دانشکده علوم ریاضی

پایان‌نامه جهت اخذ درجه کارشناسی ارشد

در رشته آمار ریاضی

عنوان :
رفتار مجانبی برآوردگر ناپارامتری تابع رگرسیون در مدل سانسور تصادفی با داده‌های α-آمیزنده

استاد راهنما :

جناب آقای دکتر وحید فکور

استاد مشاور :

جناب آقای دکتر حسن دوستی

نگارش :

سمیه باوفا

پاییز ۱۳۹۰

فهرست مطالب

پیش گفتار

وقتی که صحبت از تحلیل رگرسیونی می‌شود، معمولاً منظور برازش یک الگوی ریاضی به داده‌ها به عنوان الگوی وابستگی متغیرها، بررسی نمودار باقیمانده‌ها به عنوان انحرافات از الگو، برآورد و آزمون فرض درباره پارامترها، پیش‌بینی و مانند آن می‌باشد. تحلیل رگرسیونی یکی از ابزارهای آماری است که در سطح بسیار وسیعی کاربرد دارد.

در رگرسیون پارامتری لازم است یک شکل از قبل تعیین شده مانند چندجمله‌ای را برای منحنی رگرسیون فرض کنیم. اما برای بسیاری از مجموعه داده‌ها در مورد شکل ارتباطی بین متغیرها دانش درستی نداریم.

در این مواقع بهتر است به جای تحلیل یک الگوی پارامتری به داده‌ها از فنونی استفاده کنیم که به داده‌ها اجازه می‌دهند رفتار الگوی خود را بروز دهند. رگرسیون ناپارامتری مجموعه‌ای از فنون برای برآورد منحنی رگرسیون است بدون آن که درباره شکل تابعی منحنی رگرسیون فرض‌های قوی بگذاریم.

ریشه رگرسیون ناپارامتری را باید در برآوردگر ناپارامتری تابع چگالی احتمال جستجو کرد.

برآورد ناپارامتری تابع چگالی احتمال، با استفاده از مجموعه‌ی متناهی از مشاهدات هم‌توزیع، یکی از مسائل بنیادی و مهم در استنباط آماری است که نقش بسیار مهمی در تشخیص مدل و الگوی احتمال

جامعه دارد. روزن بلات و پارزن پیشگامان نظریه‌ی برآوردگرهای هسته‌ی چگالی می‌باشند. پس از آن دو، مسأله برآورد تابع چگالی مورد توجه بسیاری از محققان بوده است.

همان طور که گفتیم ریشه رگرسیون ناپارامتری را باید در برآوردگر ناپارامتری تابع چگالی احتمال جستجو کرد. اولین نوشته‌ی رسمی در زمینه‌ی رگرسیون ناپارامتری در سال **۱۹۶۴** منتشر شد. در آن سال نادارایا و واتسون به طور جداگانه با ایده گرفتن از برآوردگر هسته برای چگالی احتمال، برآوردگری را برای منحنی رگرسیون ارائه دادند که بعدها به نام این افراد نام‌گذاری شد. نادارایا در مقاله‌ی خود شرایط کافی برای سازگاری این برآوردگر را بیان کرده و به توزیع مجانبی نرمال آن نیز اشاره نموده است.

از طرفی در اغلب مسائل آمار کاربردی، مشاهدات در دست بررسی، از جامعه‌ای نمونه‌گیری شده‌اند که توزیع آن با توزیع مورد مطالعه‌ی محقق متفاوت است. چنین مشاهداتی، نمونه‌های اریب نامیده می‌شوند، زیرا با استفاده از این داده‌ها در استنباط در مورد توزیع جامعه، اریبی به وجود می‌آورند. داده‌های سانسور از مشهورترین نمونه‌های اریب می‌باشند.

