دانشکده علوم ریاضی و آمار

پایان نامه

برای دریافت درجه کارشناسی ارشد در رشته آمار ریاضی

عنوان

تأثیر نمونه گیری در طول اریب بر متغیرهای گمگی

در برآورد تابع بقا

استاد راهنما

جناب آقای دکتر وحید فکور بافنده

استاد مشاور

جناب آقای دکتر حسنعلی آذرنوش

نگارنده

پطروس اصغری

خرداد ماه ۱۳۸۹

این پایان نامه حاصل گوشه نظر پروردگار مهربان و همراهی و تشویق بی دریغ پدر و مادر عزیزم می باشد.

و نیز بدون کمک‌های فراوان آقایان دکتر فکور و دکتر آذرنوش هیچگاه به سرانجام نمی‌رسیدم.

امیدوارم توانسته باشم گوشه‌ای از زحمات آن عزیزان را جبران کرده باشم.

این پایان نامه را تقدیم می‌کنم به برادرم حمیدرضا:

خون کدام عاشق بیدل

نفرین کدام پیر بلاکش

ترا به باد داد ای باغ سبز

چنین گفت سیب سرخ

پطروس اصغری

خرداد ماه یکهزار وسیصد وهشتاد ونه

اطلاعات مربوط به چکیده پایان نامه

شماره دانشجو : ۸۶۱۳۳۹۵۰۲۳ **نام و نام خانوادگی :**پطروس اصغری

دانشکده :علوم ریاضی **رشته :**آمار ریاضی

مقطع :کارشناسی ارشد **دوره :**روزانه

عنوان پایان نامه : تأثیر نمونه گیری در طول اریب بر متغیرهای کمکی در برآورد تابع بقا

تاریخ دفاع : ۱۳۸۹-۳-۳۱

تعداد صفحات : ۱۰۰

کلمات کلیدی : در طول اریب ; برآورد تابع بقا ; برش چپ ; سانسور از راست ; سانسور حاصلضربی

استاد راهنما : وحید فکور بافنده

استاد مشاور : حسنعلی آذرنوش

چکیده :

تحلیل داده های بقای سانسور راست و بریده شده از چپ در بسیاری از تالیفات آماری به چشم میخورد . اهمیت تحلیل این نوع داده ها از وسعت کاربرد آنها در عمل سرچشمه می گیرد که دربررسی های پزشکی در رابطه با بیماری هایی از جمله ایدز و دمانس کاربرد فراوان دارد. نکته ای که در مورد این چنین بیماری هایی پیش می آید، این است که زمان دقیق یا حتی تخمینی شروع بیماری در دست نیست و بیماران معمولا در اثر حادثه یا کاملاً اتفاقی متوجه بیماری خود می شوند چرا که بیماری هایی مانند ایدز یا تومورهای خوش خیم، مدت نسبتا زیادی در بدن فرد باقی می مانند تا به مرحله نمود خود برسند و نشانه های خود را آشکار سازند. حتی بیماری هایی نظیر دمانس که ممکن است اطرافیان زود به آن پی ببرند فرد از زمان شروع بیماری اطلاعی ندارد . موضوعی که کمتر به آن توجه می شود، اریب شدن نتایج حاصل از تحلیل این گونه داده ها می باشد، چرا که مشاهده می شود که افرادی که وارد نمونه می شوند، نسبت به بقیه افراد جامعه دارای طول عمر بیشتری هستند . این اریبی در اصطلاح، درطول اریبی نامیده می شود. حال در کنار داده های اصلی، متغیرهای کمکی نیز می توانند وا رد مدل شوند. در این پایان نامه به بررسی چگونگی ورود و نیز نحوه تأثیر در طول اریبی بر متغیرهای کمکی پرداخته می شود . اما هدف اصلی، چگونگی رفتار با این متغیرها در برآورد پارامتری تابع بقا می باشد . در این راستا دو نوع تابع درستنمایی معرفی می شود و از روش درستن‌مایی ماکسیمم، پارامترها برآورد می شوند . این دو نوع تابع درستنمایی بر پایه دو دیدگاه معرفی می شوند. دیدگاه اول از نادیده گرفتن اطلاعات موجود در متغیرهای کمکی تأثیر می گیرد و تابع درستنمایی را شرطی بر روی آنها معرفی می کند . در دیدگاه دوم تابع درستنمایی به صورت توأم با متغیرهای کمکی تولید می شود.در پایان نتیجه می شود که این دو تابع درستنمایی در حجم نمونه زیاد، تفاوت چندانی با هم نداشته، حال آنکه در حجم نمونه کم، برآوردگرهای حاصل از تابع درستنمایی توأم کارایی بیشتری نسبت به حالت شرطی دارند

Thesis Abstract Information

Name: petros asghari **Student No.:** 8613395023

Field: Mathematical Statistics **Faculty:** Mathematics

Section: M.A.

