تحلیل بقا

افشین آشفته[۱] ابوالقاسم بزرگ‌نیا[۲]

چکیده

در بسیاری از تحقیقات پزشکی نتیجهٔ مورد علاقه، زمان وقوع یک واقعه است. برای مثال در بیماری سرطان، واقعهٔ مورد نظر معمولاً مرگ و یا رویداد مجدد بیماری است. در مطالعات مربوط به پرستاری، واقعه اغلب مرخص کردن بیمار از بیمارستان است و در عملیات پیوند اعضاء، واقعه می‌تواند رد کردن عضو پیوندی باشد. در این مطالعات زمان ورود به مطالعه و معالجه تا زمان وقوع حادثهٔ مورد نظر، تنها برای بعضی واحدهای تحت مطالعه مشخص است. بعضی بیماران که به حادثهٔ مورد نظر نرسیده‌اند، به عنوان مقادیر سانسور شده هستند و در حقیقت وجود این مقادیر سانسور شده است که تحلیل بقا را متمایز می‌کند. در این مقاله سعی شده نکات مهم این تحلیل را معرفی کرده و روشهای کاربردی تحلیل بقا را مد نظر قرار دهیم که در اکثر نرم‌افزارهای آماری می‌توان آنها را مشاهده کرد.

۱. مقدمه

تحلیل بقا مجموعه‌ای شامل تکنیکهای آماری متنوع برای تجزیه و تحلیل متغیرهای تصادفی است که دارای مقادیر مثبت می‌باشند. یکی از برجسته‌ترین مشخصات آنها، زمان از کارافتادگی یک پدیدهٔ فیزیکی (مکانیکی یا الکتریکی)، زمان مرگ یک واحد بیولوژیک (انسان، حیوان و یا سلول زنده) و ... می‌باشد که گاهی ممکن است متغیر تصادفی مورد نظر مربوط به زمان نباشد، مثلاً مقادیر متغیر مربوط به مقدار هزینهٔ پرداختی یک شرکت بیمه به بیمه شدگان در یک وضعیت بخصوص.

درحالی که تجزیه و تحلیل بقا به کارهای ابتدایی روی جداول مرگ و میر که اولین بار توسط جان گرانت در سال ۱۶۶۲ و همچنین ستاره‌شناس معروف ادموندهالی (کاشف ستاره دنباله دار *Halley*) در قرن هفدهم و در سال ۱۶۹۳ انجام گرفته، نسبت داده می‌شود ولی می‌توان گفت که دورهٔ نوین آن از پنجاه سال با کاربرد در مهندسی آغاز شد. بیشتر تحقیقات آماری در علوم مهندسی روی مدلهای پارامتری متمرکز بود. با گذشت دو دهه و با افزایش تعداد آزمایشات تجربی، مدلهای ناپارامتری بقا نیز به کار گرفته شد.

در بسیاری از تحقیقات پزشکی نتیجهٔ مورد علاقه، زمان وقوع یک واقعه است. برای مثال در بیماری سرطان، واقعهٔ مورد نظر معمولاً مرگ و یا رویداد مجدد بیماری است. در مطالعات مربوط به پرستاری، واقعه اغلب مرخص کردن بیمار از بیمارستان است و در عملیات پیوند اعضا، واقعه می‌تواند رد کردن عضو پیوندی باشد. در این مطالعات زمان ورود به مطالعه و معالجه تا زمان وقوع حادثهٔ مورد نظر، تنها برای بعضی واحدهای تحت مطالعه مشخص است و بعضی بیماران که به حادثه مورد نظر نرسیده‌اند، به عنوان مقادیر سانسور شده هستند و در حقیقت وجود این مقادیر سانسور شده است که تحلیل بقا را متمایز می‌کند.

هدف دیگر بسیاری از مطالعات پزشکی، مشخص کردن این نکته است که آیا داروی جدید، روش درمانی یا روش تشخیص جدید، بر انواعی که در حال حاضر استفاده می‌شود، برتری دارد یا خیر؟

[۱] گروه آمار، دانشگاه اصفهان

[۲] گروه آمار، دانشگاه فردوسی مشهد

شرکتهای بیمه معمولاً داده‌های مقطعی را برای تعیین امید به زندگی گروههای سنی مختلف مورد استفاده قرار داده و بر اساس این اطلاعات جدول جاری عمر[۳] را به وجود می‌آورند. همچنین در اکثر مطالعات، از جدولهای عمر هم گروهی[٤] استفاده می‌شود که در این روش همان گروه افراد تحت مطالعه برای دوره زمانی مشخص، دنبال می‌شوند.

