Aktüerya Derneği

www.istatistikciler.org
İstatistikçiler Dergisi:
İstatistik&Aktüerya 6 (2013) 1-13

İstatistikçiler Dergisi:
İstatistik&Aktüerya

Yaşam eğrilerini karşılaştırmak için kullanılan skor ve ağırlıklı testler: Sayısal örnekler

Durdu Karasoy

Hacettepe Üniversitesi
Fen Fakültesi, İstatistik Bölümü
06800-Beytepe, Ankara, Türkiye
durdu@hacettepe.edu.tr

Buket Tilki

Hacettepe Üniversitesi
Fen Fakültesi, İstatistik Bölümü
06800-Beytepe, Ankara, Türkiye
buketilki@gmail.com

Özet

İki grup arasında var olan yaşamsal farklılıkların araştırmacılar tarafından tam olarak belirlenemediği ya da açıklanamadığı durumlar vardır. Bu farklılıklar tahmin edilen yaşam fonksiyonunun grafiği çizilerek ortaya konabilir. Fakat bu grafikler, dağılımlar arasındaki farka dair kabaca bir fikir verdiğinden istatistiksel bir testin kullanılması gerekir. İki yaşam eğrisinin eşitliğini test etmek için uygun olan çeşitli parametrik olmayan testler vardır. Bu testler, skor ve ağırlıklı testler olarak ikiye ayrılmıştır. Bu çalışmada, durdurulmuş gözlem içeren ve içermeyen veriler için kullanılan bu testler ele alınmış ve orantılı tehlikeler varsayımını sağlayan ve sağlamayan iki farklı yaşam verilerine uygulanmıştır.

Anahtar sözcükler: *Yaşam eğrileri; Skor ve ağırlıklı testler; Durdurulmuş gözlemler*

Abstract

Score and weighted tests used to compare survival curves: Numerical examples

There are many situations where investigators are not able to specify in advance the specific type of the survival differences that may exist between two groups. These differences can be illustrated by drawing graph of the estimated survivorship function. However, the graph gives only a rough idea of the difference between the distributions; therefore, a statistical test is necessary. There are several distribution-free methods available for testing the equality of two survival curves. These non-parametric methods are divided into score and weighted tests. This study presents these score and weighted tests that can be used for data with and without censored observations. And these tests are applied to the two different survival data sets which one of them satisfies the assumption of proportionality, whereas the other one does not..

Keywords: *Survival curves; Score and weighted tests; Censored observations.*

1. Giriş

Tanımlanan bir başlangıç noktasından bir olayın gerçekleşmesine kadar geçen süre yaşam süresi olarak tanımlanır. Durdurulmuş (censored) gözlemler içermesi nedeniyle yaşam süresi için klasik istatistiksel yöntemler uygun değildir. Durdurulmuş gözlem ise, çalışma süresi sonunda ilgilenilen olay ile karşılaşmamış ya da karşılaşıp karşılaşmadığı bilinmeyen gözlemlerdir [2].

Durdurmanın da olduğu durumlarda iki yaşam süresi dağılımlarının eşitliği testi için dağılımdan bağımsız yöntemler mevcuttur. En çok kullanılan test istatistikleri ise log-rank istatistiği ve genelleştirilmiş Wilcoxon istatistiğidir [12, 13, 14, 15, 16]. Peto ve Peto (1972) sağdan durdurulmuş bağımsız gözlem grupları arasındaki farklılıkları belirlemek için log-rank istatistiğinin büyük bir güce sahip olduğunu ve bu testin istatistiksel önemliliğine ek olarak fiziksel önemlilik de sergilediğini belirtmişlerdir. Prentice (1978) sağdan durdurulmuş verilerle

doğrusal rank testleri hakkında bilgi vermiş ve varyans tahminine permütasyon yaklaşımı ile ilgili özel durumları ele almıştır. Terzi ve Cengiz (2006) sağdan durdurulmuş verilerle iki yaşam dağılımını parametrik olmayan yöntemlerle karşılaştırmışlar ve verilerin durdurulma biçimleri, tehlike oranlarının durumu ve grupların gösterdikleri dağılış biçimi yönünden en uygun hangi testin olduğunu araştırmışlardır. Literatürde doğrusal rank testlerinin birçok sınıfı önerilmiştir. Bu testlerin hesaplama ifadeleri yeteri kadar açık değildir. Testlerin çoğu, yaşam fonksiyonunun parametrik olmayan tahminine dayalıdır. Yaşam fonksiyonunun tahmini ilk olarak Kaplan-Meier tarafından yapılmıştır. Yaşam eğrilerini karşılaştırmak için kullanılan testler, skor ve ağırlıklı testler olmak üzere ikiye ayrılmıştır. Skor testleri varyansın permütasyon tahminini, ağırlıklı testler ise varyansın hipergeometrik dağılıma dayalı tahminini kullanır [14].

1.1. Yaşam çözümlemesi

Yaşam çözümlemesi, pozitif tanımlı rastlantı değişkenlerinin çözümlenmesi için kullanılan istatistiksel teknikler bütünü olarak da tanımlanır. Rastlantı değişkeninin değeri, bir makine parçasının başarısızlık zamanı, biyolojik bir birimin (hasta, hayvan, hücre) ölüm zamanı olabilir. İyi tanımlanmış herhangi bir olayın gerçekleşme ya da gözlenme süresinin çözümlenmesi, yaşam çözümlemesi teknikleri ile yapılabilir. Söz konusu olayın gerçekleşmesi başarısızlık olarak tanımlanır [7].

