جمهوری اسلامی ایران
مرکز آمار ایران
پژوهشکده‌ی آمار

پژوهشکدهٔ آمار

معرفی متغیرهای نیم‌پیوسته در آمارگیری‌های بزرگ

جانهی مقادیر گم‌شده‌ی متغیرهای پیوسته، نیم‌پیوسته و رسته‌ای

با روش مونت‌کارلوی زنجیر مارکوفی و تحلیل آماری این گونه داده‌ها

تابستان ۱۳۸۵

بسمه تعالی

پژوهشکدهٔ آمار

گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی

معرفی متغیرهای نیم‌پیوسته در آمارگیری‌های بزرگ، جانهی مقادیر گم‌شده‌ی متغیرهای پیوسته، نیم‌پیوسته و رسته‌ای با روش مونت‌کارلوی زنجیر مارکوفی و تحلیل آماری این‌گونه داده‌ها

مجتبیٰ گنجعلی (مجری)

محمد رضا مشکانی

امید حمیدی

حسن رنجی

مرجان نورینی

فهرست مندرجات

فصل ۱

معرفی انواع متغیرهای رسته‌ای، پیوسته و نیم پیوسته با گم شدگی

در این فصل در ابتدا انواع متغیرهای رسته‌ای، پیوسته و نیم پیوسته تعریف شده‌اند. در ادامه مسئله گم شدگی و جانهی در داده‌هایی با متغیرهای آمیخته (شامل رسته‌ای، پیوسته و نیم پیوسته) بررسی شده و مکانیسم‌های گم شدگی و روش‌های موجود جانهی برای این‌گونه داده‌ها آورده شده‌اند. برخی تعاریف لازم مورد استفاده در فصل‌های بعدی نیز در انتها ذکر شده‌اند.

۱ — ۱ متغیرهای رسته‌ای، پیوسته و آمیخته‌ای از آنها (نیم پیوسته)

در این بخش ابتدا تعریف متغیرهای رسته‌ای، ترتیبی و پیوسته آمده است. سپس توابع توزیع آمیخته و انواع آن تعریف شده اند.

۱—۱—۱ متغیرهای رسته‌ای و پیوسته

متغیرها را با چهار نوع مقیاس، اندازه گیری می‌کنند. متغیرهای کیفی که بدون هیچ ترتیب خاصی دسته بندی شده‌اند، اسمی نامیده می‌شوند. از این نوع متغیرها می‌توان مذهب (مسلمان و غیر مسلمان)، نوع وسیله‌ی نقلیه مورد استفاده (اتومبیل، دوچرخه، موتورسیکلت، مترو، اتوبوس و...) و نوع محل اقامت (آپارتمان، خانه و...) را نام برد. برای متغیرهای اسمی ترتیب فهرست کردن رسته‌ها مهم نیست و تحلیل آماری وابسته به آن ترتیب نیست. برخی متغیرها در مقیاس خود دارای نوعی ترتیب هستند چنین متغیرهایی ترتیبی نامیده می‌شوند. برای مثال وضعیت اجتماعی (بالا، متوسط و پایین)، شرایط بیمار (خوب، متوسط و خیلی بد) متغیرهای ترتیبی هستند. در این مقیاس فاصله‌ی بین مقادیر معلوم نیست، هرچند شخصی که برای مثال وضعیت اجتماعی متوسط دارد، وضعیت اجتماعی وی بهتر از شخصی است که وضعیت اجتماعی پایین دارد، اما هیچ مقدار عددی نمی‌تواند فاصله‌ی بین متوسط و بد را اندازه‌گیری کند. یک متغیر بازه‌ای متغیری است که بین هر دو مقدار آن بازه‌ای از اعداد وجود دارد ولی این متغیرها لزوماً یک مبدأ خاص تعریف شده ندارند. برای مثال سطح فشار خون و درجه حرارت از این نوع متغیرها هستند. متغیرهای نسبتی متغیرهایی هستند که در مقیاس بازه‌ای تعریف شده‌اند اما مقیاس آنها دارای یک مبدا تعریف شده است برای مثال درجه حرارت در مقیاس بازه‌ای اندازه گیری می‌شود ولی دارای صفر تعریف شده نیست. ولی متغیرهای طول عمر، درآمد و قد در مقیاس نسبتی اندازه گیری می‌شوند و مبدا در آنها تعریف شده است (به طور عام صفر).

طریقه‌ای که یک متغیر را اندازه می‌گیرد نوع آن را مشخص می‌کند. برای مثال آموزش متغیری اسمی است اگر به صورت نوع آموزشگاه (انتفاعی و غیر انتفاعی) رسته بندی شود، اگر متغیر آموزش به صورت بی‌سواد، سیکل، دیپلم و... ثبت شود متغیری ترتیبی است و چنانچه این متغیر طول آموزش فرد را نشان دهد متغیری نسبتی خواهد بود. مقیاس اندازه گیری متغیر است که تعیین می‌کند چه روش آماری برای تحلیل در مورد آن متغیر مناسب است. روش‌های آماری مورد استفاده برای یک متغیر با مقیاس اندازه‌گیری خاص، نمی‌تواند برای یک متغیر با مقیاس اندازه‌گیری دیگر استفاده شود. متغیرهای اسمی و ترتیبی از نوع رسته‌ای و متغیرهای پیوسته از نوع بازه‌ای یا نسبتی هستند. در این طرح متغیرهایی را مورد بررسی قرار می‌دهیم که از نوع اسمی یا بازه‌ای یا نسبتی باشند، متغیرهایی را

نیز در نظر خواهیم گرفت که از نوع نیم پیوسته هستند (به تعریف آن در بخش بعد توجه کنید).

