
بسم الله الرحمن الرحيم

دانشگاه تهران

دانشکده ریاضی، آمار و علوم کامپیوتر

آزمون نیکویی برازش کای– دو برای داده‌های به تصادف سانسور شده

نگارش: لیلا آذرنگ

استادان راهنما: دکتر حمید پزشک و دکتر فیروزه حقیقی

پایان نامه برای دریافت درجه کارشناسی ارشد
در
آمار ریاضی

دی ۱۳۸۷

University of Tehran

College of Science

Chi- square goodness- of- fit test for randomly censored data

By: Leila Azarang

Under Supervision of

Dr. Hamid Pezeshk and Dr. Firoozeh Haghighi

A thesis submitted to the Graduate Studies Office
in partial fulfillment of the requirements for
the degree of M.Sc. in

Mathematical Statistics

January 2009

قدردانی

خداوند مهربان را سپاس می‌گویم که مرا با مهر خود آفرید و نعمت آموختن به من عطا فرمود و معرفت خود را به من الهام کرد تا محبت او را در دل داشته باشم.

از پدر و مادر مهربانم که در تمام سالهای زندگی‌ام دلسوز و پشتیبان من بودند متشکرم. همچنین از اساتید بزرگوارم، جناب آقای دکتر حمید پزشک و خانم دکتر فیروزه حقیقی که در تمامی مراحل انجام این پایان نامه نهایت همکاری را با بنده داشتند کمال تشکر را دارم و از درگاه خداوند متعال تندرستی و موفقیت ایشان را خواستارم.

در پایان از جناب آقای دکتر احمد پارسیان که در طول این دوره دلسوزانه مرا راهنمایی کردند تشکر و قدردانی می‌کنم.

لیلا آذرنگ

دی ۱۳۸۷

تقدیم به

پدر و مادر عزیزم

چکیده

در آنالیز بقا اغلب، داده‌ها کامل نیستند. یعنی ممکن است زمان مرگ یا از کار افتادگی واحد مورد بررسی خارج از دوره زمانی مورد مطالعه باشد. در این صورت تعداد دقیق زمانهای مرگ یا از کار افتادگی که در یک بازه می‌افتند معلوم نیست. بنابراین آزمون نیکویی برازشی که برای داده‌های کامل به کار می‌رود در حضور سانسور قابل استفاده نمی‌باشد.

در این پایان نامه آزمون نیکویی برازش با استفاده از آزمونهای پیشنهادی در حضور سانسور تصادفی مورد بررسی قرار گرفته است. با برازش توزیعی جدید برای داده‌های سانسور شده از نوع تصادفی با نرخ مخاطره تک مدی این آزمونها را به کار برده‌ایم و نشان داده‌ایم که توزیع مذکور توزیع مناسبی برای برازش به این داده‌ها می‌باشد.

پیشگفتار

مشهورترین آزمون نیکویی برازش، آزمون کای– دو است. این آزمون به وسیله آماره‌ای انجام می گیرد که دارای توزیع حدی کای– دو است. چنین آزمونی را پیرسن، آماردان انگلیسی در سال ۱۹۰۰ پیشنهاد کرد. در حضور انواع سانسور آزمون نیکویی برازش پیچیده می شود.

برای سانسور نوع دوم میهالکو[۱] و مور[۲] در سال ۱۹۸۰ آزمونی را برای برازش داده های سانسور شده نوع دوم پیشنهاد کردند که آماره این آزمون دارای توزیع حدی کای– دو است.

حبیب[۳] و توماس[۴] در سال ۱۹۸٦ بر پایه فرآیند گوسی آزمونی را برای داده های سانسور شده تصادفی ارائه کردند.

سپس آکریتاس[۵] در سال ۱۹۸۸ آماره کای– دو ای را برای داده‌های سانسور تصادفی معرفی کرد که این آماره بر پایه مشاهدات سانسور نشده در هر بازه است.