روشهای معمول تحلیل اطلاعات برای اندازه‌گیری طول مدت زنده ماندن به دلایل زیر مناسب نیست:

۱. پژوهشگران اغلب باید تجزیه وتحلیل داده‌ها را پیش از آن که همهٔ واحدهای مورد مطالعه، شامل پیشامد مورد نظر شده باشند، انجام دهند. به عبارت دیگر برای مقایسهٔ روشها و تعیین این که کدام روش بهتر است، به چندین سال وقت نیاز است. به همین علت معمولاً تحلیل بقا در زمانی انجام می‌شود که پیشامد مورد نظر برای تعدادی از واحدها رخ نداده است. این واحدها را مشاهدات سانسور شده[٥] می‌گوییم. زیرا از مدت زمان لازم برای به پایان رسیدن آنها بی اطلاع هستیم.

۲. دلیل دیگر این است که زمان شروع مطالعه برای تمام واحدها یکنواخت نبوده و واحدها به طور همزمان در مطالعه قرار نمی‌گیرند. هنگامی که قرار گرفتن واحدها در مطالعه همزمان نیست و تحلیل در زمانی صورت گیرد که همچنان تعدادی از واحدها، تحت مطالعه باشند، این داده‌ها را سانسور شده تصاعدی[٦] می‌نامیم.

۲. داده‌های سانسور شده

داده‌های سانسور شده باعث تمایز تحلیلی بقا از سایر تحلیلهای آماری می‌شوند. این مجموعه داده، آنهایی هستند که ممکن است در طول مطالعه به طور کامل شرکت نداشته و یا تا پایان مطالعه از کار نیفتاده باشند. باید توجه کرد که داده‌های سانسور شده را باید با استفاده از آخرین اطلاع از واحدها دقیقاً ثبت کرد تا در تحلیل داده‌ها مورد استفاده قرار گیرند. در تحلیل بقا اغلب از علامت بعلاوه بعد از داده برای مشخص کردن سانسور بدون آن استفاده می‌شود. برای مثال، ۱۲، ۱۱+، ١٤ و ۲۳+. این سانسور به سه صورت قابل تصور است:

۱. سانسور شده از راست:

واحدهایی که تا زمان معین t_c به پایان خود نرسیده‌اند. $(t > t_c)$

۲. سانسور شده از چپ:

واحدهایی که تا زمان معین t_c به علتی بجز حادثهٔ مورد نظر به پایان خود رسیده‌اند. $(t < t_c)$

۳. سانسور شدهٔ فاصله‌ای:

واحدهایی که بین دو زمان مشخص به از کارافتادگی خود رسیده اند.

فرض کنید $T_1, T_2, \dots, T_n$ متغیرهای تصادفی زمان بقا باشند که مستقل از یکدیگر و هم توزیع هستند. (فرد i ام درنمونهٔ n تایی دارای زمان ناتوانی T_i است). در این صورت در مورد انواع داده‌های سانسور شده داریم:

نوع اول:

t_c را زمان ثابت سانسور در نظر می‌گیریم (که در اختیار محقق است). حال به جای متغیرهای تصادفی $T_1, T_2, \dots, T_n$ می‌توانیم متغیرهای $Y_1, Y_2, \dots, Y_n$ را قرار دهیم که در آن Y_i به صورت زیر تعریف می‌شود:

$$Y_i = \begin{cases} T_i & ; T_i \le t_c \\ t_c & ; T_i > t_c \end{cases}$$

نوع دوم:

فرض کنید اگر تمام مشاهدات بعد از r امین از کارافتادگی را سانسور کنیم ($r > 0$ و ثابت) و $T_{(1)} < T_{(2)} < \dots < T_{(n)}$ آماره‌های مرتب $T_1, T_2, \dots, T_n$ باشند، در این صورت خواهیم داشت:

$$Y_1 = T_{(1)}, Y_{(2)} = T_{(2)}, \dots, Y_r = T_{(r)}, \dots, Y_n = T_{(n)}$$

نوع اول و نوع دوم داده‌های سانسور شده، در علوم مهندسی کاربرد فراوان دارد. مثلاً فرض کنید محققی بخواهد طول عمر لامپهای کارخانه‌ای را بررسی کند. در این صورت ابتدا نمونه‌ای از این لامپها را اختیار و در زمان $t = 0$ همهٔ آنها را روشن کرده و زمان سوختن هر کدام را ثبت می‌کند. گاهی اوقات ممکن است که نخواهد منتظر سوختن همهٔ لامپها بماند و به همین دلیل زمانی را مشخص می‌کند و طول عمر تمام لامپها را بعد از این زمان، به عنوان دادهٔ سانسور شدهٔ نوع اول در نظر می‌گیرد. حال اگر شخص نتواند زمان مشخصی را برای انتظار در نظر بگیرد، فرض می‌کند که لازم است r/n از لامپها بسوزند و از آن پس، طول عمر بقیه لامپها را سانسور شده در نظر می‌گیرد که در این حالت نوع دوم داده‌های سانسور شده را خواهیم داشت.

[۳] Current Life Table

[٤] Cohort Life Table

[٥] Censored Observations

[٦] Progressively Censored

نوع سوم: سانسور تصادفی داده‌ها[۷]

فرض کنید $C_1, C_2, \ldots, C_n$ مستقل و هم توزیع باشند که در آن C_i زمان سانسور متغیر تصادفی T_i می‌باشد. در این صورت مشاهدات شامل $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ هستند که در آن $Y_i = \min(T_i, C_i)$ و داریم:

$$\delta_i = \begin{cases} 1 & ;T_i \le C_i \quad \text{بدون سانسور:} \\ 0 & ;T_i > C_i \quad \text{با سانسور:} \end{cases}$$

دراینجا $Y_1, Y_2, \ldots, Y_n$ مستقل و هم توزیع می‌باشند و δ_i برای نشان دادن اطلاعات سانسور شده به کار رفته است. در این حالت اگر $C_i = t_c$، یعنی مقدار ثابتی باشد، داده‌های سانسور شدهٔ نوع اول را خواهیم داشت.

سانسور تصادفی در مطالعات پزشکی، اپیدمیولوژیکی و مهندسی کاربرد فراوان دارد. چرا که در این مطالعات حالتهای مختلفی از داده‌های سانسور شده را می‌توانیم داشته باشیم. مثلاً در یک بررسی طول عمر، در مورد یک بیماری یا یک دستگاه ممکن است برخی از واحدها تا پایان دورهٔ مطالعه زنده بمانند و یا به علتی غیر از عامل مورد نظر، از مطالعه خارج شوند.

۳. زمان از کارافتادگی

تحلیل بقا به روشهای تحلیل زمان خاتمهٔ پیشامدها اشاره می‌کند. در تحلیل بقا، تمرکز بر روی گروه یا گروههایی از مشاهدات است که برای هر یک از آنها از کارافتادگی (پایانی) تعریف شده و پس از مدت زمان معینی رخ می‌دهند. این زمان را زمان ازکارافتادگی خاتمه یا پایان می‌نامند. برای تعیین زمان دقیق خاتمهٔ پیشامدها، سه چیز لازم است:

۱. تعریف دقیق مبدأ زمانی
۲. مقیاس اندازه‌گیری گذشت زمان
۳. مفهوم روشنی دربارهٔ از کارافتادگی (خاتمهٔ) پیشامد

٤. توزیع زمان از کارافتادگی

درجامعه‌ای از افراد با زمان ازکارافتادگی مشخص، یعنی مدت زمان فوت فرد و یا به نحوی خارج شدن فرد از مطالعه، متغیر تصادفی غیر منفی T که مبدأ و مقیاس اندازه‌گیری آن مشخص است، را معرف زمان ازکارافتادگی در این جامعه در نظر می‌گیریم که ممکن است گسسته یا پیوسته و یا ترکیبی از این دو باشد. تابع توزیع احتمال T را می‌توان به هر یک از سه روش زیر محاسبه کرد:

- تابع بقا[۸]
- تابع چگالی احتمال[۹]
- تابع خطر[۱۰]