Thesis title : The effect of Length Biased sampling on Covariates in estimating the survival function

Date of thesis defense : 31-3-1389

Page No. : 100

Keywords : length bias ; estimating the survival function ; left truncation ; right censored ; multiplicative censoring

---

Thesis Supervisor : Vahid Fakoor Bafandeh

Thesis Advisor : Hasan Ali Azarnoosh

Abstract :

Analysis of the right censored and left truncated survival data is seen in many statistical publications. The importance of the analysis of these data is a result of the wide usage of them in practice especially in medical surveys such as studies on some diseases such as AIDS and Dementia. The point about these kinds of data is this that the initiating time of the disease is not known and the patients would accidentally determine that they have the disease. Because these kinds of diseases stay potential in the patient's body for almost a long time until they show their sings and infect the person. Even diseases such as Dementia, which the relatives may recognize it soon, there is no accurate information about the time when the disease has started.
The subject that is ignored mostly is this, that these kinds of data are biased. As a matter of fact, it can be seen that the people whom they are in the sample, have a longer life time in comparison to whom they are not or cannot be in the sample. This biasness in statistics is called Length Bias. Beside these data, covariates can be entered into the model. In this thesis, the most concern is on how to enter covariates into the model, and how to make behave with these covariates in estimating the survival function. In this sequel, two kinds of likelihood functions are introduced on the basis of two viewpoints. In the end, it is concluded that these two types of likelihood functions, do not differ a lot when dealing with huge sample volumes. Although in low sample volumes the estimate of the survival function of one of the likelihood functions is more efficient. And that function is the function that takes into account the covariates.

فهرست مندرجات

نمادها

تابع توزیع در طول اریب	F_{LB}
کمینه بین دو مقدار	$\wedge$
متناسب بودن با	$\propto$
همگرایی قریب به یقین	$\xrightarrow{a.s.}$
همگرایی در توزیع	$\xrightarrow{D}$

پیشگفتار

تحلیل داده های بقای سانسور راست و بریده شده از چپ در بسیاری از تا لیفات آماری به چشم میخورد. اهمیت تحلیل این نوع داده ها از وس عت کاربرد آنها د ر عمل سرچشمه می گیرد که دربررسی های پزشکی در رابطه با بیماری هایی از جمله ایدز و دمانس کاربرد فراوان دارد . نکته ای که در م ورد این چنین بیماری هایی پیش می آید، این است که زمان دقیق یا حتی تخمینی شروع بیماری در دست نیست و بیماران معمولا در اثر حادثه یا کاملاً اتفاقی متوجه بیماری خود می شوند چرا که بیماری‌هایی مانند ایدز یا تومور های خوش خیم، مدت نسبتا زیادی در بدن فرد باقی می‌مانند تا به مرحله نمود خود برسند و نشانه‌های خود را آشکار سازند. حتی بیماری‌هایی نظیر دمانس که ممکن است اطرافیان زود به آن پی ببرند فرد از زمان شروع بیماری اطلاعی ندارد .

مثال دیگر، در مورد تحلیل اطلاعات مربوط به معتادین می باشد. مشکلی که در اینجا پیش می آید، این است که افراد مورد بررسی تمایل زیادی به بیان نحوه و چگونگی ابتلا و زمان شروع اعتیاد، ندارند .

معمولا برای جمع آوری اطلاعات مربوط به این موضوع ها، به آسایشگاه ها و یا بیمارستان های محل نگهداری آنها مراجعه می‌شود. مشکل بعدی در اینجا پیش می‌آید که زمان های شروع بیماری ها یا به طور کلی زمان های تجربه پیشامد مورد بررسی ما توسط افراد، با هم متفاوت است و نیز اینکه عده بسیاری قبل از ورود به آسایشگاه یا بیمارستان دچار مرگ شده یا پیشامد پای ان دهنده مطلوب ما که می تواند مرگ یا بهبود بیماری باشد را تجربه کرده‌اند .

این مشکلات و بسیاری دیگر ، اهمیت تحلیل و کار بر روی اینگونه داده ها را روشن می کند و باعث می‌شود که روشی خاص برای جمع‌آوری و تحلیل اینگونه داده ها به کار بریم .

یک روش منطقی در جمع‌آوری اینگونه داده، تقسیم بندی جامعه به سه قسمت می‌باشد.

۱_ جامعه هدف : این جامعه شامل تمام افرادی ا ست که ویژگی مورد نظر ما را دارا می باشند و هدف تحقیق و تعمیم نتایج به این جامعه است .