Yaşam çözümlemesinde temel olan, gözlenen başarısızlık sürelerinin incelenmesidir. Bu nedenle bu değişkenin iyi tanımlanması gerekir [2]. Yaşayan bir organizmanın ya da cansız bir nesnenin belirli bir başlangıç zamanı ile başarısızlığı arasında geçen zamana "yaşam süresi" ya da "başarısızlık süresi" adı verilir ve genellikle T ile gösterilir. Her bir birime ait yaşam süresi T, tanımı gereği sürekli ve pozitif bir değere sahiptir. Başarısızlık süresine örnek olarak, makine bileşenlerinin yaşam süreleri, işçilerin grev süreleri ya da ekonomide işsizlik dönemleri, psikolojik bir deneyde deneğin belirlenen görevi tamamlama süresi ve klinik bir deneyde hastaların yaşam süreleri gösterilebilir [9].

Yaşam çözümlemesini diğer çözümleme tekniklerinden ayıran en önemli özellik durdurulmuş gözlemlerin kullanılabilmesidir [1].

1.2. Yaşam çözümlemesinde kullanılan fonksiyonlar

Yaşam çözümlemesinde kullanılan üç temel fonksiyon bulunmaktadır. Bunlar; yaşam fonksiyonu, olasılık fonksiyonu ve tehlike (hazard) fonksiyonudur. Bu fonksiyonlar birbiri ile ilişkili olan fonksiyonlardır [2].

Yaşam süresi, T ile gösterilen rastgele bir değişkendir. Bir bireyin belli bir t zamanından daha fazla yaşaması olasılığına yaşam fonksiyonu denir ve yaşam fonksiyonu;

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx, \qquad 0 < t < \infty$$

biçiminde ifade edilir.

Yaşam fonksiyonu monoton azalan soldan sürekli bir fonksiyondur ve t=0 iken S(0)=1, $t = \infty$ iken $S(\infty) = 0$ olur.

T rastlantı değişkeninin olasılık yoğunluk fonksiyonu, küçük bir zaman aralığında bir bireyin başarısız olma olasılığının limitidir. Bu fonksiyon;

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}, \quad 0 < t < \infty$$

biçiminde ifade edilir.

Tehlike fonksiyonu h(t), t zamanına kadar yaşayan bir birimin (t+Δt) zamanına kadar yaşamının sona ermesi riskidir. Birimin ilgilenilen özellik bakımından başarısızlık eğiliminin bir ölçüsüdür. h(t), başarısızlık hızı (failure rate), ani ölüm hızı (instantaneous death rate) ya da ölümlülük gücü (force of mortality) olarak da ifade edilir. Tehlike fonksiyonu;

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

biçimindedir. Tehlike fonksiyonu bir zaman aralığında var olan başarısızlık riskinin tanımıdır ve koşullu başarısızlık oranı olarak da tanımlanmaktadır. Tehlike fonksiyonu bir olasılık fonksiyonu değil bir orandır. Olasılık değerleri gibi (0, 1) aralığında değil (0, ∞) aralığında yer almaktadır. Yaşam fonksiyonunun sahip olduğu dağılıma göre tehlike fonksiyonu farklı biçimde olabilir. Bu fonksiyon bir olasılık fonksiyonu olmadığından zamana göre artabilir, azalabilir veya sabit kalabilir [2, 10, 13].

Yaşam sürelerinin olasılık dağılımının belirli bir biçimi olmadığı varsayımı ile tehlike fonksiyonu,

$$h(t;\mathbf{x})=h_0(t)\exp(\boldsymbol{\beta}\mathbf{x})$$

biçimindedir. Bu model Cox regresyon modeli olarak ifade edilir. Burada **x** açıklayıcı değişkenler vektörü, **β** regresyon katsayıları vektörü, $h_0(t)$ ise temel tehlike fonksiyonudur [2].

2. Yaşam fonksiyonunun parametrik olmayan tahminleri

Yaşam çözümlemesi için parametrik olmayan tahmin yöntemleri ve parametrik olmayan test istatistikleri oldukça kullanışlıdır. Çünkü bu tür analizlerin ve testlerin örneklemlerin seçildikleri kitlelerin dağılımına ilişkin varsayımları sağlamaları gerekmez. Yaşam süresi için herhangi bir teorik dağılım varsayımı yapılmadığında, parametrik olmayan tahminler oldukça etkilidir. Yaşam fonksiyonunu tahmin etmede kullanılan pek çok yöntem vardır. En yaygın kullanılan yöntem ise Kaplan-Meier yöntemidir. Bu yöntemle yaşam fonksiyonunun tahmini

$$\hat{S}_{KM}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \quad k \le n, \quad t_{(j)} \le t < t_{(j+1)}$$

biçiminde tanımlanır. Burada,

d_j : t_j'deki başarısızlıkların sayısını,

n_j: t_j'de riskte olan birimlerin sayısını, yani t_j'den hemen önce durdurulmamış ve yaşayan birimlerin sayısını,

k: sıralı gözlem sayısını,

n: toplam birim sayısını

gösterir. Çizelge 1'de yaşam fonksiyonlarının tahmininde kullanılan çeşitli parametrik olmayan testler verilmiştir [5, 14].

Çizelge 1. Yaşam fonksiyonunun parametrik olmayan tahminleri

Kaplan-Meier (1958)	$S_j^{KM} = \prod_{i=1}^{j}\left(\frac{n_i - d_i}{n_i}\right)$
Altshuler (1970)	$S_j^{ALT} = \prod_{i=1}^{j}\exp\left(-\frac{d_i}{n_i}\right)$
Prentice (1978)	$S_j^{*PREN} = \prod_{i=1}^{j}\left(\frac{n_i}{n_i+1}\right)$
Prentice-Marek (1979)	$S_j^{*} = \prod_{i=1}^{j}\left(\frac{n_i - d_i + 1}{n_i+1}\right)$
Harris-Albert (1991)	$S_j^{**} = \prod_{i=1}^{j}\left(\frac{n_i + d_i - 1}{n_i+d_i}\right)$
Fleming-Harrington (1991)	$S_j^{FH} = \exp\left(-\sum_{i=1}^{j}\frac{d_i}{n_i}\right)$
Moreau et al. (1992)	$S_j^{PREN} = \prod_{i=1}^{j}\left(\frac{n_i}{n_i+d_i}\right)$

