

در آنالیز داده‌ها بهترین حالت زمانی است که توزیع داده‌ها معلوم و نمونه به اندازه کافی بزرگ باشد ولی اگر توزیع معلوم باشد و داده‌ها کم باشد از شبیه‌سازی استفاده می‌کنیم اما اگر توزیع جامعه نامعلوم باشد و نمونه به اندازه کافی بزرگ باشد ما از آنالیز مجانبی استفاده می‌کنیم اما بدترین حالت زمانی اتفاق می‌افتد که هم توزیع جامعه نامعلوم باشد و هم داده‌ها کافی نباشد.

زمانی که کمترین وسیله را در اختیار داریم یعنی نه توزیع جامعه معلوم است و نه اینکه داده‌ها به اندازه کافی هستند، ولی سعی می‌کنیم استنباط خوب و معتبر ارائه دهیم می‌توان روش باز نمونه‌گیری (بوت استرپ) استفاده نمود نمونه‌گیری بوت استرپ^۱ توسط افرون در سال ۱۹۷۹ ارائه شد و مدتها روی آن کار شد اسم بوت استرپ از ماجرای "بارن مانچوسن" نوشته رادلف اریک راسیه در اواخر قرن هیجده گرفته شده است که وقتی خودش را در ته دریاچه بدون هیچ کمک و وسیله‌ای می‌بیند با استفاده از بند پوتین‌اش خود را بالا می‌کشد، بدون هیچ کلکی!

در اینجا نیز موقعی که کمترین وسیله را در اختیار داریم یعنی نه توزیع جامعه معلوم است و نه اینکه داده‌ها به اندازه کافی هستند؛ ولی سعی می‌کنیم استنباط خوب و معتبر ارائه دهیم.

نمونه‌گیری بوت استرپ زمینه ریاضی قوی دارد اما برپایه کامپیوتر و روشهای استنباط آماری است که می‌تواند به خیلی از سئوالات آمار بدون فرمول جواب دهد. در اینجا 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 درواقع 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.

نمونه‌گیری بوت استرپ دارای کمترین فرضیات است، در اینجا ممکن است بدترین شرایط را داشته باشیم یعنی اینکه هم توزیع جامعه نامعلوم است و هم داده‌ها خیلی کم است یعنی ما اطلاعات زیادی از جامعه نداریم؛ فقط چیزی که باید خیلی توجه کنیم این است که نمونه‌ها باید یک نمایش خوب از جامعه شناخته نشده است.

ایده روش بوت استرپ، یک ایده قدیمی است و از آنجا نشأت می‌گیرد که به عنوان مثال، اگر بخواهیم میانگین جامعه

$$\mu = \int x dF(x), \quad (1 - 11)$$

که تابعی از تابع توزیع جامعه‌ی F است را برآورد کنیم باید به جای F ، از تابع توزیع تجربی، $\hat{F}$ ، استفاده کنیم که در مورد این مثال

$$\bar{X} = \int x d\hat{F}(x), \quad (2 - 11)$$

یا میانگین نمونه می‌باشد. لازم به ذکر است که توزیع تجربی، اندازه‌ی احتمالی است که به یک مجموعه، اندازه‌ای برابر با نسبتی از مقادیر نمونه‌ای که در آن مجموعه واقع هستند را اختصاص می‌دهد. اما این استدلال همیشه قابل اجرا نیست. یک مثال نقض برای آن، تابع چگالی احتمال است که تابعی از F می‌باشد. با این حال استفاده از روش بوت استرپ، پیش از آنکه

^۱ . Bootstrapping

افرون (1979) ویژگی ها و خواص قابل توجه آن را بدست آورد و آن را به این اسم بنامد به گستردگی اکنون نبوده است. با اصلاح و بیان دقیق مفهوم بوت استرپ، این روش در بسیاری از موارد، از جمله در مورد توابع چگالی احتمال نیز به کار می رود.

