بسم الله الرحمن الرحيم

دانشگاه شهید باهنر کرمان

دانشکده ریاضی و کامپیوتر

گروه آمار

پایان نامه تحصیلی برای دریافت درجه کارشناسی ارشد آمار ریاضی

بوت استرپ در رگرسیون فازی

استاد راهنما:

دکتر علیرضا عربپور

مؤلف:

فرزانه مرادی

خردادماه ۸۸

دانشگاه شهید باهنر کرمان

این پایان نامه

به عنوان یکی از شرایط احراز کارشناسی ارشد

به

بخش ریاضی – دانشکده ریاضی و کامپیوتر

دانشگاه شهید باهنر کرمان

تسلیم شده است و هیچگونه مدرکی به عنوان فراغت از تحصیل دوره مزبور شناخته نمی شود .

دانشجو : فرزانه مرادی

استاد راهنما : دکترعلیرضا عربپور

داور ۱ : دکترماشااله ماشین چی

داور ۲ : دکترمحمدعلی یعقوبی

نماینده تحصیلات تکمیلی دانشگاه : دکترسید ناصر حسینی

تقدیم به:

پدر و مادر دلسوزم که در تمامی مراحل پشتیبانم بودند.

با تشکر از:

زحمات بی‌دریغ استاد گرامی جناب آقای دکتر عرب‌پور

و تمام عزیزانی که در تهیه این پایان‌نامه مرا یاری نمودند.

همچنین از قطب سیستم‌های فازی به خاطر حمایت مالی

جزئی تشکر می‌کنم.

چکیده

به منظور پیدا کردن رابطه خطی مناسب بین یک متغیر وابسته و چند متغیر مستقل در یک محیط فازی مدل‌های رگرسیون خطی فازی به کار برده می‌شوند. از طرفی برای برازش مدل وقتی با داده‌های کم با جامعه نامعلوم روبه رو هستیم از بازنمونه‌گیری استفاده می‌کنیم. یکی از کارآمدترین روش‌های بازنمونه‌گیری، بوت‌استرپ است. در این پایان‌نامه، ابتدا پارامترهای مدل رگرسیون خطی فازی را با روش کمترین مربعات و برنامه‌ریزی ریاضی برآورد کرده‌ایم سپس بوت‌استرپ را به عنوان یک روش بازنمونه گیری برای بهبود برآوردهای مدل رگرسیون خطی فازی به کار برده‌ایم و با چند مثال عددی کارایی این روش را با سایر روش‌های رگرسیون کمترین مربعات فازی مورد مقایسه قرار داده‌ایم. همچنین اجرای آزمون فرض و به دست آوردن فاصله اطمینان به روش بوت‌استرپ برای پارامترهای مدل رگرسیون خطی فازی شرح داده شده است.

در فصل اول این پایان‌نامه مفاهیم و تعاریف مقدماتی موردنیاز آورده شده است. در فصل دوم مروری بر تاریخچه رگرسیون فازی داریم و به بیان چند روش کمترین مربعات خطا و برنامه‌ریزی ریاضی می‌پردازیم، همچنین یک روش عددی نیز برای رگرسیون فازی معرفی می‌شود. در فصل سوم پایان‌نامه استفاده از روش بوت‌استرپ در مدل‌های رگرسیون خطی فازی با ذکر مثال‌هایی شرح داده می‌شود. و در نهایت فصل چهارم اختصاص دارد به آزمون فرض به روش بوت‌استرپ و مثال‌های مربوط به آن، همچنین به دست آوردن فاصله اطمینان بوت‌استرپ برای پارامترهای مدل رگرسیون فازی نیز با ذکر مثال‌هایی بیان شده است.

لازم به ذکر است مطالبی که ساخته ذهن نگارنده است با [*] نشان داده شده است.