۲- جامعه نمونه گیری : چنانچه نمونه گیری از جامعه هدف ممکن نباشد، از جامعه ای که قابل نمونه گیری است (و مسلماً کوچکتر از جامعه هدف می باشد) نمونه گیری می کنیم. این جامعه قابل دسترسی برای نمونه گیری را جامعه نمونه گیری گویند .

۳- نمو نه : که از جامعه نمونه گیری به صورت زیر انتخاب می شود .

از جامعه نمونه گیری ، به صورت مقطعی در زمان کوتاهی که هیچ فرد جدیدی در آن زمان وارد جامعه نمونه گیری نشود، نمونه می گیریم. لازم به ذکر است که این روش ، تحت عنوان روش خاصی در آمار مطرح نمی شود و تنها یک راه منطقی در جمع آوری داده‌ها است .

بدیهی است فقط افرادی وارد نمونه می شوند که تا زمان نمونه گیری زنده مانده اند و کسانی که قبلا فوت کرده یا بهبود یافته اند وارد نمونه نمی شوند. (دچار برش چپ می شوند) به طور شهودی، تا اینجا دیده می شود شانس افرادی که طول عمر بیشتری دارند، برای ورود به نمونه بیشتر است و این باعث اریب شدن نتایج حاصل از تحلیل اینگونه داده ها می باشد .

اریبی که از آن یاد شد ، تحت عنوان در طول اریبی [1] در مقالات آماری بیان می شود و کمتر مورد توجه قرار می‌گیرد، حال آنکه عدم در نظر گرفتن آن باعث تغییر بسیار زیادی در نتایج می‌شود و بیشتر منجر به یک بیش برآوردی می گردد به عنوان مثال، ولفسن و همکاران (۲۰۰۱) نشان داده اند که برآورد واقعی میانه زمان بقا در یک سری داده عملی (با در نظر گرفتن در طول اریبی) ۳.۳ می باشد حال آنکه بدون در نظر گرفتن در طول اریبی این مقدار ۶.۶ برآورد شده بود. این مطلب، اهمیت در نظر گرفتن در طول اریبی را می رساند .

در زمینه تحلیل این نوع داده ها ، افرادی مانند واردی(۱۹۸۲)، واردی (۱۹۸۵) و گیل، واردی و ولنر (۱۹۸۸) و واردی (۱۹۸۹) در برآورد نا پارامتری تابع بقا با داده های در طول اریب به بررسی خواص حدی آن پرداخته اند .

همچنین پاتیل و رائو (۱۹۸۷) و پاتیل، رائو و زلن (۱۹۸۸) حالات کلی تری از اریبی را در نظر گرفته اند .

اصغریان و ولفسن (۲۰۰۱) و اصغریان و ولفسن (۲۰۰۵) به خواص مجانبی برآوردگر ناپارامتری تابع بقای داده های از راست سانسور شده و در طول اریب و نااریب پرداخته اند.

همینطور، ونگ و همکاران (۱۹۹۳)، آلیوم و کامن جس (۱۹۹۶)، نان و رایان (۱۹۸۹) و گیل (۱۹۸۱) متغیر های کمکی در مدل خود وارد کرده اند.

کریستوبال و آلکلا (۲۰۰۰) و کریستوبال و همکاران (۲۰۰۴) نیز روشهایی ناپارامتری برای رگرسیون داده های در طول اریب ارایه کرده اند.

---

[1] Length bias

اما موضوعی که تا کنون کمتر مورد بررسی قرار گرفته است تأثیر در طول اریبی بر متغیر های کمکی با وجود سانسور راست و برش چپ می باشد. که هدف اصلی این پایان نامه می باشد.

این پایان نامه مشتمل بر ۴ فصل می باشد.

در فصل اول تعاریف اولیه مورد نیاز را ارایه می کنیم.

در فصل دوم به مفهوم در طول اریبی و چگونگی ورود آن به مدل را مورد بررسی قرار می دهیم . در فصل سوم چگونگی ورود متغیرهای کمکی به مدل های گوناگون داده های بقا را بیان می کنیم. همچنین نحوه تأثیر در طول اریبی بر متغیر های کمکی مطرح می شود.

در فصل چهارم نحوه بدست آوردن برآوردگر های تابع بقا در حالت پارامتری با معرفی توابع درستنمایی شرطی بر روی متغیر های کمکی و توأم با آنها را شرح می دهیم و چند خاصیت مجانبی توابع درستنمایی و برآوردگرهای آنها بیان می شود کارایی دو دسته برآوردگر را در راستای یک تحقیق عملی مورد بررسی قرار می گیرد.