روش تابع بقا

فرض کنیم $F(t)$، تابع توزیع تجمعی برای متغیر تصادفی T، به صورت زیر باشد:

$$F(t) = P(T < t) = \int_0^t f(x)dx$$

$$S(t) = 1 - F(t)$$

تابع چگالی احتمال

تابع چگالی احتمال T، یعنی احتمال این که فرد در زمان t فوت کند، عبارت است از:

$$f(t) = \lim_{\Delta t \to 0} \frac{p(t \le T < t + \Delta t)}{\Delta t}$$

$$= \frac{dF(t)}{dt} = \frac{d(1 - S(t))}{dt} = -S'(t)$$

تابع خطر

تابع خطر عبارت است از احتمال از کارافتادگی شخص بیمار در فاصلهٔ $(t, t + \Delta t)$، وقتی بدانیم تا زمان t زنده بوده است. به این تابع، میزان مرگ و میر فوری، میزان شرطی از کارافتادگی یا شدت مرگ و میر نیز گفته می‌شود و در حالات پیوسته به صورت زیر تعریف می‌شود:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{p(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

---

۷ *Random Censoring Data*

۸ *Survival Function*

۹ *Probability Density Function*

۱۰ *Hazard Function*

در شکل (۱) می‌توان ارتباط بین تابع بقا، تابع چگالی احتمال و تابع خطر را مشاهده کرد. همان طور که دیده می‌شود با استفاده از روابط مثلثی بالا با داشتن یکی از این سه رابطه، بقیه قابل محاسبه خواهند بود.

۵. تحلیل پارامتری بقا

به دست آوردن توزیع احتمال دقیق t (زمان از کارافتادگی)، به چند دلیل از اهمیت برخوردار است:

- محاسبهٔ احتمال بقا و احتمال خطر به صورت دقیق امکان پذیر است، چرا که از روشهای دقیقی مانند ML برای برآورد پارامترها می‌توان استفاده کرد.
- از توزیع احتمال در مطالعات بعدی، برای روشهای بیزی می‌توان استفاده کرد.
- از روشهای رگرسیون پارامتری، برآوردهای دقیقتری برای پارامترها نسبت به روشهای ناپارامتری، حاصل می‌شود.
- با استفاده از شکل توزیع و خصوصیات آن از جمله میانگین، واریانس و ... نتایج جالبی را می‌توان استنباط کرد.
- با داشتن احتمال بقا یا قابلیت اعتماد دقیق، میانگین زمان بقا نیز دقیقتر برآورد می شود.

توزیعهایی مانند: نمایی، وایبول، گامبل، گاما، لگ نرمال، لوژستیک و لگ لوژستیک درتحلیل بقا کاربرد فراوان دارند. برای مثال،توزیع وایبول، توزیعی است که در تحلیل بقا کاربرد فراوان دارد و در آن تابع خطر برحسب توانی از زمان تغییر می‌کند و حالتی تعمیم یافته از مدل نمایی است که دارای دو پارامتر بوده و مکرراً برای مدلسازی داده‌های بقا به کار می‌رود و در آن:

$$S(t)=\exp[(-\lambda\times t)^{\alpha}];\quad \alpha>0, t>0$$

تابع خطر به صورت:

$$\lambda(t)=\alpha\times\lambda\times(\lambda\times t)^{\alpha-1}$$

تابع چگالی احتمال به فرم:

$$f(t)=\lambda(t)\times S(t)$$
$$=\alpha\times\lambda\times(\lambda\times t)^{\alpha-1}\times\exp[-(\lambda\times t)^{\alpha}]$$

می‌باشد. در توزیع نمایی تابع خطر وقتی $\alpha<1$ باشد، به طور یکنواخت نزولی و وقتی $\alpha>1$، تابع خطر صعودی و حالتی که $\alpha=1$ باشد، تابع خطر به فرم تابع ثابت تبدیل می‌شود.

۶. انتخاب توزیع مناسب و برآورد پارامترها

روشهای ترسیمی کاربردهای زیادی دارند که در اینجا نیز می‌توانند مورد استفاده قرار داد. در واقع رگرسیون خطی به تشخیص توزیع مناسب به ما کمک کرده و به وسیله برآورد پارامترهای رگرسیون می‌توان پارامترهای توزیع را نیز برآورد کرد. برای روشن شدن موضوع به مثال زیر توجه کنید.