۱–۱–۲ توابع توزیع آمیخته

فرض کنید X متغیری تصادفی است که نتیجه‌ی یک آزمایش را نشان می‌دهد و مقادیری که می‌تواند اختیار کند از مجموعه‌ای مانند S زیر مجموعه‌ای از $\mathbf{R}$ است. متغیر X دارای تابع توزیع آمیخته است اگر بتوان S را به دو زیر مجموعه مانند C و D افراز کرد به گونه‌ای که:

۱) D شمارش پذیر بوده و $\circ < P(X \in D) < ۱$

۲) $P(X = x) = \circ$ برای هر x متعلق به C

۳) $P(X \in D) + P(X \in C) = ۱$.

بنابراین بخشی از جرم احتمال X روی نقاط متعلق به مجموعه‌ی گسسته D متمرکز می‌شود و بقّیه‌ی آن به طور پیوسته روی C پخش می‌شود.

فرض کنید $p = P(X \in D)$ و $\circ < p < ۱$، در این صورت می‌توانیم یک تابع روی D تعریف کنیم. این تابع را تابع احتمال گسسته‌ی جزئی[۱] نامند. فرض کنید به ازای هر $x \in D$، $g(x) = P(X = x)$ باشد. در این صورت:

$$\begin{aligned} g(x) &> \circ, \quad \forall x \in D \\ \sum_{x \in D} g(x) &= p \\ P(X \in A) &= \sum_{x \in A} g(x) \ , \quad A \subseteq D \end{aligned}$$

قسمت پیوسته‌ی توزیع به وسیله‌ی یک تابع چگالی جزئی تعریف می‌شود. برای این فرض کنید یک تابع نامنفی مثل h روی C وجود داشته باشد به گونه‌ای که:

$$P(X \in A) = \int_A h(x)dx, \qquad A \subseteq C$$

در این صورت:

[۱] Partial Discrete Probability Function

$$\int_C h(x)dx = ۱ - p$$

دقت کنید که به ازای همه‌ی مقادیر X متعلق به D داریم $h(x) = ۰$ و به ازای همه‌ی مقادیر x متعلق به C، داریم $g(x) = ۰$.
فرض کنید که $A \subseteq S$ در این صورت:

$$P(X \in A) = \sum_{x \in A} g(x) + \int_A h(x)dx$$

به آسانی نتیجه می‌شود:

$$\begin{aligned} f_۱(x \mid X \in D) &= g(x)/p \ , \quad \forall x \in D \\ f_۲(x \mid X \in C) &= h(x)/(۱ - p) \ , \quad \forall x \in C \end{aligned}$$

که در آن $f_۱$ و $f_۲$ چگالی‌های شرطی x هستند. بنابراین توزیع x آمیخته‌ای از یک توزیع گسسته و یک توزیع پیوسته است.

متغیرهای بریده[۱]

متغیرهای آمیخته معمولاً زمانی رخ می‌دهند که یک متغیر تصادفی با توزیع پیوسته از نقطه‌ای به بعد بریده شود. برای مثال فرض کنید T طول عمر یک دستگاه و به ازای $t > ۰$، $f_T(t)$ تابع چگالی آن باشد. هنگام آزمایش این دستگاه نمی‌توان مدت زمان بی‌نهایتی منتظر ماند، بنابراین ممکن است یک عدد ثابت مثبت مانند a را انتخاب کنیم و مقدار متغیر زیر را ثبت کنیم. اگر مقادیر متغیر تصادفی U را

[۱] Truncated

با استفاده از متغیر تصادفی T به صورت

$$U = \begin{cases} T, & T < a \text{ اگر} \\ a, & T \geqslant a \text{ اگر} \end{cases}$$

تعریف کنیم در این صورت U یک توزیع آمیخته دارد. لذا داریم:

$$D = \{a\}, \quad g(a) = \int_{\{t:t>a\}} f_T(t)dt$$
$$C = (\circ, a), \quad h(t) = f_T(t), \quad \circ < t < a$$

فرض کنید X یک توزیع پیوسته روی $\mathbf{R}$ (مجموعه‌ی اعداد حقیقی) با تابع چگالی f_X باشد. اگر متغیر تصادفی X در a و b، $(b > a)$، بریده شود متغیر جدید (Y) با مقادیر زیر به وجود می‌آید:

$$Y = \begin{cases} a, & X \leq a \text{ اگر} \\ X, & a < X < b \text{ اگر} \\ b, & X \geq b \text{ اگر} \end{cases}$$