3. Yaşam eğrilerinin karşılaştırılması için kullanılan parametrik olmayan yöntemler

İki yaşam eğrisinin eşitliğini test etmek için kullanılan çeşitli parametrik olmayan testler vardır. Bu testler, skor ve ağırlıklı testler olarak ikiye ayrılmıştır. Skor testlerinde ve ağırlıklı testlerde kullanılan kavramlar aşağıda verilmiştir:

n_1 birimli G1 ile gösterilen ve n_2 birimli G2 ile gösterilen iki grup olsun ($n_1+n_2=n$). n birimin yaşam süreleri bilinmekte (durdurulmuş ya da değil) ve durdurulmamış birimler için sıralı yaşam süreleri $t_1 < t_2 <\ldots< t_k$ ile gösterilmektedir. Her bir t_j (j=1, 2, … , k) zamanında G1'de n_{1j} birim, G2'de n_{2j} birim ($n_{1j}+n_{2j}=n_j$) riskte olsun. $[t_j, t_{j+1})$ aralığındaki bir durdurulmuş veri, t_j'de riskte, t_{j+1}'de riskte değil olarak alınır. $[t_j, t_{j+1})$ aralığında G1'deki durdurulmuş gözlemler l_{1j}, G2'deki durdurulmuş gözlemler ise l_{2j} olmak üzere $l_j=l_{1j}+l_{2j}$ olur. G1'de t_j'deki başarısızlık sayısı d_{1j}, G2'de t_j'deki başarısızlık sayısı ise d_{2j} olmak üzere $d_j=d_{1j}+d_{2j}$ olur. Bu durumda $n_{j+1}=n_j-d_j-l_j$ olduğu açıktır. Diğer terimler;

$$D_{1j}=\sum_{i=1}^{j}d_{1i}\ ;\ j=1,2,\ldots,k \qquad D_{10}=0, \qquad D_{2j}=\sum_{i=1}^{j}d_{2i}\ ;\ j=1,2,\ldots,k \qquad D_{20}=0,$$

$$L_{1j}=\sum_{i=1}^{j}l_{1i}\ ;\ j=1,2,\ldots,kL_{10}=0, \qquad L_{2j}=\sum_{i=1}^{j}l_{2i}\ ;\ j=1,2,\ldots,k \qquad L_{20}=0,$$

$$D_j=D_{1j}+D_{2j}, \qquad L_j=L_{1j}+L_{2j}$$

$$p_{1j}=\frac{d_{1j}}{n_{1j}}, \qquad p_{2j}=\frac{d_{2j}}{n_{2j}}, \qquad \overline{p}_j=\frac{d_j}{n_j}, \qquad \overline{q}_j=1-\overline{p}_j$$

biçimindedir.

Her bir başarısızlık zamanı için aşağıdaki gibi 2x2 boyutlu k tane çapraz tablo oluşturulur [14]:

	G1	G2	Toplam
t_j'deki başarısızlıkların sayısı	d_{1j}	d_{2j}	d_j
t_j'nin ötesinde yaşayanların sayısı	n_{1j}-d_{1j}	n_{2j}-d_{2j}	n_j-d_j
t_j öncesinde riskte olanların sayısı	n_{1j}	n_{2j}	n_j

Bu testler arasındaki ayrım, varyans tahminleri ile ilgilidir. Skor testleri varyansın permütasyon tahminini kullanırken, ağırlıklı testler hipergeometrik dağılıma dayalı tahmin kullanmaktadır. Durdurma gruplarda farklılık gösteriyorsa permütasyon varyans uygun değildir. Permütasyon varyans, gruplar ve durdurma arasında bağımsızlık varsayımını kullanırken, hipergeometrik varyansın herhangi bir varsayımı yoktur. Hipergeometrik varyans tüm durumlarda geçerlidir. Durdurmanın olduğu ve olmadığı durumlarda hipergeometrik ve permütasyon varyanslar eşit değildir. Hipergeometrik varyans daima permütasyon varyanstan daha küçüktür. Bu nedenle ağırlıklı testler, skor testlerinden daha çok tercih edilir [14]. Bu testler aşağıda verilmiştir:

3.1. Skor testleri

Skor testlerin temel yöntemleri, asimptotik olarak etkin testler veren dönüşüm fonksiyonu kavramını kullanan Peto-Peto (1972) ve yerel olarak en güçlü testleri veren doğrusal modelleri kullanan Prentice (1978) tarafından tanımlanmıştır. Skor testleri, tüm örneklemleri birleştirerek her bir sıralı gözleme skorlar (durdurulmamış gözlemler için c_j ve durdurulmuş gözlemler için ise C_j) verir. Bu test istatistiği ve özellikleri aşağıdaki gibi ifade edilebilir:

$$S=\sum_{j=1}^{k} d_{1j}c_j+\sum_{j=1}^{k} l_{1j}C_j=\sum_{j=1}^{k}\left(d_{1j}c_j+l_{1j}C_j\right),$$

$$E(S)=0,$$

$$V(S)=\frac{n_1n_2}{n(n-1)}\left(\sum_{j=1}^{k} d_jc_j^2+\sum_{j=1}^{k} l_jC_j^2\right),$$

$$\chi^2=\frac{S^2}{V(S)}\sim\chi_1^2.$$

Burada; d_j: t_j'deki başarısızlıkların sayısını ve l_j: t_j'deki durdurulmuşların sayısını gösterir. Skor testleri Çizelge 2'de verildiği gibi özetlenebilir [13, 14, 21].

