کانون کارگزاران بورس و اوراق بهادار

دوره آمادگی آزمون گواهینامه های حرفه ای بازار سرمایه

روش های کمی پیشرفته

فصول دوم، سوم و چهارم

تدوین: حسین توکلیان

خرداد ۱۳۹۰

www.Seba.ir

فصل دوم

تابع توزیع احتمال

متغیر تصادفی گسسته و پیوسته و توزیع احتمال

کمیتی که حاصل از انجام یک آزمایش بوده و در هر آزمایش در زمان‌های مختلف می‌تواند مقادیر مختلفی را به خود بگیرد متغیر تصادفی نامیده می‌شود. یک متغیر تصادفی گسسته متغیری است که می‌توان تعداد برآمدهای آن را شمرد و برای هر برآمد آن ممکن یک احتمال مثبت و قابل اندازه‌گیری وجود دارد. برای مثال تعداد روزهای برفی در یک سال یک متغیر تصادفی گسسته است. یک متغیر تصادفی پیوسته متغیری است که تعداد برآمدهای ممکن آن بینهایت است حتی اگر کمترین و بیشترین مقدار آن نیز مشخص باشد. برای مثال مقدار بارش برف بین ۱ تا ۱۰ سانتیمتر یک متغیر تصادفی است زیرا میزان بارش برف می‌تواند بینهایت مقدار در این فاصله به خود بگیرد.

یک توزیع احتمال، احتمال برآمدهای ممکن برای یک متغیر تصادفی را توضیح می‌دهد. مجموع احتمال وقوع کلیه برآمدهای ممکن باید برابر با یک باشد. برای تشکیل یک توزیع احتمال ابتدا فضای نمونه را بدست می‌آوریم، سپس مقادیر ممکن برای به همراه احتمال وقوع آن‌ها را بر اساس فضای نمونه بدست می‌آوریم. در واقع توزیع احتمال همانند جدول فراوانی است که در آن احتمال وقوع هر برآمد همان فراوانی نسبی آن است. در مورد پرتاب دو سکه و در نظر گرفتن تعداد خطهای بدست آمده به عنوان متغیر تصادفی X می‌توان گفت

فضای نمونه={(خ‌خ) و (خ‌ش) و (ش‌خ) و (ش‌ش)}

x	حالات مساعد	تعداد حالات مساعد	احتمال وقوع x ($p(x)$)
۲	(خ‌خ)	۱	۰/۲۵
۱	(خ‌ش)، (ش‌خ)	۲	۰/۵
۰	(ش‌ش)	۱	۰/۲۵

توزیع احتمال مربوط به یک متغیر تصادفی گسسته یک توزیع احتمال گسسته و برای یک متغیر تصادفی پیوسته یک توزیع احتمال پیوسته است. تفاوت بین دو نوع توزیع احتمال عبارت است از

- در یک توزیع احتمال گسسته زمانی که x نمی‌تواند اتفاق بیافتد $P(x)=0$ و برای زمانی که می‌تواند اتفاق بیافتد $P(x)>0$.
- در یک توزیع احتمال پیوسته همیشه $P(x)=0$ حتی اگر x بتواند اتفاق بیافتد. در توزیع احتمال پیوسته تنها می‌توان احتمال قرار گرفتن متغیر تصادفی پیوسته X در یک فاصله را بدست آوریم. یعنی تنها می‌توان $P(x_1 \le X \le x_2)$ را بدست آورد که در آن x_1 و x_2 مقادیر واقعی هستند. بنابراین احتمال اینکه متغیر تصادفی پیوسته یک مقدار مشخص را به خود بگیرد همیشه برابر با صفر خواهد بود.

در مباحث مالی گاهی اوقات برخی از توزیع‌های گسسته را پیوسته در نظر می‌گیرند زیرا تعداد برآمدهای ممکن بسیار زیاد است. برای مثال افزایش یا کاهش قیمت سهام معامله شده در بورس تهران بر حسب ریال ثبت می‌شود. احتمال تغییر دقیقاً به اندازه ۱۲/۶۵ یا هر تغییر خاص دیگر تقریباً برابر با صفر است. بنابراین در این موارد بهتر است در مورد یک دامنه تغییر قیمت صحبت شود. زیرا می‌توان گفت احتمال تغییر قیمت یک سهم خاص در فاصله ۱۵ ریال تا ۲۵ ریال در یک روز بزرگتر از صفر است اما احتمال تغییر قیمت به اندازه ۱۷/۳۴ ریال تقریباً صفر است.

تابع احتمال (probability function)

یک تابع احتمال که آن را با $P(x)$ نشان می‌دهیم احتمال وقوع وضعیت‌های مختلف یک متغیر تصادفی را نشان می‌دهد. به عبارت دیگر $P(x)$ احتمال این است که متغیر تصادفی X مقدار x را به خود بگیرد. این تابع دارای ویژگی‌های زیر می‌باشد

- $0 < P(x) < 1$
- مجموع احتمال‌های کلیه برآمدهای ممکن، x، برای متغیر تصادفی گسسته X برابر با ۱ است یعنی $\sum P(x) = 1$

تابع چگالی احتمال (probability density function (pdf))

یک تابع چگالی احتمال (pdf) که با نشان داده می‌شود تابعی است که می‌تواند احتمال برآمدهای یک متغیر تصادفی پیوسته در یک دامنه خاص را تولید کند. این تابع معادل تابع احتمال برای یک متغیر تصادفی گسسته است.

- در مورد متغیر تصادفی پیوسته سطح زیر منحنی تابع چگالی احتمال برابر با ۱ می‌باشد یعنی $\int P(x)dx = 1$.
- در مورد متغیر تصادفی پیوسته احتمال اینکه $\alpha < X < \beta$ باشد، برابر با سطح زیر منحنی چگالی در بازه α و β است $P(\alpha < X < \beta) = \int_{\alpha}^{\beta} P(x)dx$
- بنابراین احتمال اینکه متغیر تصادفی پیوسته X مقدار مشخص α را به خود بگیرد برابر است با $P(X = \alpha) = \int_{\alpha}^{\alpha} P(x)dx = 0$.

⇦توجه:

تست‌هایی که در این زمینه طرح می‌شوند را می‌توان بر اساس ویژگی‌های ارائه شده در بالا به راحتی حل کرد. برای مثال اگر تابع چگالی احتمال داده شده و در آن یک ضریب ثابت به صورت مجهول وجود داشته باشد به راحتی می‌توان این ضریب ثابت را با استفاده از ویژگی $P(\alpha < X < \beta) = \int_{\alpha}^{\beta} P(x)dx = 1$ بدست آورد که در آن α و β حد بالا و پایین توزیع هستند.

مثال۱: به ازای کدام مقدار a تابع $x = 0,1,2,3,4$ $p(X = x) = \frac{\binom{4}{x}}{3a+1}$ یک تابع احتمال است؟ (حسابداری و مدیریت ۸۷)

۱) ۶ ۲) ۵ ۳) ۴ ۴) ۳

جواب:

با استفاده از این ویژگی تابع احتمال گسسته که $\sum P(x) = 1$ داریم

$$\sum P(x)=1 \rightarrow \sum_{x=0}^{4} \frac{\binom{4}{x}}{3a+1}=1 \rightarrow \frac{2^4}{3a+1}=1 \rightarrow 16=3a+1 \rightarrow a=5$$

نکته: $\sum_{x=0}^{n}\binom{n}{x}=2^n$

مثال ۲: مقدار m در تابع $f(x)=\frac{m}{\sqrt{x}}$ برای $0\le x\le 1$ چقدر باشد تا $f(x)$ یک تابع چگالی احتمال باشد؟ (اقتصاد ۸۸)

۱) ۰/۵ ۲) ۱ ۳) ۱/۵ ۴) ۲

جواب:

با استفاده از این ویژگی تابع احتمال پیوسته که $\int_{\alpha}^{\beta} P(x)dx=1$ داریم $p(\alpha < X < \beta)$

$$\int_0^1 \frac{m}{\sqrt{x}}dx=1 \rightarrow m[2x^{\frac{1}{2}}]_0^1=1 \rightarrow m=\frac{1}{2}$$

امید ریاضی یا کمیت انتظاری یک متغیر تصادفی (میانگین)

امید ریاضی در واقع شاخص مرکزی یک متغیر تصادفی است که به عنوان کمیت انتظاری یک متغیر تصادفی نیز شناخته می‌شود. در واقع امید ریاضی میانگین وزنی یک متغیر تصادفی است که در آن وزن‌های داده شده به متغیر مانند فراوانی نسبی است که در اینجا همان احتمال وقوع متغیر تصادفی است. امید ریاضی متغیر تصادفی X را با $E(X)$ نشان می‌دهیم. به عبارت دیگر

- در مورد متغیر تصادفی گسسته $E(X)=\sum_{i=1}^{n} x_i P(x_i)=x_1P(x_1)+x_2P(x_2)+\ldots+x_nP(x_n)$
- در مورد متغیر تصادفی پیوسته $E(X)=\int_{-\infty}^{+\infty} xP(x)dx$
- خواص امید ریاضی عبارت است از

$$E(\alpha)=\alpha$$
$$E(x\pm\alpha)=E(x)\pm\alpha$$
$$E(\beta x)=\beta x$$
$$E(E(x))=E(x)$$
$$E(x-E(x))=0$$

- امید مجذور تفاضلات از میانگین همیشه حداقل است $E[(x-E(x))^2]\le E[(x-\alpha)^2]$
- امید ریاضی تابعی از x عبارت است از

برای متغیر تصادفی گسسته $E[g(x)] = \sum g(x)\, f(x)$

برای متغیر تصادفی پیوسته $E[g(x)] = \int_{-\infty}^{+\infty} g(x)\, p(x)\, dx$

مثال ۳: امید ریاضی متغیر تصادفی X با تابع چگالی احتمال زیر چقدر است؟ (اقتصاد ۸۶)

$$f(x) = \begin{cases} \frac{1}{2\sqrt{x}} & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

۱) ۰/۵ ۲) ۱/۳۳ ۳) ۰/۶۷ ۴) ۰/۳۳

جواب:

$$E(x) = \int_0^1 xf(x)dx = \int_0^1 x\frac{1}{2\sqrt{x}}dx = \left[\frac{2}{6}x^{\frac{3}{2}}\right]_0^1 = 0.33$$

واریانس

متوسط انحرافات از میانگین یک متغیر تصادفی را واریانس آن متغیر تصادفی می‌گویند. با توجه به اینکه میانگین یک متغیر تصادفی در واقع همان امید ریاضی متغیر تصادفی است می‌توان واریانس متغیر تصادفی X را به صورت زیر بنویسیم

$$\sigma^2 = E(x-\mu)^2 = E(x-E(x))^2 = E[x^2 - 2xE(x) + (E(x))^2] = E(x^2) - E(x)^2$$