فصل اول

تعاریف و مقدمات

۱-۱ مقدمه ای در مفاهیم بقا

در این بخش، تعاریف و مقدماتی در مدل بندی داده های بقا ارائه می‌کنیم که در فصل‌های آتی به آنها احتیاج خواهیم داشت.

فرض کنید X زمان تا پیشامد مشخصی باشد. این پیشامد می تواند مرگ، بروز تومور، پیشرفت یک بیماری در یک بیمار، پایان مهمات در جنگ و یا بسیاری از اتفاقات دیگری که برای یک محقق مهم است، باشد. بعلاوه این اتفاق می‌تواند اتفاقی مثبت مانند درمان یک بیمار یا ترک سیگار نیز باشد. به طور دقیق تر در این فصل X یک متغیر تصادفی نامنفی از یک جامعه همگن می‌باشد. چهار تابع می توانند بیان کننده نوع توزیع X باشند. مهم‌تر از همه تابع بقا می‌باشد که احتمال بقای یک فرد تا زمان x است، تابع نرخ خطر یا تابع خطر که شانس این است که یک فرد با سن x در یک لحظه کوتاه بعدی پیشامد مورد بررسی را تجربه کند. تابع چگالی احتمال (یا تابع جرم احتمال) در x که احتمال اتفاق افتادن پیشامد مورد بررسی در زمان x می‌باشد. میانگین باقیمانده طول عمر در زمان x، که میانگین زمان تا پیشامد مورد بررسی می‌باشد به شرط اینکه پیشامد تا زمان x اتفاق نیافتاده باشد.

اگر هر کدام از این چهار تابع برای ما معلوم باشد، بقیه آنها را به طور یکتا می توان بدست آورد. در عمل این چهار تابع به همراه کمیت دیگری به نام تابع خطر تجمعی، برای بیان مشخصات مختلف توزیع X به کار می-روند.

۱-۱-۱ تابع بقا

کمیت اصلی که برای بررسی زمان تا پیشامد به کار می رود تابع بقا می باشد. احتمال اینکه یک فرد (عنصر) بیش از زمان x، عمر کند و باقی بماند را تابع بقا نامند و به صورت زیر بدست می‌آید :

$$S(x) = \Pr(X > x)$$

در مواردی که زمان تا نقص فنی قطعات تولیدی یا پایان ملزومات یا مهمات انجام کاری، مورد بررسی قرار می‌گیرد، $S(x)$ را به عنوان یک تابع اطمینان در نظر می‌گیرند. اگر X یک متغیر تصادفی پیوسته باشد، آنگاه $S(x)$ نیز یک تابع پیوسته و اکیداً نزولی است.

تابع بقا، متمم تابع توزیع تجمعی می باشد. یعنی :

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

$$F(x) = \Pr(X \le x)$$

بعلاوه تابع بقا، از انتگرال گیری تابع چگالی احتمال $f(x)$ نیز بدست می آید. یعنی :

$$S(x) = \Pr(X > x) = \int_x^{\infty} f(t)dt$$

$$f(x) = -\frac{dS(x)}{dx}$$

نکته اینکه $f(x)dx$ می تواند به عنوان احتمال تقریبی اینکه پیشامد در زمان x اتفاق بیفتد، در نظر گرفته شود.

نمودار تابع $S(x)$، اشکال مختلفی را می‌تواند به خود بگیرد اما نکته مهم اینجا است که همه این اشکال، خواص اساسی یکسانی دارند. همه آنها یکنوا و ناصعودی می‌باشند و نیز در نقطه صفر برابر با یک بوده و با پیشرفت زمان به بی نهایت، مقدار تابع به صفر می‌رود.

اگر X یک متغیر تصادفی گسسته باشد، موضوع به کلی فرق کرده و از فنون متفاوتی برای رفتار با متغیر های تصادفی گسسته استفاده می‌شود. متغیر های تصادفی گسسته معمولاً در مواردی مانند وقتی که داده‌های حاصل از یک نمونه گیری را گرد می‌کنیم یا وقتی که زمان های از کار افتادن را گرد می‌کنیم یا هنگامی که زمان های بقا از اعداد صحیح مربوط به اجزا، بدست می‌آیند، نمایان می‌شوند.