---

[۱] Mihalko

[۲] Moor

[۳] Habib

[۴] Thomas

[۵] Akritas

نیکولین[۱] و سولو[۲] در سال ۱۹۹۹ آمارهٔ کای– دوای را بر اساس فرآیند گوسی برای داده‌های سانسور شده دو طرفه پیشنهاد کردند.

ژانگ[۳] در سال ۱۹۹۹ آماره معرفی شده توسط نیکولین ، رائو[۴]، رابسن[۵] و مور راتوسعه داد.

هدف از انجام این پایان نامه آزمون نیکویی برازش در حضور سانسور تصادفی می‌باشد. برای این منظور از آماره‌های معرفی شده توسط حبیب و توماس و آکریتاس استفاده کرده‌ایم.

فصل اول به تعاریف و مفاهیمی از آنالیز بقا و همچنین مفاهیمی از نظریه احتمال که در پایان نامه استفاده شده است پرداخته است. در این فصل از میان انواع سانسور سانسور از راست را مورد بررسی قرار داده‌ایم و به بیان قضایایی که در فصل دوم مورد استفاده قرار گرفته است، پرداخته ایم. در فصل دوم آزمون نیکویی برازش برای داده‌های سانسور شده که بر پایه مقاله های [۱] و [۵] می‌باشد به طور کامل تشریح شده است.

در فصل سوم با اشاره به یک توزیع جدید به نام وایبل توانی تعمیم یافته و برازش این توزیع به داده‌های سرطان سر و گردن از آزمونهای پیشنهاد شده توسط حبیب و توماس استفاده شده است. همچنین اصلاحاتی در آماره معرفی شده توسط حبیب و توماس برای به کار بردن این آماره برای داده‌هایی که بزرگترین مشاهده آنها سانسور شده می باشد، انجام شده است.

در فصل چهارم یک بررسی بیزی با استفاده از شبیه سازی MCMC روی خانواده وایبل توانی تعمیم یافته انجام شده است. در این فصل نشان داده‌ایم که مقدار برآورد پارامترها به روش بیزی به مقادیر واقعی نزدیک است.

---

[۱] Nikulin

[۲] Solev

[۳] Zhang

[۴] Rao

[۵] Rabson

فهرست مندرجات

لیست اشکال

فصل ۱

پیشنیازها

در این فصل به بیان پیشنیازهایی از قابلیت اعتماد و آنالیز بقا می‌پردازیم در ابتدا مفاهیم آنالیز بقا را ارائه می‌دهیم سپس به مفاهیمی از احتمال می‌پردازیم و در پایان مفاهیمی از آزمون نیکویی برازش بیان خواهد شد. در فصلهای آتی تعمیم آزمون نیکویی برازش را برای داده‌های سانسور شده به کار خواهیم برد.

۱.۱ مفاهیمی از آنالیز بقا

آنالیز آماری که بر داده‌های زمان زندگی، طول عمر و یا زمان شکست اشاره دارد در بسیاری از زمینه‌ها مانند پزشکی، مهندسی و علوم اجتماعی یک موضوع مهم برای مطالعه و تحقیق می‌باشد. آنالیز بقا از تحقیق روی اقلام تولید شده توسط یک خط تولید تا مطالعه بیماریهای بشر و درمان آن مورد استفاده قرار می‌گیرد.

بسیاری از روشهای مرتبط با داده‌های طول عمر کاملا قدیمی هستند. اما تقریبا از سال ۱۹۷۰

تئوری و کارهای کاربردی مرتبط با طول عمر به سرعت گسترش یافت و بسته های نرم افزاری برای آنالیز داده‌های طول عمر از سال ۱۹٦۸ به طور گسترده در دسترس قرار گرفت.

۱.۱.۱ مفهوم طول عمر

در زمینه‌های مختلف آنالیز بقا، اصطلاح طول عمر(که به آن مدت زندگی یا زمان شکست نیز می‌گویند) معانی متفاوتی دارد که در ادامه با چند مثال این مفهوم را توضیح خواهیم داد.