- برای متغیر گسسته داریم $\sigma^2 = E(x^2) - E(x)^2 = \sum xf(x) - \sum x^2 f(x)$
- برای متغیر پیوسته داریم $\sigma^2 = E(x^2) - E(x)^2 = \int xf(x)dx - \int x^2 f(x)dx$
- خواص واریانس عبارت است از

$$\sigma^2(\alpha) = 0$$
$$\sigma^2(\beta x) = \beta^2\sigma^2(x)$$
$$\sigma^2(\alpha \pm \beta x) = \beta^2\sigma^2(x)$$

- خواص انحراف معیار عبارت است از

$$\sigma(\alpha) = 0$$
$$\sigma(\beta x) = |\beta|\sigma(x)$$
$$\sigma(\alpha \pm \beta x) = |\beta|\sigma(x)$$

مثال ۴: اگر $\text{var}(-2x+1) = 5$ و $E(x) = 1.5$ باشند، $E(x-2)^2$ کدام است؟ (برنامه‌ریزی شهری ۸۸)

۱) ۱/۲۵ ۲) ۱/۵ ۳) ۱/۷۵ ۴) ۲/۲۵

جواب:

$$\text{var}(-2x+1) = 5 \rightarrow 4\,\text{var}(x) = 5 \rightarrow \text{var}(x) = 1.25$$
$$E(x^2) = \text{var}(x) + E(x)^2 = 1.25 + 2.25 = 3.5$$
$$E(x-2)^2 = E(x^2) - 4E(x) + 4 = 3.5 - 6 + 4 = 1.5$$

تابع توزیع توأم (joint distribution function)

زمانی که یک پدیده تصادفی به وسیله مجموعه‌ای از متغیرهای تصادفی تفسیر شود، تغییرات متغیرهای تصادفی در ارتباط با یکدیگر و به صورت توأم (مشترک) بررسی می‌شوند. برای مثال سرمایه‌گذاران و مدیران سرمایه‌گذاری اغلب علاقه‌مند به رابطه بین بازدهی‌های داراییهای مختلف هستند.

چنانچه شرایط زیر برقرار باشد، تابع احتمال دو متغیر تصادفی X و Y با $f(x, y)$ نشان داده می‌شود

حالت گسسته

- $0 \le f(x, y) \le 1$
- $\sum_x \sum_y f(x, y) = 1$

حالت پیوسته

- $f(x, y) \ge 0$
- $\iint f(x, y)dxdy = 1$

تابع توزیع تجمعی (cumulative distribution function (cdf))

تابع توزیع تجمعی یا تابع توزیع (که با F نشان می‌دهیم)، احتمال اینکه متغیر تصادفی X مقداری کمتر یا برابر با یک مقدار مشخص x به خود بگیرد را ارائه می‌کند. این تابع مجموع یا مقدار تجمعی احتمال برآمدهایی کمتر یا برابر با یک برآمد خاص را نشان می‌دهد. بنابراین $F(x) = P(X \le x)$.

ویژگی‌های تابع توزیع تجمعی

- $0 \le F(x) \le 1$
- $P(X \le x) + P(X > x) = 1 \rightarrow P(X > x) = 1 - P(X \le x) = 1 - F(x)$
- $F(-\infty) = P(X \le -\infty) = 0, F(+\infty) == P(X \le +\infty) = 1$
- $\alpha < \beta \rightarrow F(\alpha) \le F(\beta)$
- تابع توزیع تجمعی گسسته $F(x) = P(X \le x) = \sum_{X \le x} P(x)$
- $F(x_i) = P(X \le x_i) = P(x_1) + P(x_2) + \ldots + P(x_{i-1}) + P(x_i)$
- تابع توزیع تجمعی پیوسته (α حد پایین تابع چگالی احتمال است) $F(x) = P(X \le x) = \int_\alpha^x f(x)dx$

مثال ۵: تابع توزیع کمیت تصادفی پیوسته X (طول زمان کار دستگاه تا وقتی که از کار بیافتد) به قرار ذیل می‌باشد. احتمال اینکه دستگاه در طول زمان $X \geq T$ از کار بیافتد چقدر است؟ (اقتصاد ۸۷) $F(x) = 1 - \exp(-\frac{x}{T}) \quad 0 < x \leq \infty$

۱) e ۲) e^{-1} ۳) $1 - e^{-1}$ ۴) $\frac{e^{-1}}{T}$

جواب:

$$P(X \geq T) = 1 - P(X < T) = 1 - F(T) = 1 - (1 - e^{-\frac{T}{T}}) = e^{-1}$$

نحوه محاسبه تابع توزیع تجمعی

- متغیر تصادفی گسسته

$$f(x) = \begin{cases} 0 & x < x_1 \\ f_1(x) & x_1 \leq x \leq x_2 \\ f_2(x) & x_2 \leq x \leq x_3 \\ \vdots & \\ f_{n-1}(x) & x_{n-1} \leq x \leq x_n \\ f_n(x) & x \geq x_n \end{cases} \rightarrow P(X \leq x_i) = F(x_i) = \begin{cases} 0 & x < x_1 \\ F(x_1) = f_1(x_1) & x_1 \leq x \leq x_2 \\ F(x_2) = f_1(x_1) + f_2(x_2) & x_2 \leq x \leq x_3 \\ \vdots & \\ F(x_{n-1}) = f_1(x_1) + \ldots + f_{n-1}(x_{n-1}) & x_{n-1} \leq x \leq x_n \\ F(x_n) = f_1(x_1) + \ldots + f_n(x_n) & x \geq x_n \end{cases}$$

با توجه به این رابطه داریم

$$F(x) = \begin{cases} F(x_1) = f(x_1) \\ F(x_i) = f(x_1) + \ldots + f(x_i) & i = 2,3,\ldots,n \\ F(x_n) = 1 \end{cases}$$

- متغیر تصادفی پیوسته

$$f(x) = \begin{cases} h(x) & \alpha < x < \beta \\ k(x) & \beta < x < \gamma \\ 0 & otherwise \end{cases} \rightarrow F(x) = \begin{cases} 0 & x < \alpha \\ \int_{\alpha}^{x} h(x)dx & \alpha < x < \beta \\ \int_{\alpha}^{\beta} h(x)dx + \int_{\beta}^{x} k(x)dx & \beta < x < \gamma \end{cases}$$

توزیع دوجمله‌ای (binomial distribution)

یک متغیر تصادفی دوجمله‌ای به صورت تعداد موفقیت‌ها در یک آزمایش که در آن برآمدها می‌توانند به صورت موفقیت یا شکست باشند، تعریف می‌شود. احتمال موفقیت، p، برای هر آزمایش ثابت بوده و آزمایش‌ها مستقل هستند. متغیر تصادفی دوجمله‌ای که در آن تعداد آزمایش‌ها برابر ۱ است متغیر *تصادفی برنولی* نامیده می‌شود. بنابراین اگر یک آزمایش برنولی مستقل n بار تکرار شود آنگاه متغیر تصادفی تعداد موفقیت در n بار آزمایش مستقل برنولی دارای توزیع دوجمله‌ای خواهد بود. چنانچه متغیر تصادفی x به صورت تعداد موفقیت‌ها در n بار آزمایش مستقل برنولی تعریف شود آنگاه احتمال موفقیت، $p(x)$ برابر خواهد بود با

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!\,x!} p^x (1-p)^{n-x}$$

امیدریاضی و واریانس متغیر تصادفی دوجمله‌ای

اگر n آزمایش انجام شود، احتمال موفقیت در هر آزمایش برابر است با p، بنابراین انتظار np موفقیت در این n آزمایش را داریم. لذا $E(x) = np$

واریانس یک متغیر تصادفی دوجمله‌ای عبارت است از

$$\sigma^2 = np(1-p) = npq \rightarrow \sigma = \sqrt{npq}$$

مثال ۶: ده درصد از تراشه‌های تولیدی کارخانه‌ای معیوب است. اگر یک نمونه تصادفی ۳ تایی از این تراشه‌ها انتخاب شود، احتمال مشاهده حداقل یک تراشه معیوب چقدر است؟ (اقتصاد ۸۶)

۱) ۲۳ ۲) ۲۷ ۳) ۷۳ ۴) ۷۷

جواب:

$$n = 3,\ p = 0.1,\ q = 0.9$$

$$p(x \geq 1) = 1 - p(x = 0) = 1 - \binom{n}{x} p^x q^{n-x} = 1 - \binom{3}{0} (0.1)^0 (0.9)^3 = 0.271$$

مثال ۷: در یک توزیع دوجمله‌ای $E(x) = 9$ و $\text{var}(x) = 6$ است. مقدار $p(x \geq 1)$ برابر است با: (اقتصاد ۸۸)

۱) $\frac{2}{3}$ ۲) $1 - \left(\frac{1}{3}\right)^{27}$ ۳) $\frac{1}{3}$ ۴) $1 - \left(\frac{2}{3}\right)^{27}$

جواب:

$$\left.\begin{aligned} \text{var}(x) = npq = 6 \\ E(x) = np = 9 \end{aligned}\right\} \Rightarrow \frac{npq}{np} = q = \frac{6}{9} = \frac{2}{3} \rightarrow p = \frac{1}{3} \rightarrow n = 9 \times 3 = 27$$

$$p(x \geq 1) = 1 - p(x < 1) = 1 - p(x = 0) = 1 - \binom{27}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^{27} = 1 - \left(\frac{2}{3}\right)^{27}$$

مثال ۸: احتمال موفقیت در یک آزمایش برنولی $\frac{3}{5}$ است. اگر x تعداد موفقیت‌ها در هر ۲۴ مشاهده باشد، انحراف معیار x کدام است؟ (برنامه‌ریزی شهری ۸۷)

۱) ۱/۲ ۲) ۱/۴۴ ۳) ۲/۴ ۴) ۲/۵۶

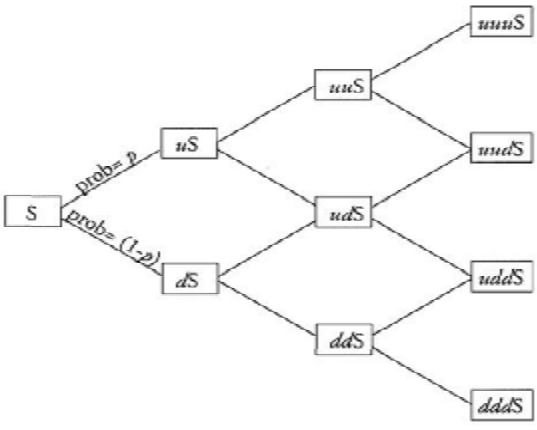
جواب:

$$n=24,\ p=\frac{3}{5},\ q=1-p=\frac{2}{5}\rightarrow \text{var}(x)=npq=24\times\frac{3}{5}\times\frac{2}{5}=\frac{144}{25}\rightarrow\sigma=\frac{12}{5}=2.4$$

کاربرد توزیع دوجمله‌ای در تحلیل نوسانات قیمت سهام

می‌توان یک مدل دوجمله‌ای را برای تحلیل نوسانات قیمت سهام بکار برد. تنها کافی است که دو برآمد ممکن و احتمال وقوع هر برآمد را تعریف کنیم. یک سهم با قیمت کنونی S را در نظر بگیرید که در دوره بعد می‌تواند به اندازه ۱ درصد افزایش یا به اندازه ۱ درصد کاهش پیدا کند (تنها دو برآمد ممکن). احتمال تعییر به سمت بالا (احتمال تغییر به سمت بالا، u) برابر است با p و تغییر به سمت پایین (احتمال تغییر به سمت پایین، d) برابر است با $(1\text{-}p)$. برای این مثال فاکتور تغییر به سمت بالا (U) 1.01 و فاکتور تغییر به سمت پایین (D) $\frac{1}{1.01}$. بنابراین در دوره بعد به احتمال p قیمت سهام به $S(1.01)$ و به احتمال $(1\text{-}p)$ به $S/1.01$ تغییر خواهد کرد.