فهرست مطالب

عنوان صفحه

فصل ۱

مفاهیم و تعاریف مقدماتی

۱-۱ مجموعه‌های فازی

مفهوم مجموعه و نظریه مجموعه ها ابزارهای قوی در ریاضیات هستند. در نظریه مجموعه‌ها، مجموعه به صورت گردایه ای از اشیاء کاملاً مشخص تعریف می شود و عضویت یا عدم عضویت یک شیء در مجموعه قطعی است، به عنوان مثال مجموعه افرادی در یک کلاس که قد آنها بالای ۱۸۰ سانتیمتر است. در بسیاری از مسائل واقعی حدود مجموعه ها کاملاً مشخص نیست مانند مجموعه افراد قدبلند در یک کلاس. در سال ۱۹۶۵ پرفسور زاده برای مجموعه‌هایی که حدودشان کاملاً مشخص نیست، مفهوم مجموعه فازی و عضویت جزئی را مطرح کرد. او مفهوم مجموعه فازی را به صورت گردایه‌ای از اشیاء معرفی کرد که با درجه‌ای بین ۰ و ۱ متعلق به مجموعه هستند که درجه ۱ بیانگر عضویت کامل و درجه ۰ بیانگر عدم عضویت کامل در مجموعه است. این تعریف با به کارگیری مفهوم تابع عضویت انجام شد که به هر شیء عددی از بازه [۱و۰] نسبت می‌دهد که بیانگر درجه عضویت آن شیء در مجموعه فازی است [۷].

فرض کنید X مجموعه مرجع و A یک زیرمجموعه از X باشد. زیرمجموعه A از X را می توان با استفاده از مفهوم تابع مشخصه بیان کرد. به عبارتی تابع مشخصه A، به صورت تابعی از X به {۱و۰} است که اینگونه تعریف می شود:

$$\mu_A(x) = \begin{cases} 1 & , x \in A \\ 0 & , x \notin A \end{cases}$$

که در آن برای هر $x \in X$، $\mu_A(x)$ تنها یکی از مقادیر ۰ یا ۱ را خواهد گرفت. حال اگر برد تابع μ_A را از مجموعه {۱و۰} به بازه [۱و۰] توسیع دهیم، تابعی به دست خواهیم آورد که به هر $x \in X$، عددی را در بازه [۱و۰] نسبت می دهد. در چنین وضعیتی در مورد عضویت یا عدم عضویت x در A با عدم قطعیت مواجهیم. بنابراین در اینجا به نوعی مفهوم عضویت یک عنصر را گسترش داده ایم. حال برای بیان کلی این موضوع تعاریف زیر را ارائه می دهیم:

تعریف ۱: فرض کنید X مجموعه ای ناتهی باشد. هر زیرمجموعه فازی $\widetilde{A}$ از X توسط یک تابع عضویت $\mu_{\widetilde{A}} : X \to [0,1]$ مشخص می شود که در آن برای هر $x \in X$ مقدار $\mu_{\widetilde{A}}(x)$ در بازه [0,1] میزان عضویت x را در $\widetilde{A}$ نشان می دهد.

از این به بعد برای راحتی زیرمجموعه فازی $\widetilde{A}$ از X را با تابع عضویت $\widetilde{A} : X \to [0,1]$ و درجه عضویت x در $\widetilde{A}$ را با $\widetilde{A}(x)$ نشان می دهیم. همچنین مجموعه تمام زیرمجموعه‌های فازی X را با F(X) نشان می دهیم. به عبارتی دیگر:

$$F(X) = \{\widetilde{A} \mid \widetilde{A} : X \to [0,1]\}$$

تعریف ۲: فرض کنید $\widetilde{A} \in F(X)$. تکیه گاه $\widetilde{A}$ که با $Supp(\widetilde{A})$ نشان داده می شود به صورت زیر تعریف می شود:

$$Supp(\widetilde{A}) = \overline{\{x \in X \mid \widetilde{A}(x) > 0\}}$$

تعریف ۳: فرض کنید $\widetilde{A} \in F(X)$. برای هر $0 < h \leq 1$، h-برش $\widetilde{A}$ که با $[\widetilde{A}]^h$ نشان داده می شود به صورت زیر تعریف می شود:

$$[\widetilde{A}]^h = \{x \in X \mid \widetilde{A}(x) \geq h\}$$

تعریف ٤: کمیت فازی $\widetilde{A}$ را عدد فازی می گوییم هرگاه در سه شرط زیر صدق کند:

۱) دقیقاً یک $x \in \Re$ وجود داشته باشد که $\widetilde{A}(x) = 1$.

۲) $\widetilde{A}$ نیم پیوسته بالایی باشد یعنی همه h-برشهای آن بازه بسته باشند.

۳) تکیه گاه $\widetilde{A}$ کراندار باشد.