Bu testlerin dışında literatürde karşılaşılan diğer skor testleri Cox-Mantel, Cox'un F ve Mantel-Haenszel testleridir.

Çizelge 2: Skor testleri (S)

Test S	c_j	C_j
Gehan S_G	n_j-D_j	-D_j
Peto-Peto S_{PP}	$S_j^{KM} + S_{j-1}^{KM} - 1$	$S_j^{KM} - 1$
Prentice S_{PREN}	$2S_j^{PREN} - 1$	$S_j^{PREN} - 1$
LR Altshuler S_{LRALT}	$1 + \mathrm{Ln}S_j^{ALT}$	$\mathrm{Ln}S_j^{ALT}$
Tarone-Ware S_{TW}	$\sqrt{n_j} - \sum_{i=1}^{j} \frac{d_i}{\sqrt{n_i}}$	$-\sum_{i=1}^{j} \frac{d_i}{\sqrt{n_i}}$

Cox-Mantel testi aşağıda verildiği gibi ifade edilir:

$$S = \sum_{j=1}^{k} d_{2j} - \sum_{j=1}^{k} d_j A_j,$$

$$E(S)=0,$$

$$V(S) = \sum_{j=1}^{k} \frac{d_j (n_j - d_j)}{n_j - 1} A_j (1 - A_j),$$

$$\chi^2 = \frac{S^2}{V(S)} \sim \chi_1^2.$$

Burada; $A_j=n_{2j}/n_j$ biçimindedir.

Cox'un F testi ise aşağıdaki gibi ifade edilir:

j. gözlemin beklenen değeri t_{jn} ile gösterildiğinde t_{jn};

$t_{jn} = \frac{1}{n} + \ldots + \frac{1}{n-j+1}$ (j=1, 2, …, n) biçimindedir. Başarısızlıklar için t_{1n}, t_{2n}, …, t_{pn} ($p=d_1+d_2$) skorları, durdurulmuş gözlemler için ise $t_{(p+1)n}$, …, t_{qn} ($q=n_1+n_2-d_1-d_2$) skorları kullanılır. İki grup için $\bar{t}_1$ ve $\bar{t}_2$ ortalama skorları hesaplanır. Birinci grup için ortalama skor,

$$\bar{t}_1 = \frac{d_1 \bar{t}_1' + (n_1 - d_1)\bar{t}_2'}{d_1}$$

biçiminde elde edilir. Burada, $\bar{t}_1'$: başarısızlıkların ortalama skoru ve $\bar{t}_2'$: durdurulmuşların ortalama skorudur.

İkinci grup ortalama skoru $\bar{t}_2$ da benzer biçimde elde edilir. Bu iki ortalama skor değerleri kullanılarak test istatistiği aşağıdaki gibi bulunur:

$$F = \frac{\bar{t}_1}{\bar{t}_2} \sim F_{2d_1, 2d_2}.$$

Mantel-Haenszel testi ise özellikle diğer bir faktöre göre düzeltme olduğunda, iki grubu karşılaştırmada kullanılır. Örneğin, kalp hastalığı için sigara içenlerde ve içmeyenlerde, yüksek kolesterolü olanlarla olmayanların yaşam eğrilerini karşılaştırmada kullanılabilir.

Mantel-Haenszel testinde, veriler değişkene göre tabakalandırır ve her bir tabaka için aşağıdaki gibi 2x2 tablo oluşturulur:

Grup	Başarısızlıkların sayısı	Durdurulmuşların sayısı	Toplam
1	d_{1i}	n_{1i}-d_{1i}	n_{1i}
2	d_{2i}	n_{2i}-d_{2i}	n_{2i}
Toplam	D_i	S_i	T_i

p_{ij}=P(başarısızlık | j. grup, i. tabaka) olmak üzere H_0:$p_{11}=p_{12}$; $p_{21}=p_{22}$; …; $p_{s1}=p_{s2}$ (s, tabaka sayısı) hipotezini test etmek için

$$\chi^2 = \frac{\left[\sum_{i=1}^{s} d_{1i} - \sum_{i=1}^{s} E(d_{1i})\right]^2}{\sum_{i=1}^{s} V(d_{1i})} \sim \chi_1^2$$

test istatistiği kullanılır. Burada;

$$E(d_{1i}) = \frac{n_{1i}D_i}{T_i}, \text{ (i=1, 2, …, s)}$$

$$V(d_{1i}) = \frac{n_{1i}n_{2i}D_iS_i}{T_i^2(T_i-1)}, \text{ (i=1, 2, …, s)}$$

biçimindedir [13].

3.2. Ağırlıklı testler

Ağırlıklı testleri oluşturmak için temel alınan esaslar, asimptotik olarak etkin testler veren dönüşüm fonksiyonu kavramını kullanan Peto-Peto (1972) ve Radhakrishna'nın metodolojisine (asimptotik etkinliği maksimize eder) dayanan Tarone-Ware (1977) tarafından tanımlanmıştır. Ağırlıklı testler, her bir durdurulmamış gözlem süresi için tanımlanan w_j ağırlıklarına dayanır. Bu testler, gözlenen değerler (d_{1j}) ile beklenen değerler (e_{1j}) arasındaki farklara dayanır ve aşağıdaki gibi ifade edilebilir:

$$U = \sum_{j=1}^{k} w_j(d_{1j} - e_{1j}) = \sum_{j=1}^{k} w_j\left(d_{1j} - d_j\frac{n_{1j}}{n_j}\right)$$