در سانسور، برخی از مشاهدات به طور کامل مشخص نمی‌شوند و فقط این اطلاع در اختیار محقق قرار می‌گیرد که این مشاهدات عضوی از یک مجموعه می‌باشند. به عنوان مثال فرض کنید متغیر X نشان‌دهنده‌ی زمان رخ‌دادن یک پیشامد باشد. برای پیشامدی که تا زمان t رخ نمی‌دهد، مقدار X نامشخص است و فقط این اطلاع در مورد آن وجود دارد که مقدار این متغیر دربازه‌ی (t,∞) قرار دارد. به عبارت دیگر یک مشاهده درمجموعه‌ی A سانسور می‌شود اگر به جای مشاهده‌ی X فقط $X \in A$ مشخص باشد. انتخاب‌های متفاوت از مجموعه‌ی A گونه‌های مختلفی از سانسور را به وجود می‌آورند. یکی از مهمترین انواع سانسور، سانسور تصادفی است.

بنابراین وجود سانسور در مشاهدات را نمی‌توان نادیده گرفت.

بران (۱۹۸۱) یک کلاس از برآوردگرهای ناپارامتری تابع رگرسیون در مدل سانسور شده را معرفی کرد. بعدها داباروسکا (۱۹۸۷)، کاربونز (۱۹۹۵) و کهلر (۲۰۰۲) نیز به معرفی برآوردگرهای مختلف تابع رگرسیون در شرایط سانسور، همراه با بیان ویژگی‌های آن‌ها پرداختند.

بدیهی است که پذیرفتن استقلال بین متغیرهای تصادفی همیشه امکان‌پذیر نیست و در برخی رویدادهای واقعی، نوعی از وابستگی بین متغیرهای تصادفی دیده می‌شود. وابستگی $\alpha-$ آمیزنده، یکی از ضعیف‌ترین وابستگی‌های موجود می‌باشد که تاکنون مورد توجه بسیاری قرار گرفته است. این نوع وابستگی توسط روزن بلات (۱۹۵۶) معرفی شد. وی قضیه‌ی حد مرکزی را برای متغیرهای تصادفی $\alpha-$ آمیزنده ثابت نمود. از آن زمان تاکنون توجه به متغیرهای تصادفی $\alpha-$ آمیزنده، از اهمیت بسیاری برخوردار بوده است. کاربرد این متغیرها در نظریه‌ی آمار و احتمال بسیار گسترده می‌باشد.

دوکن (۱۹۹۴)، بوسک (۱۹۹۸) و لیبچر (۲۰۰۱) به بررسی ویژگی‌های تابع رگرسیون تحت شرایط $\alpha-$ آمیزنده پرداختند.

گوسوم و اولدسعید (۲۰۰۹) به معرفی برآوردگرهای نادارایا-واتسون تحت شرایط سانسور و نیز $\alpha-$ آمیزنده پرداختند و ویژگی‌های مجانبی آن را بیان کردند.

ما نیز باتوجه به نتایج به دست آمده، به بررسی رفتار مجانبی برآوردگرهای تعریف شده توسط آن‌ها در این مدل می‌پردازیم و برای نشان دادن رفتار برآوردگر در نمونه‌های با حجم کم، به مطالعه‌ی شبیه‌سازی انجام شده می‌پردازیم.

در فصل اول پایان‌نامه، تعاریف و مفاهیم موردنیاز در فصل‌های آتی، ارائه خواهد شد. در این فصل، ابتدا مروری بر برآوردگر ناپارامتری تابع چگالی احتمال داریم. پس از آن، با مفهوم متغیرهای سانسور تصادفی آشنا می‌شویم که پایه و اساس ورود به فصل دوم می‌باشد. در ادامه، تعریف وابستگی ضعیف و

متغیرهای تصادفی آمیزنده مورد مطالعه قرار می‌گیرد. سپس به معرفی برخی از انواع رگرسیون ناپارامتری، قضایا و نابرابری‌های لازم می‌پردازیم.