مثال ۱.۱.۱ در آزمایش اقلام تولید شده کارخانه‌ای برطبق اطلاعات در مورد دوام این اقلام آنها را در آزمایشگاهی در معرض آزمایش قرار می‌دهند و تا زمان خرابی تحت نظر می‌گیرند. در اینجا مدت زندگی بر اساس زمان خراب شدن محصول بیان می‌شود. زمانی که یک محصول آنگونه که مد نظر ماست کار نکند می‌گوییم آن محصول خراب شده است.

مثال ۲.۱.۱ آمار گیران جمعیت و متخصصان علوم اجتماعی علاقه‌مند به بررسی طول عمر برخی وقایع از وضعیت زندگی بشر می‌باشند. به عنوان مثال ازدواج افراد یک جامعه را در نظر بگیرید ازدوجهایی که در طول سال ۱۹۸۰ در یک کشور خاصی صورت گرفته است. در اینجا طول عمر یک ازدواج ممکن است مدت دوام آن باشد. ممکن است یک ازدواج به دلیل طلاق یا مرگ به پایان برسد.

مثال ۳.۱.۱ در یک آزمایش می‌خواهند تاثیر یک دارو را که موجب پیدایش غده‌ای در بدن می

شود بر روی موجود زنده‌ای مورد بررسی قرار دهند. در اینجا زمان پیدایش غده یک متغیر تصادفی است و منظور از طول عمر فاصله بین زمان استفاده از دارو تا زمان پیدایش غده می‌باشد.

مثال ۴.۱.۱ در مطالعات پزشکی مرتبط با بیماری‌های وخیم و کشنده طول عمر افراد بیمار مورد مطالعه قرار می‌گیرد که از تاریخ تشخیص بیماری یا مراجعه افراد یا بروز بیماری یا شروع درمان محاسبه می‌شود. به عنوان مثال برای درمان یک بیماری، توزیع طول عمر بیماران مختلف با روشهای درمان مختلف را با هم مقایسه می‌کنند. در این مثال طول عمر مدت زمان زندگی شخص از لحظه تشخیص یا بروز بیماری تا زمان مرگ است و منظور از زمان شکست یا مدت زندگی زمان مرگ افراد از مبدا در نظر گرفته شده می‌باشد. در اینجا زمان زندگی را متغیر تصادفی پیوسته در نظر گرفته و با T نشان می دهیم و فرض می کنیم T یک متغیر تصادفی نامنفی است که نشان دهنده زمان زندگی افراد در برخی جوامع می‌باشد. فرض می‌کنیم $f(t)$ و$F(t)$ به ترتیب نشان دهنده تابع چگالی احتمال و تابع توزیع احتمال باشند. با این مفروضات آماده‌ایم تا چند تعریف اساسی در آنالیز بقا و قابلیت اعتماد را بیان کنیم.

تعریف ۱.۱.۱ فرض کنید T یک متغییر تصادفی پیوسته و نامنفی باشد که مدت زندگی اشخاص را نشان می‌دهد. احتمال اینکه شخص بعد از زمان t زنده بماند، با تابع بقا [1] نشان داده می‌شود که به صورت زیرتعریف می شود:

$$S(t) = Pr(T \geq t) = \int_t^{\infty} f(x)\,dx$$

[1] Survival function

به $S(t)$ تابع قابلیت اعتماد نیز گفته می‌شود.

یک مفهوم بسیار مهم در آنالیز بقا و قابلیت اعتماد تابع مخاطره یا نرخ مخاطره[۱] است که به صورت زیر نشان داده می‌شود:

$$h(t) = \frac{f(t)}{S(t)} = lim_{\Delta t \longrightarrow ۰} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

تابع مخاطره نرخ آنی مرگ در آنالیز بقا و نرخ آنی شکست در قابلیت اعتماد را در زمان t با دانستن اینکه شخص بعد از زمان t زنده می‌ماند را مشخص می کند.

