{n-r_1-n_1-r_2}} \qquad (3.1.3.3)$$

(3.1.3.2) ve (3.1.3.3) numaralı eşitlikleri birleştirerek (3.1.3.1) 'de gösterilen olabilirlik fonksiyonu, (3.1.3.5) ile gösterilen haliyle elde edilir. (Lawlees 2003)

$$\frac{n!}{(n-r_1)!}f(t_{(1)})\ldots f(t_{(r_1)})\left[S(t_{(r_1)})\right]^{n-r_1}\frac{(n-r_1-n_1)!}{(n-r_1-n_1-r_2)!}f(t^*_{(1)})\ldots f(t^*_{(r_2)})\frac{\left[S(t_{r_2})\right]^{n-r_1-n_1-r_2}}{\left[S(t_{r_1})\right]^{n-r_1-n_1-r_2}} \qquad (3.1.3.4.)$$

$$\frac{n!}{(n-r_1)!}\frac{(n-r_1-n_1)!}{(n-r_1-n_1-r_2)!}f(t_{(1)})\ldots f(t_{(r_1)})\left[S(t_{(r_1)})\right]^{n_1}f(t^*_{(1)})\ldots f(t^*_{(r_2)})\left[S(t_{(r_2)})\right]^{n-r_1-n_1-r_2} \qquad (3.1.3.5)$$

3.1.4 Rasgele sansürleme

Burada, her bireyin yaşam süresi T ve sansürleme zamanı C bağımsız rasgele değişkenlerdir.

$S(t) \to T\,rd.'nin\ yaşam\ fonksiyonu$
$f(t_i) \to T\,rd.'nin\ olasılık\ yoğunluk\ fonksiyonu$
$G(t) \to C\,rd.'nin\ yaşam\ fonksiyonu$
$g(t_i) \to C\,rd.'nin\,olasılık\ yoğunluk\ fonksiyonu$

ve $t_i = \min\left(T_i, C_i\right)$ ve $\delta_i = I(T_i \le C_i)$ olmak üzere,

$\left(t_i, \delta_i\right)$ için birleşik olasılık yoğunluk fonksiyonu,

$$P(t_i,\delta_i) = \left[f(t_i)G(t_i)\right]^{\delta_i}\left[g(t_i)S(t)\right]^{1-\delta_i} = \left[f(t_i)^{\delta_i}S(t)^{1-\delta_i}\right]\left[g(t_i)^{1-\delta_i}G(t_i)^{\delta_i}\right] \qquad (3.1.4.1)$$

olarak elde edilir.

Rasgele sansürlemede, her birey rasgele olarak sansürlenir. Bazı çalışmalarda, bireyler ilgilenilen olay dışında çalışmadan ayrılmalarına neden olacak bir takım olaylarla (competing event) karşılaşabilirler. Böyle durumlarda, ilgilenilen olay gözlenmeyecektir. Bu durum rasgele sansürleme olarak adlandırılır. Bu olaylara örnek olarak, kazara ölümler, göçler, ilgilenilen olay dışındaki ölümler verilebilir

$$t_i = \min(T_i, C_i) \quad \delta_i = I(T_i \le C_i)$$

$$\begin{aligned} P(t_i, \delta_i = 0) &= P(t_i, T_i > C_i) = P(t_i = C_i, T_i > C_i) \\ &= P(C_i = t, T_i > C_i) = P(C_i = t)P(T_i > C_i) \\ &= g(t)S(t) \end{aligned}$$

$$\begin{aligned} P(t_i, \delta_i = 1) &= P(t_i, T_i \le C_i) = P(t_i = t, C_i > T_i) \\ &= f(t)G(t) \end{aligned}$$

$P(t_i, \delta_i) = \left[f(t)G(t)\right]^{\delta_i}\left[g(t)S(t)\right]^{1-\delta_i} = \left[f(t_i)^{\delta_i} S(t)^{1-\delta_i}\right]\left[g(t_i)^{1-\delta_i} G(t_i)^{\delta_i}\right]$ olarak elde edilir.

$G(t)$ ve $g(t)$ fonksiyonları $f(t)$ fonksiyonunda ki parametreleri içermediği için bu iki fonksiyon çıkarılabilir ve olabilirlik fonksiyonu (3.1.1.2) 'deki biçime dönüşür.

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(C_i)^{1-\delta_i}$$

3.2 Soldan Sansürleme

Çalışmadaki bir bireyin yaşam süresi $T < C_l$ ise bu bireyin yaşam süresinin soldan sansürlenmiş olduğu düşünülür. Yani, söz konusu olay (event) C_l zamanında gözlenen bu birey için C_l 'den daha önce gerçekleşmiş demektir (Klein and Moeschberger 1997).

$$\varepsilon_i = I(T_i > C_l)$$
$$t_i = \max\left(T_i, C_l\right)$$
$$\left\{\begin{array}{l}\varepsilon = 0 \; ise \; T_i \le C_l \; ve \; T \; sansürlenmiştir. \\ \varepsilon = 1 \; ise \; T_i > C_l \; ve \; T \; gözlenmiştir. \end{array}\right\}$$

Soldan sansürlemede, gözlenmek istenen olay bazı bireyler için, verilen zamana gelmeden önce gerçekleşmiştir. Örneğin, dişi tavşanların cinsel olgunluk çağı verileri ele alındığında kayıt tutmaya hayatın 21. gününde başlanılır. Ancak, hayvanların bir kısmı 21. günden önce cinsel olgunluğa ulaşmıştır. Böyle bir hayvan için elde edilen veri soldan sansürlüdür.

3.3 İkili Sansürleme

Bazı uygulamalarda, yaşam süresi T_i iki olay arasında geçen zaman olabilir. Örneğin, bir kimsenin hastalı yapan virüs tarafından enfekte edilmesinden hastalığın tanısının konulmasına kadar geçen süre ilgilenilen değişken olabilir. Burada hastanın virüs tarafından enfekte edilmesi ilk olay (initial event), tanının konulması ikinci olay (second event) olarak düşünüldüğünde, aşağıdaki tanımlar yapılabilir.

L_i^*: Soldan sansürleme zamanı

R_i^*: Sağdan sansürleme zamanı

U_i^*: ilk olayın gerçekleşme zamanı. $U_i^* \in \left(L_i^*, R_i^*\right]$

y_i: Gözlenen sansürleme yada başarısızlık zamanı

T_i için başarısızlık yada sansürleme zamanı, $t_i = y_i - u_i^*$ olacaktır. Bu değişken için, $y_i - R_i^* \le t_i < y_i - L_i^*$ olduğu bilinmektedir. Bu yapı ikili sansürleme olarak adlandırılmaktadır.

$g_i(u)$ U_i^* değişkeninin olasılık yoğunluk fonksiyonu olarak düşünüldüğünde ve T_i değişkeni U_i^* değişkeninden bağımsız ise olabilirlik yapısı aşağıdaki şekilde kurulabilir.

$$P\left(y_i, \delta_i \mid U_i^* \in \left(L_i^*, R_i^*\right]\right) = \int_{L_i^*}^{R_i^*} g_i(u) f_i(y_i - u)^{\delta_i} S_i(y_i - u)^{1-\delta_i} du \qquad (3.3.1)$$

Bu yapıyı elde etmekteki zorluk $g_i(u)$ fonksiyonunu belirlemenin gerekliliği olacaktır. İkili sansürleme, sağdan ve soldan sansürlemenin birleştirilmiş şeklidir. (Lawlees 2003)

Bu sansürleme türü için şöyle bir örnek verilebilir. Bir çocuğun okumayı öğrenme süresi (T), ilkokul birinci sınıfa başlama ve bitiş süresince elde edilir. Çocuk okumayı ilkokul birinci sınıfa başlamadan öğrenebileceği gibi, birinci sınıfın bitimininde de hala öğrenememiş olabilir.