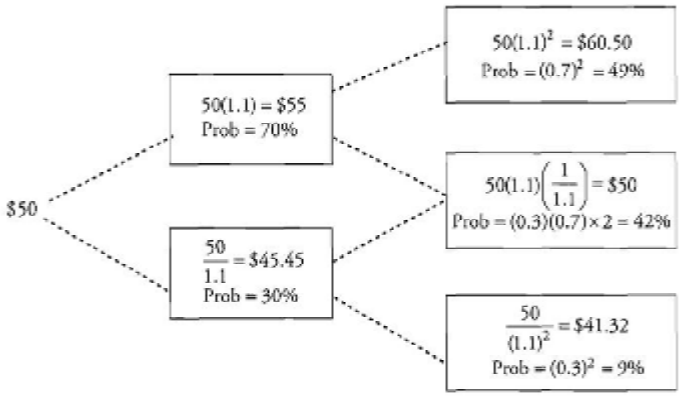
با نشان دادن کلیه ترکیبات ممکن تغییر به سمت بالا و تغییر به سمت پایین در تعدادی از دوران متوالی موفقیت درخت دوجمله-ای ساخته می‌شود. برای دو دوره این ترکیبات عبارتند از UU، UD، DU و DD. دو ترکیب UD و DU بعد از دو دوره قیمت سهام یکسانی را بدست خواهد داد زیرا $S=S(1.01)/(1.01)$ و ترتیب برآمدها نتیجه را تغییر نمی‌دهد. هر کدام از مقادیر ممکن در یک شاخه دوجمله‌ای یک گره است. درخت دوجمله‌ای برای سه دوره در نمودار زیر رسم شده است.



با قیمت اولیه سهام S=50، U=1.01، $D=\frac{1}{1.01}$ و $Prob(U)$=0.6 می‌توان قیمت‌های احتمالی سهام بعد از دو دوره را بدین صورت محاسبه نمود:

- $uuS = 1.01^2 \times 50 = 51.01$ با احتمال $(0.6)^2 = 0.36$
- $udS = 1.01(1/1.01) \times 50 = 50$ با احتمال $(0.6)(0.4) = 0.24$
- $duS = (1/1.01)1.01 \times 50 = 50$ با احتمال $(0.4)(0.6) = 0.24$
- $ddS = (1/1.01)^2 \times 50 = 49.01$ با احتمال $(0.4)^2 = 0.16$

از آنجایی که قیمت ۵۰ می‌تواند هم از تغییر UD و هم از تغییر DU بدست آید احتمال قیمت سهام معادل ۵۰ بعد از دو دوره (مقدار میانی) عبارت است از $2 \times (0.6)(0.4) = 48\%$. درخت دوجمله‌ای با S=50، U=1.1 و $Prob(U)$=0.7 در نمودار زیر نشان داده شده است. توجه داشته باشید که مقدار میانی بعد از دو دوره (یعنی ۵۰) برابر با مقدار اولیه است. احتمال اینکه قیمت سهام بعد از دو دوره پایین بیاید (<۵۰) برابر است با احتمال دو دوره کاهش در قیمت، $(1-0.7)^2 = 9\%$.



یکی از کاربردهای مهم مدل دوجمله‌ای قیمت سهام در قیمت‌گذاری اختیارات است. می‌توان درخت دوجمله‌ای قیمت داراییها را با کاهش طول دوره و افزایش تعداد دوره‌ها و برآمدهای ممکن به مسائل واقعی نزدیک‌تر کرد.

خطای دنبال کردن (tracking error)

خطای دنبال کردن عبارت است از تفاضل بین بازدهی کل یک پورتفو و بازدهی کل معیاری که عملکرد پورتفو با آن مقایسه می‌شود. برای مثال اگر یک پورتفو از سهام بازار بورس تهران دارای بازدهی کل ۴٪ در یک دوره باشد و در مقابل بازدهی کل شاخص بازار بورس تهران برابر با ۷٪ باشد، خطای دنبال کردن این پورتفو ۳-٪ است.

توزیع احتمال یکنواخت (uniform distribution function)

حالت گسسته

اگر در اثر انجام آزمایش x به N نتیجه $x_1,...,x_N$ با احتمال وقوع یکسان دست یابیم، آنگاه می‌گوییم متغیر تصادفی x دارای توزیع یکنواخت گسسته است که احتمال وقوع هر یک از وضعیت‌ها برابر با $\frac{1}{N}$ است. یعنی

$$\sum_{i=1}^{N} p(x_i)=1 \rightarrow p(x_1)+p(x_2)+...+p(x_N)=1 \xrightarrow{p(x_1)=...=p(x_N)} p(x_i)=\frac{1}{N}; i=1,2,...,N \rightarrow f(x)=\frac{1}{N}$$

ویژگی‌های توزیع احتمال یکنواخت گسسته

$$E(x)=\frac{\sum x_i}{N} \quad -$$

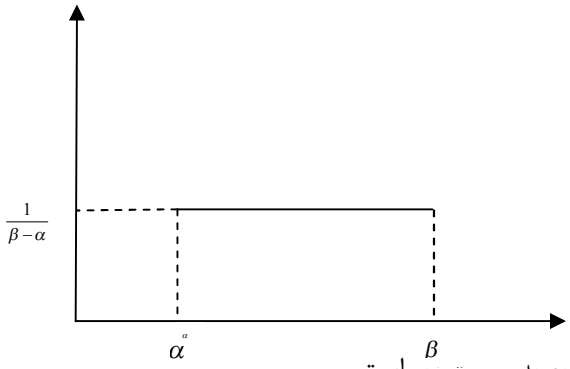
$$\sigma_x^2=\frac{\sum (x_i-E(x))^2}{N}=\frac{\sum x_i^2}{N}-\left(\frac{\sum x_i}{N}\right)^2 \quad -$$

حالت پیوسته

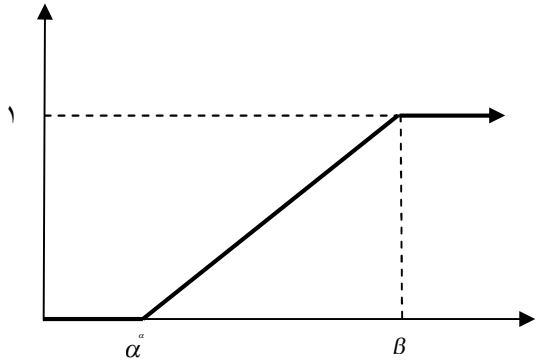
چنانچه احتمال وقوع متغیر تصادفی X در هر نقطه از فاصله یکسان باشد، آنگاه متغیر تصادفی X دارای توزیع یکنواخت پیوسته است. به عبارت دیگر متغیر تصادفی پیوسته X در بازه $\alpha < x < \beta$ دارای چگالی یکنواخت است اگر و فقط اگر تابع چگالی آن به ازای هر مقدار X برابر با $\frac{1}{\beta-\alpha}$ باشد. از آنجایی که این توزیع پیوسته است با اینکه $\alpha < x < \beta$ باشد اما احتمال وقوع یک مقدار مشخصی از X برابر با صفر است یعنی $p(X=x_0)=0$.

ویژگی‌های توزیع یکنواخت پیوسته

- $f(x)=\frac{1}{\beta-\alpha},\alpha<x<\beta$
- $p(x_0\le X\le x_1)=\frac{x_1-x_0}{\beta-\alpha}$
- $p(X<\alpha)=p(X>\beta)=0$
- $E(x)=\frac{\alpha+\beta}{2}$
- $\sigma^2=\frac{(\beta-\alpha)^2}{12}$
- نمودار یک توزیع یکنواخت پیوسته در بازه $\alpha<x<\beta$ به شکل زیر است



- نمودار تابع توزیع تجمعی این توزیع به صورت ریر است



مثال ۹: توزیع $f(x)=\frac{1}{\beta-\alpha},\alpha<x<\beta$ را در نظر بگیرید، $E(x^2)$ عبارت است از: (اقتصاد ۸۷)

۱) $\frac{\beta^3-\alpha^3}{3(\beta-\alpha)}$ ۲) $\frac{\beta^3-\alpha^3}{(\beta-\alpha)}$ ۳) $\frac{(\alpha-\beta)^2}{12}$ ۴) $\frac{(\alpha+\beta)^2}{4}$

جواب:

$$E(x^2)=\int_\alpha^\beta x^2\,f(x)dx=\int_\alpha^\beta x^2\frac{1}{\beta-\alpha}dx=\frac{1}{\beta-\alpha}\left[\frac{1}{3}x^3\right]_\alpha^\beta=\frac{\beta^3-\alpha^3}{3(\beta-\alpha)}$$

مثال ۱۰: اگر $f(x)=\begin{cases}\frac{1}{3} & 1<x<4\\ 0 & otherwise\end{cases}$ تابع چگالی متغیر تصادفی x باشد، واریانس این متغیر تصادفی کدام است؟

(برنامه‌ریزی شهری ۸۶)

۱) ۰/۲۵ ۲) ۰/۷۵ ۳) ۱/۲۵ ۴) ۱/۵

جواب:

$$\alpha=1,\beta=4$$

$$f(x)=\frac{1}{4-1}=\frac{1}{3}\qquad 1<x<4$$

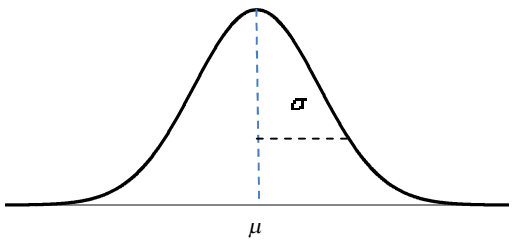
$$\sigma^2=\frac{(\beta-\alpha)^2}{12}=\frac{(4-1)^2}{12}=\frac{3}{4}=0.75$$