$E(U)=0$

$$V(U) = \sum_{j=1}^{k} w_j^2 V(d_{1j}) = \sum_{j=1}^{k} w_j^2 v_{1j} = \sum_{j=1}^{k} w_j^2 \frac{n_{1j}n_{2j}d_j(n_j-d_j)}{n_j^2(n_j-1)}$$

$$\chi^2 = \frac{U^2}{V(U)} \sim \chi_1^2$$

Bazı yazarlar, gözlenen ve beklenen frekanslar arasındaki farkları ağırlıklandırma yerine, iki gruptaki oranlar arasındaki farkları w_j^* ile ağırlıklandırmayı tercih etmektedirler. w_j^* aşağıdaki gibi tanımlanır:

$$w_j = w_j^*\left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}}\right) = w_j^*\left(\frac{n_j}{n_{1j}n_{2j}}\right)$$

U ve V(U) ifadeleri w_j^* ile yeniden yazılırsa,

$$U = \sum_{j=1}^{k} w_j^*(p_{1j} - p_{2j})$$

$$V(U) = \sum_{j=1}^{k} w_j^{*2} \frac{n_j}{n_{1j}n_{2j}} \frac{n_j}{n_j - 1} \overline{p}_j \overline{q}_j$$

biçiminde olur.

Ağırlıklı testlerde kullanılan ağırlıklar Çizelge 3'de verildiği gibi özetlenebilir [4, 5, 14]:

Çizelge 3: Ağırlıklı testler (U) için kullanılan ağırlıklar

Test U	w_j
Gehan U_G	n_j
Peto-Peto U_{PP}	S_{j-1}^{KM}
Prentice U_{PREN}	S_j^{PREN}
LR Altshuler U_{LRALT}	1
Tarone-Ware U_{TW}	$\sqrt{n_j}$
Flemington-Harrington U_{FH}	$\left[S_{j-1}^{KM}\right]^p\left[1 - S_{j-1}^{KM}\right]^q$

3.3.Testlerin karşılaştırılması

Yaşam eğrilerini karşılaştırmada kullanılan testi doğru seçmek oldukça önemlidir. Farklı test, farklı sonuca götürebilmektedir.

Skor teslerinden log-rank testindeki S istatistiği değeri Cox-Mantel testindeki değerle hemen hemen aynıdır. Yuvarlamadan dolayı küçük farklılıklar olabilmektedir [13].

Örneklem büyüklükleri küçük olduğunda (n_1, $n_2 \leq 50$) ve eğer örneklem dağılımları üstel ya da weibull ise Cox'un F testinin Gehan testinden daha güçlü olduğu belirtilmektedir [6].

Örneklemler üstel dağılımdan örneklemler ise Cox-Mantel ve log-rank testlerinin Gehan ve Peto-Peto testlerinden daha güçlü ve daha etkili olduğu belirtilmektedir [11].

Tehlike oranı sabit olmadığında (orantılı tehlikeler varsayımı sağlanmadığında) Gehan ve Peto-Peto testlerinin diğer testlerden daha güçlü olduğu, log-rank testinin ise böyle durumlarda uygun olmadığı, tersi durumda ise log-rank testinin en yüksek güce sahip olduğu ifade edilmektedir [20].

Log-rank testi tüm başarısızlıklara eşit ağırlık verirken Gehan ve Peto-Peto testleri erken görülen başarısızlıklara daha fazla ağırlık vermektedir. Bu nedenle, Gehan ve Peto-Peto testlerinin iki yaşam dağılımlarındaki erken farklılıkları belirlemesi daha olası iken, log-rank testi sağ kuyruktaki farklılıklar için daha duyarlı olmaktadır [13, 19].

İki dağılım farklı, fakat onların tehlike ya da yaşam fonksiyonları çakışıyorsa log-rank ve Gehan çok güçlü değildir. Bu durumda Tarone-Ware gibi diğer testleri incelemek gerekmektedir [21].

Peto-Peto ağırlığı olan S_{j-1}^{KM} ile Prentice ağırlığı olan S_j^{PREN} 'in çok benzer sonuçlar verdiği belirtilmektedir [14].

Flemington-Harrington testi, p ve q değerlerinden dolayı oldukça esnek bir testtir. p=0, q=0 olduğunda log-rank teste, p=1, q=0 olduğunda ise Peto-Peto testine dönüşmektedir [5, 8]. Bu test, q=0, p>0 olduğunda,p=0.5, q=0.5 olduğunda erken meydana gelen başarısızlıklara, p=0, q>0 olduğunda, p=1, q=1 olduğunda, p=0.5, q=2 olduğunda geç meydana gelen başarısızlıklara daha çok ağırlık vermektedir **[**8, 22].

4. Sayısal örnekler

İlk örnek için kullanılan veriler Kleinbaum'dan (1996) alınmış ve Çizelge 4'de verilmiştir. Bu veriler 42 lösemi hastasına ait verilerdir. Bu hastalardan 21'ine belirli bir tedavi verilirken 21'ine placebo verilmiştir. İyileşmenin olmaması başarısızlık olarak alınmıştır. Bunun dışındaki veriler ise durdurulmuş olarak incelemeye alınmıştır. Yaşam süresinin rankı ile Schoenfeld artıkları arasındaki ilişki testi (p=0.79>0.05) ve log-log yaşam eğrileri yöntemleri kullanılarak orantılı tehlikeler varsayımı incelendiğinde tedavi değişkeninin orantılı tehlikeler varsayımını sağladığı görülmüştür.