Böyle bir durumda, okumayı öğrenme süresi ilk durumda soldan sansürlüdür, ikinci durumda ise sağdan sansürlüdür.

3.4 Aralık Sansürlemesi

Genelleştirilmiş bir sansürleme çeşidi olan aralık sansürlemesi, yaşam süresinin bir aralık içinde olduğu durumlarda meydana gelir. Bu sansürleme tipinde beklenilen olay iki zaman noktası arasında gerçekleşmiştir ve gerçekleşme anı hakkında bilgi yoktur.

Örneğin, bir klinik denemesindeki bir hastanın periyodik bir takip süresi vardır ve hasta için söz konusu olayın meydana gelme süresi $(\mathrm{L_i},\mathrm{R_i}]$ aralığında olacağı bilinir.

$L_i < t < R_i$

n tane bağımsız bireyin oluşturduğu örneklemden elde edilen olabilirlik fonksiyonu (observed likelihood) aşağıdaki şekilde olacaktır.

$$L = \prod_{i=1}^{n} \left[F_i(R_i) - F_i(L_i) \right] \tag{3.4.1}$$

Burada, $F_i(t)$ fonksiyonu T_i değişkeninin dağılımıdır ve $F_i(0) = 0$ olduğu kabul edilir. Aralık sansürlemesi, sağdan sansürlemenin ve soldan sansürlemenin genelleştirilmiş halidir. Çünkü, eğer sol sınır noktası 0 ve sağ sınır noktası C_l alınırsa soldan sansürlemeye, sol sınır noktası C_r ve sağ sınır noktası sonsuz (∞) alınırsa sağdan sansürlemeye ulaşılır (Lawlees 2003).

4. SAYMA SÜREÇLERİ VE YAŞAM MODELLERİ

Sansürlenmiş ve budanmış veriler için, çıkarsama (inference) yöntemlerinin geliştirilmesinde alternatif bir yaklaşım, sayma sürecine ilişkin yöntemlerin kullanılmasıdır. Bu yaklaşım, ilk olarak 1975 yılında Aalen tarafından, stokastik integrasyon elemanlarının, sayma süreci ve sürekli zaman martingale teorisinin bir metodoloji olarak birleştirilmesi sonucunda geliştirilmiştir (Klein and Moeschberger 1997).

Bir sayma süreci, $[0,t)$ zaman aralığında meydana gelen belli tipteki olayları sayan, sağdan sürekli bir stokastik süreçtir. Yaşam analizinde, bu süreç her birey için verilen zaman aralığında meydana gelen ölümleri (failure) sayar. (0 yada1 değerini alır.) T_i, $i.$bireyin yaşam süresi olmak üzere, $N_i(t) = I(T_i \le t)$ ile gösterilir (Klein and Moeschberger 1997).

Sağdan sansürlenmiş bir veri seti tanımlanırken, gözlenmiş yaşam süreleri için sayma süreci gösterimi aşağıdaki şekilde olacaktır.

$N_i(t) = I(T_i \le t, \delta \quad)$

$dN_i(t) = N_i(t+dt)^{-} - N_i(t)^{-}$ olarak ifade edilir.

n tane bireyin $t = 0$ anından, her bir birey için başarısızlık durumu meydana gelene kadar yada bireyler sansürlenene kadar takip edildiği düşünülsün. Yaşam ve sansürleme sürelerinin kesikli olduğu varsayımı altında, bu değişkenlerin her bir birey için aldığı değerler $t = 0,1,...$ olsun. Açıklayıcı değişkenlerin, zamanla değişmediği (sabit) düşünülerek $h_i(t)$ ve $S_i(t)$ sırasıyla $i.$ birey için, gözlenen açıklayıcı değişkenlerin varlığı altında, hazard ve yaşam fonksiyonudur.
$dC_i(t) = Y_i(t)\, I(i.\ birey\ t\ zamanında\ sansürlenmiştir.)$ tanımı yapıldığında; herhangi bir birey için, $\{dN_i(t),\ dC_i(t),\ t \ge 0\}$ değerlerinden sadece bir tanesinin sıfırdan farklı

olacağı açıktır.

$d\boldsymbol{N}(t)=\{dN_1(t),...,dN_n(t)\}$, $d\boldsymbol{C}(t)=\{dC_1(t),...,dC_n(t)\}$ vektörlerinden oluşan

$$\mathcal{H}(\mathrm{t})=\{(d\boldsymbol{N}(s),d\boldsymbol{C}(s)),\quad s=0,1,...,t-1\} \tag{4.1}$$

(4.1) ifadesi, başarısızlık ve sansürleme süreçlerinin t zamanındaki geçmişi (history) olarak adlandırılır ve $(t-1)$ zamanına kadar meydana gelen tüm başarısızlık ve sansürlemeye ilişkin bilgileri içerir. (4.1) ifadesindeki tüm olasılıklar açıklayıcı değişkenlere koşulludur fakat karmaşayı önlemek adına bu durum gösterimde belirtilmemiştir (Lawlees 2003).

Açıklayıcı değişkenler göz ardı edildiğinde, gözlenen veriler aşağıdaki şekilde ifade edilir.

$$Data=(d\boldsymbol{N}(t),d\boldsymbol{C}(t);\quad t=0,1,2,...)$$

Yaşam analizinde kabul edilen, yaşam süresinin potansiyel başarısızlık zamanından bağımsız olması varsayımı altında, $Pr(Data)$ aşağıdaki biçimde ayrıştırılabilir.

$$Pr(Data)=\prod_{t=0}^{\infty}Pr(d\boldsymbol{N}(t)\,|\,\mathcal{H}(t))\,Pr(d\boldsymbol{C}(t)\,|\,d\boldsymbol{N}(t),\mathcal{H}(t)) \tag{4.2}$$

Bu tanımlarda sonra; T rasgele değişkeninin kesikli olduğu varsayımını göz önüne alalım, daha önceki tanımlardan, $dN_i(t)$ 0 ve 1 değerlerini alan rasgele değişkendir ve $dN_i(t)\,|\,\mathcal{H}(t)\sim Bern(h_i(t))$ yazılabilir (Fleming and Harrington 1991).

Bu durumda, $d\boldsymbol{N}(t)\,|\,\mathcal{H}(\mathrm{t})$ fonksiyonu aşağıdaki şekilde elde edilebilir ve bu fonksiyonun, Bernoulli dağılımına sahip bağımsız rasegele değişkenlerin birleşik olasılık fonksiyonu olduğu açıktır.

$$Pr\left(d\boldsymbol{N}\left(t\right)|\mathcal{H}\left(t\right)\right)=\prod_{i=1}^{n}h_i\left(t\right)^{dN_i(t)}\left[1-h_i\left(t\right)\right]^{Y_i(t)\left(1-dN_i(t)\right)} \tag{4.3}$$

Eğer, sansürleme süreci bağımsız ise, N_i sayma süreci için yoğunluk (intensity) modelinin aşağıdaki gibi olacağı daha önceden verilmişti.

$$Pr\left(dN_i\left(t\right)=1|H\left(t\right)\right)=Y_i\left(t\right)h_i\left(t\right),\quad i=1,2,...,n \tag{4.4}$$

(3.4) koşulu, t zamanındaki sansürlemenin ve başarısızlığın koşullu bağımsızlığını göstermektedir. Bir başka deyişle, t zamanındaki sansürlemenin t zamanındaki yada sonrasındaki başarısızlıkla bir ilgisi yoktur (Lawlees 2003). Bu koşulu sağlayan mekanizmalar sıklıkla bağımsız sansürleme mekanizmaları olarak adlandırılırlar.