آنگاه Y دارای توزیع آمیخته است و نتایج زیر را داریم:

$$\begin{aligned} D &= \{a, b\}, \quad g(a) = \int_{\{x:x<a\}} f_X(x)dx, \quad g(b) = \int_{\{x:x>b\}} f_X(x)dx \\ C &= (a, b), \quad h(x) = f_X(x), \quad a < x < b. \end{aligned}$$

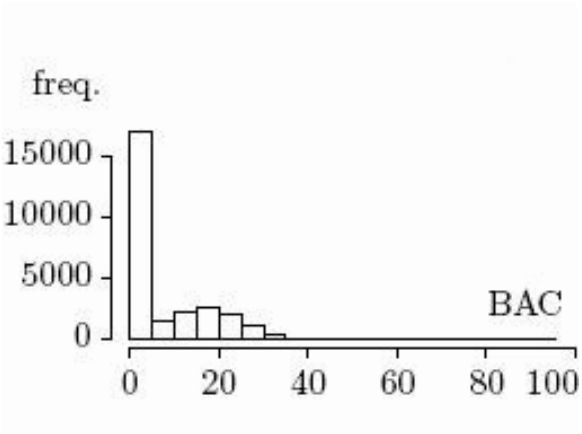
متغیرهای نیم پیوسته

متغیر نیم پیوسته متغیری است که مقداری تک (معمولا صفر) را با احتمالی مثبت و مقادیری پیوسته در یک بازه را طبق یک چگالی پیوسته می‌تواند اختیار کند. این متغیرها متفاوت از متغیرهای از چپ سانسور شده[۱] یا بریده هستند، چرا که در آن‌ها احتمالی که به صفر نسبت داده می‌شود به جهت پاسخ‌های منفی یا داده‌های گم شده نیست. چنین داده‌هایی در زمینه‌های مختلف به وفور یافت

[۱] Left Censord

می‌شوند. متغیرها در بررسی‌های اقتصادی و مربوط به درآمد یا هزینه‌ها اکثراً نیم پیوسته هستند، چون بسیاری از افراد یک نوع خاص از هزینه را ندارند. در مطالعات مصرف مواد مخدر و یا الکل، بسیاری از افراد در طول دوره‌ی مطالعه، از این مواد استفاده نمی‌کنند.

بافت نگار نشان داده شده در شکل ۱—۱ را در نظر بگیرید. این بافت نگار نتایج آزمایش مقدار الکل خون (BAC)[۲] از ۲۷۶۳۳ راننده وسایل نقلیه و عابران پیاده در تصادف‌های شدید در بزرگراه‌های امریکا در سال ۱۹۹۳ را نشان می‌دهد، این بررسی که توسط شی‌فر و اولسن[۳] (۱۹۹۹) انجام شده است، نشان می‌دهد که برای ۵۷/۳ درصد از موارد نشان داده شده در شکل ۱—۱، مقدار BAC صفر است که نمایانگر عدم مصرف الکل است، سایر مقادیر که از ۱ تا ۹۴ تغییر می‌کنند، نشان دهنده‌ی سطوح الکل خون از ۰/۰۱ تا ۰/۹۴ گرم در هر دسی لیتر (g⁄ dl) است.



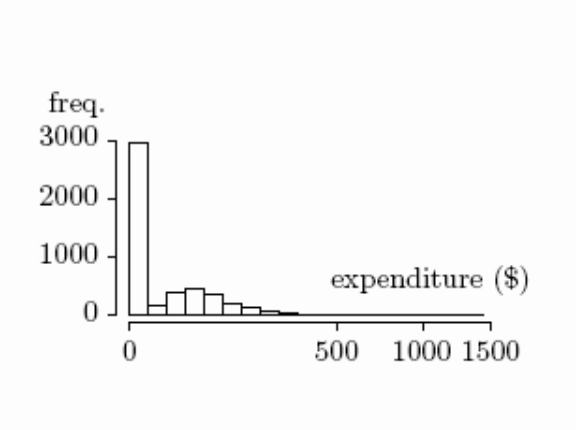
نمودار ۱—۱: بافت نگار مقدار الکل خون در بین رانندگان

بافت نگار شکل ۱—۲ نشان دهنده‌ی هزینه‌ی خرید کفش در فصل اول سال ۱۹۹٦ به وسیله‌ی ۴۸۷٦ واحد مصرف کننده است که توسط اداره‌ی آمار آمریکا در بررسی هزینه‌ی مصرف کننده جمع آوری شده است (شی‌فر و اولسن، ۱۹۹۹). یک واحد مصرف کننده یک خانوار یا گروهی از افراد است که در آمد مشترک دارند. برای کاهش چولگی ظاهر شده، جذر مقادیر هزینه رسم شده اند.

[۲] Blood alchol content

[۳] Schafer and Olsen

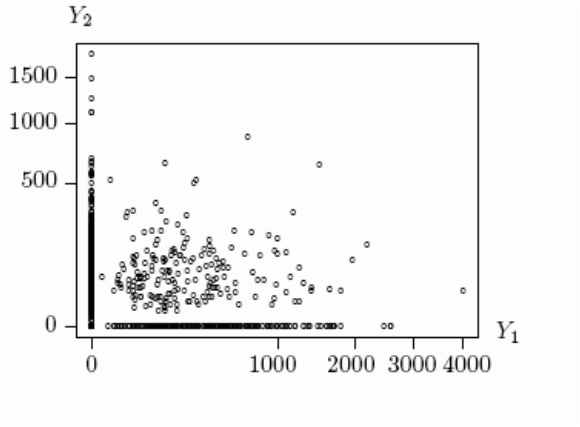
نزدیک ٦١ درصد از مقادیر مشاهده شده صفر است و بقیه بین ۱ تا ۱۳۱۵ دلار توزیع شده‌اند.