تعریف ٥: کمیت فازی $\widetilde{A}$ را عدد فازی $L-R$ گوییم هرگاه تابع عضویت $\widetilde{A}$ به صورت زیر باشد:

$$\widetilde{A}(X) = \begin{cases} L(\frac{a-x}{\alpha}) & , x \le a \\ R(\frac{x-a}{\beta}) & , x \ge a \end{cases}$$

که در آن $\alpha, \beta > 0$ و $L, R : [0,\infty) \to [0,1]$ توابعی پیوسته، نزولی و روی بازه $[0,1]$ معکوس پذیر هستند، $L(0) = R(0) = 1$ و $L(1) = R(1) = 0$ به توابع L و R توابع مرجع می‌گوییم. همچنین به a, α, β به ترتیب مرکز، پهنای چپ و پهنای راست عدد فازی $\widetilde{A}$ می‌گوییم. برای سادگی $\widetilde{A}$ را با $\widetilde{A} = (a, \alpha, \beta)_{LR}$ نشان می‌دهیم. اگر $L = R$ و $\alpha = \beta$، $\widetilde{A}$ را عدد فازی متقارن می گوییم و آن را با $\widetilde{A} = (a, \alpha)_L$ نشان میدهیم. به عدد فازی با توابع مرجع $L(x) = R(x) = \max\{0, 1-x\}$ عدد فازی مثلثی می گوییم و آن را با $\widetilde{A} = (a, \alpha, \beta)$ نشان می دهیم [۷].

تعریف زیر به دلیل اهمیت زیادی که در فرهنگ ریاضیات فازی دارد به اصل گسترش شهرت یافته است [۲].

اصل گسترش: فرض کنید X و Y دو مجموعه ناتهی و $f : X \to Y$ یک تابع و $A \in F(X)$ باشند. در این صورت $B = f(A) \in F(Y)$ را به صورت زیر تعریف می‌کنیم:

$$B(y) = \sup_{y=f(x)} A(x)$$

که در آن سوپریمم روی مجموعه تهی صفر تعریف می‌شود.

تعریف ٦ : فرض کنید $\widetilde{A}_i \in F(X_i), i = 1,...,n$ و $X = X_1 \times ... \times X_n$ حاصل ضرب دکارتی X_iها باشد. در این صورت حاصل ضرب دکارتی $\widetilde{A}_1 \times ... \times \widetilde{A}_n$ نیز یک زیرمجموعه فازی از X است که به صورت زیر تعریف می شود:

$$(\widetilde{A}_1 \times ... \times \widetilde{A}_n)(x_1,...,x_n) = \min\{\widetilde{A}_1(x_1),...,\widetilde{A}_n(x_n)\}$$

با توجه به تعریف فوق و اصل گسترش می توان اعمال دوتایی R را به F(R) به صورت زیر گسترش داد:

(به جای sup و inf به ترتیب از نمادهای $\vee$ و $\wedge$ استفاده می کنیم.)

تعریف ۷ : اگر o یک عمل دوتایی روی R باشد، آنگاه این عمل را روی F(R) به صورت زیر تعریف می کنیم:

$$AoB(x) = \sup_{aob=x} \{\inf(A(a), B(b))\}$$
$$= \vee_{aob=x} \{A(a) \wedge B(b)\}$$

که در آن سوپریمم روی مجموعه تهی صفر تعریف می‌شود [۲]. به عنوان مثال داریم:

$$(A + B)(x) = \vee_{a+b=x} \{A(a) \wedge B(b)\}$$
$$(A.B)(x) = \vee_{a.b=x} \{A(a) \wedge B(b)\}$$

دقت کنید که با توجه به اصل گسترش می‌توان دید که

$$(-A)(x) = \vee_{y=-x} \{A(y)\} = A(-x)$$

همچنین به آسانی می‌توان اثبات کرد که

$$[A + (-B)](x) = (A - B)(x)$$

عدد فازی مثلثی را که قبلاً به طور خلاصه معرفی شد، می‌توان اینگونه تعریف کرد:

تعریف ۸ : یک عدد فازی مثلثی یک عدد فازی است با تابع عضویت زیر:

$$A(x)=\begin{cases} 0 & if \quad x<a \\ \dfrac{x-a}{b-a} & if \quad a\le x\le b \\ \dfrac{x-c}{b-c} & if \quad b<x\le c \\ 0 & if \quad c<x \end{cases}$$

در این تعریف عدد فازی مثلثی را با سه تایی (a,b,c) بیان می کنند که به ترتیب نقطه پایینی، مرکز و نقطه بالایی عدد فازی مثلثی را نشان می دهند. برای راحت تر جمع بستن دو عدد فازی مثلثی، قضیه زیر را می آوریم.