Eğer; $Pr\left(d\boldsymbol{C}\left(t\right)|d\boldsymbol{N}\left(t\right),\mathcal{H}\left(t\right)\right)$ terimleri $h_i\left(t\right)$ 'yi belirleyen bir parametre içermiyorsa bu terimler bilgi verici olamayacaktır ve olabilirlik fonksiyonundan çıkarılabilir. Bu terimleri fonksiyondan çıkartarak ve (4.3) ifadesini (4.2) ifadesinde yerleştirerek aşağıdaki olabilirlik fonksiyonu elde edilir (Lawlees 2003).

$$L=\prod_{i=1}^{n}\prod_{t=0}^{\infty}h_i\left(t\right)^{dN_i(t)}\left[1-h_i\left(t\right)\right]^{Y_i(t)\left(1-dN_i(t)\right)} \tag{4.5}$$

Çalışmadaki her birey, t zamanında ya sansürlenmiş olarak yada ölmüş olarak gözlenir. t zamanında ölmüş olarak gözlendiğinde, $dN_i\left(t=1\right)$ ve $Y_i\left(s\right)=I\left(s\le t\right)$; sansürlenmiş olarak gözlendiğinde ise, $dN_i\left(t\right)=0$ ve $Y_i\left(s\right)=I\left(s\le t\right)$ olacaktır.

$S_i\left(t\right)=\prod_{s=0}^{t-1}\left(1-h_i\left(s\right)\right)$ ve $f_i\left(t\right)=h_i\left(t\right)S_i\left(t\right)$ olduğundan;

$\left(t_i,\delta_i\right)$ gösterimi kullanılarak, yukarıda elde edilen (4.5) olabilirlik fonksiyonu

$$L = \prod_{i=1}^{n} f_i\left(t_i\right)^{\delta_i} S_i\left(C_i\right)^{1-\delta_i} \tag{4.6}$$

ifadesine dönüşür.

Sürekli ve karşık (mix) dağılımların olduğu durumlarda, olabilirlik fonksiyonunu elde etmek için, $\left[t, t+dt\right)$ zaman aralığı olabildiği kadar küçük aralıklara (partition) bölünür. (4.3) ve (4.4) ifadelerinde $h_i\left(t\right)$ yerine $dH_i\left(t\right)$ yazılarak, kesikli durumda elde edilen sonuçların özü değişmeden aşağıdaki sonuçlar elde edilir (Lawlees 2003).

$$\begin{aligned} L &= \prod_{i=1}^{n} \prod_{(0,\infty)} dH_i\left(t\right)^{dN_i(t)} \left[1 - dH_i\left(t\right)\right]^{Y_i(t)\left(1-dN_i(t)\right)} \\ &= \prod_{i=1}^{n} f_i\left(t_i\right)^{\delta_i} S_i\left(t_i\right)^{1-\delta_i} \end{aligned} \tag{4.7}$$

5. SAĞDAN SANSÜRLENMİŞ VERİLER İÇİN PARAMETRİK OLMAYAN YAŞAM FONKSİYONU TAHMİNİ

Grafikler ve tanımlayıcı istatistikler verileri tanımlamak ve analiz etmek için kullanılan önemli araçlardır. Sıklık tabloları, histogramlar, ampirik dağılım fonksiyonları verileri tanımlamak için kullanılan en bilinen araçlardan bazılarıdır. (Lawlees 2003)

Sansürlemenin varlığı, yaşam süresi verilerini tanımlamak ve analiz etmek için yukarıda sözü edilen araçlar üzerinde bazı uyarlamalar yapmayı zorunlu kılar.

Yaşam verilerinin analizinde önemli olan, yaşam fonksiyonunun bu fonksiyona ait varyansın ve ilgili hazard fonksiyonunun tahmin edilmesidir.

Çalışmanın bundan sonraki bölümünde, ilk olarak sansürlemenin olmadığı durumda daha sonra da sansürlemenin varlığında yaşam fonksiyonu tahmini yapılmıştır.

5.1 Kaplan-Meier Tahmini (Çarpım-Limit Tahmini)

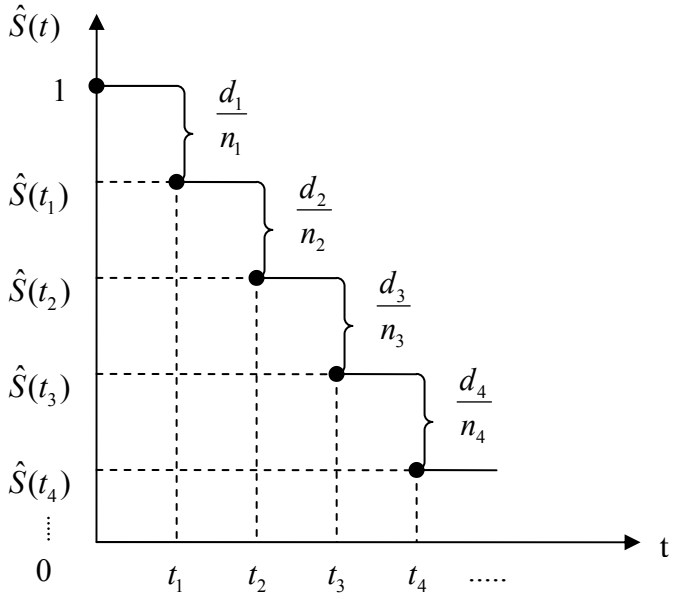
$t_1, t_2, ..., t_n$ şeklinde gösterilen bir rasgele örneklemi anlatmak için en uygun yol, ampirik yaşam fonksiyonunun yada ampirik dağılım fonksiyonunun grafiğini çizmektir. Bu yöntem, dağılımın parametrik olmayan tahminini gerektirmektedir (Lawlees 2003).

n birimlik bir rasgele örneklemde sansürlenmiş verilerin olmadığı durumda ampirik yaşam fonksiyonunun tahmini;

$$\hat{S}(t) = \frac{t \text{ zamanında yaşayan birey sayısı}}{n} = \frac{N_t}{n} \quad , \ t > 0 \qquad (5.1.1)$$

olarak tanımlanır (Lawlees 2003, London 1988).

Ampirik yaşam fonksiyonunun grafiği azalan bir basamak fonksiyonudur .
Genel olarak, yaşam süresi *t*'ye eşit *d* tane yaşam süresi var ise, ampirik yaşam fonksiyonu her t_i zamanında d_i / n_i kadar azalır (Şekil 5.1), (Lawlees 2003).



Şekil 5.1 Ampirik yaşam fonksiyonu grafiği

Ampirik Yaşam Fonksiyonunun tahmininde,

N_t: t zamanında yaşayan birey sayısını göstermektedir.

N_t parametreleri n ve $S(t)$ olan Binom dağılımına sahiptir.

$\frac{N_t}{n}$ ifadesinin burada dağılımın parametresi olan $S(t)$'nin tahmini olduğu açıktır. (Binomial proportion)

Yukarıdaki, açıklamalardan hareketle, aşağıdaki ifadeler yazılabilir.

$$E[\hat{S}(t)] = S(t) \tag{5.1.2}$$

$$V\left[\hat{S}(t)\right] = \frac{S(t).(1-S(t))}{n} \tag{5.1.3}$$

Yaşam fonksiyonun varyansının tahmini de (5.1.4) ile gösterildiği gibi olacaktır (Chap 1997).

$$\hat{Var}\left[\hat{S}(t)\right] = \frac{\hat{S}(t).(1-\hat{S}(t))}{n} \qquad (5.1.4)$$

Sansürlenmiş gözlemlerin olduğu örneklemlerde ise, yaşam süresi t'den büyük gözlemlerin sayısı kesin olarak bilinemeyecektir. Bu durum, sansürlemenin olmadığı durumda elde edilen yaşam fonksiyonunun tahmininin sansürlemenin olduğu duruma uyarlanmasını gerektirir. Bu uyarlama sonucu elde edilen yaşam fonksiyonunun tahmini "Çarpım Limit (Ç-L) tahmini" ya da "Kaplan-Meier (K-M) tahmini" olarak bilinir. (Lawlees 2003)

Yaşam fonksiyonu için, Ç-L tahmini aşağıdaki belirtilen aşamalara dayanır. (Kaplan and Meier 1958)

a) Zaman ölçeği $[0,t)$ uygun olarak seçilen alt aralıklara bölünür. $[u_0 = 0, u_1)$, $[u_1, u_2)$, ..., $[u_{k-1}, u_k = t)$ gibi.

b) Her bir, (u_{j-1}, u_j) aralığı için $p_j = \frac{P_j}{P_{j-1}}$ oranı tahmin edilir. Bu oran, u_{j-1}'den sonra yaşadığı bilinen bir bireyin, u_j'den sonrada yaşama olasılığı olarak tanımlanan koşullu olasılıktır.