درتوزیع وایبول داریم:

$$S(t)=\exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \quad \text{و} \quad \log S(t)=-\left(\frac{t}{\alpha}\right)^{\beta}$$

$$\log[-\log S(t)]=\beta\log t-\beta\log\alpha$$

اگر توزیع وایبول به داده‌ها برازنده باشد، در این صورت نمودار X و Y که به صورت زیر تعریف می‌شوند، باید یک خط راست را مشخص کنند:

$$X=\log t_{(i)} \quad \text{و} \quad Y=\log[-\log S(t)]$$

برای تشخیص مقدار برآورد R می‌توان از برآورد ناپارامتری هموار شدهٔ کاپلان- مایر[11] استفاده کرد. با توجه به معادلهٔ خط بالا مشخص است که ضریب زاویه، برآوردی برای β خواهد بود و داریم:

$$\hat{\alpha}=\exp\left(\frac{-INTERCEPT}{\hat{\beta}}\right)$$

مثال:

داده‌های زیر توسط من و فرتیج[12] [۶] برای تست زمان بقای یک وسیلهٔ هواپیما به دست آمده است که بعد از کارافتادگی ۱۰ ام آزمایش متوقف شده است:

۳/۰۰، ۲/۵۰، ۱/۷۶، ۱/۵٤، ۱/۳۳، ۱/۳۲، ۱/۰۰، ۰/۸۸، ۰/۵۰، ۰/۲۲

در این آزمایش سه مشاهدهٔ سانسور شده از راست در زمان ۳/۰۰ وجود دارد. به وسیلهٔ نرم افزار مینی تب[13] محاسبات لازم انجام شده و نتایج به صورت زیر به دست آمده‌اند:

---

[11] Kaplan - Meire Estimator

[12] Mann and Fertig

[13] Minitab

ضریب زاویه برابر است با ۱/۳۹ که به عنوان برآوردی برای β استفاده می‌شود. همچنین داریم:

$$\hat{\alpha} = \exp\left(\frac{-(-۱/۱۴)}{۱/۳۹}\right) = ۲/۲۷$$

میزان ضریب تعیین[14] برابر با ۰/۹۸ است که مقداری قابل قبول و نشاندهندهٔ برازنده بودن مدل است.

تعدادی از مدلهای آشنا را می‌توان مانند بالا به دست آورد و جدول (۱) را برای تحقیق مدلهای دیگر بکار برد.

۷. برآورد بقا از طریق روش حد حاصلضربی کاپلان ـ مه یر

محاسبهٔ تابع بقا به روش (PL)[15] که به آن برآورد کنندهٔ کاپلان ـ مه یر نیز گفته می‌شود. یکی از قدیمی‌ترین روشهای ناپارامتری است که می‌توان توسط آن، در تحلیل بقا، طول عمر بیماران را برآورد کرد به طوری که اثر داده‌های سانسور شده را نیز به حساب آورد.

اگر نمونه‌ای n تایی از زمانهای ازکارافتادگی مجزا بدون دادهٔ سانسور شده در یک جامعه همگن داشته باشیم، در این حالت تابع بقا یک تابع پله‌ای است که با مشاهدهٔ هر زمان ناتوانی، به اندازهٔ $1/n$ کاهش می‌یابد.

با این حال، همان گونه که قبلاً بیان شد داده‌های بقا اغلب دارای موارد سانسور شده می‌باشند و لذا روش خاصی را برای برآورد تابع بقا، ایجاب می‌کنند. این روش هنگامی به کار می‌رود که توزیع زمان بین دو پیشامد که یکی از آنها الزاماً رخ نداده است، مورد نظر باشد. (در واقع تعیین احتمال وقوع حادثه، تا پیشامد بعدی اندازه‌گیری می‌شود). پیشامد دوم که الزاماً رخ نداده است، همان باقی ماندن (زنده ماندن) مشاهده است. (مثلاً، برآورد نسبت کارکنانی که پس از استخدام در یک شرکت باقی می‌مانند، در هر زمان معین که یک فرد شرکت را ترک می‌کند، نسبت افراد باقی مانده دوباره محاسبه می‌شود).

در حقیقت روش کاپلان ـ مه یر برای برآورد بقا، با روش تحلیل جدول عمر یکسان است، بجز این که در این روش مدت زمان سپری شده از ورود شخص به مطالعه، تا زمان انجام تحلیل به فاصله‌های زمانی، تقسیم نمی‌شود. به همین دلیل، روش کاپلان ـ مه یر در مطالعاتی که بر روی تعداد کمی از بیماران می‌باشد، مناسب است. بعلاوه تعداد محاسبات، نسبت به روش جدول عمر کمتر است. چون بقای هر زمان که در آن یک پیشامد مورد نظر رخ می‌دهد، محاسبه می‌شود و بدین ترتیب واحدهای کنار گذاشته شده در نظر گرفته نمی‌شوند، همچنین به جای محاسبهٔ توزیع زمانهای بقای مشاهده شده در قالب یک جدول بقا، مستقیماً زمانهای مرگ یا بقای پیوسته محاسبه می‌شود.