یادآوری: وقتی که توزیع جامعه نامعلوم باشد از توزیع تجربی استفاده می کنیم. توزیع تجربی Y به صورت زیر تعریف می شود:

$$\hat{F}_n(y) = \frac{\text{number}(obs \leq y)}{n} = \frac{\sum_{i=1}^n I(obs \leq y)}{n} \quad (3-11)$$

که نسبت مشاهداتی است که از y کوچکتر یا مساوی با آن باشند.

همچنین افرون (1979) نشان داد که در موارد پیچیده، که در آن محاسبه ی مقدار آماره های بوت استرپ ساده نیست می توان با استفاده از باز نمونه گیری بوت کارلو، مقدار آنها را تقریب زد. یعنی نمونه هایی با حجم یکسان از نمونه ی اصلی، استخراج کرد که به هر کدام، باز نمونه گفته می شود. سپس مقدار آماره برای هر باز نمونه محاسبه شده و در نتیجه آماره ی بوت استرپ، با متوسط گیری از یک تابع مناسب از این مقادیر، تقریب زده می شود. این تقریب با افزایش تعداد باز نمونه ها بهتر می شود. چون در اکثر کاربردهای بوت استرپ، کمیت مورد بحث در بوت استرپ ممکن است مانند امید ریاضی شرطی روی نمونه، یا به طور معادل، به صورت انتگرال نسبت به تابع توزیع نمونه بیان شود این روش عملی می باشد. گسترش اساسی روش های بوت استرپ، مبنی بر استفاده از این خاصیت می باشد.

بوت استرپ را می توان روشی برای تقریب زدن معیارهای دقیق بر آوردگرایی دانست که توزیع آنها مجهول است. معیارهایی مانند واریانس، انحراف معیار و فاصله های اطمینان.

در دیدگاه سنتی، اگر $\mathbf{X}_n = (x_1, \dots, x_n)$ نمونه تصادفی iid از توزیع F باشند، واریانس $T_n(\mathbf{X}_n)$ به صورت زیر است:

$$Var(T_n) = \int (T_n(\mathbf{x}_n) - \int T_n(\mathbf{y}_n) d \prod_{i=1}^n F(y_i))^2 d \prod_{i=1}^n F(x_i) \quad (4-11)$$

که در آن $\mathbf{x}_n = (x_1, \dots, x_n)$ و $\mathbf{y}_n = (y_1, \dots, y_n)$

اگر $T_n = \bar{X}_n$ (میانگین نمونه) آنگاه،

$$Var(T_n) = \frac{1}{n} Var(X_1) \quad (5-11)$$

و اگر F متعلق به خانواده ی پارامتری نباشد، آنگاه $Var(X_1)$ را می توان به صورت زیر بر آورد کرد:

$$\widehat{Var}(X_1) = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (6-11)$$

اما همواره T_n مانند میانگین نمونه، ساده نیست. بنابراین تعیین یک فرمول دقیق و صریح برای $Var(T_n)$ در رابطه ی (۱۱)-۴ کار مشکلی است. این مشکل به دو دلیل می تواند ایجاد شود:

۱. F مجهول است.

۲. انتگرال گیری رابطه ی (۱۱)-۴ ساده نباشد.

۱۱-۲ بر آورد بوت استرپ واریانس

چنانچه در (4)، F را به وسیله $\hat{F}$ برآورد کنیم، آنچه حاصل می شود را برآورد کننده ی بوت استرپ واریانس T_n می نامند و آن را با $Var^*(T_n(X_n^*)|X_n)$ نشان می دهند. یعنی:

$$Var^*(T_n(X_n^*)|X_n) = \int (T_n(x_n) - \int T_n(y_n) d \prod_{i=1}^n \hat{F}(y_i))^2 d \prod_{i=1}^n \hat{F}(x_i) \quad (7-11)$$

که در آن $\{X_1^*, \dots, X_n^*\}$ یک نمونه ی تصادفی از $\hat{F}$ است و به آن بازنمونه گویند. برای برآورد F یعنی $\hat{F}$ می توان از F_n (توزیع تجربی) استفاده کرد و به این ترتیب انتخاب نمونه ی $\{X_1^*, \dots, X_n^*\}$ از F_n به معنی انتخاب یک نمونه ی تصادفی با جایگذاری از $\{x_1, \dots, x_n\}$ است.