توزیع نرمال (normal distribution)

در علم آمار اثبات می‌شود که بسیاری از پدیده‌های طبیعی دارای توزیع نرمال (یا توزیع زنگوله‌ای) بوده و بسیاری از توزیع‌ها در حد دارای تقریب نرمال هستند. این توزیع یک توزیع پیوسته و متقارن است که دارای کاربردهای بسیار زیادی است. تابع چگالی متغیر تصادفی x که دارای توزیع نرمال با میانگین μ و واریانس σ^2 است عبارت است از:

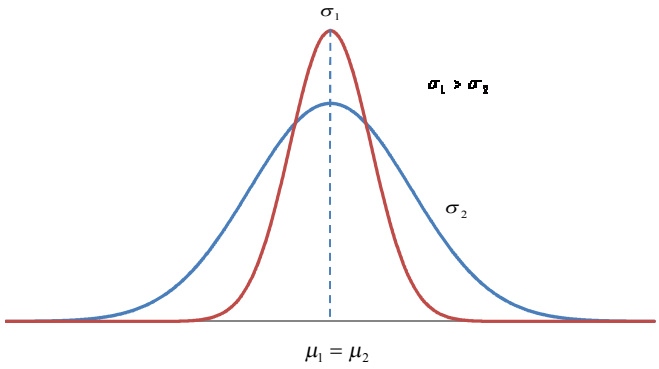
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};-\infty<x<+\infty$$

شکل عمومی توزیع نرمال همانگونه که در زیر نشان داده شده به دو پارامتر میانگین μ و واریانس σ^2 بستگی دارد.

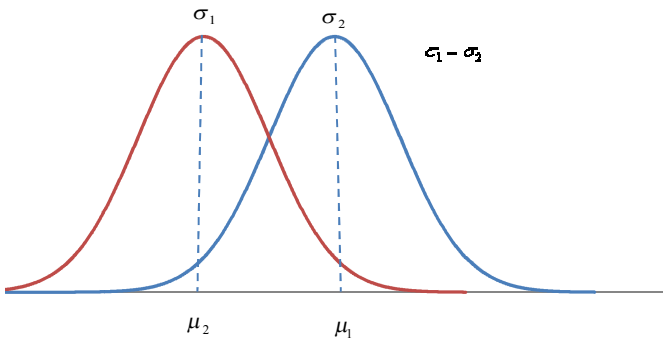


بنابراین انتظار داریم که شکل منحنی توزیع نرمال با تغییر میانگین و واریانس آن تغییر کند. با توجه به این نکته حالت‌های زیر ممکن است اتفاق بیافتد.

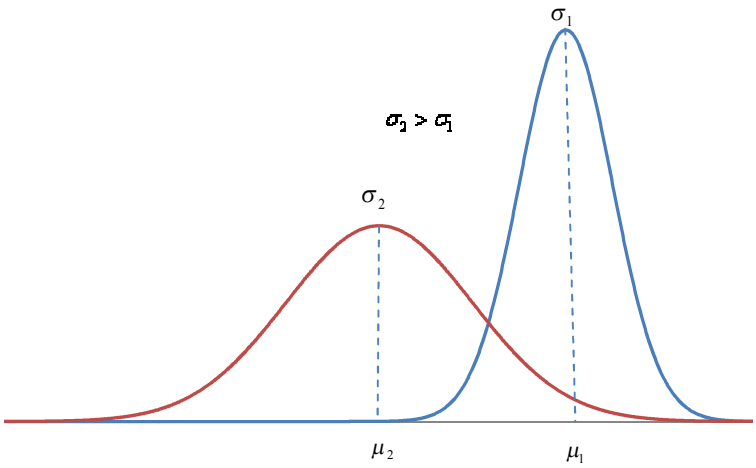
- میانگین دو توزیع برابر اما انحراف معیار آنان متفاوت باشد



- واریانس دو توزیع یکسان اما میانگین آنان متفاوف باشد



- میانگین و واریانس دو توزیع متفاوت باشند



ویژگی‌های اصلی توزیع نرمال

- از آنجایی که توزیع نرمال متقارن است بنابراین مقدار چولگی آن برابر صفر است. لذا $P(X \le \mu) = P(\mu \le X) = 0.5$ و میانگین = میانه = نما.
- کشیدگی توزیع نرمال برابر است با ۳. معیار کشیدگی هر توزیع با عدد ۳ یعنی کشیدگی توزیع نرمال مقایسه می‌شود.
- یک ترکیب خطی از متغیرهای با توزیع نرمال نیز دارای توزیع نرمال است. برای حالتی که متغیرهای تصادفی از هم مستقل باشند داریم

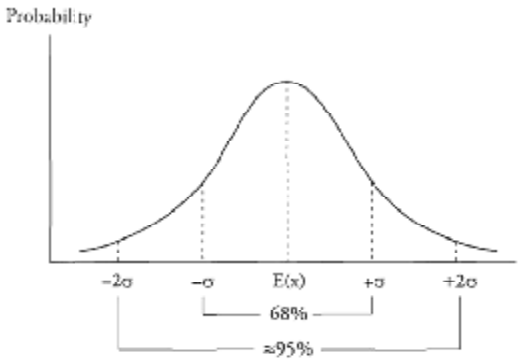
$$if \quad x_i \approx N(\mu,\sigma^2) \quad \forall i \rightarrow \sum_{i=1}^{n} x_i \approx N(n\mu, n\sigma^2)$$

- احتمال برآمدهای بالاتر و پایین‌تر از میانگین کوچکتر و کوچکتر می‌شوند اما به صفر میل نمی‌کنند. یعنی دنباله سری بسیار نازک می‌شود اما تا بینهایت ادامه پیدا می‌کند.

تعیین احتمال قرار گرفتن یک متغیر با توزیع نرمال در یک فاصله معین

فاصله اطمینان (confidence interval) دامنه‌ای از مقادیر حول برآمد مورد انتظار است که انتظار داریم برآمد واقعی در درصد مشخصی از تعداد دفعات انجام آزمایش در آن دامنه قرار گیرد. برای مثال فاصله اطمینان ۹۵٪ دامنه‌ای است که ما انتظار داریم متغیر تصادفی در ۹۵٪ از اوقات در آن قرار گیرد. برای یک توزیع نرمال این فاصله اطمینان بر اساس مقدار انتظاری (امید ریاضی) متغیر تصادفی، که گاهی اوقات برآورد نقطه‌ای نامیده می‌شود و نوسانات آن که با انحراف معیار آن متغیر اندازه‌گیری می‌شود، می‌باشد.

همانگونه که در نمودار زیر نشان داده شده برای هر متغیر تصادفی با توزیع نرمال ۶۸٪ برآمدها در فاصله یک انحراف معیار بالاتر و پایین‌تر از امید ریاضی و تقریباً ۹۵٪ از برآمدها در فاصله دو انحراف معیار بالاتر و پایین‌تر از امید ریاضی قرار دارند.



در عمل معمولاً مقادیر واقعی میانگین و انحراف مشخص نیستند اما می‌توان آنها را به صورت $\overline{X}$ و s برآورد نمود. سه فاصله اطمینان که بیشتر کاربرد دارند عبارتند از

- فاصله اطمینان ۹۰٪ برای X عبارت است از $\overline{X}-1.65\,s$ تا $\overline{X}+1.65\,s$.
- فاصله اطمینان ۹۵٪ برای X عبارت است از $\overline{X}-1.96\,s$ تا $\overline{X}+1.96\,s$.
- فاصله اطمینان ۹۹٪ برای X عبارت است از $\overline{X}-2.58\,s$ تا $\overline{X}+2.58\,s$.

مثال ۱۱: بازدهی متوسط یک صندوق سرمایه‌گذاری ۱۰/۵٪ در سال و انحراف معیار بازدهی متوسط آن ۱۸٪ است. چنانچه بازدهی تقریباً نرمال باشد، فاصله اطمینان ۹۵٪ برای بازدهی سال بعد چقدر است؟

$$\overline{X} \pm 1.96s = 10.5 \pm 1.96(18) = -24.78\% \text{ تا } 45.78\%$$

$$P(-24.78\% < R < 45.78) = 95\%$$

یعنی انتظار می‌رود که بازدهی سالانه ۹۵٪ از اوقات یا ۹۵ سال از ۱۰۰ سال در این فاصله باشد.

نقش همبستگی در توزیع نرمال چندمتغیره

می‌توان از توزیع نرمال چندمتغیره برای توضیح متغیرهای تصادفی پیوسته استفاده کرد اگر کلیه متغیرهای تصادفی دارای توزیع نرمال باشند. همانگونه که قبلاً اشاره شد یک ترکیب خطی از متغیرهای تصادفی با توزیع نرمال نیز دارای توزیع نرمال خواهد بود. برای مثال اگر بازدهی هر سهم در یک پورتفو دارای توزیع نرمال باشد بازدهی پورتفو نیز دارای توزیع نرمال خواهد بود.

همانند توزیع نرمال یک‌متغیره یک توزیع نرمال چندمتغیره را نیز می‌توان با میانگین و واریانس متغیرهای تصادفی توضیح داد. اما در توضیح توزیع نرمال چندمتغیره همبستگی دو به دوی بین متغیرها مهم است. همبستگی ویژگی است که توزیع چندمتغیره را از توزیع یک متغیره تفکیک می‌کند. همبستگی در واقع بیانگر وجود رابطه خطی بین دو متغیر تصادفی است.