فرض کنید X یک متغیر تصادفی گسسته با تابع جرم احتمال زیر باشد :

$$P(x_j) = \Pr(X = x_j) \quad j = 1, 2, \ldots$$

تابع بقا برای این متغیر تصادفی گسسته بدین صورت بدست می‌آید :

$$S(x) = \Pr(X > x) = \sum_{x_j > x} p(x_j)$$

۱-۱-۲ تابع خطر

یک کمیت اصلی و زیر بنایی در تحلیل بقا، تابع خطر می‌باشد. این تابع به نرخ شکست شرطی نیز در مباحث قابلیت اعتماد، معروف می‌باشد و یا در جمعیت شناسی به شدت میرایی و در فر آیندهای تصادفی به تابع شدت و در علم همه گیرشناسی به نرخ شکست و معکوس نسبت میل در اقتصاد معروف می‌باشد. همه اینها در حقیقت همان نرخ شکست می‌باشد که به صورت زیر تعریف می‌شود :

$$h(x) = \lim_{\Delta x \to 0} \frac{\Pr(x \le X < x + \Delta x \mid X \ge x)}{\Delta x}$$

اگر X یک متغیر تصادفی پیوسته باشد، در این صورت :

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d \ln[S(x)]}{dx}$$

یک کمیت وابسته به نرخ خطر، تابع خطر تجمعی می‌باشد که به صورت زیر تعریف می شود :

$$H(x)=\int_0^x h(u)du=-\ln[S(x)]$$

پس برای طول عمرهای پیوسته داریم :

$$S(x)=\exp[-H(x)]=\exp[-\int_0^x h(u)du]$$

از تعریف تابع نرخ خطر، می‌توان مشاهده کرد که $h(x)\Delta x$، احتمال تقریبی این است که یک فرد با سن x پیشامد مورد بررسی را در لحظه ای بعد، تجربه کند. این تابع به خصوص در مشخص کردن توزیع شکست مناسب با استفاده از اطلاعات کیفی در مورد نوع شکست یا مرگ مورد استفاده قرار می گیرد.

بر خلاف تابع بقا، برای تابع نرخ خطر اشکال مختلفی ممکن است بدست آید. نرخ خطر می تواند صعودی، نزولی و یا حتی ثابت ویا غیر یکنوا باشد و نیز می تواند مشخصات دیگری را که نحوه مرگ یا شکست را بیان می کنند، به خود بگیرد.

مدل هایی با نرخ خطر صعودی، ممکن است بیان کننده این باشد که با یک سیربقای طبیعی سر و کار داریم. مانند وقتی پیشامد مورد بررسی ما بررسی مرگ در اثرپیری در بین جوامع انسانی باشد.

توابع خطر نزولی، در عمل کمتر پیش می آیند و ممکن است موقعی پیش آید که احتمال شکست یا مرگ یا به طور کلی احتمال اتفاق پیشامد مورد بررسی، بسیار زیاد باشد . به عنوان مثال در بعضی لوازم الکتریکی بسیار حساس که با کوچکترین نوسانی در برق می سوزد یا در بعضی موارد از عمل جراحی پیوند اعضا در بیماران، کهاحتمال عفونت بعد از عمل زیاد باشد، پیش می آید. تابع نرخ خطر ثابت در مواردی مشابه بعضی لوازم تولیدی که ممکن است در اثر نقص در بعضی از قسمت هایشان ، به پایان عمر خود برسند، پیش می‌آید.

در انتها نرخ خطری که زود صعود کند و یک باره شروع به نزول کند ، یعنی جهت تقعر آن همیشه رو به پایین باشد، بیشتر در مواردی مانند مدل بندی داده های بقا مربوط به اعمال جراحی که بعد از مدتی از عمل ، فرد بیمار دچار ریسک زیادی می شود که می تواند در اثر عفونت ، یا یک عمل جراحی دیگر باشد، به وجود می‌آید.

مثال ۱-۱-۱ توزیعی که به اندازه کافی انعطاف پذیر است و می‌تواند دارای نرخ خطر صعودی ، نزولی یا ثابت باشد، توزیع وایبل می‌باشد. توزیع وایبل زیر را در نظر بگیرید :

$$f_X(x)=\lambda\alpha x^{\alpha-1}e^{-\lambda x^{\alpha}} \qquad \alpha>0,\lambda>0,x>0$$

در این صورت تابع بقا بدین صورت می‌شود :

$$S(x) = \exp(-\lambda x^{\alpha})$$

و نیز تابع نرخ خطر نیز بدین صورت می‌شود :

$$h(x) = \alpha\lambda x^{\alpha-1}$$

مشاهده می شود که تابع برای مقادیر مختلف پارامتره‌ایش، بسیار انعطاف پذیر می‌باشد. □

وقتی X یک متغیر تصادفی گسسته باشد، تابع نرخ خطر به صورت زیر تعریف می‌شود :

$$P(x_j) = \Pr(X = x_j) \quad j = 1,2,...$$

$$h(x_j) = \Pr(X = x_j \mid X > x_j) = \frac{p(x_j)}{S(x_{j-1})} \qquad j = 1,2,...$$

که در آن، $S(x_o) = 1$.