۳-۱۱ الگوریتم بوت استرپ

اصول کار بوت استرپ به این صورت است که B بار از نمونه ای که در دست داریم نمونه باجایگذاری بر می داریم و پارامتر مورد نظر را برای هر یک از نمونه ها محاسبه کرده و میانگین آن برآوردها را به عنوان برآورد بوت استرپ بپذیریم. الگوریتم این روش به صورت زیر است:

۱. به اندازه n از توزیع تجربی $\hat{F}_n$ نمونه تولید می کنیم. (x^*)
۲. $\hat{\theta}^*$ را برای این مرحله محاسبه کنید. $(\hat{\theta}^*$ هر بار پارامتر دلخواه از جامعه نامعلوم است)
۳. مرحله های (۱) و (۲) را B مرتبه تکرار کنید.

میانگین این B نمونه به عنوان برآورد بهتری در اختیار ماست. معمولاً برای کاهش اریبی این برآوردگر بعد از نمونه گیری بوت استرپ از روش جک نایف استفاده می کنیم.

۴-۱۱ فاصله ی اطمینان بوت استرپ

فاصله ی اطمینان بوت استرپ در خیلی مواقع به آسانی بکار می رود و درک مفهوم فاصله ی اطمینان در این مورد خیلی راحت و ملموس می باشد، برای ساختن فاصله ی اطمینان باید $se(\hat{\theta})$ را بدانیم. برای بررسی سازگاری در فاصله ی اطمینان بوت استرپ باید B را افزایش دهیم. این فاصله ی اطمینان حاصله اغلب متقارن نیست زیرا این فاصله براساس توزیع تجربی می باشد. همانطور که می دانیم فاصله ی اطمینان رابطه ی زیادی با آزمون فرض دارد به عبارت دیگر ناحیه پذیرش به عنوان فاصله ی اطمینان بوت استرپ بکار می رود. با فرض نرمال بودن $\hat{\theta}^*$ و خطای استاندارد بوت استرپ داریم:

$$\left(\hat{\theta} - Z_{1-\frac{\alpha}{2}} se_B(\hat{\theta}), \hat{\theta} + Z_{1-\frac{\alpha}{2}} se_B(\hat{\theta}) \right) \quad (8-11)$$

که Z همان توزیع نرمال استاندارد می باشد و خطای استاندارد بوت استرپ عبارتست از:

$$\widehat{se}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^B \left(\hat{\theta}^*(b) - \hat{\theta}^*(.) \right)^2 \right\}^{1/2} \quad (9-11)$$

$$\hat{\theta}^*(.) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^*(b)$$

اما اگر فرض نرمال بودن لازم نباشد:

$$(\hat{\theta} - \zeta_{1-\frac{\alpha}{2}} se_B(\hat{\theta}), \hat{\theta} + \zeta_{1-\frac{\alpha}{2}} se_B(\hat{\theta})) \quad (10 - 11)$$

که در آن ζ_α عبارتست از α امین چندک مقادیر α_i و

$$\zeta_i = \frac{\hat{\theta}_i^* - \hat{\theta}}{se(\hat{\theta}_i^*)} \quad (11 - 11).$$

همیشه بخاطر داشته باشید که:

قبل از بکار بردن بوت استرپ باید مطمئن شوید نمونه ها از برآورد خوبی از جامعه شناخته نشده باشند در غیر اینصورت بوت استرپ بکار نمی رود.