در فصل دوم با برآوردگر هسته‌ای تابع رگرسیون در مدل سانسور تصادفی درحالت یک متغیره آشنا می‌شویم و سازگاری قوی را همراه با یک نرخ برای برآوردگر معرفی شده و نرمال بودن مجانبی آن نشان می‌دهیم. همچنین با توجه به نتایج کسب شده، یک بازه‌ی اطمینان برای برآوردگر معرفی شده ارائه می‌کنیم. در نهایت به منظور روشن ساختن نتایج و نشان دادن ویژگی نرمال بودن مجانبی برآوردگر در نمونه‌های متناهی، شبیه‌سازی را گنجانده‌ایم.

در فصل سوم به بررسی برآوردگر هسته‌ای تابع رگرسیون با شرط $\alpha-$ آمیزنده بودن متغیرهای تصادفی مورد بررسی در حالت چندمتغیره می‌پردازیم و همچنین سازگاری قوی را با نرخی برای برآوردگر معرفی شده نشان می‌دهیم.

لازم به ذکر است که در طول این پایان‌نامه همواره سعی بر آن بوده است که تمامی قضایا و نتایج موردنیاز در همین مجموعه آورده شود تا خواننده بتواند بدون نیاز به مراجع، برهان قضایای اصلی را دنبال کند. همچنین برای جلوگیری از حجیم شدن این گردایه، از آوردن برخی برهان‌ها اجتناب کرده و تنها به ذکر نام مرجع بسنده کرده‌ایم. علی‌رغم این که نمادهای مربوط به هر فصل، در همان قسمت معرفی می‌شوند، در ادامه فهرستی از نمادهای استفاده شده در کل پایان‌نامه ارائه می‌شود تا سبب سهولت در امر آشنایی با نمادها شود. به منظور حفظ یکپارچگی و انسجام در نوشتار پایان‌نامه و نیز جلوگیری از ترکیب کلمات انگلیسی با متن فارسی، واژه‌نامه‌ای را در انتها گردآوری نمودیم تا خوانندگان بتوانند به صورت یکجا و آنی، برگردان بیشتر واژه‌های مطرح شده در پایان‌نامه را مشاهده نمایند. در برگردان علائم، اسامی و ترجمه کلمات کلیدی از کتاب واژه‌ها و اصطلاحات آماری چاپ مرکز آمار ایران استفاده گردیده است.