با استفاده از بازدهی داراییها به عنوان متغیر تصادفی توزیع نرمال استاندارد برای بازدهی n دارایی را می‌توان با استفاده از سه مجموعه از پارامترهای زیر به طور کامل توضیح داد

- n میانگین n سری از بازدهی‌ها $(\mu_1, \mu_2, ..., \mu_n)$
- n واریانس n سری از بازدهی‌ها $(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$
- $0.5n(n-1)$ همبستگی دو به دو

توزیع نرمال استاندارد (standard normal distribution)

توزیع نرمال استاندارد توزیع نرمالی است که به نحوی استاندارد شده که میانگین آن صفر و واریانس آن ۱ است ($N\sim(0,1)$). برای استاندارد کردن یک مشاهده از یک توزیع نرمال معین، مقدار Z مشاهده باید محاسبه شود. مقدار Z تعداد انحراف معیارهایی را نشان می‌دهد که یک مشاهده داده شده از میانگین جامعه فاصله دارد. استانداردسازی فرایندی است که در آن مقدار مشاهده شده یک متغیر تصادفی به مقدار Z خود تبدیل می‌شود. نحوه استاندارد کردن متغیر تصادفی X که $X\sim N(\mu, \sigma)$ به صورت زیر می‌باشد

$$z = \frac{\text{میانگین جامعه} - \text{مقدارمشاهده شده}}{\text{انحراف معیار}} = \frac{x-\mu}{\sigma} \rightarrow z\sim N(0,1)$$

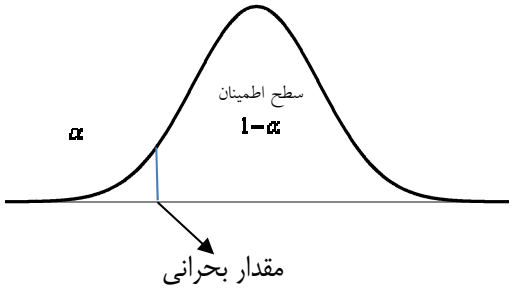
برای محاسبه احتمال در هر بازه زمانی برای متغیر تصادفی X با توزیع احتمال نرمال با میانگین μ و واریانس σ^2، ابتدا با استفاده از $\frac{x-\mu}{\sigma}$ فرم نرمال استاندار را بدست آورده و سپس با استفاده از جدول نرمال استاندارد مقدار احتمال مورد نظر را بدست می‌آوریم:

$$P(x<a) = P(\frac{x-\mu}{\sigma} < \frac{a-\mu}{\sigma}) = P(Z < \frac{a-\mu}{\sigma})$$

$$P(x>a) = P(\frac{x-\mu}{\sigma} > \frac{a-\mu}{\sigma}) = P(Z > \frac{a-\mu}{\sigma}) = 1 - P(Z < \frac{a-\mu}{\sigma})$$

$$P(a<x<b) = P(\frac{a-\mu}{\sigma} < \frac{x-\mu}{\sigma} < \frac{b-\mu}{\sigma}) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = 1 - P(Z < \frac{a-\mu}{\sigma})$$

جدول توزیع نرمال یا جدول Z جدولی است که شامل مقادیر تولید شده با استفاده از تابع توزیع تجمعی توزیع نرمال استاندارد یا $F(Z)$ است. بنابراین مقادیر جدول Z احتمال مشاهده مقدار Z است که کمتر از یک مقدار مشخص است یعنی $P(Z<z)$. توجه داشته باشید که $F(-Z) = 1-F(Z)$. مقدار Z جدول را مقدار بحرانی می‌نامند.



مثال ۱۲: در یک توزیع نرمال با میانگین ۳۲ و واریانس ۴، تقریباً چند درصد از داده‌ها بین دو عدد ۳۸ و ۲۶ قرار می‌گیرند؟ ($S_{-\infty}^{-3} = 0.0013$) (حسابداری و مدیریت ۸۵)

۱) ۸۹/۶ ۲) ۹۲/۳ ۳) ۹۵/۴ ۴) ۹۹/۷

جواب:

$$P(Z<-3) = P(Z>3) = S_{-\infty}^{-3} = 0.0013$$
$$P(26<x<38) = P(\frac{26-32}{2} < Z < \frac{38-32}{2}) = P(-3<Z<3) = 1-P(Z<-3)-P(Z>3)$$
$$= 1-2P(Z<-3) = 1-2P(Z>3) = 1-2\times 0.0013 = 99.74\%$$

مثال ۱۳: در ۱۲۰ داده آماری دسته‌بندی شده نمودار بافتنگار فراوانی مطلق متقارن مجموع این داده‌ها ۸۴۰ و مجموع مربعات آنها ۶۱۵۰ می‌باشد. تقریباً ۹۵ درصد از داده‌ها در کدام بازه قرار می‌گیرند؟ (حسابداری و مدیریت ۸۷)

۱) (۴و۹) ۲) (۵و۹) ۳) (۴و۱۰) ۴) (۵و۱۰)

جواب:

بر اساس تعریف ماینگین و واریانس

$$\mu = \frac{\sum x_i}{n} = \frac{840}{120} = 7$$
$$\sigma^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2 = \frac{6150}{120} - \left(\frac{840}{120}\right)^2 = 2.25 \rightarrow \sigma = 1.5$$
$$P(\mu \pm 2\sigma) = 0.95 \rightarrow \mu \pm 2\sigma = 7 \pm 2\times 1.5 = (4,10)$$

مثال ۱۴: در یک توزیع نرمال با انحراف معیار ۵ داریم $P(x \geq 9.8) = 0.67$ و $P(Z<-0.44) = 0.33$. میانگین این توزیع کدام است؟ (GIS ۸۶)

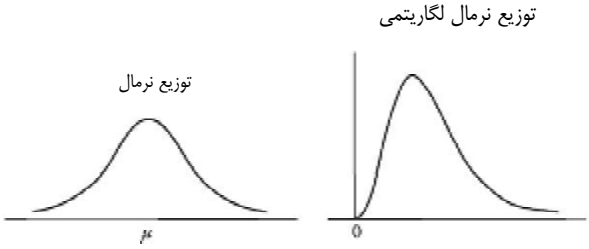
۱) ۸ ۲) ۹ ۳) ۱۱ ۴) ۱۲

جواب:

$$P(x \geq 9.8) = 0.67 \rightarrow P(\frac{x-\mu}{\sigma} \geq \frac{9.8-\mu}{5}) = P(Z \geq \frac{9.8-\mu}{5}),$$
$$P(Z > -0.44) = 1 - P(Z < -0.44) = 0.67 \rightarrow \frac{9.8-\mu}{5} = 0.44 \rightarrow \mu = 12$$

توزیع نرمال و توزیع لگاریتمی نرمال

توزیع لگاریتمی نرمال با استفاده از تابع e^x تولید می‌شود که در آن x دارای توزیع نرمال است. از آنجایی که $\ln(e^x) = x$، لگاریتم یک متغیر تصادفی لگاریتم نرمال دارای توزیع نرمال است. تفاوت توزیع نرمال و توزیع لگاریتمی نرمال در شکل زیر نشان داده شده است



بنابراین

- توزیع نرمال لگاریتمی به سمت راست چوله است
- توزیع نرمال لگاریتمی از پایین به صفر محدود است. لذا این توزیع در مدل‌سازی قیمت داراییها که هرگز مقدار منفی به خود نمی‌گیرند مورد استفاده قرار می‌گیرد.

چنانچه از یک توزیع نرمال بازدهی‌ها برای مدل‌سازی قیمت داراییها در طول زمان استفاده شود، احتمال بازدهی کمتر از ۱۰۰-٪ قبول شده که به معنی قبول امکان کمتر از صفر بودن قیمت دارایی است. استفاده از توزیع نرمال لگاریتمی در مدل‌سازی نسبت قیمت‌ها این مشکل را از بین می‌برد. نسبت قیمت عبارت است از قیمت دارایی در انتهای دوره تقسیم بر قیمت دارایی در ابتدای دوره (S_1/S_0) و برابر است با (بازدهی دوره نگهداری + ۱). برای رسیدن به قیمت دارایی در پایان دوره می‌توان به سادگی نسبت قیمت را به قیمت دارایی در ابتدای دوره ضرب کرد. از آنجایی که توزیع نرمال لگاریتمی دارای حداقل مقدار صفر است، قیمت دارایی در انتهای دوره نمی‌تواند کمتر از صفر باشد. نسبت قیمت برابر با صفر متناظر با بازدهی دوره نگهداری برابر با ۱۰۰-٪ است. در بحث درخت دوجمله‌ای ما از اصطلاح (ضرایب) تغییر به سمت بالا و تغییر به سمت پایین برای نسبت قیمت استفاده کردیم.

تفاوت بین نرخ بازده مرکب گسسته و پیوسته

بازده مرکب گسسته بازدهی مرکبی است که با آن آشنا هستیم است که درآن دوره مرابحه گسسته است برای مثال دوره شش ماهه یا سه ماهه. هر چه قدر دوره زمانی مرابحه بیشتر باشد، بازده مؤثر سالانه بیشتر خواهد بود. برای نرخ بازده سالانه ۱۰ درصد، بازده مرکب شش ماهه بازده مؤثری برابر با $\left(1+\frac{0.10}{2}\right)^{2}-1=10.25\%$ و بازده مرکب ماهانه بازده مؤثری برابر با $\left(1+\frac{0.10}{12}\right)^{12}-1=10.47\%$ بدست می‌دهد. بازده مرکب روزانه یا حتی ساعتی بازده مؤثر بالاتری را تولید خواهد کرد. در حالت حدی زمانی که دوره‌ها کوتاه‌تر و کوتاه‌تر می‌شوند بازده مرکب گسسته تبدیل به بازده مرکب پیوسته می‌شود. نرخ مؤثر سالانه بر اساس مرابحه مرکب با نرخ سالانه R_{cc} را می‌توان بر اساس رابطه زیر محاسبه نمود

$$\text{نرخ مؤثر سالانه} = e^{R_{cc}}-1$$

با استفاده از تابع ln می‌توان نرخ بازده مرکب پیوسته را از نرخ مؤثر سالانه بدست آورد. می‌توان از این روش برای پیدا کردن نرخ بازده مرکب پیوسته‌ای که نرخ دوره نگهداری مشخصی را ایجاد می‌کند استفاده کرد. از آنجایی که (S_1/S_0) = (بازدهی دوره نگهداری + ۱)، می‌توان محاسبات را مستقیماً از نسبت قیمت انجام داد. چنانچه نرخ دوره نگهداری را با HPR نشان دهیم، نرخ بازده مرکب پیوسته عبارت است از

$$\ln\left(\frac{S_1}{S_0}\right)=\ln(1+HPR)=R_{cc}$$

در حالت کلی بازده دوره نگهداری بعد از T سال با نرخ بازده مرکب سالانه R_{cc} عبارت است از:

$$HPR_T=e^{R_{cc}\times T}-1$$

فصل سوم

نمونه‌گیری و برآورد

نمونه‌گیری تصادفی ساده (simple random sampling) و خطای نمونه‌گیری (sampling error)

نمونه‌گیری تصادفی ساده روش انتخاب نمونه به روشی است که هر فرد یا جزء از جامعه مورد مطالعه احتمال یکسانی در انتخاب شدن در نمونه را داشته باشند. روش دیگر نمونه‌گیری تقریبی نمونه‌گیری تصادفی منظم یا سیستماتیک است که در آن در هر مرحله nامین عضو از جامعه انتخاب می‌شود. هر ویژگی یک نمونه با آماره نشان داده می‌شود در حالی که هر ویژگی جامعه با پارامتر نشان داده می‌شود. برای مثال میانگین نمونه یک آماره است در حالی که میانگین جامعه یک پارامتر است. مقدار آماره در هر نمونه ممکن است متفاوت باشد بنابراین آماره خود یک متغیر تصادفی است در حالی که پارامتر یک جامعه کمیت ثابتی است. خطای نمونه‌گیری تفاوت بین آماره نمونه و پارامتر جامعه متناظر با آن است.