برای برآورد تابع بقا از روش حد حاصلضربی، فرض کنید $Y_{(1)} < Y_{(2)} < \dots < Y_{(n)}$ آماره‌های مرتب زمانهای بقای $Y_1, Y_2, \dots Y_n$ باشند. در این صورت:

$$P_j = 1 - q_j = \begin{cases} 1 - \dfrac{1}{n_j} & \text{اگر حادثه رخ دهد؛ } \delta_j = 1 \\ & d_j = 1 \\ 1 & \text{اگر داده سانسور شده باشد؛ } \delta_j = 0 \end{cases}$$

که در آن n_j تعداد افرادی است که در لحظهٔ وقوع حادثه مربوط به Y_j، زنده هستند، P_j نسبت بقا در فاصلهٔ I_j است به شرط آن که افراد در شروع فاصله زنده باشند و d_j تعداد مرگها مربوط به زمان $Y_{(j)}$ می‌باشد.

در اینجا $S(t)$، تابع بقا در زمان t، به صورت زیر است و n تعداد کل مشاهدات:

$$S(t) = \prod_{Y(j) \le t} \left[\frac{n-j}{n-j+1}\right]^{\delta(j)}$$

$$= \prod_{Y(j) \le t} \left[1 - \frac{1}{n_j}\right]^{\delta(j)} \quad ; n_j = n - j + 1$$

و $\delta_{(j)}$ تابعی است که با ضابطهٔ زیر تعیین می‌شود:

$$\delta_{(j)} = \begin{cases} 1 & \text{مشاهده } j \text{ ام ازبین رفته باشد؛} \\ 0 & \text{مشاهده } j \text{ام از باقیمانده باشد؛} \end{cases}$$

مثال:

داده‌های زیر برگرفته از لاو لس[16] [5] در مورد زمانهای بهبود متأثر از دارویی خاص است. اعداد مشخص شده با ستاره، سانسور شده از

---

[14] *R-Squared*

[15] *Product- Limit*

[16] *Lawless*

راست هستند:

۶,۶,۶,۶*,۷,۹*,۱۰,۱۰*,۱۱*۱۳,۱۶

۱۷*,۱۹*,۲۰*,۲۲,۲۳,۲۵*,۳۲*,۳۲*,۳۴*,۳۵*

جدول (۲) می‌تواند با استفاده از داده‌های بالا به دست آید.

۸. برآوردکننده‌های نلسن و فلمینگ- هارینگتون

برای محاسبهٔ تابع توزیع تجمعی خطر، از برآورد نلسن می‌توان استفاده کرد:

$$\hat{H}(t)=\sum_{t_i<t}\frac{d(t_i)}{r(t_i)},\quad \hat{H}(t)=\sum_{i=1}^{n}\int_{0}^{t}\frac{dN_i(s)}{r(s)}$$

که در آن $d(t_i)$ تعداد مرگها و $r(t_i)$ تعداد در معرض خطر می‌باشد.

برآورد نلسن یک تابع پله‌ای است و از صفر شروع شده و طول هر پله به ازای هر مرگ، به اندازهٔ $d(t_i)/r(t_i)$ می‌باشد. یک مشکل این برآورد این است که بسیار مستعد پیوند دادن داده‌ها می‌باشد. به این مفهوم که اگر سه اتفاق در سه زمان بسیار نزدیک به هم مانند t_1، t_2 و t_3 اتفاق بیافتند و ما ۷ واحد در معرض خطر دیگر داشته باشیم، طبق فرمول، مقدار برآورد نلسن باید برابر ۱/۸ +۱/۹+۱/۱۰ شود ولی اگر این تقارب از نظر دور بماند، مقدار برآورد برابر ۳/۱۰ خواهد شد که با مقدار واقعی تفاوت دارد. راه حل در نظر گرفته شده توسط نلسن و فلمینگ- هارینگتون درسال ۱۹۸٤، در نظر گرفتن بیشترین تعداد زمانهای مرگ بود. رابطهٔ $H(t)=-\log S(t)$ که برای توزیعهای پیوسته کاربرد دارد توسط نلسن و فلمینگ- هارینگتون در سال ۱۹۸٤ به برآورد زیر منجر شد:

$$\hat{S}(t_j)=\exp[-\hat{H}(t_j)]$$

در نمونه‌های به اندازهٔ کافی بزرگ، برآورد های FH و KM بسیار به هم نزدیک هستند.