سمیه باوفا

پاییز ۱۳۹۰

نمادها

امید ریاضی ..... E

تابع توزیع مورد بررسی ..... F

تابع چگالی مورد بررسی ..... f

تابع توزیع متغیر سانسور تصادفی ..... g

تابع نشانگر مجموعه‌ی A ..... $I(A)$

تابع هسته ..... K

پهنای باند ..... h_n

ثابت عمومی ..... C_1, C_2, D

همگرایی تقریباً حتمی ..... $\xrightarrow{a.s}$

همگرایی در توزیع ..... $\xrightarrow{d}$

همگرایی در احتمال ..... $\xrightarrow{P}$

سیگما میدان ..... $\mathcal{A}, \mathcal{B}, \mathcal{F}$

ضریب آمیزنده ..... $a(n)$

مجموعه‌ی اعداد حقیقی ..... $\mathbb{R}$

مجموعه‌ی اعداد طبیعی ..... $\mathbb{N}$

مجموعه‌ی اعداد صحیح ..... $\mathbb{Z}$

بزرگ‌ترین عدد صحیح کوچک‌تر مساوی n ..... $[n]$

مجموعه‌ی تهی ..... $\emptyset$

متمم مجموعه‌ی A ..... A^c

متغیر مورد بررسی ..... Y

مینیمم بین Y و T ..... $Y \wedge T$

مشتق j–ام تابع چگالی ..... $f^{(j)}$

مشتق j–ام تابع هسته ..... $K^{(j)}$

واریانس ..... $\texttt{var}$

مجموعه‌ی همه‌ی توابع با تغییر کراندار روی $[a,b]$ ..... $BV([a,b])$

$\sup\{y, \overline{F}(y) > ۰\}$ ..... τ_F

$\inf\{z, F(z) > ۰\}$ ..... a_F

تعریف می‌کنیم ..... $:=$

فصل ۱

تعاریف و مفاهیم موردنیاز

۱.۱ برآورد تابع چگالی

۲.۱ تحلیل بقا و سانسور

۳.۱ وابستگی ضعیف و متغیرهای تصادفی آمیزنده

۴.۱ رگرسیون ناپارامتری

۵.۱ سایر تعاریف، قضایا و نابرابری‌های مورد نیاز

در این فصل با تعاریف و ویژگی‌های مقدماتی و برخی مفاهیم دیگر که در فصل‌های آینده مورد استفاده قرار می‌گیرند آشنا می‌شویم. به این منظور ابتدا در نخستین بخش به بررسی برآورد تابع چگالی و بعضی از انواع آن می‌پردازیم.

در بخش دوم مروری بر تحلیل بقا، تابع بقا و سانسور خواهیم داشت. سپس در بخش بعد وابستگی ضعیف و متغیرهای تصادفی آمیزنده مورد مطالعه قرار می‌گیرند. در بخش چهارم به معرفی رگرسیون ناپارامتری و برخی از انواع مهم آن می‌پردازیم و سرانجام در بخش پنجم به بیان سایر تعاریف، قضایا و نابرابری‌های لازم در برهان‌های ارائه شده، می‌پردازیم.

۱.۱ برآورد تابع چگالی احتمال

در این قسمت به بررسی برآورد تابع چگالی و انواع آن توسط یک نمونه‌ی تصادفی مانند X_1، ...، X_n که دارای چگالی f هستند می‌پردازیم.

ابتدا فرض کنید $a < x < b$، برای برآورد تابع چگالی در نقطه‌ی x ، یعنی برآورد $f(x)$ ، چنانچه فاصله‌ی a تا b یعنی $h = b - a$ به اندازه کافی کوچک باشد آن‌گاه می‌توان گفت

$$f(x) \sim \frac{\int_a^b f(u)du}{b-a} \tag{۱.۱}$$

از طرفی اگر فاصله‌ی a تا b یعنی h به اندازه‌ی کافی بزرگ باشد به طوری که مقدار قابل ملاحظه‌ای از مشاهدات X_1، ...، X_n در آن قرار بگیرند آن‌گاه

$$\int_a^b f(u)du = P(a < X < b)$$

را می‌توان توسط فراوانی نسبی تعداد مشاهداتی که در بازه‌ی (a, b) قرار می‌گیرند تقریب زد. فرض کنید

متغیر تصادفی Y تعداد مشاهداتی باشد که در بازه‌ی (a,b) قرار می‌گیرند. پس

$$P(a<X<b)=\int_a^b f(u)du\sim\frac{Y}{n} \qquad (۲.۱)$$

توسط (۱.۱) و (۲.۱) داریم:

$$f(x)\sim\frac{\int_a^b f(u)du}{b-a}\sim\frac{Y}{n(b-a)}$$

بنابراین برآوردی برای $f(x)$ را می‌توان $\frac{Y}{n(b-a)}$ در نظر گرفت که آن را با $f_n(x)$ نمایش می‌دهیم.

در ادامه به تعریف بعضی از انواع برآوردگرهای تابع چگالی می‌پردازیم.

تابع چگالی را می‌توان به وسیله‌ی روش‌های مختلفی برآورد کرد، از جمله برآوردگر هیستوگرام، برآوردگر باکس کار، برآوردگر هسته‌ی تابع چگالی و ... که در ادامه به بعضی از آن‌ها اشاره می‌کنیم.