$p_j = \Pr\left(T > u_j \mid T > u_{j-1}\right) = \frac{\Pr\left(T > u_j\right)}{\Pr\left(T > u_{j-1}\right)} = \frac{P_j}{P_{j-1}}$ olarak ifade edilir.

c) Eğer, t bir bölme noktası ise; kitle için $\mathrm{P}(t) = S(t) = \Pr(T > t)$ oranı t'den önceki tüm aralıklar için tahmin edilen p_j'lerin çarpımı olarak tahmin edilir.

Başlagıçta, hiç bir zaman aralığının aynı zamanda sansürleme ve ölümü durumunu aynı anda içermediği kabul edilmektedir.

$$P = \Pr\left(T > u_j\right) = S\left(u_j\right), \qquad j = 1,....,k+1$$

$j=0$ için , $S(u_0)=S(0)=P_0=1$ olur.

$p_j=\dfrac{P_j}{P_{j-1}}$ olduğundan;

$j=1$ için $\quad p_1=\dfrac{P_1}{P_0}=P_1\rightarrow P_1=p_1$

$j=2$ için $\quad p_2=\dfrac{P_2}{P_{1\,1}}\rightarrow P_2=p_2p_1$

$j=3$ için $p_3=\dfrac{P_3}{P_2}\rightarrow P_3=p_3p_2p_1$

.

.

.

$P_j=\prod_{j=1}^{k}p_j$ olarak bulunur.

$n_j:\left[u_{j-1},\,u_j\right)$ zaman aralığında risk altında olan birey sayısı

$d_j:\left[u_{j-1},u_j\right)$ zaman aralığında gözlenen ölüm sayısı

olarak tanımlanırsa;

$$\hat{P}_j=\frac{n_j-d_j}{n_j}=\frac{n_j^{'}}{n_j} \tag{5.1.5}$$

olarak tahmin edilir. Bu tahmin K-M tahmini olarak bilinir.

Bu durumda, Ç-L (K-M) tahmini;

$$\hat{S}(t)=\hat{P}(t)=\prod_{j=1}^{k}\frac{n_j^{'}}{n_j}\quad,\quad u_k=t\,,\;\; n_j^{'}=n_j-d_j \tag{5.1.6}$$

olacaktır (Kaplan and Meier 1958).

Bununla birlikte, herhangi bir alt zaman aralığı yalnızca sansürleme zamanlarını içeriyorsa, $\hat{p}_j=1$ olacaktır.

Eğer, en büyük gözlem zamanı t^*, bir sansürleme zamanı ise, (5.1.6) formülü $t > t^*$ için kullanılamayacaktır. Bu durmda, $\hat{P}(t)$ tahmini 0 ile $\hat{P}(t^*)$ arasında tanımlı olacaktır ve $\hat{P}(t^*)$ ise, net bir biçimde tanımlı olmayacaktır (Kaplan and Meier 1958).

(5.1.5) ifadesinde, $n_1 = n$ ve $n_j = n_{j-1} - \delta_{j-1}$, $j = 1,2,...k$ yazılırsa, sansürlemenin olmadığı durumda elde edilen, ampirik yaşam fonksiyonu tahminine (5.1.1) indirgenir.

İlgilenilen zaman uzunluğu, gözlenen yaşam sürelerine göre bölünebilir. $t'_{(1)} \le t'_{(2)} \le ... \le t'_{(N)}$ sıralanmış yaşam süreleri olsun. Bu durumda, $\hat{S}(t)$ aşağıdaki şekilde tahmin edilir.

$$\hat{S}(t) = \hat{P}(t) = \prod_r \frac{N-r}{N-r+1} \qquad (5.1.7)$$

$t'_r \le t$ için, r' pozitif bir tamsayıdır ve t'_r, r. ölümün gözlendiği (sansürlemenin değil) zamandır.

Örnek:

Gözlem zamanları: 0.8 3.1 5.4 9.2 ay
Sansürleme zamanları: 1.0 2.7 7.0 12.1 ay

$\lambda_j = n'_j - n_{j+1}$ olarak tanımlanırsa,

u_j	n_j	n'_j	λ_j	$\hat{P}(u_j)$
0.8	8	7	2	7/8
3.1	5	4	0	7/10
5.4	4	3	1	21/40
9.2	2	1	0	21/80
(12.1)	1	1	1	21/80

$$\hat{P}(t>0.8)=\frac{8-1}{8}=\frac{7}{8}$$

$$\hat{P}(t>3.1)=\frac{7}{8}\left(\frac{4}{5}\right)=\frac{7}{10}$$

$$\hat{P}(t>5.4)=\frac{7}{10}\left(\frac{3}{4}\right)=\frac{21}{40}$$

$$\hat{P}(t>9.2)=\frac{21}{40}\left(\frac{1}{2}\right)=\frac{21}{80}$$

$$\hat{P}(t>12.1)=\frac{21}{80}$$

olarak hesaplanır.

Son satırda, gösterilen (12.1) değeri için bir sansürleme zamanı olduğu için, $\hat{P}(t)$ tahminin tanımsız olduğu zamandır.

5.2 En Çok Olabilirlik Tahmin Edicisi Olarak Çarpım Limit Tahmini

Ç-L tahmini, $S(t)$ yaşam fonksiyonunun parametrik olmayan en çok olabilirlik tahmini olarak türetilebilir (Lawlees 2003).

$T_1,...,T_n$ bağımsız yaşam süresi değişkenleri $S(t)$ yaşam fonksiyonu ve $h(t)$ hazard fonksiyonu ile kesikli bir dağılıma sahip olsun, genelleştirmeden vazgeçmeyerek, $t=0,12,...$ olarak alınsın. $h(t)$ hazard fonksiyonu, dağılımın parametresi olduğu düşünülsün.

Sansürleme için kabul edilen varsayımlar altında ve $h_i(t)=h(t)$ alınarak olabilirlik fonksiyonu aşağıdaki gibi olacaktır (Lawlees 2003).

$$L=\prod_{i=1}^{n}\prod_{t=0}^{\infty}h(t)^{dN_i(t)}\left[1-h(t)\right]^{Y_i(t)(1-dN_i(t))} \qquad (5.2.1)$$

t_i, i. birey için yaşam süresini yada sansürleme zamanını göstermek üzere, $\delta_i = I\left(t_i \ bir \ yaşam \ süresidir\right)$, $Y_i(t)=I(t_i \geq t)$ ve $dN_i(t)=I(t_i=t,\delta_i=1)$ tanımları yapıldıktan sonra (5.2.1) ifadesi (5.2.2) ifadesine dönüşecektir.

$$L=\prod_{t=0}^{\infty}h(t)^{d_t}\left[1-h(t)\right]^{n_t-d_t} \qquad (5.2.2)$$

Bu ifade de,

$d_t=\sum_{i=1}^{n}dN_i(t)$ ve $n_t=\sum_{i=1}^{n}Y_i(t)$ şeklindedir.

d_t, t zamanındaki gözlenmiş yaşam süresi saysısı (ölen birey sayısı) ve n_t yine t zamanında risk altında bulunan birey sayısıdır.