۹. مقایسه منحنیهای بقا

اگرچه در برخی موارد آمار بقا برای فقط یک گروه گزارش می‌شود ولی پژوهشگران در بسیاری از موارد تمایل دارند که بقا را بین دو نمونه مقایسه کنند.

اگر تحلیل جدول عمر و برآوردهای حد حاصلضربی کاپلان- مه یر را برای هر دو نمونه انجام داده و نمودارهای مربوطه را رسم کنیم، بالاتر بودن منحنی بقا نشان دهندهٔ این است که در هر مرحله‌ای از زمان، نسبت بالاتری از واحدها دارای بقا می‌باشند. به هر حال ممکن است تغییرات در نمونه‌ها، فقط به طور شانسی باشد و سوال منطقی این است که آیا تفاوتهای موجود بزرگتر از مقدار مورد انتظار بر اساس شانس می‌باشد یا خیر. برای آزمایش این فرضیه، به روشهایی برای مقایسهٔ توزیعهای بقا نیاز داریم.

معرفی روشهای مقایسه بطور کلی

فرض کنید که بخواهیم توزیع زمان بقا p گروه مختلف را با هم مقایسه کنیم. یک روش، ایجاد جدولی $2\times p$ برای هر زمان که مرگی رخ می دهد می‌باشد. پس به تعداد زمانهایی که مرگی در آن اتفاق می‌افتد، جدولهای جداگانه‌ای خواهیم داشت (مشخص است که حجم محاسبات بسیار زیاد خواهد شد).

جدول (۳) حاکی از یک آزمایش چندگانه ساده با d پیشامد در N آزمایش می‌باشد. تعداد مرگ مورد انتظار در هر گروه برابر است با dn_i/N، با یک ماتریس واریانس چندگانهٔ استاندارد مانند V.

در هر جدول برای هر گروه، تعداد مرگها و تعداد مورد انتظار را به دست آورده و آنها را با هم جمع می‌کنیم تا برای هر جدول یک O_i و E_i داشته باشیم. حال تفاضل این دو مقدار را در هر جدول محاسبه می‌کنیم. برداری از این تفاضلها با در نظر گرفتن کل جداول به دست می‌آید که دارای ماتریس واریانس $\sum V_k$ می‌باشد. مطالب بالا را می‌توان با در نظر گرفتن وزن W_i برای هر زمان مرگ تعمیم داد. در این صورت بردار وزنی ما برابر با $\sum w_k(O_k-E_k)$ خواهد بود که در آن O_k مقدار مشاهده شده و E_k مقدار مورد انتظار در هر جدول می‌باشد.

مقدار واریانس نیز برابر با $\sum W_k^2V_k$ خواهد بود. حال داریم:

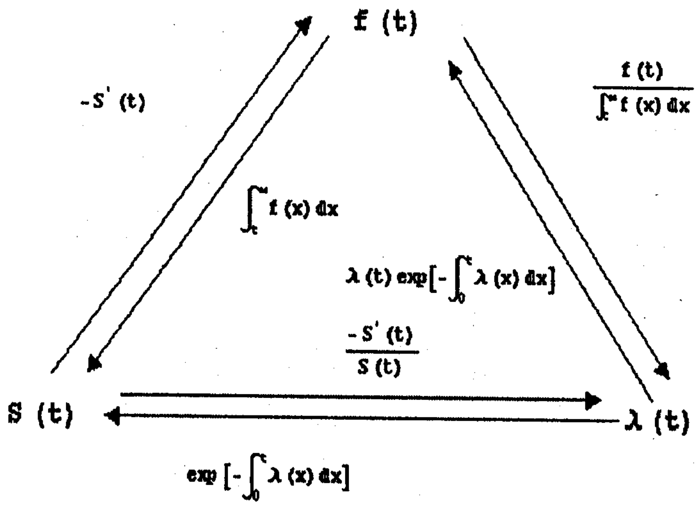
- اگر $W_k=1$ ، روش بالا تبدیل به روش مانتل- هنزل[۱۷] یا روش *Log-Rank Test* خواهد شد.
- اگر $W_k=n_k$ ، آزمون گهان- ویلکاکسون[۱۸] را خواهیم داشت.
- اگر $W_k=S_{KM}(t_k)$ ، آنگاه اصلاح پتو- پتو[۱۹] برای آزمون ویلکاکسون را داریم.