۱.۱.۱ برآوردگر هیستوگرام

ساده‌ترین برآوردگر تابع چگالی برآوردگر هیستوگرام است. فرض کنید f در فاصله‌ی دلخواهی تعریف شده باشد. بدون از دست دادن کلیت فرض می‌کنیم که f روی فاصله‌ی $[۰,۱]$ تعریف شده باشد. اگر m یک عدد صحیح باشد بازه‌هایی را به صورت زیر تعریف می‌کنیم

$$B_۱=\left[۰,\frac{۱}{m}\right],B_۲=\left[\frac{۱}{m},\frac{۲}{m}\right],...,B_m=\left[\frac{m-۱}{m},۱\right]$$

$h=\frac{۱}{m}$ را پهنای باند تعریف می‌کنیم. فرض کنیم Y_j، تعداد مشاهدات که در بازه‌ی B_j قرار می‌گیرند، باشد. قرار می‌دهیم

$$\hat{P}_j=\frac{Y_j}{n}\quad,\quad P_j=\int_{B_j}f(u)du$$

تعریف ۱.۱.۱ برآوردگر هیستوگرام تابع چگالی عبارت است از

$$\hat{f}_n(x) = \sum_{j=۱}^{m} \frac{\hat{P}_j}{h} I(x \in B_j)$$

توجه کنید که برای هر $x \in B_j$ و h کوچک

$$E(\hat{f}_n(x)) = \frac{E(\hat{P}_j)}{h} = \frac{P_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{h} = f(x)$$

بنابراین برآوردگر هیستوگرام تابع چگالی تقریباً ناریب می‌باشد.

□

۲.۱.۱ برآوردگر باکس‌کار

فرض کنید Y_j در قسمت قبل را با $K_n(x - \frac{h}{۲}, x + \frac{h}{۲})$ تعویض کنیم که $K_n(x - \frac{h}{۲}, x + \frac{h}{۲})$ برابر است با تعداد مشاهدات دربازه‌ی $(x - \frac{h}{۲}, x + \frac{h}{۲})$. با توجه به این که

$$x - \frac{h}{۲} < X_i < x + \frac{h}{۲} \Longleftrightarrow \left| \frac{x - X_i}{h} \right| < \frac{۱}{۲}$$

و با در نظر گرفتن تابع

$$e(t) = \begin{cases} ۱ & |t| < \frac{۱}{۲} \\ ۰ & |t| > \frac{۱}{۲} \end{cases}$$

به تعریف زیر دست می‌یابیم.

تعریف ۱.۱.۱ برآوردگر تابع چگالی باکس‌کار عبارت است از

$$\hat{f}_n(x) = \frac{K_n(x - \frac{h}{۲}, x + \frac{h}{۲})}{nh} = \frac{۱}{n} \sum_{i=۱}^{n} \frac{۱}{h} e\left(\frac{x - X_i}{h} \right)$$

□

۳.۱.۱ برآوردگر هسته‌ای تابع چگالی

به علت هموار نبودن هیستوگرام‌ها، اکنون برآوردگرهای هسته‌ی تابع چگالی که هموارتر می‌باشند و سریع‌تر به یک چگالی صحیح همگرایند، را بیان می‌کنیم. قبل از تعریف و بیان ویژگی‌های برآوردگرهای هسته‌ای به تعریف هسته و بیان انواع آن می‌پردازیم.

تعریف ۳.۱.۱ تابع نامنفی K را هسته گویند هرگاه شرایط زیر را داشته باشد:

۱. $\int K(x)dx = ۱$

۲. $\int xK(x)dx = ۰$

۳. $(\sigma^۲ > ۰)\ \sigma^۲ = \int x^۲K(x)dx \neq ۰$

□

کلمه‌ی هسته به هر تابع هموار K که شرایط بالا را در بر داشته باشد اطلاق می‌شود. هسته‌ها به دو گروه تقسیم می‌شوند.