نمودار ۱ — ۲: بافت نگار هزینه‌ی کفش

در هر یک از این مثال‌ها کمیت مورد نظر نیم پیوسته است، یعنی آمیخته‌ای از صفر و مقادیر مثبتی است که به طور پیوسته توزیع شده‌اند. متغیرهای نیم پیوسته برای محققی که پیوند بین این متغیرها را با متغیرهای دیگر ارزیابی می‌کند مشکلاتی به وجود می‌آورد، زیرا این روابط نسبتا پیچیده هستند. برای نشان دادن انواع روابطی که یافت می‌شوند، در شکل ۱ — ۳ دو متغیر دیگر در همان دوره‌ی سه ماهه در بررسی هزینه‌ی مصرف کننده رسم شده‌اند که y_1 جذر هزینه‌های خرید لوازم بزرگ منزل (مثل یخچال) و y_2 جذر هزینه‌های خرید لوازم کوچک منزل است. در بین ۴۸۷٦ واحد مصرف کننده که مورد بررسی قرار گرفته‌اند ۷۰/۳ درصد هیچ هزینه‌ای نداشتند ($y_1 = y_2 = 0$)، ۵/۳ درصد لوازم بزرگ خریداری کرده‌اند اما لوازم کوچک خریداری نکرده‌اند ($y_1 > 0, y_2 = 0$)، ۲۰/۵ درصد لوازم کوچک خریداری کرده‌اند اما لوازم بزرگ خریداری نکرده‌اند ($y_1 = 0, y_2 > 0$) و ۳/۹ درصد هر دو را خریداری کرده‌اند ($y_1 > 0, y_2 > 0$). بین آنهایی که لوازم کوچک خریداری کرده‌اند ($y_2 > 0$) همبستگی بین $y_2^{\frac{1}{3}}$ و نشانگر دودویی برای $y_1 > 0$ برابر $r = +0/11$ است. میزان مخارج لوازم کوچک دارای یک رابطه‌ی معنی دار مثبت با خرید لوازم بزرگ است. از بین آنهایی که لوازم بزرگ خریداری کرده‌اند ($y_1 > 0$)، میزان مخارج لوازم بزرگ $y_1^{\frac{1}{3}}$ دارای رابطه‌ی معنی داری منفی با نشانگر دودویی برای خرید لوازم کوچک است ($r = -0/12$). بالاخره بین آنهایی که هر دو را خریداری

کرده‌اند ($y_۱ > ۰, y_۲ > ۰$) همبستگی بین $y_۱^{\frac{۱}{۴}}, y_۲^{\frac{۱}{۴}}$ تقریباً صفر است ($r = +۰/۰۳$).



نمودار ۱ـ۳: نمودار پراکنش هزینه‌های لوازم بزرگ و لوازم کوچک منزل

تابع توزیع تجمعی متغیر نیم پیوسته

تابع توزیع تجمعی $F^{sc}(x)$ برای متغیر نیم پیوسته را می‌توان به صورت

$$F^{sc}(x) = pF^{dg}(x) + (۱ - p)F^{ac}(x) \qquad ۰ < p < ۱, \tag{۱.۱}$$

بیان کرد که $F^{dg}(.)$ و $F^{ac}(.)$ به ترتیب تابع توزیع تجمعی بخش گسسته و پیوسته است. $F^{dg}(.)$ تباهیده در نقطه‌ی مفروض a و $F^{ac}(.)$ مطلقاً پیوسته است. تابع توزیع $F_X^{dg}(.)$ در نقطه‌ی x به صورت

$$F_X^{dg}(x) = \begin{cases} ۰ & x < a \text{اگر} \\ ۱ & x \geq a \text{اگر} \end{cases}$$

است. $F^{ac}(x)$ برای مقادیر کوچکتر یا مساوی a برابر صفر است.

مثال ۱.۱ ***متغیر نیم پیوسته‌ی نمایی.*** مدت توقف راننده‌ی یک وسیله‌ی نقلیه را در مقابل تابلوی ایست یک خیابان یک طرفه در نظر بگیرید. اگر X متغیری تصادفی باشد که مدت توقف او را تا پایان

ایست نشان می‌دهد احتمال معینی وجود دارد که هیچ وسیله‌ی نقلیه‌ای در حال گذر از تقاطع نباشد و این راننده بتواند بدون توقف به حرکت خود ادامه دهد. از سوی دیگر اگر مجبور باشد توقف کند وقفه‌ی او می‌تواند هر مقدار از زمان‌های ممکن باشد. این آزمایش را با فرض این که X دارای تابع توزیع معلوم به صورت $F_X^{sc}(x) = [۱ - (۱ - p)e^{-\lambda x}]$ که $x \geq ۰$ می‌باشد می‌توان مدل‌بندی کرد. این $F_X^{sc}(x)$ در $x = ۰$ دارای جهشی به اندازه‌ی p است اما برای $x > ۰$ پیوسته با توزیع نمایی است.