رابطه دیگری برای تابع نرخ خطر متغیرهای تصادفی گسسته بدین صورت بدست می‌آید :

$$h(x_j) = 1 - \frac{S(x_j)}{S(x_{j-1})} \qquad j = 1,2,...$$

نکته اینکه تابع بقای متغیر تصادفی گسسته را می توان به صورت ضرب احتمالات بقای شرطی نیز نوشت :

$$S(x) = \prod_{x_j \le x} \frac{S(x_j)}{S(x_{j-1})}$$

پس تابع بقای مربوط به متغیر تصادفی گسسته را می‌توان بدین صورت نیز نوشت :

$$S(x) = \prod_{x_j \le x} [1 - h(x_j)]$$

۱-۱-۳ تابع میانگین باقیمانده عمر و میانه طول عمر

چهارمین کمیت مهم و اصلی در تحلیل داده های بقا، تابع میانگین باقیمانده عمر در زمان x است. برای هر فرد با طول عمر x، این تابع میزان انتظار از باقیمانده طول عمر او را اندازه گیری می کند.

$$mrl(x) = E(X - x \mid X > x)$$

به راحتی می توان بدست آورد که میانگین باقیمانده طول عمر، ناحیه زیر منحنی بقا در سمت راست نقطه x تقسیم بر $S(x)$ می‌باشد.

در این راستا، میانگین طول عمر را با نماد μ نشان داده و برابر است با :

$$\mu = mrl(0)$$

که این مقدار تمام ناحیه زیر منحنی بقا می باشد.

برای یک متغیر تصادفی پیوسته داریم :

$$mrl(x) = \frac{\int_x^\infty (t-x) f(t)dt}{S(x)} = \frac{\int_x^\infty S(t)dt}{S(x)}$$

$$\mu = \mathrm{E}(X) = \int_0^\infty t f(t)dt = \int_0^\infty S(t)dt$$

و همین طور واریانس X بدین صورت بدست می‌آید :

$$Var(X) = 2\int_0^\infty tS(t)dt - [\int_0^\infty S(t)dt]^2$$

و در همین راستا ، چندک مرتبه p ام توزیع X، کوچکترین x_p می‌باشد که در رابطه زیر صدق کند.

$$S(x_p) \le 1-p$$

یعنی :

$$x_p = \inf\{t : S(t) \le 1-p\}$$

اگر X یک متغیر تصادفی پیوسته باشد، چندک مرتبه p ام از حل معادله زیر بدست می‌آید :

$$S(x_p) = 1-p$$

و میانه طول عمر ، در حقیقت همان چندک مرتبه ۰.۵ ام می باشد. یعنی میانه طول عمر یک متغیر تصادفی پیوسته باید در معادله زیر صدق کند :

$$S(x_{0.5}) = 0.5$$

۱-۲ مدل های پارامتری رایج برای داده‌های بقا

با اینکه روشهای ناپارامتری و نیمه پارامتری در تحلیل داده های بقا بسیار پرکاربرد هستند، مدل های پارامتری نیز در خور توجه می باشند.در این بخش نگاهی به مدل ها و توزیع هایی که بیشتر برای داده های بقا مورد استفاده می‌گیرند و دلایل استفاده از آنها، می‌اندازیم. این مدل‌ها نه به خاطر محبوبیت و شهرت در بین محققان بلکه به دلیل خواصی که به توابع مربوط به بقا و نیز پارامترهای مختلف جامعه، القا می کنند، مورد استفاده زیادی دارند. که مهمترین این توابع تابع نرخ خطر می باشد. بعضی از مدل های مهمی که بیشتر مورد

استفاده قرار می گیرند عبارتند از توزیع های نمایی، وایبل، گاما، لگ نرمال، لگ لجستیک، نرمال، توانی نمایی، گمپرتز، گوسی معکوس، پارتو و گامای تعمیم یافته.

ابتدا به دلیل سادگی کار توزیع نمایی را مورد بررسی قرار می‌دهیم. تابع بقای توزیع نمایی را مورد بررسی قرار می دهیم.