۱. هسته‌های متقارن

۲. هسته‌های نامتقارن

بعضی از انواع هسته‌های متقارن و نامتقارن را در زیر بیان می‌کنیم:

هسته‌های متقارن

هسته‌ی باکس‌کار: $K(x) = \frac{۱}{۲}I(x)$

هسته‌ی گوسی: $K(x) = \frac{۱}{\sqrt{۲\pi}} e^{\frac{-x^۲}{۲}}$

هسته‌ی اپانچنیکوف: $K(x) = \frac{۳}{۴}(۱ - x^۲)I(x)$

هسته‌ی تریکوب: $K(x) = \frac{۷۰}{۸۱}(۱ - |x|^۳)^۳ I(x)$

هسته‌های نامتقارن

هسته‌ی گاما: $K_{\frac{x}{b+۱}}, b(t) = \frac{t^{\frac{x}{b}} e^{\frac{-t}{b}}}{b^{\frac{x}{b+۱}} ۲(\frac{x}{b+۱})}$

هسته‌ی معکوس گوسی:

$$K_{IG}(m,d)ty = \frac{\sqrt{d}}{\sqrt{۲\pi y^۳}} \exp(-\frac{d}{۲m}(\frac{d}{m} - ۲ + \frac{m}{y})) \quad y > ۰$$

یا

$$K(M,h) = \frac{۱}{\sqrt{۲\pi h x^۳}} e^{-(\frac{۱}{۲h})(x-M)^۲/xM^۲} \quad x>۰$$

هسته‌ی عکس معکوس گوسی:

$$K_{RIG}(m,d)^{(۲)} = \frac{\sqrt{d}}{\sqrt{۲\pi z}} \exp(-\frac{d}{۲m}(mz - ۲ + \frac{۱}{mz})) \quad z > ۰$$

تعریف ٤.١.١ فرض کنید هسته‌ی متقارن K و عدد مثبت h که پهنای باند نامیده می‌شود داده شده است، در این صورت برآوردگر چگالی هسته به صورت زیر تعریف می‌شود:

$$\hat{f}_n(x) = \frac{۱}{n} \sum_{i=۱}^{n} \frac{۱}{h} K\left(\frac{x - X_i}{h}\right)$$

□

معمولاً انتخاب K و انتخاب پهنای باند یعنی h مهم‌تر است. مثلاً ممکن است با پهنای باند کوچک یک برآورد بسیار ناهمواری به دست آید، در حالی که با پهنای باند بزرگ‌تر، برآورد هموارتری داشته باشیم.

تحت شرایطی برآوردگر هسته‌ی تابع چگالی، برآوردگری سازگار (ضعیف) برای f می‌باشد که این مطلب در قضیه‌ی زیر بیان شده است.

قضیه ۱.۱.۱ فرض کنید f در x پیوسته باشد و وقتی $n \longrightarrow \infty$ ، $h \longrightarrow ۰$ و $nh \longrightarrow \infty$. در این‌صورت داریم:

$$\hat{f}_n(x) \xrightarrow{P} f(x)$$

یعنی برآوردگر $\hat{f}_n(x)$ در احتمال به $f(x)$ میل می‌کند و یا

$$\lim_{n\longrightarrow\infty} P(|\hat{f}_n(x) - f(x)| > \epsilon) = ۰ \quad \forall \varepsilon > ۰$$

تعریف ۵.۱.۱ (برآوردگر هسته‌ای تابع چگالی در حالت چندمتغیره)

فرض کنید $\mathbf{X}_۱$، ...، $\mathbf{X}_n$ یک نمونه‌ی تصادفی d بعدی با چگالی f باشد. همچنین

$$\mathbf{X}_i = (X_{i۱}, \ldots, X_{id})^T$$

دوهاولز (۱۹۷۷) در حالت چندمتغیره، برآوردگر هسته‌ای تابع چگالی احتمال را به صورت زیر معرفی کرد.

$$\hat{f}(\mathrm{x}, \mathbf{H}) = n^{-۱} \sum_{i=۱}^{n} K_{\mathbf{H}}(\mathrm{x} - \mathbf{X}_i)$$