قضیه ۱: برای دو عدد فازی مثلثی $\widetilde{A}=(a,b,c)$ و $\widetilde{B}=(d,e,f)$ خواهیم داشت:

$$\widetilde{A}+\widetilde{B}=(a+d;b+e,c+f)$$

بدین معنی که مجموع دو عدد فازی مثلثی، یک عدد فازی مثلثی است که مرکز و پهناهای چپ و راست آن به ترتیب از جمع بستن مراکز و پهناهای چپ و راست آن دو عدد به دست می آید [۳۸].

برای ضرب آسانتر اعداد فازی $L-R$ قضیه زیر مورد استفاده خواهد بود.

قضیه ۲: برای دو عدد فازی $\widetilde{A}=(a,\alpha,\beta)_{LR}$ و $\widetilde{B}=(b,\gamma,\delta)_{LR}$ که به ترتیب مرکز، پهنای چپ و پهنای راست این دو عدد را نشان می‌دهد، اگر $\widetilde{A}>0,\widetilde{B}>0$ به طور تقریبی خواهیم داشت:

$$\widetilde{A}.\widetilde{B}\approx(ab,a\gamma+b\alpha,a\delta+b\beta)$$

از آنجایی که اعداد فازی $L-R$ تعمیمی از اعداد فازی مثلثی هستند، از قضیه بالا می‌توان این نتیجه را گرفت که حاصلضرب دو عدد فازی مثلثی به طور تقریبی برابر با یک عدد فازی مثلثی است که مرکز، مقدار پایینی و مقدار بالایی آن بر حسب مراکز، مقادیر پایینی و مقادیر بالایی آن دو عدد به دست می‌آید. در حالی که حاصلضرب دقیق دو عدد فازی مثلثی لزوماً یک عدد فازی مثلثی نخواهد شد [۲۰].

در ادامه این فصل به طور خلاصه مدل‌های رگرسیون ساده و چندگانه را برای داده‌های جقیقی معرفی می‌کنیم و به بررسی برخی ویژگی‌های مدل رگرسیونی می‌پردازیم.

۱-۲ رگرسیون

بسیاری از پیشامدها و پدیده‌هایی که در جهان اتفاق می‌افتند، تابع برخی از پیشامدهای دیگر می‌باشند. یکی از هدف های مهم در تحقیقات علمی، کشف وجود یا عدم وجود رابطه بین پدیده هاست. یکی از روش‌های آماری که به طور گسترده برای این منظور به کار می رود روش تحلیل رگرسیون می‌باشد [۳۷]. در واقع این روش برای بررسی رابطه بین دو یا چند متغیر استفاده می‌شود به طوری که یک متغیر از روی دیگری یا بقیه می‌تواند پیش‌بینی شود. در صورتی که در یک مدل رگرسیون y به عنوان متغیر وابسته و x به عنوان متغیر مستقل فرض شود، معادله خط مستقیمی که این دو متغیر را به یکدیگر ارتباط می‌دهد به صورت زیر است:

$$y = a_0 + a_1 x \qquad (۱-۱)$$

رابطه (۱-۱) را زمانی می توان به کار برد که تمام داده ها روی یک خط راست واقع شده باشند. از آنجا که در عمل چنین نیست، رابطه (۱-۱) باید تعدیل شود. اختلاف بین y (متغیر وابسته) و

خط مستقیم $a_0 + a_1 x$ مقدار خطای آماری را بیان می‌کند و با ε نمایش داده می‌شود. این خطا ضعف مدل در برازش دقیق داده‌ها را نشان می‌دهد. مدل اصلاح شده به صورت زیر است:

$$y = a_0 + a_1 x + \varepsilon$$

که مدل رگرسیون خطی ساده نامیده می‌شود. مدل های مشابه نیز که تنها دارای یک متغیر مستقل است، مدل رگرسیون ساده نامیده می‌شوند و مدل:

$$y = a_0 + a_1 x_1 + ... + a_k x_k + \varepsilon$$

که شامل k متغیر مستقل است، مدل رگرسیون خطی چندگانه نامیده می‌شود. عبارت خطی نشان می‌دهد که مدل بر حسب پارامترها $(a_0, a_1, ..., a_k)$ خطی است نه اینکه y یک تابع خطی از xها باشد. مدل های بسیاری وجود دارند که y را به صورت غیرخطی به xها ارتباط می‌دهند و می‌توان آنها را تا زمانی که بر حسب پارامترها خطی می‌باشند، به صورت مدل‌های رگرسیون خطی بیان نمود. برای مثال مدل‌های زیر، مدل رگرسیون خطی ساده هستند:

$$y = a_0 + a_1 \log x + \varepsilon$$
$$y = a_0 + a_1 x^2 + \varepsilon$$

صحت یک مدل رگرسیونی به برقرار بودن شرایط زیر بستگی دارد:

۱- میانگین خطاها برابر صفر و واریانس خطاها برابر یک مقدار ثابت باشد.

$$E(\varepsilon_i) = 0 \quad , \quad Var(\varepsilon_i) = \sigma^2$$

۲- خطاها نسبت به هم ناهمبسته باشند.