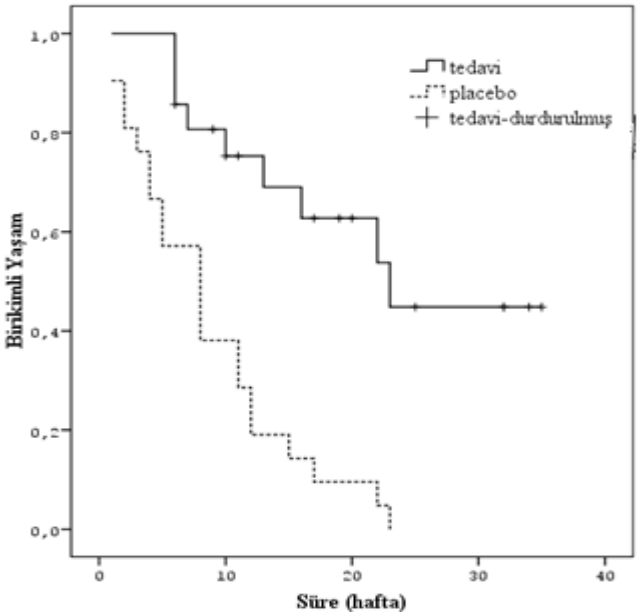
Çizelge 4: Lösemi hastalarına ilişkin veriler

Tedavi (n=21)	6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+
Placebo (n=21)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

+; durdurulmuş gözlemler

Bu veriler için Kaplan-Meier yaşam eğrisi grafiği Şekil 1'de verilmiştir. Süre hafta olarak alınmıştır.

3 aylık yaşam olasılığı tedavi gören grupta 0.75 iken placebo verilen grupta 0.29; 6 aylık yaşam olasılığı tedavi gören grupta 0.45 iken placebo verilen grupta 0.0 olarak elde edilmiştir. Şekil 1'de de görüldüğü gibi tedavi gören grupta yaşam olasılığı daha yüksektir. Ancak aralarında anlamlı fark olup olmadığını görsel olarak söylemek uygun değildir. Test ederek buna karar verilmelidir. Bu amaçla skor ve ağırlıklı testler uygulanmış ve elde edilen sonuçlar Çizelge 5 ve Çizelge 6'da özetlenmiştir. Çizelge 5 ve Çizelge 6'daki testlerde kullanılan ifadelere ilişkin değerler ise Çizelge 7'de verilmiştir.



Şekil 1. Lösemi hastalarına ilişkin Kaplan-Meier yaşam eğrisi

Çizelge 5: Skor testlerin sonuçları

Skor testi	S	V(S)	χ^2	p
Gehan	271	5644.39	13.011	0.00
Peto-Peto	6.9	3.401	13.99	0.00
Prentice	6.2	2.92	13.16	0.00
LR Altshuler	10.1	6.87	14.85	0.00
Tarone-Ware	51.2	174.84	14.99	0.00
Cox-Mantel	10.26	6.253	16.84	0.00
Cox'un F	-	-	F=5.52	0.00
Mantel-Haenszel	-	-	16.4	0.00

Çizelge 6: Ağırlıklı testlerin sonuçları

Ağırlıklı test	U	V(U)	χ^2	p
Gehan	271	5498.03	13.35	0.00
Peto-Peto	6.9	3.296	14.44	0.00
Prentice	6.2	2.8	13.73	0.00
LR Altshuler	10.1	6.18	16.51	0.00
Tarone-Ware	51.2	173.01	15.15	0.00
Flemington-Harington	3.37	0.8619	13.18	0.00

Çizelge 5 ve Çizelge 6 incelendiğinde tüm testler için $p<0.05$ bulunmuştur. Buradan, tedavi gören grupla placebo alan grup arasında yaşam olasılıkları açısından farklılığın anlamlı olduğu %95 güvenle söylenebilmektedir. Tüm testlerde, skor testlerinde kullanılan permütasyon varyansının, ağırlıklı testlerde kullanılan hipergeometrik varyansdan daha büyük olduğu görülmektedir.

Tedavi gören grup referans olarak alınıp Cox regresyon çözümlemesi yapıldığında $h(t)=h_0(t)\exp(1.509\text{placebo})$ modeli elde edilmiştir. Buradan placebo alan grubun tedavi gören gruptan 4.5 kat ($\exp(1.509)=4.523$) daha riskli olduğu sonucu elde edilmiştir. Bu riskin 2.027 ile 10.094 arasında olduğu %95 güven düzeyinde söylenebilmektedir. Bilgisayar yazılımlarında tüm skor ve ağırlıklı testler mevcut değildir. En çok test ise Stata/SE 8.0 da mevcuttur. Lösemi verileri için Stata/SE 8.0 daki test sonuçları Çizelge 8'de verilmiştir. Elde edilen değerlerle Çizelge 6'da verilen değerler arasındaki farklılıklar yuvarlamalardan kaynaklanmaktadır.

Çizelge 8: Lösemi verileri için Stata/SE 8.0 sonuçları

Stata/SE 8.0 testleri	χ^2	P
Log-rank	16.79	0.0000
Wilcoxon (Gehan)	13.46	0.0002
Tarone-Ware	15.12	0.0001
Peto-Peto	14.08	0.0002
Flemington-Harington (p=0, q=1)	13.84	0.0002
Flemington-Harington (p=1, q=1)	12.82	0.0003
Flemington-Harington (p=1, q=3)	12.26	0.0005
Flemington-Harington (p=3, q=1)	8.99	0.0027