که در آن $\mathbf{H}$ یک ماتریس $d \times d$ ، مثبت معین و متقارن است که ماتریس پهنای باند نامیده می‌شود و

$$K_{\mathbf{H}}(\mathrm{x}) = |\mathbf{H}|^{-\frac{۱}{۲}} K(\mathbf{H}^{-\frac{۱}{۲}}\mathrm{x})$$

که در آن K یک تابع هسته d متغیره است که شرط زیر را برقرار می‌کند

$$\int_{\mathbb{R}^d} K(x_۱, \ldots, x_d) dx_۱ \ldots dx_d = ۱$$

معمولاً به دو روش، با استفاده از هسته‌ی یک متغیره و متقارن K، هسته‌ی چندمتغیره در نظر گرفته می‌شود

$$K^P(\mathrm{x}) = \prod_{i=۱}^{d} \kappa(x_i) \quad , \quad K^s(\mathrm{x}) = c_{\kappa,d}\kappa\{(\mathrm{x}^T\mathrm{x})^{۱/۲}\}$$

که

$$c_{\kappa,d}^{-۱} \int \kappa\{(\mathrm{x}^T\mathrm{x})^{۱/۲}\}d\mathrm{x} \quad , \quad \mathrm{x} = (x_۱, ..., x_d)$$

$K^P(X)$ را هسته‌ی حاصلضربی و $K^S(X)$ را هسته‌ی کروی نامند.

یک حالت خاص از برآوردگر هسته‌ای تابع چگالی در حالت چندمتغیره را می‌توان تابع زیر را که در اکثر کاربردها استفاده می‌شود، در نظر گرفت (با فرض ماتریس قطری $\mathbf{H} = \mathtt{diag}\,(h_۱^۲, \ldots, h_d^۲)$)

$$\hat{f}(x_i h) = n^{-۱} \left(\prod_{l=۱}^{d} h_l\right)^{-۱} \sum_{i=۱}^{n} K\left(\frac{x_۱ - X_{i۱}}{h_۱}, ..., \frac{x_d - X_{id}}{h_d}\right)$$

و با در نظر گرفتن پهنای باند یکسان، خواهیم داشت

$$\hat{f}(\mathrm{x}; h) = n^{-۱}h^{-d}\sum_{i=۱}^{n} K\{(\mathrm{x} - X_i)/h\}$$

□

۲.۱ تحلیل بقا و سانسور

تعریف ۱.۲.۱ تحلیل بقا عبارت است از گردایه‌ی از روش‌های آماری که برای تجزیه و تحلیل داده‌های نامنفی به کار می‌رود. معمولاً این داده‌های نامنفی (داده‌های بقا)، داده‌های مربوط به زمان شروع تا رسیدن به یک پیشامد خاص می‌باشند.

پیشامد خاص را به عنوان مثال می‌توان پیشامد رسیدن به یک بیماری خاص، پیشامد رسیدن مرگ بر اثر بیماری خاص، پیشامد انتشار یک تومور و ... در نظر گرفت. در ادامه ما از نماد T برای نشان دادن متغیر تصادفی بقا استفاده می‌کنیم. همچنین فرض کنید متغیر تصادفی T دارای تابع چگالی $f(t)$ و تابع توزیع $F_T(t)$ باشد.

معمولاً به دو دلیل عمده‌ی زیر، تجزیه و تحلیل داده‌های بقا از روش‌های استاندارد آماری استفاده شده در آنالیز داده‌ها تبعیت نمی‌کند:

۱. داده‌های بقا دارای توزیع نرمال نمی‌باشند.

۲. دلیل دیگری که باعث می‌شود روش‌های استاندارد در مورد داده‌های بقا مناسب نباشد، این است که عموماً سانسور می‌شوند. که در ادامه مفهوم سانسور بیان خواهد شد.

در خلاصه کردن داده‌های بقا، دو تابع به نام‌های تابع بقا و تابع خطر نقش اساسی دارند.