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad , i \neq j$$

۳- خطاها دارای توزیع نرمال با میانگین صفر و واریانس ثابتی باشد.

$$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$$

در ادامه مدل رگرسیون خطی ساده را تحت فرض‌های فوق مورد بررسی قرار می‌دهیم [۳۷].

با فرض وجود n جفت داده به صورت : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ و با استفاده از روش حداقل مربعات خطا می‌توان پارامترهای مدل را به صورت زیر برآورد کرد:

خطای هر مشاهده به صورت زیر به دست می‌آید:

$$\varepsilon_i = y_i - a_0 - a_1 x_i \qquad , i = 1,2,\ldots,n$$

بنابراین می‌توان مجموع مربعات خطا را به فرم زیر نوشت:

$$s(a_0, a_1) = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

در صورتی که $\hat{a}_1, \hat{a}_0$ برآورد پارامترهای a_1, a_0 به روش کمترین مربعات خطا باشند، آنگاه باید داشته باشیم:

$$\left.\frac{\partial s}{\partial a_0}\right|_{(\hat{a}_0,\hat{a}_1)} = -2\sum_{i=1}^{n}(y_i - \hat{a}_0 - \hat{a}_1 x_i) = 0$$
$$\left.\frac{\partial s}{\partial a_1}\right|_{(\hat{a}_0,\hat{a}_1)} = -2\sum_{i=1}^{n}(y_i - \hat{a}_0 - \hat{a}_1 x_i)x_i = 0$$

با ساده‌سازی روابط بالا داریم:

$$n\hat{a}_0 + \hat{a}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$\hat{a}_0 \sum_{i=1}^{n} x_i + \hat{a}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

از حل دستگاه معادلات فوق که معادلات نرمال نامیده می‌شوند، برآورد پارامترها به صورت زیر به دست می آیند:

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x} \quad , \quad \hat{a}_1 = \frac{S_{xy}}{S_{xx}}$$

که در آن:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$$
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$$

و

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad , \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

پس از برآورد پارامترها، مدل رگرسیونی به صورت رابطه زیر در می‌آید:

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x$$

بعد از برازش خط به داده‌ها به روش مینیمم مربعات خطا، پرسش‌های زیر مطرح می‌گردد:

۱- کیفیت برازش مدل چگونه است؟

۲- آیا مدل به دست آمده برای پیش‌بینی مناسب است؟

۳- آیا فرضیات اولیه (مانند ثابت بودن واریانس و ناهمبسته بودن خطاها) برقرارند و اگر چنین نیست، نقض آنها تا چه حد جدی است؟

در این رابطه چند روش برای بررسی موارد فوق توضیح داده می‌شود:

- آزمون فرض (با انجام آزمون فرض درباره پارامترهای مدل می توان نتایجی را به دست آورد که از آنها برای بررسی فرضیات فوق استفاده می‌شود.)
- ضریب تعیین R^2 (که میزان مشارکت رگرسیون را مشخص می‌کند.)

- رسم نمودارهای مانده (پس از برازش مدل، رسم نمودارهای مربوط به مانده‌ها می‌تواند برای روشن کردن کاستی‌های موجود در آن مفید باشد.)

با انجام آزمون فرض درباره پارامترها می توان نتایجی را به دست آورد که از آنها برای بررسی فرضیاتی که در مورد کیفیت برازش مدل مطرح شده‌است، استفاده می‌شود.

به عنوان مثال، در رگرسیون خطی ساده با معادله

$$y_i = a_0 + a_1 x_i + \varepsilon_i \quad , \ i = 1,2,...,n$$

آزمون اعتبار رگرسیون می تواند به صورت زیر باشد :

$$\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases} \qquad (۱-۴)$$

در صورتی که فرض H_0 رد نشود، یعنی فرض $a_1 = 0$ پذیرفته شود، می‌توان نتیجه گرفت که بین x و y رابطه خطی وجود ندارد. البته رد فرض H_0 نیز به این مفهوم نیست که حتماً بین x و y ارتباط خطی وجود دارد زیرا در برخی حالت‌های غیرخطی هم این فرض رد می‌شود.