$\boldsymbol{h}=\left(h(0),h(1),...\right)$ vektörü , yaşam süresi dağılımında parametre olarak göz önüne alındığında, (4.6) olabilirlik fonksiyonu $L(\boldsymbol{h})$ olacaktır ve bu fonksiyonun maksimum yapan değerin,

$\hat{h}(t)=\dfrac{d_t}{n_t} \quad (t=0,1,...\tau) \qquad \tau=\max\left(t:n_t>0\right)$

olacağı kolaylıkla bulunur.

$S(t)$'nin en çok olabilirlik tahmini, $t=0,1,...\tau$ için, aşağıdaki şekilde olacaktır.

$$\begin{aligned}\hat{S}(t)&=\prod_{s=0}^{t-1}\left[1-\hat{h}(s)\right]\\&=\prod_{s=0}^{t-1}\left(1-\frac{d_s}{n_s}\right)\end{aligned} \qquad (5.2.3)$$

5.3 Kaplan Meier Tahmini İçin Ortalama ve Varyans Tahmini

$n_j^{'} = n_j - d_j$ olarak tanımlanmıştı.

i). $n_j^{'} \sim Binom(n_j, p_j)$ ve

ii). Her $\{n_j^{'}\}$ için, $\hat{p}_i$ tahminleri bağımsız kabul edilirse;

$$\hat{p}_j = \frac{n_j^{'}}{n_j} \Rightarrow E\left[\hat{p}_j \mid \{n_j^{'}\}\right] = p_j \text{ olur.}$$

$$E\left[\hat{S}(t) \mid \{n_j^{'}\}\right] = E\left[\hat{p}_1 ... \hat{p}_k \mid \{n_j^{'}\}\right] = \prod_{j=1}^{k} E\left[\hat{p}_j \mid \{n_j^{'}\}\right] = p_1 ... p_k = S(t)$$

olduğundan, kabul edilen varsayımlar altında, $\hat{S}(t)$, $S(t)$ için yansız bir tahmin edicidir. Aynı zamanda bu tahmin, ihmal edilebilir bir yanlılığa sahiptir.

$$Var\left[\hat{S}(t) \mid \{n_j^{'}\}\right] = E\left[\hat{S}^2(t) \mid \{n_j^{'}\}\right] - \left(E\left[\hat{S}(t) \mid \{n_j^{'}\}\right]\right)^2$$

$$= E\left[\hat{p}_1^2 ... \hat{p}_k^2 \mid \{n_j^{'}\}\right] - \left(E\left[\hat{p}_1 ... \hat{p}_k \mid \{n_j^{'}\}\right]\right)^2$$

$$= \prod_{j=1}^{k} E\left(\hat{p}_j^2 \mid \{n_j^{'}\}\right) - \left(\prod_{j=1}^{k} E\left(\hat{p}_j \mid \{n_j^{'}\}\right)\right)^2$$

$Var\left(\hat{p}_j\right) = \frac{p_j q_j}{n_j}$ ve $E\left(\hat{p}_j\right) = p_j$ olduğu açıktır. (London 1988)

$Var(X) = E\left(X^2\right) - \left[E(X)\right]^2$ olduğundan,

$$\text{Var}\left[\hat{S}(t) \mid \{n_j^{'}\}\right] = \prod_{j=1}^{k}\left(\frac{p_j q_j}{n_j} + p_j^2\right) - \left(\prod_{j=1}^{k} p_j\right)^2$$

$$= \prod_{j=1}^{k} p_j^2 \left(1 + \frac{q_j}{p_j n_j} \right) - \left(\prod_{i=1}^{k} p_j \right)^2$$

$$= \prod_{j=1}^{k} p_j^2 \prod_{j=1}^{k} \left(1 + \frac{q_j}{p_j n_j} \right) - \left(\prod_{i=1}^{k} p_j \right)^2$$

$$= \left(\prod_{j=1}^{k} p_j \right)^2 \prod_{j=1}^{k} \left(1 + \frac{q_j}{p_j n_j} \right) - \left(\prod_{j=1}^{k} p_j \right)^2$$

$$Var\left[\hat{S}(t) \mid \{n_j'\} \right] = \left(\prod_{j=1}^{k} p_j \right)^2 \left[\prod_{j=1}^{k} \left(\frac{q_j}{p_j q_j} + 1 \right) - 1 \right]$$

$$= \left[S(t) \right]^2 A$$

$$A = \left(1 + \frac{q_1}{p_1 n_1} \right) \left(1 + \frac{q_2}{p_2 n_2} \right) \ldots \left(1 + \frac{q_k}{p_k n_k} \right)$$

$$A = \left(1 + \frac{q_1}{p_1 n_1} \right) + \left(1 + \frac{q_2}{p_2 n_2} \right) + \ldots + \left(1 + \frac{q_k}{p_k n_k} \right) + \left(\textit{yüksek dereceli terimler} \right)$$

Yüksek dereceli terimler ihmal edildiğinde,

$A = \left(1 + \frac{q_1}{p_1 n_1} \right) + \left(1 + \frac{q_2}{p_2 n_2} \right) + \ldots$ bulunur ve,

$$\mathrm{Var}\left[\hat{S}(t) \right] \approx \left[S(t) \right]^2 \sum_{j=1}^{k} \frac{q_j}{p_j n_j} \qquad (5.3.1)$$

olarak hesaplanan varyans Greenwood formülü olarak bilinir (London 1998).
Bu varyansın tahmini;

$\hat{q}_j = \frac{d_j}{n_j}$ ve $\hat{p}_j = \left(1 - \frac{d_j}{n_j} \right)$ tahminleri formülde yerine yazıldığında ve, $S(t)$ için K-M tahmini yazıldığında;

$$\mathrm{Var}\left[\hat{S}(t) \right] \approx \left[S(t) \right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j \left(n_j - d_j \right)} \qquad (5.3.2)$$

olarak elde edilir.

Sansürlenmiş gözlemlerin olmadığı durumlarda bu varyans tahmini,

$$\widehat{\mathrm{Var}}\left[\hat{S}(t)\right] = \frac{\hat{S}(t).(1-\hat{S}(t))}{n} \qquad (5.3.3)$$

ifadesine dönüşür.

Greenwood formülü, delta yöntemi yardımıyla da elde edilebilir. Bu yöntem, bir parametre vektörünün en çok olabilirlik tahmin edicisinin sürekli bir fonksiyonunun taylor açılımının bulunmasına dayanır. Şöyle ki;

$f(X)$, X rasgele değişkeninin bir fonksiyonu ve $E(X) \to c$ olamak üzere,

$$f(X) \approx f(c) + f'(c)(X-c)$$

olduğu varsayılır. Buradan,

$$E\left[f(X)\right] \approx f(c) + f'(c)\left(E(X)-c\right)$$

$$\mathrm{Var}\left[f(X)\right] \approx f'(c)^2 \,\mathrm{Var}(X)$$

Bu yöntemi kullanarak;

$X = \hat{S}(t)$ ve $f(X) = \log \hat{S}(t)$ alınırsa;

$$f(X) = \log \hat{S}(t) = \sum \log\left[1-\hat{h}_j(t)\right] = \sum \log\left[1-\frac{d_j}{n_j}\right] \text{ yazılabilir.}$$

$$\widehat{\mathrm{Var}}\left[\sum \log\left(1-\hat{h}_j(t)\right)\right] = \frac{1}{\hat{S}(t)^2}\widehat{\mathrm{Var}}\left[\hat{S}(t)\right]$$

$$\widehat{\mathrm{Var}}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum \widehat{\mathrm{Var}} \log\left[1-\hat{h}_j(t)\right]$$

Yine delta yöntemi yaklaşımını kullanarak,

$Y = 1 - h_i(t) \quad f(Y) = \log\left[1 - \hat{h}_j(t)\right]$ alınırsa ve

$d_j \sim Binom(n_j, p_j)$ olduğu kabul edildiğinde $\hat{p}_j = \frac{d_j}{n_j}$ olduğu bulunur.