روش مدل بندی داده‌های نیم پیوسته نشان دادن آنها به صورت یک آمیخته‌ی دو بخشی از یک توزیع پیوسته و یک جرم نقطه‌ای تباهیده می‌باشد. فرض کنید مشاهدات نیم پیوسته $y_۱, y_۲, \ldots, y_n$ به صورت دو متغیر (w_i, z_i) $i = ۱, ۲, \ldots, n,$ دوباره کدگذاری شده باشند که

$$w_i = \begin{cases} ۱ & y_i \neq ۰ \text{ اگر} \\ ۰ & y_i = ۰ \text{ اگر} \end{cases} \qquad (۲.۱)$$

و

$$z_i = \begin{cases} g(y_i) & w_i = ۱ \text{ اگر} \\ \text{نامرتبط} & w_i = ۰ \text{ اگر} \end{cases} \qquad (۳.۱)$$

و g تابع یکنوای افزایشی (برای مثال لگاریتم) باشد که برای نرمال کردن تقریبی مقادیر غیر صفر انتخاب شده است. فرض می‌شود نشانگرهای دودویی w_i دارای توزیع برنولی با $p(w_i = ۱) = \pi$ و توزیع شرطی z_i به شرط $w_i = ۱$ نرمال با پارامترهای μ و $\sigma^۲$ باشد و وقتی $w_i = ۰$ مقدار $y_i = ۰$ بوده و در توزیع z_i $(z_i > ۰)$ نقشی نداشته باشد در این صورت گوییم z_i نامرتبط است. تحت این فرض‌ها، درستنمایی برای $\theta = (\pi, \mu, \sigma^۲)$ به توابع مجزا برای π و $(\mu, \sigma^۲)$ به صورت زیر تجزیه می‌شود:

$$L(\theta) \propto \prod_{i=۱}^{n} \pi^{w_i}(۱ - \pi)^{۱-w_i} \prod_{i:y_i \neq ۰} \sigma^{-۱} \exp\left\{-\frac{(z_i - \mu)^۲}{۲\sigma^۲}\right\}, \qquad (۴.۱)$$

و برآوردهای ML به فرم بسته برای پارامترها به صورت زیر به دست می‌آید:

$$\begin{aligned} \hat{\pi} &= n^{-۱} \sum_{i=۱}^{n} w_i \\ \hat{\mu} &= n_۱^{-۱} \sum_{i:y_i \neq ۰} z_i \\ \hat{\sigma}^۲ &= n_۱^{-۱} \sum_{i:y_i \neq ۰} (z_i - \hat{\mu})^۲ \end{aligned}$$

که در آن $n_{۱} = \sum_{i=۱}^{n} w_i$.

۱ — ۲ مسئله‌ی گم شدگی و جانهی در داده‌هایی با متغیرهای رسته‌ای، پیوسته و نیم پیوسته

تحلیل آماری داده‌های گم شده غالباً پیچیده است، به خصوص در بررسی‌هایی که موارد عدم پاسخ فراوانند. روش عمومی در برخورد با چنین داده‌هایی جانهی آنها قبل از انجام هرگونه تحلیل آماری است. این روش، تحلیل داده‌ها را به دو مرحله تقسیم می‌کند. نخست، به وسیله جانهی داده‌های گم شده به مجموعه‌ای از داده‌های کامل تبدیل می‌شود. دوم، در مرحله تحلیل، مجموعه‌ی داده‌های جانهی شده، همانند وقتی که داده‌ی گم شده وجود ندارد، تحلیل می‌شوند. معمولاً جانهی با کمک شخصی که داده‌ها را جمع آوری کرده و نیز کمک گرفتن از اطلاعات اضافی موجود و استفاده از ابزارهای آماری موجود (مثل، روش‌های مونت کارلو) انجام می‌شود. مزیت اصلی جانهی داده‌های گم شده این است که از پیچیدگی‌ها در تحلیل‌های آماری که به وسیله‌ی داده‌های گم شده به وجود می آیند، جلوگیری می‌کند. بنابراین از نقطه نظر محاسباتی، جانهی مرحله‌ی تحلیل را ساده می‌کند، به طوری که در تحلیل تنها روش‌های استاندارد (روش‌هایی مربوط به تحلیل داده‌های بدون مقادیر گم شده) مورد نیاز است. مهم تر این که، این روش از دانش و اطلاعات شخص جمع آوری کننده‌ی داده‌ها در جانهی استفاده می‌کند.

متاسفانه، قطعی بودن کامل جانهی نسبتاً غیر معمول است، یعنی حتی با اطلاعات جمع آوری کننده ماهر داده‌ها، در جانهی عدم اطمینان وجود دارد. استنباط معتبر بایستی این عدم اطمینان را در نظر بگیرد. جانهی چندگانه (MI)[۱] (رابین[۲] ، ۱۹۸۷) یک روش پذیرفته شده برای به دست آوردن چنین استنباط معتبر با وجود داده‌های گم شده می‌باشد (مروری بر جانهی چندگانه در فصل ۲، ارائه خواهد شد).