نمونه‌گیری تصادفی طبقه‌ای یا گروهی (stratified random sampling)

نمونه‌گیری تصادفی طبقه‌ای از یک سیستم طبقه‌بندی شده برای تفکیک کردن جامعه به گروه‌های کوچکتر بر اساس یک یا چند ویژگی متمایز استفاده می‌کند. برای هر زیرگروه یا طبقه یک نمونه تصادفی انتخاب شده و نمونه‌ها با هم جمع می‌شوند. حجم هر نمونه از هر طبقه بر اساس نسبت طبقات به حجم جامعه تعیین می‌شود.

به دلیل مشکلات و هزینه مربوط به گزارش کامل کلیه انواع اوراق قرضه، نمونه‌گیری طبقه‌ای معمولاً برای شاخص‌بندی اوراق قرضه استفاده می‌شود. در این مورد اوراق قرضه در یک جامعه بر اساس بالاترین عامل ریسک ورقه قرضه مانند دوره، سررسید، نرخ کوپن و مانند اینها طبقه‌بندی می‌شوند. آنگاه نمونه‌ها از هر طبقه مجزا استخراج شده و با هم ترکیب شده تا نمونه نهایی تعیین شود.

توزیع نمونه‌گیری

تشخیص اینکه آماره نمونه خود یک متغیر تصادفی است و بنابراین دارای توزیع احتمال است اهمیت دارد. توزیع نمونه‌گیری آماره نمونه توزیع احتمال کلیه آماره‌های ممکن است که از مجموعه‌ای از نمونه‌های با حجم یکسان که به صورت تصادفی از یک جامعه انتخاب شده‌اند، بدست می‌آید. برای مثال می‌توان توزیع نمونه‌ای میانگین یا انحراف معیار بدست آورد.

تفاوت بین داده‌های سری زمانی (time-series data) و داده‌های مقطعی (cross-sectional data)

داده‌های سری زمانی شامل مشاهداتی هستند که در طول یک دوره زمانی و در فاصله زمانی مشخص و برابر بدست آمده‌اند. برای مثال بازدهی کل بورس اوراق بهادار تهران در دوره زمانی فرودین ۱۳۸۰ تا فروردین ۱۳۹۰ یک سری زمانی را تشکیل می‌دهد.

داده‌های مقطعی نمونه‌ای از مشاهداتی هستند که در یک نقطه خاص از زمان گرفته بدست آمده‌اند. برای مثال نمونه گزارش P/E کلیه شرکت‌های NASDAQ در ۳۱ دسامبر ۲۰۱۰ یک نمونه با داده‌های مقطعی است.

قضیه حد مرکزی

قضیه حد مرکزی بیانگر این است که برای نمونه‌های تصادفی ساده با حجم n از یک جامعه با میانگین μ و واریانس محدود σ^2، با افزایش حجم نمونه توزیع نمونه‌ای میانگین نمونه $\bar{x}$ به توزیع احتمال نرمال با میانگین μ و واریانس $\frac{\sigma^2}{n}$ میل می‌کند. قضیه حد مرکزی از کاربرد بسیاری برخوردار است زیرا بکارگیری توزیع نرمال در آزمون فرضیه و ساختن فاصله اطمینان بسیار آسان است. تا زمانی که حجم نمونه به اندازه کافی بزرگ است (که معمولاً به معنی $n \geq 30$ است)، بدون توجه به توزیع جامعه بر اساس میانگین نمونه می‌توان استنباط‌های خاصی در مورد میانگین جامعه انجام داد.

ویژگی‌های مهم قضیه حد مرکزی عبارت است از

- اگر حجم نمونه به اندازه کافی بزرگ باشد($n \geq 30$)، توزیع نمونه‌ای میانگین نمونه‌ها تقریباً نرمال خواهد بود. توجه داشته باشید که مفهوم این قضیه چیست. نمونه‌های تصادفی با حجم n به طور مداوم از یک جامعه انتخاب می‌شوند، هر کدام از این نمونه‌های تصادفی میانگین خود را دارند که خود یک متغیر تصادفی است و این مجموعه میانگین نمونه‌ها دارای توزیعی است که تقریباً نرمال است.
- میانگین جامعه μ، و میانگین توزیع کلیه میانگین‌های نمونه برابر است.
- واریانس توزیع میانگین نمونه برابر است با $\frac{\sigma^2}{n}$ یعنی واریانس جامعه تقسیم بر حجم نمونه.
- خطای استاندارد میانگین نمونه عبارت است از انحراف معیار توزیع میانگین نمونه‌ها. زمانی که انحراف معیار جامعه، σ مشخص است، خطای استاندارد میانگین نمونه عبارت است از $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- معمولاً انحراف معیار جامعه مشخص نیست. در این حالت از خطای استاندارد نمونه به جای انحراف معیار جامعه استفاده می‌شود بنابراین در این حالت $s_{\bar{x}}^2 = \frac{s^2}{n} \rightarrow s_{\bar{x}} = \frac{s}{\sqrt{n}}$
- با افزایش حجم نمونه میانگین نمونه به میانگین جامعه نزدیکتر می‌شود. همچنین با افزایش حجم نمونه خطای استاندارد کاهش می‌یابد.

مثال ۱: توزیع میانگین‌های نمونه یک جامعه نامحدود با میانگین ۱۰ و انحراف معیار ۲ دارای واریانس ۱ خواهد بود اگر تعداد نمونه عبارت باشد از: (مدیریت ۷۴)

۱) ۴ ۲) ۱۰ ۳) ۱۶ ۴) ۳۳

جواب:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \rightarrow 1 = \frac{2}{\sqrt{n}} \rightarrow n = 4$$

مثال ۲: اگر بخواهیم انحراف معیار میانگین نمونه‌ای ($\sigma_{\bar{x}}$) بر اساس حجم نمونه n = ۶۴ تایی از جامعه‌ای که دارای انحراف معیار ۶ است به نصف کاهش یابد، حجم نمونه باید چند تا شود؟ (اقتصاد ۸۲)

۱) ۱۲۸ ۲) ۱۸۲ ۳) ۲۵۶ ۴) ۳۲۰

جواب:

$$\sigma_{\bar{x}}' = \frac{1}{2}\sigma_{\bar{x}}'' \rightarrow \frac{1}{2}\frac{\sigma}{\sqrt{n}} \rightarrow \frac{\sigma}{\sqrt{4n}} \rightarrow 4n = 4\times 64 = 256$$

برآورد نقطه‌ای و برآورد فاصله‌ای (برآورد فاصله اطمینان)

در برآورد نقطه‌ای مقدار آماره نمونه برای برآورد پارامتر جامعه مورد استفاده قرار می‌گیرد. فرمول مورد استفاده برای محاسبه برآورد نقطه‌ای برآوردگر نامیده می‌شود. برای مثال میانگین نمونه $\bar{x}$ یک برآوردگر میانگین جامعه μ است و با استفاده از فرمول زیر محاسبه می‌شود. $\bar{x} = \frac{\sum x}{n}$

فاصله اطمینان دامنه‌ای از مقادیری است که انتظار می‌رود پارامتر جامعه با احتمال مشخص $1-\alpha$ در آن قرار گیرد. α *سطح اطمینان* برای فاصله اطمینان و احتمال $1-\alpha$ درجه *اطمینان* نامیده می‌شود. فاصله اطمینان با کم و زیاد کردن یک مقدار مناسب به برآورد نقطه‌ای پارامتر جامعه بدست می‌آید. در حالت کلی فاصله اطمینان به شکل زیر است

برآوردکننده نقطه‌ای $\pm$ (ضریب اطمینان $\times$ خطای استاندارد)

که در آن ضریب اطمینان عددی است که به توزیع نمونه‌ای برآوردکننده نقطه‌ای و احتمالی که برآوردکننده نقطه‌ای در فاصله اطمینان قرار می‌گیرد ($1-\alpha$) بستگی دارد. برای مثال در مورد فاصله اطمینان میانگین داریم

$$\bar{x} \pm \varepsilon \rightarrow \bar{x} - \varepsilon < \mu_x < \bar{x} + \varepsilon$$

ویژگی‌های لازم برای یک برآوردکننده مطلوب

فارغ از نوع برآورد (نقطه‌ای یا فاصله‌ای)، ویژگی‌های آماری خاصی وجود دارند که برخی از برآوردکننده‌ها را از بقیه مطلوب‌تر می‌سازند. این ویژگی‌های مطلوب عبارتند از *ناارییی*، *کارایی* و *سازگاری*.

- برآوردکننده ناریب برآوردکننده‌ای است که امیدریاضی آن برابر با پارامتری است که سعی دارد آن را برآورد کند. برای مثال امیدریاضی میانگین نمونه برابر است با میانگین جامعه $E(\bar{x}) = \mu$ ، بنابراین میانگین نمونه یک برآوردکننده ناریب میانگین جامعه است. به عبارت دیگر برآوردکننده $\hat{\lambda}$ یک برآوردکننده ناریب پارامتر λ است اگر $E(\hat{\lambda} - \lambda) = 0$. بنابراین میزان اریب یا تورش برآوردکننده $\hat{\lambda}$ برابر است با $E(\hat{\lambda} - \lambda) = E(\hat{\lambda}) - \lambda$.
- یک برآوردکننده ناریب کارا نیز هست اگر واریانس توزیع نمونه‌ای آن برآوردکننده کمتر از کلیه برآوردکننده‌های ناریب دیگر پارامتر مورد نظر باشد. برای مثال میانگین نمونه یک برآوردکننده ناریب و کارای میانگین جامعه است.

کارایی نسبی دو برآوردکننده نااریب $\hat{\lambda}_1$ و $\hat{\lambda}_2$ به صورت نسبت واریانس $\hat{\lambda}_2$ به $\hat{\lambda}_1$ تعریف می‌شود یعنی $\frac{\text{var}(\hat{\lambda}_2)}{\text{var}(\hat{\lambda}_1)}$

. بنابراین $\hat{\lambda}_1$ کاراتر از $\hat{\lambda}_2$ است اگر $\frac{\text{var}(\hat{\lambda}_2)}{\text{var}(\hat{\lambda}_1)} > 1$.

- یک برآوردکننده سازگار برآوردکننده‌ای است که صحت برآورد پارامتر در آن با افزایش حجم نمونه افزایش می‌یابد. با افزایش حجم نمونه خطای استاندارد میانگین نمونه کاهش می‌یابد و توزیع نمونه‌ای حول میانگین جامعه جمع‌تر و فشرده‌تر می‌شود. در واقع زمانی که حجم نمونه به سمت بینهایت میل می‌کند، خطای استاندارد به صفر میل می‌کند. در این حالت توزیع نمونه‌ای تباهیده شده و میانگین نمونه برابر با میانگین جامعه می‌شود.