Çizelge 7: Lösemi verisine ilişkin (n=42) değerler

t_j	d_{1j}	d_{2j}	d_j*	n_{1j}	n_{2j}	n_j+	l_{1j}	l_{2j}	l_j†	D_{1j}	D_{2j}	D_j‡	A_j•	L_{1j}	L_{2j}	L_j◦	e_{1j}▪	e_{2j}▫	S_{j-1}^{KM}	S_j^{PREN}	S_j^{ALT}
1	0	2	2	21	21	42	0	0	0	0	2	2	0.5	0	0	0	1	1	0.95	0.95	0.95
2	0	2	2	21	19	40	0	0	0	0	4	4	0.475	0	0	0	1.05	0.95	0.91	0.90	0.90
3	0	1	1	21	17	38	0	0	0	0	5	5	0.447	0	0	0	0.55	0.45	0.88	0.88	0.88
4	0	2	2	21	16	37	0	0	0	0	7	7	0.432	0	0	0	1.14	0.86	0.83	0.83	0.83
5	0	2	2	21	14	35	0	0	0	0	9	9	0.4	0	0	0	1.2	0.80	0.79	0.79	0.78
6	3	0	3	21	12	33	1	0	1	3	9	12	0.364	1	0	1	1.91	1.09	0.71	0.72	0.71
7	1	0	1	17	12	29	0	0	0	4	9	13	0.414	1	0	1	0.59	0.41	0.69	0.70	0.69
8	0	4	4	16	12	28	1	0	1	4	13	17	0.428	2	0	2	2.29	1.71	0.59	0.61	0.60
10	1	0	1	15	8	23	1	0	1	5	13	18	0.348	3	0	3	0.65	0.35	0.57	0.58	0.57
11	0	2	2	13	8	21	1	0	1	5	15	20	0.381	4	0	4	1.24	0.76	0.51	0.53	0.52
12	0	2	2	12	6	18	0	0	0	5	17	22	0.33	4	0	4	1.33	0.67	0.46	0.48	0.47
13	1	0	1	12	4	16	0	0	0	6	17	23	0.25	4	0	4	0.75	0.25	0.43	0.45	0.44
15	0	1	1	11	4	15	0	0	0	6	18	24	0.267	4	0	4	0.73	0.27	0.40	0.42	0.41
16	1	0	1	11	3	14	0	0	0	7	18	25	0.214	4	0	4	0.79	0.21	0.37	0.39	0.38
17	0	1	1	10	3	13	3	0	3	7	19	26	0.231	7	0	7	0.77	0.23	0.34	0.36	0.35
22	1	1	2	7	2	9	0	0	0	8	20	28	0.222	7	0	7	1.56	0.44	0.27	0.29	0.28
23	1	1	2	6	1	7	5	0	5	9	21	30	0.143	12	0	12	1.71	0.29	0.19	0.23	0.21

*: başarısızlık sayısı, +: risktekilerin sayısı, †:durdurulmuş gözlem sayısı, ‡: birikimli başarısızlık sayısı, •: n_{2j}/n_j, ◦: birikimli durdurulmuş gözlem sayısı, ▪: $n_{1j}xd_j/n_j$, ▫: $n_{2j}xd_j/n_j$

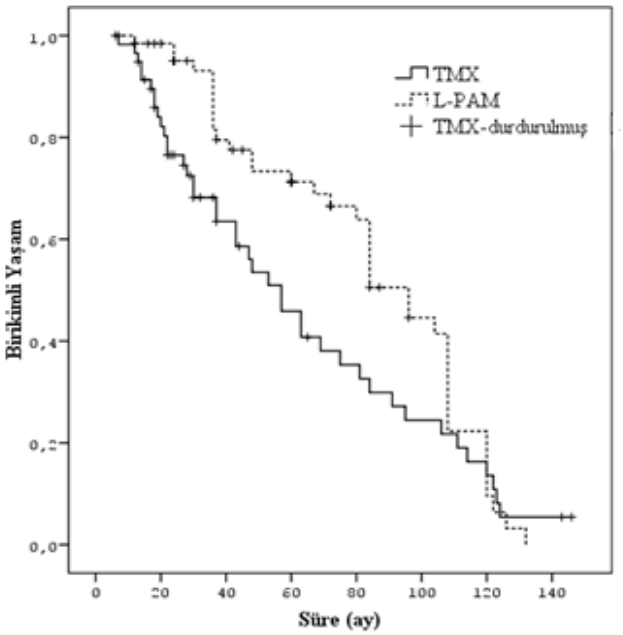
İkinci örnek için Erdoğan (1993) tarafından kullanılan veriler incelenmiştir. Bu veriler, TMX ve L-PAM türü ilaç tedavisi uygulanan meme kanserli 124 hastaya ait verilerdir. Bu hastalardan 59'u (41'i ölmüş, 18'i durdurulmuş) TMX, 65'i (39'u ölmüş, 26'sı durdurulmuş) de L-PAM türü ilaç ile tedaviye alınmıştır. Bu veriler için Kaplan-Meier yaşam eğrisi grafiği Şekil 2'de verilmiştir. Süre ay olarak alınmıştır.

TMX türü tedavi grubunda ortanca yaşam süresi 57 ay, L-PAM türü tedavi grubunda ise ortanca yaşam süresi 96 ay olarak elde edilmiştir. TMX türü tedavi grubunda 2 yıllık yaşam olasılığı 0.77, 5 yıllık yaşam olasılığı ise 0.46 olarak, L-PAM türü tedavi grubunda ise 2 yıllık yaşam olasılığı 0.95, 5 yıllık yaşam olasılığı ise 0.71 olarak elde edilmiştir.

Yaşam olasılıkları açısından iki grup arasında anlamlı fark olup olmadığı Stata/SE 8.0 kullanılarak test edilmiş ve elde edilen sonuçlar Çizelge 9'da verilmiştir. Çizelge 9 incelendiğinde Wilcoxon, Tarone-Ware, Peto-Peto ve Flemington-Harington (p=3, q=1) testleri sonucunda gruplar arası fark anlamlı bulunurken ($p<0.05$), diğer testlerde fark anlamlı bulunmamıştır ($p>0.05$).

Yaşam süresinin rankı ile Schoenfeld artıkları arasındaki ilişki testi ($p=0.0029<0.05$) ve log-log yaşam eğrileri yöntemleri kullanılarak orantılı tehlikeler varsayımı incelendiğinde tedavi değişkeninin orantılı tehlikeler varsayımını sağlamadığı görülmüştür. Orantısız tehlikeler için

Gehan ve Peto-Peto testlerinin diğer testlerden daha uygun olduğu literatürde belirtildiğinden bu testlerin sonuçlarının yorumlanması doğru olacaktır. Buna göre iki tedavi türü arasındaki farkın anlamlı olduğu söylenebilmektedir.