معمولاً در جانهی چندگانه یک روش، بر اساس مدل بندی، برای تولید m جانهی استفاده می‌شود

[۱] Multiple Imputation

[۲] Rubin

که برآمدهای تصادفی تکرار شده از یک مدل را برای عدم پاسخ نشان می‌دهد. در مرحله تحلیل، هر یک از m مجموعه داده‌ی جانهی شده همانند یک مجموعه داده‌ی کامل، تحلیل می‌شوند و m مجموعه‌ی برآوردها و خطاهای مربوط بر طبق قانون‌های ساده ترکیب می‌شوند. توان جانهی چندگانه در توانایی ارائه‌ی استنباط‌های معتبر می‌باشد که تنها نیاز به روش‌های آماری استاندارد در مرحله‌ی تحلیل دارد. یعنی این روش تغییرات اضافی در نتیجه‌ی مقادیر گم شده تحت مدل جانهی را منعکس می‌کند.

البته مهم است که مدل جانهی برای مجموعه داده‌های خاص مناسب باشد. برای انواع معین داده‌ها، مدل‌های موجود می‌توانند مدل‌های جانهی مناسبی باشند. به عنوان مثال مدل مکانی عام (GLoM)[۱] (اولکین و تیت[۲]، ۱۹۶۱ و شی‌فر، ۱۹۹۷) یک مدل جانهی عام است که برای جانهی داده‌های ناقص با متغیرهای پیوسته و رسته‌ای استفاده می‌شود. این مدل فرض می‌کند که به شرط سطوح مختلف متغیرهای رسته‌ای، متغیرهای پیوسته یا تبدیلی از آنها دارای توزیع نرمال چند متغیره با میانگینی است که وابسته به سطح متغیرهای رسته‌ای می‌باشد ولی واریانسی دارد که در سطوح مختلف متغیرهای رسته‌ای ثابت است. اگر این فروض برقرار باشند این مدل، مدل مناسبی برای استفاده در جانهی است. متغیرهای نیم پیوسته با مقادیر گم شده نیز جمع آوری کننده داده‌ها را که نیاز به جانهی مقادیر گم شده این متغیرها دارد دچار مشکل می‌کنند. روش‌های جانهی که برای داده‌های نیم پیوسته مناسب نیست، ممکن است به طور جدی توزیع کناری متغیرها یا روابط آن‌ها با سایر کمیت‌ها را تحریف کند. یک روش جانهی معقول بایستی شکل توزیعی همانند آنچه در شکل‌های ۱—۱ و ۱—۲ نشان داده شده است را حفظ کند. روش‌های جانهی جدید بر اساس مدل سازی شامل جانهی چندگانه برای داده‌های چند متغیره پیوسته و رسته‌ای توسط شی‌فر (۱۹۹۷) گسترش داده شده است.

توجه به دلایلی که موجب گم شدن داده‌ها می‌شوند اهمیت زیادی دارد زیرا با استفاده از آن‌ها می‌توان روش‌های تحلیل واقعی را مشخص کرد. دلایل گوناگونی برای داده‌های گم شده وجود دارند. برخی داده‌ها ممکن است به طور کاملاً تصادفی گم شده باشند در حالی که برخی ممکن است به طور تصادفی و یا غیر تصادفی گم شده باشند. اغلب راهی برای حصول اطمینان از دلایل گم شدن

---

[۱] The General Location Model

[۲] Olkin and Tate

داده‌ها وجود ندارد. در مطالعات طولی، ممکن است افراد از همکاری در آزمایش انصراف دهند و یا در دوره‌های جمع آوری داده‌ها در دسترس نباشند. وقتی داده‌ها با استفاده از پرسشنامه جمع آوری می‌شوند، ممکن است افراد برخی از سؤال‌ها را به علت نداشتن وقت و یا علاقه پاسخ ندهند و کامل نکنند. ممکن است افراد به جواب دادن به سؤال‌هایی که آنها را مضطرب یا شرمنده کند تمایل نداشته باشند. ممکن است برخی سؤال‌ها خوب بیان نشده باشند و پاسخ گو نداند که چطور پاسخ دهد. حتی اگر پرسشنامه بسیار معتبر باشد و مشکلات فوق را نداشته باشد، ممکن است برخی سؤال‌ها در مورد بعضی از افراد مصداق نداشته باشد.

در تحلیل بعضی از داده‌ها عواملی که منجر به داده‌های گم شده می‌شوند، صراحتاً دخالت داده نمی‌شوند که در چنین مواردی فرضی تحت عنوان «قابل چشم پوشی» در نظر گرفته می‌شود. البته ممکن است با در نظر گرفتن یک متغیر نشانگر پاسخ (یک برای داده‌های ثبت شده و صفر برای داده‌های گم شده) عاملی را برای داده‌های گم شده در مدل وارد کنند.