۱۱-۵ به دست آوردن استنباطهای پارامتریک (مطالعه موردی جامعه کارمندان)

در هنگام جمع آوری داده ها، اغلب به ویژگی های جامعه که از آن نمونه گیری کرده اید، علاقه مند می شوید. شما درباره ی پارامترهای جامعه استنباطهایی می کنید و در همان حال برآوردهایی را از نمونه به دست می آورید. مثلاً درمجموعه داده های Employee data.sav که شامل نمونه تصادفی تولید شده از جامعه بزرگ کارمندان می باشد، میانگین نمونه ای \$۳۴۴۱۹,۵۷ برای حقوق فعلی، برآوردی از میانگین حقوق فعلی برای جامعه کارمندان است .

۱۱-۶ آماده سازی داده ها

برای به دست آوردن استنباطهای پارامتریک در باره متغیر Current salary مراحل زیر را انجام دهید

۱- برای باز کردن فایل داده Employee data.sav آدرس زیر را دنبال کنید

C:\Program Files\IBM\Spss\Statistics\19\Samples\English\Employee data.sav

*Employee data.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

16 : salary 40800

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	v
1	1	m	02/03/1952	15	3	\$57,000	\$27,000	98	144	0	
2	2	m	05/23/1958	16	1	\$40,200	\$18,750	98	36	0	
3	3	f	07/26/1929	12	1	\$21,450	\$12,000	98	381	0	
4	4	f	04/15/1947	8	1	\$21,900	\$13,200	98	190	0	
5	5	m	02/09/1955	15	1	\$45,000	\$21,000	98	138	0	
6	6	m	08/22/1958	15	1	\$32,100	\$13,500	98	67	0	
7	7	m	04/26/1956	15	1	\$36,000	\$18,750	98	114	0	
8	8	f	05/06/1966	12	1	\$21,900	\$9,750	98	0	0	
9	9	f	01/23/1946	15	1	\$27,900	\$12,750	98	115	0	
10	10	f	02/13/1946	12	1	\$24,000	\$13,500	98	244	0	
11	11	f	02/07/1950	16	1	\$30,300	\$16,500	98	143	0	
12	12	m	01/11/1966	8	1	\$28,350	\$12,000	98	26	1	
13	13	m	07/17/1960	15	1	\$27,750	\$14,250	98	34	1	
14	14	f	02/26/1949	15	1	\$35,100	\$16,800	98	137	1	
15	15	m	08/29/1962	12	1	\$27,300	\$13,500	97	66	0	
16	16	m	11/17/1964	12	1	\$40,800	\$15,000	97	24	0	
17	17	m	07/18/1962	15	1	\$46,000	\$14,250	97	48	0	
18	18	m	03/20/1956	16	3	\$103,750	\$27,510	97	70	0	
19	19	m	08/19/1962	12	1	\$42,300	\$14,250	97	103	0	
20	20	f	01/23/1940	12	1	\$26,250	\$11,550	97	48	0	
21	21	f	02/19/1963	16	1	\$38,850	\$15,000	97	17	0	
22	22	m	09/24/1940	12	1	\$21,750	\$12,750	97	315	1	
23	23	f	03/15/1965	15	1	\$24,000	\$11,100	97	75	1	
24	24	f	03/27/1933	12	1	\$16,950	\$9,000	97	124	1	
25	25	f	07/01/1942	15	1	\$21,150	\$9,000	97	171	1	
26	26	m	11/08/1966	15	1	\$31,050	\$12,600	96	14	0	
27	27	m	03/19/1954	19	3	\$60,375	\$27,480	96	96	0	
28	28	m	04/11/1963	15	1	\$32,550	\$14,250	96	43	0	
29	29	m	01/28/1944	19	3	\$135,000	\$79,980	96	199	0	
30	30	m	09/17/1961	15	1	\$31,200	\$14,250	96	54	0	
31	31	m	02/24/1964	12	1	\$36,150	\$14,250	96	83	0	