Buradan,

$$\text{Var}\left[\log\left(1 - \hat{h}_j(t)\right)\right] \approx \frac{1}{\left(1 - \hat{h}_j(t)\right)^2} \text{Var}\left[1 - \hat{h}_j(t)\right]$$

$$\approx \frac{1}{\left(1 - \frac{d_j}{n_j}\right)^2} \frac{n_j p_j (1 - p_j)}{n_j^2}$$

$$\approx \frac{n_j p_j (1 - p_j)}{\left(n_j - d_j\right)^2}$$

olarak bulunur.

Bulunan ifadeler yerine yazıldığında,

$$\text{Var}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum \frac{n_j p_j (1 - p_j)}{\left(n_j - d_j\right)^2}$$

$$\text{Var}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum \frac{d_j}{n_j \left(n_j - d_j\right)} \tag{5.3.4}$$

olarak elde edilir.

Bazı τ'lar için $t \geq 0$ iken $S(t) = 0$ kabul edildiğinde, Greenwood formülü, en çok olabilirlik geniş örneklem teorisi yardımıyla elde edilir.

Bilgi matrisi $I(\boldsymbol{h})$'nin köşegen elemanları;

$I_{rr}(\boldsymbol{h}) = \dfrac{\partial^2 LogL}{\partial h(r)^2} = \dfrac{n_r}{\{h(r)[1-h(r)]\}}$ şeklinde bulunur ve bu matrisin diğer elemanları 0 'a eşittir.

Geniş örneklem sonuçlarını kullanarak,

$As\,\mathrm{var}(\hat{\boldsymbol{h}}) = I(\hat{\boldsymbol{h}})^{-1}$ 'dir va asimptotik varyans formülünden, yaşam fonksiyonu için asimptotik varyans hesaplanabilir.

5.4 Asimptotik Varyansın Hesaplanması

$\theta = S(t)$ ve $T_n = \hat{S}(t)$ olmak üzere;

$\sqrt{n}[\hat{S}(t) - S(t)] \xrightarrow{D} N(0, \sigma^2)$

$g(T_n)$ ilk türeve sahip bir fonksiyon olsun.

$\sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{D} N[0, g'(\theta)\sigma^2]$ olur.

Buradan, $g(T_n)$ 'nin asimptotik varyansı;

$g[\hat{S}(t)] = \log \hat{S}(t)$ alınırsa,

$$\text{Asvar}[\log \hat{S}(t)] = \frac{1}{\hat{S}(t)^2}\,\text{Asvar}[\hat{S}(t)]$$

$\hat{S}(t)$ 'nin aranılan asimptotik varyansı,

$$Asvar[\hat{S}(t)] = \hat{S}(t)^2\, Asvar\{\log \hat{S}(t)\}$$

$$= \hat{S}(t)^2 \sum_{s=0}^{t-1} \text{Asvar}\left\{\log\left[1-\hat{h}(s)\right]\right\}$$

$$= \hat{S}(t)^2 \sum_{s=0}^{t-1} \frac{\hat{h}(s)\left[1-\hat{h}(s)\right]}{n_s} \qquad (5.3.5)$$

olarak bulunur. Elde edilen bu varyansın Greenwood formülü ile bulunan varyansla aynı olduğu görülmektedir.

6. UYGULAMA

Memeyi oluşturan süt bezleri ve kanalları döşeyen hücrelerin, kontrol dışı olarak çoğalmaları ve vücudun çeşitli yerlerine giderek çoğalmaya devam etmelerine meme kanseri denir.

Radyoterapi, meme bölgesine ve koltuk altına uygulanarak, cerrahi girişimden sonra kalma olasılığı olan kanser hücrelerinin öldürülmesini sağlamak amacı ile yapılır.

Meme kanseri bir çok ülkede, kadınların en korkulu sağlık sorunu olma özelliğini taşımaktadır. Günümüzde ABD'de, sekiz kadından birisi meme kanserine yakalanmaktadır. Bu oran Avrupa ülkelerinde on kadında birdir. Meme kanseri ile ilgili sayıları şu şekilde sıralayabiliriz;

1950-1970 yılları arasında ABD' de, 1 milyon kadın meme kanseri nedeni ile hayatını kaybetti. Bu sayı ABD'nin 2. Dünya savaşı, Kore ve Vietnam savaşlarında kaybettiği insan sayısından fazladır. 1998 yılında Avrupa'da 1 milyon kadın, meme kanserin nedeni ile tedavi görmekteydi. 2000 yılında dünyada 1 milyon kadına, yeni meme kanseri tanısı kondu. Dünyada her 11 dakikada 1 kadın, meme kanseri nedeni ile hayatını kaybediyor. Dünyada her 3 dakikada 1 kadına, yeni meme kanseri tanısı konuyor.

Türkiye' de ise sağlıklı bir istatistik bulunmuyor. Gerek beslenme, gerekse iklim açısından, ülkemiz şartlarına yakın sayabileceğimiz bir Akdeniz ülkesi olan İtalya istatistiklerini ülkemize uyguladığımızda, Türkiye' de her yıl 30 bin kadın meme kanserine yakalanmaktadır.

2001-2006 yılları arasında, Ankara Numune Eğitim ve Araştırma Hastanesi Radyasyon Onkolojisi servisinde RT (Radyoterapi) alan ileri evre meme CA tanılı 63 kadın hasta, tanı konulmasından Eylül 2006 tarihine kadar takip edilmiştir. Sözü edilen 63 kadın hastaya, *Co60* cihazı ile *SAD* (Source Axe Distance) yöntemine göre tanjansiyel alanlarda yapılan tedavi planı sonucunda meme kanseri radyoterapisi (RT)

uygulanmıştır. Hastaların, tanı konulmasından Eylül 2006 tarihine kadar olan yaşam süreleri ay olarak kaydedilmiştir.

Elde edilen yaşam süresi verileri kullanılarak, her gözlem zamanındaki yaşam fonksiyonu, yaşam fonksiyonuna ait varyans ve yaşam fonksiyonuna ait güven aralığı tahmini paket programı yardımıyla ayrı ayrı hesaplanmıştır. Aynı zamanda yaşam fonksiyonu ve hazard fonksiyonu tahminine ilişkin grafikler aynı paket programı yardımıyla çizdirilmiş ve yorumlanmıştır.