Şekil 2. Meme kanseri hastalarına ilişkin Kaplan-Meier yaşam eğrisi

Çizelge 9: Meme kanseri verileri için Stata/SE 8.0 sonuçları

Stata/SE 8.0 testleri	χ^2	p
Log-rank	3.36	0.067
Wilcoxon (Gehan)	9.4	0.0022
Tarone-Ware	7.31	0.007
Peto-Peto	9.3	0.0023
Flemington-Harington (p=0, q=1)	0.09	0.759
Flemington-Harington (p=1, q=1)	1.34	0.247
Flemington-Harington (p=1, q=3)	0.49	0.486
Flemington-Harington (p=3, q=1)	5.52	0.019

5. Tartışma ve sonuç

Yaşam çözümlemesinde iki yaşam eğrisini karşılaştırmak için kullanılan skor ve ağırlıklı testlerin varyansları farklılık göstermektedir. Ağırlıklı testlerin varyansı her durumda daha küçük elde edilmekte, bu nedenle de ağırlıklı testler skor testlere tercih edilmektedir. Yazılımlarda, sadece tüm birimlerin durdurulmamış olduğu durumlar için bazı skor testleri mevcutken, birçok ağırlıklı testler ise yer almaktadır.

İki yaşam eğrisini karşılaştırırken doğru testin seçilmesi oldukça önemlidir. Bu testleri seçerken öncelikle orantılı tehlikeler varsayımının sağlanıp sağlanmadığına dikkat edilmelidir. Daha sonra ise testlerin özelliklerine göre uygun test seçilmelidir. Oysaki literatürde bir çok çalışmada yaşam eğrilerini karşılaştırmada doğrudan log-rank testi yapılmakta ve yorumlanmakta, orantılılık varsayımı incelenmemektedir. Orantılılık varsayımı sağlandığı durumda log-rank testi doğru bir test iken bu varsayım sağlanmadığında yanlış sonuçlara götürebilmektedir.

Bu çalışmada, orantılı tehlikeler varsayımını sağlayan ve sağlamayan iki farklı sayısal örnek kullanılarak skor ve ağırlıklı testler uygulanmış ve doğru testlerin yorumlanması sonucunda lösemi hastalarında tedavi gören grupla placebo alan grup arasında yaşam olasılıkları açısından farklılığın anlamlı olduğu ve meme kanseri hastaları için iki tedavi türü arasındaki farkın anlamlı olduğu sonuçlarına varılmıştır.

Kaynaklar

[1] D. Collett, 2003, Modelling Survival Data in Medical Research, Chapman and Hall, New York.

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

[3] A. Erdoğan, 1993, Orantılı hazard modeli, Bilim Uzmanlığı Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[4] T. R. Fleming, D. P. Harrington, M. O'Sullivan, 1987, Supremum versions of the log-rank and generalized Wilcoxon statistics", Journal of the American Statistical Association, 82(397), 312–320.

[5] T. R. Fleming, D. P. Harrington, 1991, Counting Processes and Survival Analysis, John Wiley, New York.

[6] E. A. Gehan, D. G. Thomas , 1969, The performance of some two-sample tests in small samples with and without censoring. Biometrika, 56, 127-132.

[7] D. Karasoy, N. Ata, M. T. Sözer, 2005, Yaşam çözümlemesinde zamana bağlı açıklayıcı değişkenli Cox regresyon modeli, Ankara Üniversitesi Tıp Fakültesi Mecmuası, 44, 153-158.

[8] J. P. Klein, M. L. Moeschberger, 1997, Survival Analysis, Techniques for Censored and Truncated Data, Springer-Verlag, New York.

[9] D. G. Kleinbaum, 1996, Survival Analysis, Springer-Verlag, New York.

[10] J. F. Lawless, Statistical Models and Methods for Life Time Data, John Wiley, New York (1982).

[11] E. T. Lee, M. M. Desu, E. A. Gehan, 1975, A Monte-Carlo study of the power of some two-sample tests, Biometrika, 62, 425-432.

[12] J. W. Lee, 1996, Some versatile tests based on the simultaneous use of weighted log-rank statistics, Biometrics, 52, 721-725.

[13] E. T. Lee, J. W. Wang, 2003, Statistical Methods for Survival Data Analysis, John Wiley and Sons, New Jersey.

[14] E. Leton, P. Zuluaga, 2001, Equivalence between score and weighted tests for survival curves, Commun. Statist.–Theory Meth., 30(4), 591-608.

[15] N. Mantel, 1966, Evaluation of survival data and two new rank order statistics arising in its consideration, Cancer Chemotherapy Reports, 50 (3), 163–70.

[16] T. Moreau, J. Maccario, J. Lellouch, C. Huber, 1992, Weighted log-rank statistics for comparing two distributions, Biometrika, 79(1), 195–198.

[17] R. Peto, J. Peto, 1972, Asymptotically efficient rank invariant test procedures, J R Stat Soc Ser A, 135(2), 185-207.

[18] R. L. Prentice, 1978, Linear rank tests with right-censored data, Biometrika, 65, 167–179.

[19] R. L. Prentice, P. Marek, 1979, A Quantitative discrepancy between censored data rank tests, Biometrics, 35, 861-867.

[20] D. Schonfeld, 1981, The asymptotic properties of nonparametric tests for comparing survival distribution, Biometrika, 68, 316-319.

[21] R. E. Tarone, J. Ware, 1977, On distribution-free tests for equality of survival distributions, Biometrika, 64(1), 156–160.

[22] Y. Terzi, M. A. Cengiz, 2006, Sağdan sansürlü iki sağkalım dağılımının karşılaştırılmasında kullanılan parametrik olmayan yöntemlerin gerçek verilere uygulaması, 15. İstatistik Araştırma Sempozyumu, Ankara, 531-541.