عاملی که منجر به داده‌های گم شده می‌شود عموماً قابل چشم‌پوشی نیست. مثلاً در تحقیق درباره‌ی درآمد افراد ممکن است بی‌پاسخ ماندن، مربوط به مقدار خود درآمدها باشد یعنی مثلاً تمام درآمدهایی که از یک مقدار کمتر بوده‌اند مشاهده نشده‌اند و در این حالت دیگر نمی‌توان اظهار داشت که گم شدن داده‌ها تصادفی بوده بلکه عامل کم بودن درآمد باعث گم شدگی آن است و در نظر نگرفتن این عوامل در تحلیل داده‌ها می‌تواند ایجاد مشکل کند. این عوامل مکانیسم‌های گم شدگی را به وجود می آورند که در زیر آنها را مورد بحث قرار می‌دهیم.

۱–۲–۱ مکانیسم‌های گم شدگی

مجموعه‌ی داده‌ها را به صورت یک ماتریس $n \times p$ که سطرهای آن مشاهداتی از یک توزیع احتمالی چند متغیره‌اند و متغیرها از هم مستقل هستند، در نظر گیرید. این ماتریس شامل n سطر (تعداد مشاهدات) و p ستون (تعداد متغیرهای ثبت شده برای این واحدها) است.

فرض کنید ماتریس Y نشان دهنده‌ی ماتریس $n \times p$ داده‌ها باشد، که به طور کامل مشاهده نشده‌اند، و y_i سطر i ام Y برای $i = ۱, ۲, \ldots, n$ باشد. با فرض داشتن توزیع یکسان و مستقل، تابع

چگالی یا تابع احتمال داده‌های کامل را به صورت زیر نشان می‌دهیم

$$P(Y \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta), \qquad (۵.۱)$$

که f تابع چگالی یا تابع احتمال یک سطر است و θ بردار پارامترهای مجهول است. همچنین فرض کنید بخش مشاهده شده‌ی Y را به وسیله‌ی Y_{obs} و بخش گم شده‌ی Y را به وسیله‌ی Y_{mis} نشان دهیم، به طوری که $Y = (Y_{obs}, Y_{mis})$. طبق تعریف رابین (۱۹۷۶) ، چنانچه گم شدگی بستگی به Y_{obs} داشته باشد، ولی به شرط Y_{obs} به Y_{mis} بستگی نداشته باشد، گم شدگی به صورت تصادفی است.

فرض کنید U و V دو متغیر باشند. همچنین فرض کنید واحدهایی از U که مشاهده شده‌اند و مقدار مشاهده‌ی آن u باشد، مد نظر هستند. گم شدگی تصادفی V به این معنی است که بین این واحدها، توزیع V در بین بخشی از V که مشاهده شده و بخشی که مشاهده نشده به شرط $U = u$، یکسان است. یعنی احتمال ثبت V فقط به مقادیر U بستگی داشته باشد ولی به مقادیر خود V بستگی نداشته باشد.

گم شدگی تصادفی به این معنی نیست که مقادیر داده‌های گم شده به صورت نمونه‌ی تصادفی ساده از کل داده‌ها هستند. حالت بعدی گم شدگی، گم شدگی کاملاً تصادفی نامیده می‌شود. گم شدگی کاملاً تصادفی تنها حالت خاصی از گم شدگی تصادفی است. گم شدگی تصادفی محدودیت کمتری نسبت به گم شدگی کاملاً تصادفی دارد، زیرا تنها نیاز به این دارد که مقادیر گم شده همانند یک نمونه‌ی تصادفی از همه‌ی مقادیر داخل زیر رده‌های تعریف شده به وسیله‌ی مقادیر داده‌های مشاهده شده رفتار کنند. به عبارت دیگر، گم شدگی تصادفی اجازه می‌دهد که احتمال گم شدن یک داده به شرط تعداد کمیت‌هایی که مشاهده شده‌اند، بستگی به خود داده نداشته باشد.

رابین (۱۹۷۶) گم شدگی تصادفی را برحسب مدل احتمالاتی برای گم شدگی تعریف کرد. فرض کنید $R = (R_{ij})$ یک ماتریس نشانگر $n \times p$ باشد که اعضای آن به صورت زیر تعریف می‌شوند:

$$R_{ij} = \begin{cases} ۱ & \text{اگر } y_{ij} \text{ مشاهده شود} \\ ۰ & \text{در غیر این صورت} \end{cases}$$

به طور کلی نباید انتظار داشت که توزیع R با Y نامرتبط باشد، بنابراین می‌توان یک مدل احتمالاتی برای R به صورت $P(R \mid Y, \xi)$ در نظر گرفت، که بستگی به Y و پارامترهای مجهول ξ دارد. فرض گم شدگی تصادفی به این معنی است که این توزیع بستگی به Y_{mis} ندارد، یعنی

$$P(R \mid Y_{obs}, Y_{mis}, \xi) = P(R \mid Y_{obs}, \xi). \qquad (۶.۱)$$

مکانیسم گم شدگی غیر تصادفی نه حالت گم شدگی تصادفی و نه حالت گم شدگی کاملاً تصادفی است در این حالت مکانیسم گم شدن به داده‌هایی که ممکن است گم شوند وابسته است. در دو حالت گم شدگی کاملاً تصادفی و گم شدگی تصادفی می‌توان از مکانیسم گم شدن داده‌ها چشم پوشی کرد ولی در گم شدن غیر تصادفی نمی‌توان از عامل گم شدن داده‌ها چشم پوشی کرد که این حالت به مدل غیر قابل چشم پوشی موسوم است.