مثال ۳: آماره $\bar{x}$ یک آماره سازگار است چون وقتی n به سمت بینهایت میل می‌کند $\bar{x}$ به سمت ..... میل می‌کند. (مدیریت ۸۳)

۱) ∞ ۲) $N\mu_x$ ۳) μ_x ۴) صفر

جواب:

چون میانگین نمونه برآوردکننده سازگار میانگین جامعه است پس گزینه ۳ صحیح است.

توزیع کای‌دو (χ^2)

چنانچه V توزیع نرمال استاندارد به توان دو رسیده و با هم جمع شوند این مجموع دارای توزیع کای‌دو با درجه آزادی V خواهد بود.

$$Z=\left(\frac{x_i-\mu_i}{\sigma_i}\right), i=1,\ldots v, \rightarrow \chi^2_v=\sum_i^v Z_i^2=\sum_i^v\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$$

بنابراین بر اساس تعریف توزیع کای‌دو چنانچه متغیر تصادفی X دارای توزیع کای‌دو با درجه آزادی V_1 و متغیر تصادفی Y دارای توزیع کای‌دو با درجه آزادی V_2 باشد آنگاه

$$x\pm y\approx\chi^2_{V_1\pm V_2}$$

این توزیع برای $n\leq 10$ یعنی درجه آزادی کم دارای چولگی مثبت بوده و برای $n<10$ چولگی آن کاهش یافته و به سمت نرمال میل می‌کند.

توزیع F (فیشر)

توزیع F از تقسیم دو توزیع کای‌دو که هر کدام بر درجه آزادی خود تقسیم شده‌اند بدست می‌آید. این توزیع دارای دو پارامتر می‌باشد که همان درجه آزادی توزیع کای‌دو در صورت و مخرج آن هستند. بنابراین توزیع F با درجه آزادی V_1 و V_2 عبارت است از

$$F_{v_1,v_2}=\frac{\frac{\chi^2_{v_1}}{v_1}}{\frac{\chi^2_{v_2}}{v_2}}$$

ویژگی‌های توزیع F

- تنها یک نما دارد.
- دارای چولگی مثبت است.
- $F_{1-\alpha/2,v_1,v_2}=\frac{1}{F_{\alpha/2,v_2,v_1}}$
- $t_v=\frac{Z}{\sqrt{\frac{\chi^2_v}{v}}}\rightarrow (t_v)^2=\left(\frac{Z}{\sqrt{\frac{\chi^2_v}{v}}}\right)^2=\frac{Z^2}{\frac{\chi^2_v}{v}}=\frac{\frac{\chi^2_1}{1}}{\frac{\chi^2_v}{v}}=F_{1,v}$

توزیع *t*-student

توزیع *t*-student یا توزیع t یک توزیع زنگوله‌ای شکل است که نسبت به میانگین متقارن است. این توزیع، توزیع مناسب برای ساختن فاصله اطمینان بر اساس نمونه‌های کوچک ($n<30$) از جامعه‌ای با واریانس نامعلوم و توزیع نرمال یا تقریباً نرمال است. زمانی که واریانس جامعه نامعلوم و حجم نمونه به اندازه کافی بزرگ است به نحوی که بر اساس قضیه حد مرکزی توزیع نمونه‌ای تقریباً نرمال است نیز می‌توان از توزیع t استفاده نمود.

چنانچه Z دارای توزیع نرمال استاندارد و متغیر تصادفی کای‌دو با درجه آزادی v و مستقل از Z مفروض باشد آنگاه توزیع t با درجه آزادی v برابر خواهد بود با

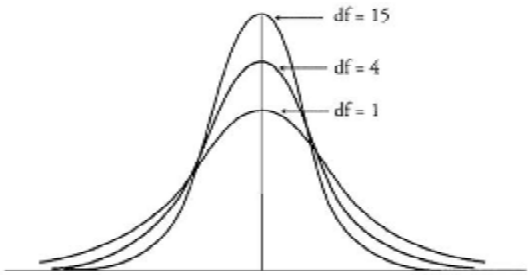
$$t_v=\frac{Z}{\sqrt{\frac{\chi^2_v}{v}}}$$

ویژگی‌های توزیع *t*

- این توزیع متقارن است.
- این توزیع با یک پارامتر تعریف می‌شود، درجه آزادی (df). درجه آزادی آن برای میانگین نمونه برابر است با تعداد مشاهدات نمونه منهای یک، n-1.
- در دنباله این توزیع نسبت به توزیع نرمال احتمال بیشتری وجود دارد. یعنی دنباله آن پهن‌تر است. بنابراین در آزمون فرضیه با استفاده از توزیع t رد فرضیه صفر نسبت به توزیع Z مشکل‌تر است.
- با افزایش درجه آزادی آن (یعنی افزایش حجم نمونه) شکل توزیع t به توزیع نرمال استاندارد نزدیک می‌شود. این موضوع در نمودار زیر نشان داده شده است.

- امید ریاضی و واریانس متغیر تصادفی x با توزیع t برابر است با

$$E(x)=0, \quad n>1; \quad \sigma^2(x)=\frac{n}{n-2}, \quad n>2$$



محاسبه و تفسیر فاصله اطمینان برای میانگین جامعه با توزیع نرمال

۱) واریانس جامعه مشخص باشد

چنانچه جامعه دارای توزیع نرمال با میانگین μ و واریانس مشخص σ^2 باشد، فاصله اطمینان برای میانگین جامعه را می‌توان به صورت زیر محاسبه نمود.

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow P(\bar{x}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}})=1-\alpha$$

که در آن $Z_{\alpha/2}$ ضریب اطمینان است. یعنی یک متغیر تصادفی نرمال استاندارد که احتمال قرار گرفتن آن در دنباله سمت راست $\alpha/2$ است. ضریب اطمینان‌هایی که بیشتر مورد استفاده قرار می‌گیرند عبارتند از

- $Z_{\alpha/2}=1.645$ برای فاصله اطمینان ۹۰٪ (سطح اطمینان ۱۰٪ و برای هر دنباله ۵٪)
- $Z_{\alpha/2}=1.960$ برای فاصله اطمینان ۹۵٪ (سطح اطمینان ۵٪ و برای هر دنباله ۲/۵٪)
- $Z_{\alpha/2}=2.757$ برای فاصله اطمینان ۹۹٪ (سطح اطمینان ۱٪ و برای هر دنباله ۰/۵٪)

برای مثال در مورد دوم این موضوع بدین معنی است که احتمال زیر منحنی نرمال استاندارد بین $Z=-1.960$ و $Z=+1.960$ برابر است با ۹۵٪. به دلیل متقارن بودن توزیع نرمال استاندارد ۲/۵ درصد احتمال قرار گرفتن در دنباله پایین‌تر $Z=-1.960$ و بالاتر از $Z=+1.960$ می‌باشد.

تفسیر آماری فاصله اطمینان به این شکل است که بعد از نمونه‌گیری مکرر از جامعه و ساختن فاصله اطمینان برای میانگین هر نمونه، در بلندمدت $1-\alpha$ درصد از فواصل اطمینان بدست آمده میانگین جامعه را در بر می‌گیرند. تفسیر کاربردی فاصله اطمینان به این شکل است که $1-\alpha$ درصد مطمئن هستیم که میانگین جامعه بین $\bar{x}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ و $\bar{x}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ قرار می‌گیرد.

۲) واریانس جامعه مشخص نباشد

چنانچه توزیع جامعه نرمال باشد اما واریانس آن مشخص نباشد، می‌توان از توزیع t برای ساختن فاصله اطمینان استفاده نمود.

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \rightarrow P(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}) = 1 - \alpha$$

که در آن $t_{\alpha/2}$ ضریب اطمینان متناظر با یک متغیر تصادفی با توزیع t با $n-1$ درجه آزادی است که n حجم نمونه است. سطح زیر دنباله توزیع t در سمت راست $t_{\alpha/2}$ برابر است با $\alpha/2$ و $\frac{s}{\sqrt{n}}$ خطای استاندارد میانگین نمونه و s انحراف معیار نمونه است. برعکس توزیع نرمال استاندارد، ضریب اطمینان توزیع t به حجم نمونه بستگی دارد بنابراین نمی‌توانیم به ضرایب اطمینان معمول تکیه کنیم. بنابراین برای بدست آوردن ضریب اطمینان توزیع t باید به جدول توزیع t مراجعه کنیم. به دلیل پهن‌تر بودن دنباله توزیع t فاصله اطمینان ساخته شده با استفاده از ضریب اطمینان t ($t_{\alpha/2}$) بیشتر از فاصله اطمینان ساخته شده با استفاده از ضریب اطمینان Z ($Z_{\alpha/2}$) است.

محاسبه فاصله اطمینان برای میانگین جامعه با استفاده از یک نمونه با حجم بزرگ و هر نوع توزیع

در این حالت حجم نمونه در ساختن فاصله اطمینان مناسب برای میانگین نمونه تأثیرگذار خواهد بود.

- چنانچه توزیع نرمال نباشد اما واریانس جامعه مشخص باشد بر اساس قضیه حد مرکزی می‌توان از آماره Z استفاده کرد اگر حجم نمونه بزرگ باشد ($n \geq 30$).
- چنانچه توزیع غیرنرمال و واریانس جامعه نامشخص باشد می‌توان از آماره t استفاده کرد اگر حجم نمونه بزرگ باشد ($n \geq 30$).
- بنابراین زمانی که از یک جامعه غیرنرمال نمونه‌گیری می‌کنیم برای یک نمونه با حجم کمتر ۳۰ نمی‌توانیم فاصله اطمینان بسازیم.
- کلیه این تحلیل‌ها به این بستگی دارد که نمونه‌ای که ما از جامعه استخراج می‌کنیم تصادفی باشد. چنانچه نمونه تصادفی نباشد، قضیه حد مرکزی قابل کاربرد نبوده و برآوردهای ما ویژگی‌های مطلوب را نخواهند داشت و نمی‌توانیم فاصله اطمینان ناریب بسازیم.