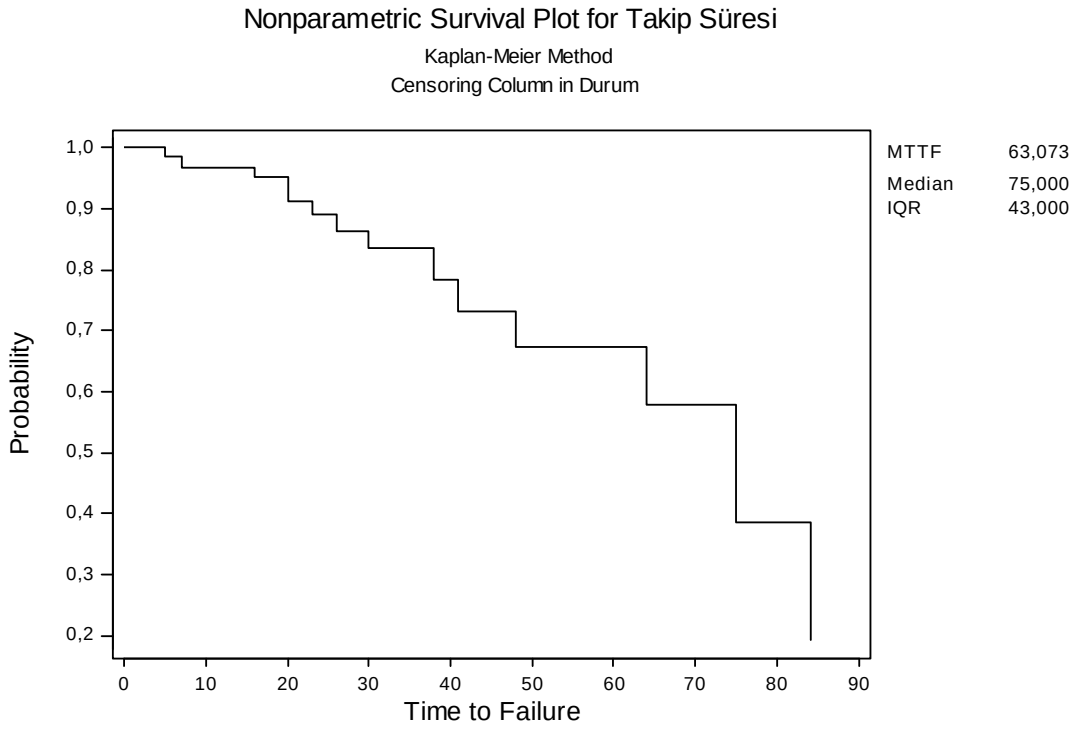
6.1 Analiz Sonuçları

Çizelge 6.1. Temel istatistikler

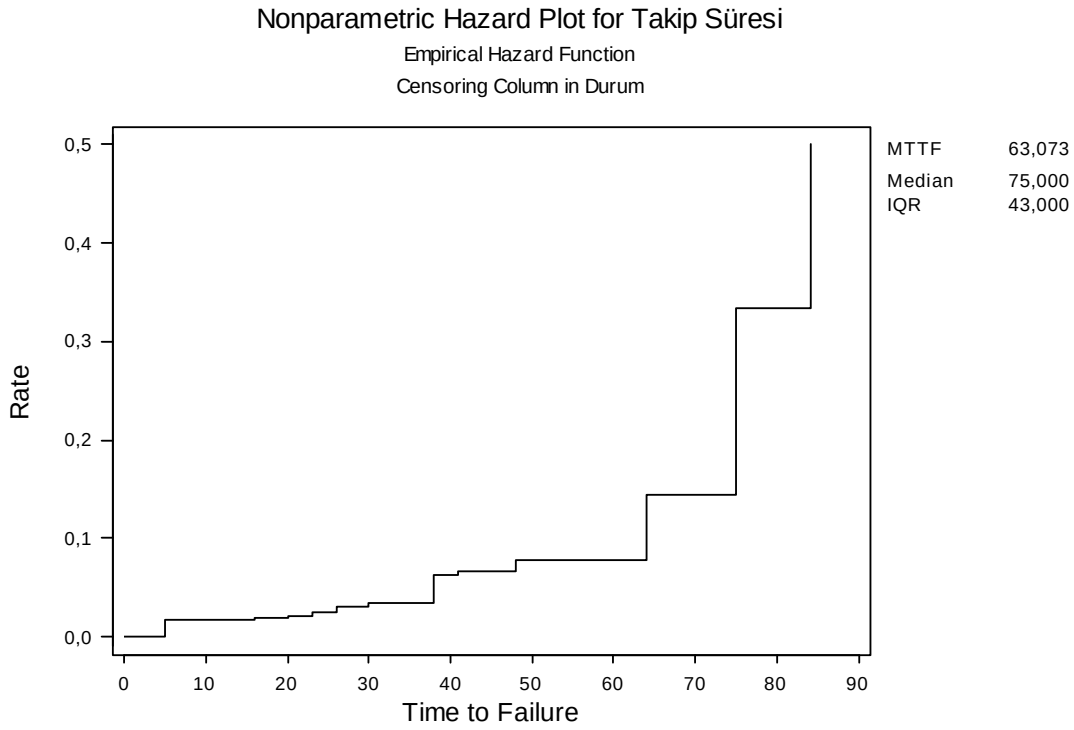
Değişken:	Takip Süresi	
Sansürlenmiş Birim Sayısı	49	
Sansürlenmemiş Birim Sayısı	14	
Ortalama	63.0703	
Standart Hata	4.7890	
% 95 Normal G.A	Alt Sınır	Üst Sınır
	53.6867	72.4594
Ortanca	75	
Q1	41	
Q3	84	

Çizelge 6.2. Kaplan - Meier tahminleri

Zaman	**Risk Altındaki Birey Sayısı**	**Ölüm Sayısı**	**Kaplan-Meier Tahmini**	**Standart Hata**	**95 % Normal G.A**	
					Alt Sınır	**Üst Sınır**
5	63	1	0.9841	0.0157	0.9533	1.0000
7	62	1	0.9683	0.0221	0.9250	1.0000
16	55	1	0.9506	0.0278	0.8961	1.0000
20	49	2	0.9118	0.0379	0.8376	0.9861
23	42	1	0.8901	0.0427	0.8064	0.9739
26	34	1	0.8640	0.0489	0.7682	0.9597
30	29	1	0.8342	0.0555	0.7254	0.9430
38	16	1	0.7820	0.0725	0.6399	0.9241
41	15	1	0.7299	0.0844	0.5646	0.8952
48	13	1	0.6737	0.0947	0.4881	0.8594
64	7	1	0.5775	0.1206	0.3412	0.8138
75	3	1	0.3850	0.1765	0.0390	0.7310
84	2	1	0.1925	0.1622	0.0000	0.5105



Şekil 6.1 Takip süresi için yaşam fonksiyonu



Şekil 6.2 Takip süresi için hazard fonksiyonu

6.2 Yorumlar

Hastalar, hastalığın tanısının konulmasından, önceden belirlenmiş bir zamana (Eylül 2006) kadar izlenmiştir. Eylül 2006 tarihinden sonra elde edilen, hastaların yaşam süreleri sansürlenmiştir. Hastalar, önceden belirlenmiş bir zamana kadar izlenmiş olduğundan bu zamandan sonraki yaşam süreleri, I. Tür sağdan sansürlemeye göre sansürlenmiştir.

Çalışmanın sonunda elde edilen, 63 yaşam süresi verisinden 14 tanesi gözlenmiş ve 49 tanesi ise sansürlenmiştir. Yaşam süresinin ortanca değeri 75 olarak bulunmuştur. Yani, hastaların yarısının yaşam süresi 75 aydan daha kısa diğer yarısının ise daha uzundur. Ortalama yaşam süresi ise, 63 ay olarak bulunmuştur. Analizde, Kaplan-Meier tahminleri her gözlem süresi için ayrı ayrı hesaplanmıştır. Her gözlem zamanında, risk altında bulunan birey sayısı, meydana gelen ölüm sayısı verilmiştir. Daha sonra Kaplan-Meier tahminleri yapılmış ve tahminlerin standart hataları Greenwood formülü ile hesaplanan varyansın karekökü alınarak elde edilmiştir. Son olarak, tahminlerin içinde bulundukları güven aralıkları elde edilmiştir.

Parametrik olmayan yaşam fonksiyonu grafiği incelendiğinde, yaklaşık 50 aya kadar düşüşün daha yavaş olduğu, daha sonra hızlandığı görülmektedir. Hazard fonksiyonu incelendiğinde ise, yaşam fonksiyonun aksine riske odaklandığından, takip edilen hasta grubu için artan bir risk olduğu görülmektedir.

Sonuçlar incelendiğinde, 5. aydan itibaren risk altında bulunan birey sayısı 63'tür ve 1 hasta ex olmuştur. Bu bilgiler doğrultusunda, 5 ay yaşamış bir hastanın 5. aydan sonra yaşama olasılığı yani, $\hat{S}(5) = \Pr(T > 5)$ olarak ifade edilen yaşam fonksiyonu tahmnini; (5.1.6) formülünden , $\hat{S}(5) = \frac{62}{63} = 0.9841$ olarak hesaplanmıştır. Bu tahminin varyans tahmini ise, (5.3.2) ile gösterilen Greenwood formülü kullanılarak bulunmuştur. Analiz sonuçlarında yer alan, standart hata ise, yukarıda sözü edilen Greenwood formülü ile hesaplanan varyansın karekökü dür.