۱-۲-۱ توزیع نمایی

تابع چگالی و نرخ خطر آن بدین صورت می‌باشند :

$$f_X(x) = \lambda \exp(-\lambda x) \quad \lambda, x > 0$$

$$h(x) = \lambda$$

مهمترین خاصیت توزیع نمایی، خاصیت بی حافظه بودن آن می‌باشد. یعنی :

$$\Pr(X \geq x + z \mid X \geq x) = \Pr(X \geq z)$$

این خاصیت در بسیاری از موارد خاصیت مثبت و مطلوبی به حساب می آید حال آنکه در بسیا ری از موارد عملی، کاربرد ندارد. مانند طول عمر قطعات صنعتی یا حتی طول عمر انسان‌ها. به دلیل خاصیت بی حافظه بودن توزیع نمایی، میانگین باقی مانده عمر در این توزیع مقداری ثابت می شود:

$$E(X - x \mid X > x) = E(X) = \frac{1}{\lambda}$$

این اتفاق به این دلیل است که در توزیع نمایی ، زمان تا اتفاق افتادن یک پیشامد ، به طول عمر قبل از این زمان بستگی ندارد . این خاصیت توزیع نمایی ، در تابع نرخ خطر آن نیز مشهود می باشد. در اینجا احتمال شرطی شکست در هر زمان ، به شرط اینکه پیشامد تا زمان t اتفاق نیفتاده باشد ، به زمان t بستگی ندارد. با اینکه توزیع نمایی توزیع محبوبی می باشد، نرخ خطر ثابت این توزیع ، محدودیت بسیاری را در کارهای عملی در صنعت و بهداشت تحمیل می‌کند.

توزیع نمایی حالت خاصی از دو توزیع گاما و وایبل می‌باشد.

۱-۲-۲ توزیع وایبل

توزیع وایبل اولین بار توسط وایبل در سال ۱۹۳۹ معرفی شد. وی این توزیع را برای طول عمر مواد شیمیایی به کار می برده است. در توزیع وایبل با شکل زیر ، $\lambda > 0$ یک پارامتر مقیاس و $\alpha > 0$ یک پارامتر شکل می‌باشد.

$$f_X(x) = \alpha\lambda x^{\alpha-1} \exp(-\lambda x^{\alpha}) \quad \alpha, \lambda, x > 0$$

یکی از خواص مهم توزیع وایبل، تابع نرخ خطر انعطاف پذیر آن می‌باشد.

$$h(x) = \lambda\alpha x^{\alpha-1}$$

همانطور که دیده می شود این تابع می تواند صعودی $(\alpha > 1)$، نزولی $(\alpha < 1)$ و یا ثابت $(\alpha = 1)$ باشد. این خاصیت تابع نرخ خطر در تابع بقا نیز دیده می‌شود.

$$S(x) = \exp(-\lambda x^{\alpha})$$

گاهی اوقات به جای استفاده از خود طول عمرها، از لگاریتم این طول عمرها استفاده می‌شود. این کار در بعضی موارد به دلیل سادگی کار انجام می‌شود. اگر قرار دهیم $Y = \ln(X)$ که X یک متغیر تصادفی وایبل است، در این صورت Y دارای تابع چگالی زیر می شود:

$$f_Y(y) = \alpha \exp\left\{\alpha\left[y - \left(\frac{-\ln(\lambda)}{\alpha}\right)\right] - \exp\left\{\alpha\left[y - \left(\frac{-\ln(\lambda)}{\alpha}\right)\right]\right\}\right\} \quad , -\infty < y < \infty$$

۱-۲-۳ توزیع لگ نرمال

متغیر تصادفی X را دارای توزیع لگ نرمال گوییم هرگاه لگاریتم طبیعی آن $Y = \ln(X)$ دارای توزیع نرمال باشد. برای داده های زمان تا یک پیشامد مطلوب ، این توزیع به دلیل رابطه ای که با توزیع نرمال دارد، مورد استفاده زیاد قرار می گیرد. دلیل دیگر کاربرد ز یاد این توزیع این است که بسیاری از محققین ، مشاهده نموده اند که این توزیع زمان های بقا یا طول عمر در زمان ش روع چند بیماری مشخص را تقریب می - زند.

$$f_X(x) = \frac{\exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)\right]}{x(2\pi)^{1/2}\sigma} \quad \mu, \sigma, x > 0$$

همانند توزیع نرمال، توزیع لگ نرمال هم کاملاً با دو پارامتر μ, σ مشخص می شود.

$$S(x) = 1 - \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$$

$$h(x) = \frac{f_X(x)}{S(x)}$$

که Φ تابع توزیع چگالی نرمال می‌باشد.

مقدار تابع نرخ خطر در صفر، صفر می باشد و در ابتدا صعود می کند و بعد با میل کردن x به بینهایت، به صفر نزول می کند. به دلیل اینکه تابع نرخ خطر این توزیع برای طول عمر های زیاد کوچک می شود، این مدل در بسیاری از موارد نامناسب عمل می کند. چرا که این خاصیت تابع نرخ خطر ، در عمل خیلی کم اتفاق می‌افتد. این مدل بیشتر در مواردی مورد استفاده دارد که مقادیر بسیار بزرگ X از اهمیت کمی برخور دار می‌باشند.