□

تعریف ۲.۲.۱ تابع بقا که آن را با $S(t)$ نشان می‌دهیم عبارت است از

$$S(t) = P(T > t)$$

این تابع نشان‌دهنده‌ی احتمال بقای یک عنصر (فرد) تا زمان t است.

□

تعریف ۳.۲.۱ تابع خطر که آن را با $h(t)$ نشان می‌دهیم به صورت زیر تعریف می‌شود

$$h(t) = \lim_{\Delta t \longrightarrow ۰} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}$$

تابع فوق نشان‌دهنده‌ی نرخ مرگ فوری برای افراد زنده مانده تا زمان t می‌باشد. اگر T متغیر تصادفی پیوسته باشد، به سادگی داریم

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{۱ - F_T(t)}$$

□

همان‌طور که قبلاً بیان شد یکی از دلایل مهمی که باعث می‌شود تحلیل داده‌های بقا متمایز از تحلیل‌های استاندارد شود، سانسور می‌باشد.

تعریف ٤.٢.١ زمان بقای یک عنصر (فرد) را سانسور شده گویند هرگاه زمان بقای آن عنصر (فرد) تا پایان دوره‌ی مطالعه نشود. به عبارت دیگر زمان بقای یک عنصر را سانسور شده گویند، هرگاه نقطه‌ی پایان پیشامد موردنظر برای آن عنصر مشاهده نشود. بنابراین داده‌ی سانسور شده یک داده‌ی ناقص است.

معمولاً دلایل عمده‌ی سانسور شدن عبارت‌اند از

۱. پایان مطالعه

۲. عدم امکان پیگیری

۳. خارج شدن از مطالعه

از دیدگاهی، سانسور را می‌توان به صورت زیر تقسیم‌بندی کرد.

۱. سانسور نوع اول:

فرض کنید C یک عدد ثابت و $X_۱$، ...، X_n یک نمونه تصادفی از توزیع F باشند. در این سانسور به جای مشاهدات $X_۱$، ...، X_n مشاهدات $Y_i = \min(X_i, C)$ یعنی

$$Y_i = \begin{cases} X_i & X_i \leqslant C \\ C & X_i \geqslant C \end{cases}$$

دیده می‌شوند.

۲. سانسور نوع دوم

فرض کنید $X_{(۱)}$، ...، $X_{(n)}$ آماره‌های ترتیبی $X_۱$، ...، X_n و $r < n$ باشد. در این نوع سانسور فرض می‌کنیم تا مشاهده شدن r امین شکست، عناصر تحت مطالعه قرار گیرند. بنابراین مشاهدات به صورت زیر به دست می‌آیند

$$Y_۱ = X_{(۱)}, Y_۲ = X_{(۲)}, \ldots, Y_r = X_{(r)}, Y_{(r+۱)} = X_{(r)}, \ldots, Y_{(n)} = X_{(r)}$$

۳. سانسور تصادفی (نوع سوم):

متغیر تصادفی X_i توسط متغیر تصادفی Y_i از (راست) سانسور تصادفی می‌شود هرگاه زوج مرتب (Z_i, δ_i) مشاهده شوند که در آن $\delta_i = I(X_i \leqslant Y_i)$ و

$$Z_i = \min(X_i, Y_i)$$

فرض می‌کنیم متغیرهای نشان‌دهنده‌ی زمان‌های سانسور یعنی $Y_i, ۱ \leqslant i \leqslant n$ مستقل و هم‌توزیع با تابع توزیع مشترک G هستند. همچنین Y_i ها مستقل از X_i ها فرض می‌شوند. از آنجایی که داده‌های سانسور بیشتر در مسائل مربوط به تحلیل بقا رخ می‌دهند، بدون کم‌شدن از کلیت مسأله X_i ها و Y_i ها نامنفی فرض می‌شوند. برای داده‌های سانسور شده δ_i ها برابر صفر و برای داده‌های غیرسانسور، δ_i ها برابر یک است.

□