برخی نکات در مورد تابع مخاطره:

تابع مخاطره مشخصه مهمی از توزیع زمان زندگی می‌باشد که نرخ شکست را با سن و زمان بیان می‌کند که بیشتر در کاربردهای آنالیز بقا مورد استفاده قرار می‌گیرد. اطلاعات قبلی در مورد تابع مخاطره می‌تواند ما را در انتخاب مدل مناسب راهنمایی کند. شکل‌های تابع مخاطره از لحاظ کیفی کاملامتفاوت می‌باشد و این اشکال به چهار صورت می‌باشند.

۱. ثابت

۲. یکنوا(نزولی یکنوا – صعودی یکنوا)

۳. $\cap$ – شکل

۴. $\cup$ – شکل

به عنوان مثال اگر افراد یک جامعه را در نظر بگیریم در این حالت تابع مخاطره (نرخ آنی مرگ) اغلب به صورت $\cup$ – شکل می‌باشد. زیرا ما با این حقیقت در جوامع بشری آشنا هستیم که بعد از دوره اولیه که مرگ ناشی از تولد بچه‌های ناقص یا بیمار است، نرخ مرگ با افزایش زمان کاهش

[۱] Hazard function

می‌یابد و نسبتا تا سن ۳۰ یا کمی بالاتر ثابت می‌ماند و بعد از آن با افزایش سن افزایش می‌یابد.

داده های طول عمر اغلب با شکل شناخته شده به صورت داده های سانسور شده می‌باشند. یک قسمت عمده آنالیز داده‌های طول عمر مربوط به سانسور می‌باشد. بخش بعدی بر اساس سانسور از راست و فرم تابع درستنمایی تحت این سانسور می‌باشد.

۲.۱ سانسور از راست و تابع درستنمایی

سانسور از راست به دلایل گوناگون رخ می‌دهد که ممکن است از قبل برنامه ریزی شده باشد (به عنوان مثال زمانی که در یک آزمایش از قبل تصمیم گرفته می‌شود که در پایان آزمایش همه افراد یا اقلام مورد آزمایش از بین نروند.) یا ممکن است برنامه ریزی نشده باشد (به عنوان مثال زمانی که یک شخص در جریان آزمایش در ادامه بررسی در دسترس نباشد به دلیل اینکه از منطقه ای که در آنجا تحت نظر بوده نقل مکان کرده و به جای دیگری رفته باشد و برای بررسی به مکان قبلی مراجعه نکرده باشد.) برای بدست آوردن تابع درستنمایی لازم است فرض کنیم که فرآیند بر اساس زمان شکست (طول عمر) و زمان سانسور رخ می‌دهد.

برای انجام چنین کاری به وضوح نیاز به مدل احتمال برای مکانیزم سانسور داریم.

ابتدا برخی نمادها را برای داده‌های سانسور شده بیان می‌کنیم.

فرض می‌کنیم طول عمر (زمان شکست) n شخص توسط متغیرهای تصادفی $T_1, \ldots, T_n$ نشان داده شود برای هر فرد مقدار t_i را مشاهده می کنیم به طوری که t_i مدت زمان زندگی یا مدت زمان سانسور می‌باشد. متغیر تصادفی $\delta_i = I(T_i = t_i)$ برای فرد iام را چنین تعریف می کنیم $\delta_i = I(T_i = t_i)$ برابر ۱ است اگر $T_i = t_i$ باشد و برابر ۰ است اگر $T_i > t_i$ باشد که تابع اخیر تابع

نشانگر سانسور یا تابع وضعیت نامیده می‌شود.