مجزا بودن پارامترها

برای کنکاش بیشتر در این زمینه، نیاز به این فرض داریم که پارامترهای مدل داده‌ها θ و پارامترهای مکانیسم گم شدگی ξ مجزا باشند. از دیدگاه فراوانی گرا، این بدان معنی است که بایستی فضای توأم پارامتر (θ, ξ) حاصل ضرب دکارتی فضاهای پارامترهای θ و ξ باشد. به این معنی که θ به صورت تابعی به ξ وابسته نباشد. از دیدگاه بیزی، این بدان معنی است که بایستی هر توزیع توأم پیشین که برای (θ, ξ) بکار برده می‌شود، حاصلضرب توزیع‌های حاشیه‌ای پیشین θ و ξ باشد. در بسیاری از شرایط، منطقی است که θ حاوی مقدار کمی اطلاعات در مورد ξ و بالعکس باشد. اگر فرض گم شدگی تصادفی و مجزا بودن همزمان اتفاق بیافتند، آنگاه گوییم مکانیسم گم شدن داده‌ها قابل چشم پوشی است (لیتل[۱] و رابین، ۲۰۰۲؛ رابین ۱۹۸۷).

۱—۲—۲ تابع درستنمایی و توزیع پسین داده‌های مشاهده شده

تابع درستنمایی داده‌های مشاهده شده

بنا بر رابین (۱۹۷۶) و لیتل و رابین (۲۰۰۲)، می‌توان نشان داد که تحت فرض گم شدگی قابل چشم پوشی، هنگامی که براساس تابع درستنمایی و یا روش بیزی در مورد θ استنباط می‌کنیم، نیازی به در نظر گرفتن مدل برای R و در نظر گرفتن پارامترهای ξ نداریم.
چون داده‌های مشاهده شده شامل Y_{obs} و R هستند، توزیع احتمالاتی داده‌های مشاهده شده به صورت

[۱] Little

زیر می‌باشد

$$P(R, Y_{obs} \mid \theta, \xi) = \int P(R, Y \mid \theta, \xi) dY_{mis} \quad (۷.۱)$$
$$= \int P(R \mid Y, \xi) P(Y \mid \theta) dY_{mis},$$

که برای توزیع‌های گسسته این انتگرال چندگانه به جمع چندگانه تبدیل می‌شود. تحت فرض گم شدگی تصادفی، رابطه‌ی فوق به رابطه‌ی زیر تبدیل می‌شود

$$P(R, Y_{obs} \mid \theta, \xi) = P(R \mid Y_{obs}, \xi) \int P(Y \mid \theta) dY_{mis} \quad (۸.۱)$$
$$= P(R \mid Y_{obs}, \xi) P(Y_{obs} \mid \theta).$$

بنابراین می‌توان تحت فرض گم شدگی تصادفی، تابع درستنمایی داده‌های مشاهده شده را به دو جزء افراز کرد که یک جزء مربوط به پارامتر مورد علاقه θ و جزء دیگر مربوط به پارامتر مزاحم ξ می‌باشد. هنگامی که دو پارامتر مجزا هستند، استنباط بر اساس درستنمایی در مورد θ تحت تأثیر ξ یا $P(R \mid Y_{obs}, \xi)$ قرار نمی‌گیرد. بنابراین برآورد ماکسیمم درستنمایی θ، آزمون‌های نسبت درستنمایی مربوط به θ و غیره را می‌توان بدون در نظر گرفتن مکانیسم گم شدن داده‌ها انجام داد؛ یعنی، از مکانیسم گم شدگی داده‌ها چشم پوشی می‌کنیم.

عامل مربوط به θ در رابطه‌ی (۸.۱) (یا به طور دقیق‌تر، هر تابع متناسب با این عامل) را همان طور که در لیتل و رابین (۲۰۰۲) ذکر شده، می‌توان به عنوان درستنمایی که از مکانیسم گم شدن داده‌ها چشم پوشی می‌کند در نظر گرفت، یعنی

$$L(\theta \mid Y_{obs}) \propto P(Y_{obs} \mid \theta) \quad (۹.۱)$$

برای اختصار، رابطه‌ی (۹.۱) را به عنوان درستنمایی داده‌های مشاهده شده در نظر می‌گیریم. به طور کلی چون فرض قابلیت چشم پوشی را در نظر می‌گیریم، نیاز به کار کردن با تابع کامل (۸.۱) نداریم.
توجه کنید که در نگاه اول، به نظر می‌رسد تجزیه به عوامل (۸.۱) شامل فرض‌های ذکر شده در مورد مکانیسم گم شدگی نیست. توزیع توأم دو متغیر تصادفی مانند، Z_1 و Z_2، را همواره می‌توان به صورت حاصل ضرب توزیع حاشیه‌ای Z_1 و توزیع شرطی Z_2 به شرط Z_1 نوشت. یک تفاوت مهم بین این