جهت آزمون فرض‌های (۴-۱) از آماره t_0 استفاده می شود [۳۷]. بدین صورت که در سطح خطای نوع اول α، فرض $H_0 : a_1 = 0$ هنگامی رد می‌شود که :

$$|t_0| > t_{(n-2,\alpha/2)}$$

که در آن :

$$t_0 = \frac{\hat{a}_1 - \hat{a}_0}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

البته در روابط بالا σ^2 که میزان تغییرات خطا را نشان می‌دهد مقداری مجهول است و می‌توان از برآورد آن که MSE نامیده می‌شود و به صورت زیر محاسبه می‌گردد استفاده نمود:

$$\hat{\delta}^2 = \frac{SSE}{n-2} = MSE$$

که در آن :

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} - \hat{a}_1 S_{xy}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

در ادامه این فصل به معرفی روش بوت‌استرپ می‌پردازیم و الگوریتمی ارائه می‌کنیم که به فهم آسانتر طرز کار روش بوت‌استرپ کمک می‌کند.

۱-۳ بوت‌استرپ[1]

در استنباط آماری معمولی ما با جامعه‌ای سروکار داریم که توزیع آن معلوم می‌باشد و داده‌ها به اندازه کافی در دسترس هستند و اگر داده‌ها کم باشد در حالی که توزیع جامعه معلوم است، از شبیه‌سازی استفاده کرده و آنالیز می‌کنیم. اما اگر توزیع جامعه نامعلوم باشد در حالی که داده‌ها به اندازه کافی زیاد هستند از آنالیز مجانبی (قضیه حد مرکزی) استفاده می‌کنیم. اما بدترین حالت زمانی اتفاق می‌افتد که هم توزیع جامعه نامعلوم باشد و هم داده‌ها کم باشند. در این حالت از بازنمونه‌گیری[2] استفاده می‌کنیم. بوت‌استرپ نوعی شبیه‌سازی است که در سال ۱۹۷۹ توسط افرون[3] [۲۳] ارائه شد.

---

[1] .Bootstrap
[2] .Resampling
[3] .Efron

اسم بوت‌استرپ از ماجرای «بارن مانچوسن[۴]» نوشته رادلف اریک راسپه[۵] در اواخر قرن ۱۸ گرفته شده است که وقتی خودش را در ته دریاچه بدون هیچ کمک و وسیله‌ای می‌بیند با استفاده از بند پوتینش خود را بالا می‌کشد. در اینجا نیز موقعی که کمترین اطلاعات را داریم یعنی نه توزیع جامعه معلوم است نه اینکه داده‌ها به اندازه کافی هستند، سعی می‌کنیم با استفاده از نمونه کوچکی که در دست داریم استنباط خوب و معتبری ارائه دهیم [۳].

در این روش آنالیز روی داده‌های واقعی است نمونه‌های کوچک ولی معتبر و موثق که قسمت مهم یک تحقیق است و در عمل بر پایه شبیه‌سازی است. اما در برخی مثال‌ها نیازی به شبیه‌سازی نیست، در واقع شبیه‌سازی از توزیع جامعه می‌باشد ولی اگر توزیع جامعه نامعلوم باشد به تعداد دلخواه نمونه از نمونه اصلی برداشته و میانگین آنها را در نظر می‌گیریم پس نیازی به توزیع جامعه نیست.

در این روش چون توزیع جامعه نامعلوم است ما از توزیع تجربی نمونه‌هایی که در دست داریم شبیه‌سازی می‌کنیم که معادل با این است که به اندازه دلخواه نمونه (باجایگذاری) از نمونه اصلی برداشته و میانگین این نمونه‌ها را به عنوان برآوردی بهتر منظور می‌کنیم. این تکنیک‌ها بدون استفاده از کامپیوتر محاسبات طولانی دارد و کار خسته کننده‌ای می‌باشد لذا در اینجا از دستورات برنامه R برای به کار بردن این روش استفاده شده است.

نمونه‌گیری بوت‌استرپ زمینه ریاضی قوی دارد اما بر پایه کامپیوتر و روش‌های استنباط آماری است که می‌تواند به خیلی از سؤالات آمار بدون فرمول جواب دهد[۱].

---

[4] .Barron Munchausen
[5] .Rudolf Erich Raspe