بنابراین داده‌های مشاهده شده به صورت زوج (t_i, δ_i) و $i = ۱, \ldots, n$ می‌باشند. برای مکانیزم سانسور تابع درستنمایی مشاهدات به صورت زیر نوشته می‌شود:

$$L = \prod_{i=۱}^{n} f(t_i)^{\delta_i} S(t_i)^{۱-\delta_i} \qquad (۱.۲.۱)$$

که بسط این تابع برای تعیین تابع درستنمایی انواع دیگر سانسور به کار می‌رود. در ادامه برخی از انواع سانسور از راست را معرفی کرده و فرم تابع درستنمایی را برای آنها نشان می‌دهیم.

۱.۲.۱ سانسور نوع ۱

سانسور نوع ۱ زمانی اتفاق می‌افتد که هر شخص یک زمان سانسور ثابت $C_i > ۰$ داشته باشد. در این صورت T_i مشاهده می‌شود اگر $T_i \leq C_i$ باشد. در غیر این صورت فقط می دانیم که $T_i > C_i$ می‌باشد. سانسور نوع ۱ اغلب زمانی اتفاق می‌افتد که یک مطالعه روی یک دوره زمانی خاص صورت گیرد. در این حالت به ازای تمامی i ها که i=1,...,n است C_i برابر مقدار ثابت C می‌باشد.

در این حالت قرار می‌دهیم:

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

تابع درستنمایی برای نمونه تصادفی سانسور نوع ۱ روی تابع چگالی احتمال (t_i, δ_i) و i=1,...,n پایه گذاری شده است. که t_i و δ_i هر دو متغیر تصادفی می‌باشند. تابع چگالی توام این دو متغیر تصادفی به صورت زیر می‌باشد :

$$h(t_i, \delta_i) = f(t_i)^{\delta_i} Pr(T_i > C_i)^{۱-\delta_i}$$

که C_i ها مقادیر ثابتی هستند و t_i ها مقادیر کمتر یا مساوی C_i ها را می‌گیرد. همچنین

$$Pr(T_i > C_i) = Pr(t_i = C_i, \delta_i = ۰), Pr(t_i, \delta_i = ۱) = f(t_i), \qquad t_i \leq C_i$$

در عبارت اخیر Pr نشان دهنده یک تابع چگالی احتمال یا تابع جرم احتمال بر حسب پیوستگی یا گسستگی تابع توزیع می‌باشد.

فرض می‌کنیم مدت زمانهای زندگی $T_۱, \ldots, T_n$ مستقل از هم باشند در این صورت تابع درستنمایی را از فرمول زیر بدست می‌آوریم:

$$L = \prod_{i=۱}^{n} f(t_i)^{\delta_i} S(t_i^+)^{۱-\delta_i}$$

در حالت کلی $S(t_i^+) = Pr(T_i > t_i)$ می‌باشد. اگر $S(t)$ در t_i پیوسته باشد در این صورت $S(t_i^+) = S(t_i)$ می‌باشد.

در سانسور نوع ۱ زمانهای سانسور C_i زمانهای ثابت خاصی هستند اما در بسیاری از زمینه‌ها زمانهای سانسور واقعا تصادفی هستند. که در ادامه این بخش به برخی از آنها اشاره می‌کنیم.

۲.۲.۱ سانسور نوع ۲

اصطلاح سانسور نوع ۲ منسوب به‌وضعیتی است که فقط r تا از کوچکترین زمان شکست $t_{(۱)} < \ldots < t_{(r)}$ در یک نمونه تصادفی n تایی مشاهده شود. که r یک عدد صحیح خاص بین ۱ و n می‌باشد.

در این نوع سانسور n شخص در زمان خاصی مورد بررسی قرار می‌گیرند تا اینکه r شکست مشاهده شود. آزمایشهایی که با سانسور نوع ۲ تنظیم می‌شوند مشکل ویژه‌ای دارند. زمان کل یعنی $t_{(r)}$ که آزمایش را تمام می‌کند تصادفی است و بنابراین در شروع آزمایش نامعلوم است. از این رو