۱-۲-٤ توزیع لگ لجستیک

متغیر تصادفی X، از توزیع لگ لجستیک تبعیت می کند، هرگاه لگاریتم طبیعی آن $Y = \ln(X)$ دارای توزیع لجستیک باشد. این توزیع شباهت بسیاری به توزیع نرمال دارد، با این تفاوت که تابع بقای منعطف تری دارد.

$$f_X(x) = \frac{\alpha x^{\alpha-1}\lambda}{\left[1+\lambda x^{\alpha}\right]^2} \quad \lambda, \alpha > 0, x \geq 0$$

$$f_Y(y) = \frac{\exp\left(\frac{y-\mu}{\sigma}\right)}{\sigma\left[1+\exp\left(\frac{y-\mu}{\sigma}\right)\right]^2} \quad -\infty < y < \infty, \mu, \sigma > 0$$

$$h(x) = \frac{\alpha x^{\alpha-1}\lambda}{1+\lambda x^{\alpha}}$$

$$S(x) = \frac{1}{1+\lambda x^{\alpha}}$$

صورت کسر تابع نرخ خطر توزیع لگ لجستیک شبیه تابع خطر وایبل اس ت و مخرج کسر آن تابع، باعث می‌شود که تابع دارای ا ین مشخصات شود که ، برای $\alpha \leq 1$ نزولی و برای $\alpha > 1$ تابع نرخ خطر تا مقدار ماکسیمم اش در زمان $\left(\frac{\alpha-1}{\lambda}\right)^{1/\alpha}$ صعود کند و بعد به صورت نزولی هنگامی که زمان به بینهایت میل می کند، به سمت صفر رود.

این توزیع شبیه به مدل های نمایی و وایبل می باشد که این شباهت در تابع بقا و تابع نرخ خ طر این توزیع ها، مشاهده می‌شود.

۱-۲-۵ توزیع گاما

توزیع گاما دارای مشخصاتی شبیه به توزیع وایبل می باشد با اینکه به اندازه توزیع وایبل انعطاف پذیر نمی باشد. این توزیع همانند وایبل، تعمیمی از توزیع نمایی می باشد. اگر X دارای توزیع گاما باشد، دارای تابع چگالی احتمال زیر می‌باشد :

$$f_X(x) = \frac{\lambda^\beta x^{\beta-1} \exp(-\lambda x)}{\Gamma(\beta)} \quad \beta, \lambda > 0, x \geq 0$$

وقتی β، پارامتر شکل توزیع گاما به سمت بینهایت میل کند، این توزیع به توزیع نرمال میل می‌کند.

$$h(x) = \frac{f_X(x)}{S(x)} = \frac{\frac{\lambda^\beta x^{\beta-1} \exp(-\lambda x)}{\Gamma(\beta)}}{\int_0^{\lambda x} \frac{u\beta - 1 \exp(-u)}{\Gamma(\beta)} du}$$

این تابع نرخ خطر برای $\beta > 1$ صعودی می‌باشد با :

$$h(0) = 0$$
$$h(x) \xrightarrow[x \to \infty]{} \lambda$$

و نیز برای $\beta < 1$ نزولی می‌باشد با :

$$h(0) = \infty$$
$$h(x) \xrightarrow[x \to \infty]{} \lambda$$

وقتی $\beta < 1$ باشد، نما ی تابع خطر در $x = \frac{\beta - 1}{\lambda}$ اتفاق می افتد.

۱-۳ سانسور و برش

داده‌های مربوط به زمان تا رخ دادن یک پیشامد، به راه ه ا و انواع گوناگونی نمود پیدا می کنند که باعث مشکلی کار با آنها می‌شود. یکی از مشخصه‌های این نوع داده‌ها که بعضی اوقات اتفاق می افتد، به سانسور معروف می باشد. که به طور کلی زمانی اتفاق می افتد که می دانیم بعضی از طول عمر ها، فقط در بازه‌های مشخصی اتفاق افتاده اند. بقیه طول عمرها را به طور کامل می بینیم. انواع مختلفی از سانسور را داریم ، مانند سانسور راست، سانسور چپ و سانسور فاصله‌ای.

برای بررسی جامع پدیده سانسور در تحلیل داده های بقا، باید طرح و روشی را که برای بدست آوردن داده‌ها به کار می بریم در نظر بگیریم.

مشخصه دیگری که در مطالعات بقا بسیار پیش می آید، و با سانسور اشتباه گرفته می شود، برش می-باشد. برش چپ زمانی اتفاق می افتد که افراد در یک سن خاص وارد مطالعه ما می شوند و نه الزاماً در زمان شروع طول عمرشان یا لحظه تجربه پیشامد آغازین (مانند شروع بیماری)، و از این به اصطلاح زمان تأخیر، تا