Son olarak, sağdan sansürlenmiş veriler için, belirlenmiş bir *t* zamanındaki yaşam fonksiyonunun güven aralığı tahmini, çeşitli yollarla yapılabilir.

Yaşam fonksiyonu için, çarpım-limit tahmin edicisi kullanıldığında;

$Z_1 = \dfrac{\hat{S}(t) - S(t)}{\hat{\sigma}_s(t)} \sim N(0,1)$ olur.

$\hat{\sigma}_s(t)$, Greenwood'un Formülü kullanılarak bulunan varyans tahminidir.

Bu açıklamalar sonrasında güven aralığının genel hali,

$$P\left[\hat{S}(t) - Z_{\alpha/2}\hat{\sigma}_s(t) \le S(t) \le \hat{S}(t) + Z_{\alpha/2}\hat{\sigma}_s(t)\right] = 1-\alpha \qquad (6.2.1)$$

biçimine dönüşür.

Sözü edilen güven aralığı, bir çok paket programında kullanılmaktadır. Sansürlenmemiş veri sayısı çok az olduğunda veya yaşam fonksiyonu *0*'a ya da *1*'e yaklaştığında Z_1 değişkeni standart normal dağılıma yaklaşmayabilir. Böyle bir durumda, farklı yöntemler kullanılarak yaşam olasılıkları için güven aralıkları oluşturulur. Analiz sonucunda $\hat{S}(5)$ tahmini için elde edilen güven aralığı yukarıda anlatılan şekliyle oluşturulmuştur.

Bu açıklamalardan sonra, diğer gözlem tahminleri için benzer yorumlar yapılabilir.

7. TARTIŞMA VE SONUÇ

Bu çalışmada öncelikle, yaşam analizine ilişkin temel tanımlar ayrıntısıyla incelendi. Daha sonra, yaşam analizinde önemli yer tutan sansürleme kavramı ve çeşitleri üzerinde durularak ve bazı sansürleme mekanizmaları için olabilirlik yapılarının elde edilişi anlatıldı.

Yaşam analizine ilişkin temel tanımların ve sansürleme kavramının anlatılmasından sonra, yaşam fonksiyonun parametrik olmayan tahmininin elde edilişi, ilk olarak sansürlü gözlemlerin olmadığı yani tam örneklem durumunda gösterildi. Daha sonra, sağdan sansürlü gözlemlerin varlığında sözü edilen tahminin, literatürdeki adıyla Kaplan-Meier tahmininin elde edilişi bir örnek ile anlatıldı. Kaplan-Meier tahmininin aynı zamanda, parametrik olmayan en çok olabilirlik tahmin edicisi olduğu gösterildi. Çalışmanın devamında, Kaplan-Meier tahmininin yansızlığı incelendi. Yaşam fonksiyonunun parametrik olmayan tahmini için Kaplan-Meier tahmini kullanıldığında, bu tahminin varyansının Greenwood formülü olarak elde edilişi anlatıldı. Bu varyansın aynı zamanda delta metodu kullanılarakta bulunabileceği gösterilmiştir. Son olarak, Greenwood formülünün asimptotik olarak elde edilişine değinildi.

Çalışmaya uygulama olarak, 2001-2006 yılları arasında Ankara Numune Eğitim ve Araştırma Hastanesi, Radyasyon Onkolojisi Servisi'nde yürütülen bir çalışma için derlenen veriler kullanılmıştır.Meme CA tanısı konulan, 63 kadın hastanın tanı konulmasından çalışmanın bitirildiği tarih olan Eylül 2006 tarihine kadar olan yaşam süreleri verileri için, paket programı yardımıyla öncelikle temel istatistikler daha sonra da, Kaplan-Meier tahminleri ve bu tahminlere ilişkin Greenwood formülü olarak bilinen varyans tahminleri elde edilerek yorumlandı. Söz konusu veri setine ilişkin, parametrik olmayan yaşam fonksiyonu grafiği, hazard fonksiyonu grafiği çizdirilerek yorumlandı.

Analiz sonuçlarından elde edilen bilgilere göre, radyoterapi alan ileri evre meme kanserli bir hastanın, 1 yıllık (12 ay) yaşama oranı yaklaşık %90-91, 5 yıllık (60 ay) yaşam oranı ise yaklaşık olarak %65-67 olarak bulunmuştur. Elde edilen sonuçlar, Radyasyon Onkologları için temel kaynak olarak kabul edilen ve kanser araştırmaları ile

ilgili daha önce yapılmış ve yayınlanmış çalışmalardan oluşan Radyasyon Onkolojisi Tedavi Kararlı ve Temel Pataloji adlı kaynaklarda yer alan sonuçlarla hemen hemen benzerlik göstermektedir. Sonuçlardaki sapmalar, örneklem büyüklüğünden ve hastalığın her hastada farklı seyrinden kaynaklanmaktadır. Sözü edilen kaynaklara göre, evre 2 meme kanserleri için beş yıllık sağ kalım oarnı %65, evre 3 meme kanserleri için ise %40'dır. Tüm meme kanserleri için ise, on yıllık sağ kalım oranı %50'den fazla değildir.

.

KAYNAKLAR

Breslow, N and Crowley, J.1974. A large Saample Study of The Life Table and Product Limit Estimates Under Random Censorship. The Annals of Statistics, Vol. 2 (No:3); 437-453.

Efron, B. 1988. Logistic Regresion, Survival Analysis and the Kaplan-Meier Curve. Journal of the American Statistical Association, (83); 414-425.

Fleming, T.R and Harrington, D.P. 1991. Counting Processes and Survival Analysis. 448, New York.

Gemici, C., Mayadağlı, A.ve Parlak, C. 2004. Radyasyon Onkolojisi: Tedavi Kararları. 706, Türkiye.

Hougaard, P. 1999. Fundamentals of Survival Data. Biometrics, Vol 55, (1); 13-12.

Kaplan, E.J. and Meier, P. 1958. Nonparametric Estimation From Incomplete Observations. Journal American Statistics Association, (53), 457-481.

Klein, J.P., Moeschberger M.L. 1997. Survival Analysis Techniques for Censored and Truncated Data. 465, New York

Kleinbaum, D.G. 1996. Survival Analysis a Self Learning Text. Springer, 197, New York.

Kleinbaum, D.G. 1996. Survival Analysis a Self Learning Text. Springer, 197, New York.

Kral, J.M., Uthoff, V. A., Harley, J. B. 1975. A set-up procedure for selecting variables associated with survival data. Biometrics, (31); 49-75.

Kumar, V., Cotran, R.G., Robbins, S.L. 1996. Basic Pathology.

Lawlees, J.F. 2003. Statistical Models and Methods for Lifetime Data, University of Waterloo, 630, New Jersey.

Le, C.T. 1997. Applied Survival Analysis. 156, New York.

London, D. 1998. Survival Models and Their Estimation. 326, Connecticut.

ÖZGEÇMİŞ

Adı Soyadı : Çiğdem TOPÇU

Doğum Yeri : Burdur

Doğum Tarihi : 02.01.1983

Medeni Hali : Bekar

Yabancı Dili : İngilizce

Eğitim Durumu

Lise : Antalya Aldemir Atilla Konuk Yabancı Dil Ağırlıklı Lise (2000)

Lisans : Hacettepe Üniversitesi, Fen Fakültesi İstatistik Bölümü (2004)

Yüksek Lisans : Ankara Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı. (Eylül 2004-Ocak 2007)

Çalıştığı Kurum

Ankara Üniversitesi Fen Fakültesi, Araştırma Görevlisi -2005

Yayınları

Topçu, Ç., Arslan, F. 2006. Sağdan Sansürlenmiş Veriler İçin Kaplan-Meier Yöntemi Yardımıyla Yaşam Fonksiyonlarının Parametrik Olmayan Tahmini, 5. İstatistik Günleri Sempozyumu 24-27 Mayıs, ANTALYA