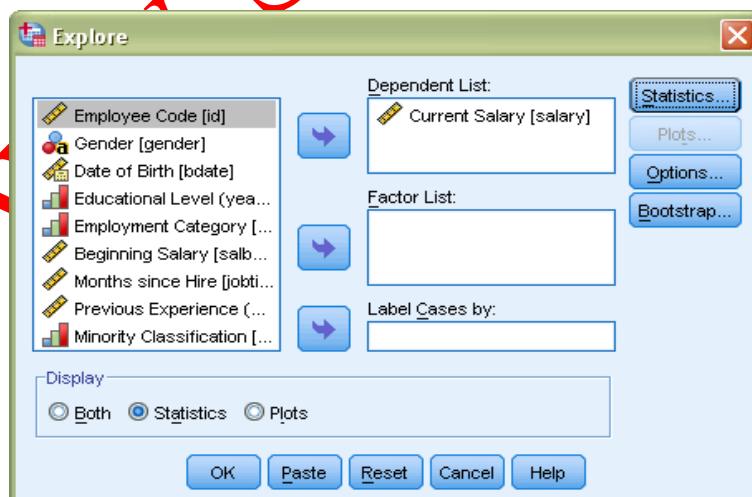
شکل ۱۱-۱: اجرای فایل داده‌های حقوق کارمندان در Spss

۲- دستور زیر را اجرا کنید

Analyze

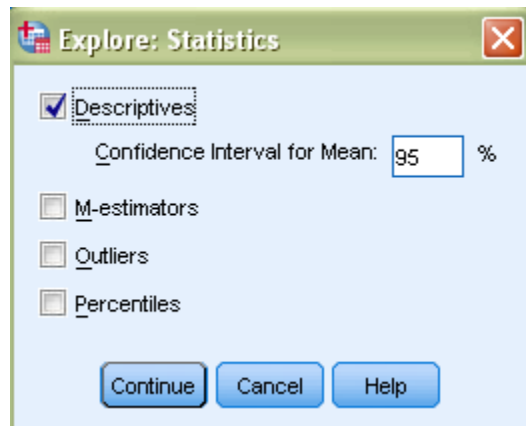
Descriptive statistics

Explore



شکل ۱۱-۲: کادر محاوره ای Explore

۳- بر روی دکمه Statistics کلیک کرده تا کادر آن باز شود.



شکل ۱۱-۳: انتخاب فاصله اطمینان از کادر محاوره ای Explore

۴- Ok نمایش دهید.

خروجی در زیر آمده است

Descriptives				Statistic	Std. Error
Current Salary	Mean			\$34,419.57	\$784.311
	95% Confidence Interval for Mean	Lower Bound		\$32,878.40	
		Upper Bound		\$35,960.73	
	5% Trimmed Mean			\$32,455.19	
	Median			\$28,875.00	
	Variance			2.916E8	
	Std. Deviation			\$17,075.661	
	Minimum			\$15,750	
	Maximum			\$135,000	
	Range			\$119,250	
	Interquartile Range			\$13,163	
	Skewness			2.125	.112
	Kurtosis			5.378	.224

شکل ۱۱-۴: خروجی آمارهای توصیفی متغیر Current Salary

این برآورد، خطای استاندارد \$784,311 با حجم نمونه ۴۷۴ دارد، پس فاصله اطمینان ۹۵٪ برای میانگین حقوق فعلی در جامعه کارمندان از \$ ۳۲۸۷۸,۴۰ تا \$۳۵۹۶۰,۷۳ است. اما این برآوردگرها چقدر معتبر هستند؟ برای جامعه‌های «معلوم» و پارامترهای مناسب که در مورد ویژگیهای برآوردهای نمونه می‌دانیم، می‌تواند در مورد این نتایج، قابل اطمینان باشد. بوت استرپ در جستجوی باز کردن اطلاعات بیشتری درباره‌ی ویژگی‌های برآوردها در جامعه‌های نامعلوم با پارامترهای نامناسب است.

منبع: کتاب تحلیل داده‌های پرسشنامه‌ای به کمک نرم افزار SPSS - ابراهیم بایزیدی و همکاران - انتشارات عابد - چاپ پنجم - ۱۳۹۱

WWW.SPSS-PASW.IR