---

[۱۷] *Mantel- Haenszel*

[۱۸] *Gehan-Wilcoxon Test*

[۱۹] *Peto-Peto Modification*



شکل (۱)

جدول (۱): تبدیلات لازم برای بررسی توزیع داده‌های بقا توسط رگرسیون

X	Y	مدل
t	$\log \hat{S}(t)$	نمایی
$\log t$	$\log \hat{S}(t)$	پارتو
t	$\log\lvert -\log \hat{S}(t)\rvert$	گامبل
$\log t$	$\log\lvert -\log \hat{S}(t)\rvert$	وایبول
t	$\phi^{-1}\lvert ۱-\hat{S}(t)\rvert$	نرمال
$Log\, t$	$\phi^{-1}\lvert ۱-\hat{S}(t)\rvert$	لگ- نرمال
$\sqrt{t}$	$\phi^{-1}\lvert ۱-\hat{S}(t)\rvert$	گاما
t	$\log\left[\frac{۱-\hat{S}(t)}{\hat{S}(t)}\right]$	لوژستیک
$\log t$	$\log\left[\frac{۱-\hat{S}(t)}{\hat{S}(t)}\right]$	لگ- لوژستیک

جدول (۲): محاسبات مربوط به برآوردهای بقا با روش حد حاصلضربی کاپلان- مایز

$T(j)$	n_j	d_j	$\frac{(n_j-d_j)}{n_j}$	$\hat{S}(t)$	$\hat{S}_{(t-1)}$	$\hat{S}_t=\frac{1}{2}(\hat{S}_{(t)}+\hat{S}_{(t-1)})$
۶	۲۱	۳	۱۸/۲۱	۰/۸۵۷	۱/۰۰۰	۰/۹۲۹
۷	۱۷	۱	۱۶/۱۷	۰/۸۰۷	۰/۸۵۷	۰/۸۳۲
۱۰	۱۵	۱	۱۴/۱۵	۰/۷۵۳	۰/۸۰۷	۰/۷۸۰
۱۳	۱۲	۱	۱۱/۱۲	۰/۶۹۰	۰/۷۵۳	۰/۷۲۲
۱۶	۱۱	۱	۱۰/۱۱	۰/۶۲۷	۰/۶۹۰	۰/۶۵۹
۲۲	۷	۱	۶/۷	۰/۵۳۸	۰/۶۲۷	۰/۵۸۳
۲۳	۶	۱	۵/۶	۰/۴۴۸	۰/۵۳۸	۰/۴۹۳

جدول (۳): اطلاعات لازم برای معرفی روشهای مقایسه منحنیهای بقا

	P	....	۲	۱	گروه
d	d_p	....	d_2	d_1	تعداد مرگ
a	a_p	....	a_2	a	تعداد در معرض خطر
N	n_p	....	n_2	n_1	تعداد کل

مراجع

[۱] آشفته، افشین، ۱۳۸۰، تحلیل بقا و بررسی بیماران سرطانی مشهد، پایان نامهٔ کارشناسی ارشد آمار، دانشگاه فردوسی مشهد.

[2] Everitt, B.S. and Dunn, G., 1998, *Statistical Analysis of Medical Data,* First Published in Great Britain by Arnold, a member of theHodder Headline Groups, Oxford University Press, New York.

[3] Fleming, T.R. and Harrington, D.P., 1984, *Nonparametric Estimation of Survival Distribution in Censored Data,* Communications in Satatistics, 13(20), 2469- 2486.

[4] Kalbfiesisch, J.D and Prentice, R.C., 1980, *The Statistical Analysis of Failure Time Data,* Wiley, NewYork.

[5] Lawless, J.F., 1982, *Statistical Models and Methods for Lifetime Data,* Wiley, New York.

[6] Mann, N.R. and Fertig, K.W., 1973, *Tables for Obtaining Confidence Bounds and Tolerance Bounds Based on Best Linear Estimates of Parameters of the Enterme Value Distribution,* Technometrics, 16, 335-346.

[7] Wolstenholme, L.C., 1999, *Reliability Modeling: A Statistical Approach,* Chapman and Hall/CRC.