مثال ۴: نمرات یک نمونه تصادفی ۳ تایی از دانشجویان کلاسی که دارای توزیع نرمال است، ۱۶، ۱۵ و ۱۷ بوده است. فاصله اطمینان ۹۰ درصد میانگین نمرات دانشجویان کدام است؟ ($t \approx 3$) (اقتصاد ۸۶)

۱) ۱۵/۳–۱۶/۷ ۲) ۱۴/۳–۱۷/۷ ۳) ۱۳/۹–۱۸/۱ ۴) ۱۳/۷–۱۸/۳

جواب:

با توجه به اینکه $n < 30$ بنابراین

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \rightarrow P(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}) = 1 - \alpha$$

$$\bar{x} = 16, \quad s^2 = 1$$

$$P(\bar{x} - t_{0.05} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{0.05} \frac{s}{\sqrt{n}}) = 0.9 \rightarrow P(16 - 3\frac{1}{\sqrt{3}} < \mu < 16 + 3\frac{1}{\sqrt{3}}) = 0.9$$

$$P(14.3 < \mu < 17.7) = 0.9$$

مثال ۵: فرض کنید یک نمونه ۲۵ تایی از یک توزیع نرمال با میانگین μ و واریانس ۱۶ انتخاب شده است و میانگین نمونه‌ای برابر با ۱۰ است. یک فاصله اطمینان ۹۵٪ برای میانگین جامعه کدام است ؟ (محیط زیست ۸۷)

۱) $10 \pm \frac{8}{5}$ ۲) $10 \pm 2 \times \frac{16}{5}$ ۳) $10 \pm \frac{32}{25}$ ۴) $\frac{10}{25} \pm \frac{16}{5}$

جواب:

از آنجایی که توزیع جامعه نرمال و واریانس آن مشخص است بنابراین

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow 10 \pm 1.96 \times \frac{4}{\sqrt{25}} = 10 \pm 1.96 \times \frac{4}{5} \approx 10 \pm 2 \times \frac{4}{5} = 10 \pm \frac{8}{5}$$

مثال ۶: یک نمونه تصادفی از ۶۴ لامپ نشان می‌دهد که عمر متوسط نمونه ۳۵۰ ساعت است. یک فاصله اطمینان ۹۵ ٪ برای متوسط طول عمر واقعی لامپ‌ها با فرض $\sigma_x = 100$ عبارت است از:(اقتصاد ۸۵)

۱) ۱۵۰ تا ۵۵۰ ۲) ۱۵۴ تا ۵۴۶ ۳) ۲۵۰/۵تا ۴۴۹/۵ ۴) ۳۲۵/۵تا ۳۷۴/۵

جواب:

از آنجایی که $n \geq 30$ داریم

$$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow 350 \pm 1.96 \times \frac{100}{\sqrt{64}} = 350 \pm 1.96 \times \frac{100}{8} \approx 350 \pm 24.5 = (325.5, 374.5)$$

تعیین حجم مناسب نمونه و انواع خطاهای نمونه‌گیری

همانگونه که ملاحظه شد هر چقدر حجم نمونه بالاتر باشد خطای نمونه‌گیری کاهش خواهد یافت زیرا با افزایش حجم نمونه خطای استاندارد برآوردهای نقطه‌ای پارامترهای جامعه کمتر شده و فاصله اطمینان کاهش می‌یابد. اما این دیدگاه که حجم نمونه بالاتر بهتر است محدودیت‌هایی وجود دارد. اولاً نمونه‌های بزرگتر ممکن است مشاهداتی از جوامع (و بنابراین توزیع‌های) مختلف را در بر گیرند. در این حالت لزوماً خطای نمونه‌گیری بهبود نمی‌یابد و حتی ممکن است صحت برآوردکننده کاهش یابد. یکی از ملاحظات دیگر هزینه نمونه‌گیری است. هزینه نمونه‌گیری بزرگتر باید نسبت به افزایش دقت حاصله از آن سنجیده شود. بنابراین ممکن است حجم نمونه بالا همیشه خوب نباشد.

خطاهای نمونه‌گیری عبارتند از:

- داده‌کاوی زمانی اتفاق می‌افتد که تحلیل‌گران مکرراً از پایگاه داده‌ای یکسانی برای رسیدن به مدل استفاده کنند تا زمانی که مدلی که به خوبی کار می‌کند کشف شود. **خطای داده‌کاوی** (**data mining bias**) اشاره به نتایجی دارد که در آن معناداری آماری الگو بیش از حد برآورد می‌شود زیرا نتایج از طریق داده‌کاوی بدست آمده‌اند.
- **خطای انتخاب نمونه (sample selection bias)** زمانی رخ می‌دهد که برخی از داده ها معمولاً به دلیل در دسترس نبودن به‌صورت سیستماتیک از تحلیل حذف می‌شوند. این عمل باعث می‌شود که نمونه مشاهده شده غیر نرمال باشد و هر استدلالی که از این نمونه می‌شود را نمی‌توان به جامعه تعمیم داد زیرا نمونه مشاهده شده و نسبتی از جامعه که مشاهده نشده‌اند متفاوت هستند.
- **خطای ابقاء (survivorship bias)** معمول‌ترین نوع خطای انتخاب نمونه است. یک مثال خوب از وجود خطای ابقاء در سرمایه‌گذاری مطالعه عملکرد صندوق سرمایه‌گذاری است. بیشتر پایگاه‌های اطلاعاتی صندوق سرمایه‌گذاری تنها شامل صندوق‌هایی است که هم اکنون وجود دارند اما صندوق‌هایی که به خاطر ادغام یا بسته شدن دیگر وجود ندارند را کنار می‌گذارد.
- **خطای پیش‌نگری (look-ahed bias)** زمانی اتفاق می‌افتد که یک مطالعه رابطه‌ای را با استفاده از داده‌های نمونه‌ای آزمون می‌کند که در زمان انجام آزمون در دسترس نیستند. برای مثال آزمون کردن قاعده معامله‌ای را در نظر بگیرید که بر اساس نسبت قیمت به ارزش دفتری در پایان سال مالی است. قیمت سهام برای همه شرکتها در آن زمان در دسترس است اما ارزش دفتری در پایان سال ممکن است تا ۳۰ تا ۶۰ روز بعد از سال مالی در دسترس نباشد. برای انجام چنین آزمونی ممکن است از برآوردی از ارزش دفتری استفاده شود.
- **خطای دوره زمانی (time-period bias)** می‌تواند به این خاطر باشد که دوره زمانی که در آن داده‌ها جمع‌آوری می‌شوند خیلی کوتاه یا خیلی طولانی باشد. چنانچه دوره زمانی خیلی کوتاه باشد ممکن است نتایج تحقیق پدیده‌ای را منعکس کند که خاص آن دوره زمانی است. اگر دوره زمانی خیلی طولانی باشد ممکن است روابط اساسی اقتصادی که پایه نتایج می‌باشند، تغییر کنند.

فصل چهارم

آزمون فرضیه

مقدمه

آزمون فرضیه ارزیابی آماری یک ادعا یا یک ایده در مورد یک جامعه بر اساس نمونه‌ای از جامعه است. بنابراین چنانچه زمانی که سرشماری صورت گرفته شده باشد، نمی‌توان آزمون فرضیه را انجام داد اما برای نمونه آماری با هر حجمی می‌توان این کار را کرد. برای مثال ادعا می‌تواند به این شکل باشد: میانگین بازده بازار سهام ایران بزرگتر از صفر است. با توجه به داده‌های بازده بازار سهام ایران می‌توان از فرایند آزمون فرضیه برای آزمون کردن صحت این ادعا در یک سطح معناداری مشخص استفاده نمود.

فرضیه آماری و مراحل انجام آن

یک فرضیه ادعا یا حدسی در مورد مقدار پارامتر جامعه است که برگرفته از اطلاعات نمونه‌گیری است. بنابراین در آزمون فرضیه دو فرضیه وجود دارد. فرضیه درست بودن ادعا و فرضیه نادرست بودن ادعا. روش آزمون فرضیه بر اساس آماره نمونه و نظریه احتمال برای بررسی درستی یا نادرستی ادعا مورد استفاده قرار می‌گیرد. فرایند آزمون فرضیه در نمودار زیر نشان داده شده است.

بیان فرضیه
↓
انتخاب آماره آزمون مناسب
↓
تعیین سطح معناداری
↓
بیان قاعده تصمیم‌گیری در مورد فرضیه
↓
جمع‌آوری نمونه و محاسبه آماره نمونه
↓
تصمیم‌گیری در مورد فرضیه
↓
تصمیم‌گیری بر اساس نتایج آزمون

فرضیه صفر و فرضیه مقابل

فرضیه صفر که آن را با H_0 نشان می‌دهند، فرضیه‌ای است که محقق به دنبال رد آن است. این فرضیه‌ای است که در واقع آزمون می‌شود و مبنای اصلی برای انتخاب آماره‌های آزمون است. فرضیه‌صفر معمولاً به صورت یک بیان یا ادعای ساده در مورد پارامتر جامعه عنوان می‌شود. معمولاً فرضیه صفر مثلاً در مورد میانگین جامعه شامل $H_0 : \mu = \mu_0$ ، $H_0 : \mu \le \mu_0$ و $H_0 : \mu \ge \mu_0$ می‌شود که μ میانگین جامعه و μ_0 میانگین مفروض برای جامعه است.

⟸**توجه:** فرضیه صفر همیشه برابری را در بر دارد.

فرضیه مقابل که با H_a یا H_1 آن را نشان می‌دهند نتیجه‌ای است که در صورت وجود شواهد کافی برای رد فرضیه صفر گرفته می‌شود. در واقع این فرضیه مقابل است که شما می‌خواهید ارزیابی کنید. چون زمانی که فرضیه صفر نامعتبر است شما هیچ موقع نمی‌توانید با آمار فرضیه صفر ثابت کنید، تفسیر این است که فرضیه مقابل صحیح می‌باشد.

آزمون فرضیه یک دامنه و دو دامنه (one-tailed and two-tailed)

فرضیه مقابل می‌تواند یک دامنه یا دو دامنه باشد. اینکه آزمون یک دامنه است یا دو دامنه به قضیه‌ای که آزمون می‌شود بستگی دارد. برای مثال چنانچه محقق به دنبال آزمون این فرضیه است که آیا بازده سهام بزرگتر از صفر است یا نه باید از آزمون یک دامنه استفاده شود. اما اگر محقق به دنبال آزمون فرضیه غیر صفر بودن بازده سهام باشد از آزمون فرضیه دو دامنه استفاده می‌شود. آزمون دو دامنه اجازه انحراف پیدا کردن از دو طرف مقدار مورد فرض را مهیا می‌سازد. در عمل معمولاً آزمونها به صورت دو دامنه هستند.

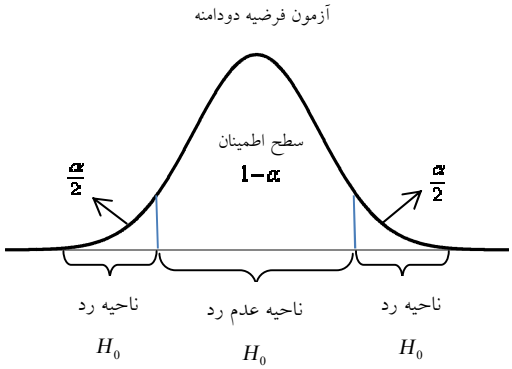
آزمون دو دامنه میانگین جامعه را می‌توان به صورت زیر ساخت:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

قاعده عمومی تصمیم‌گیری برای یک آزمون دو دامنه عبارت است از:

H_0 رد می‌شود اگر: مقدار بحرانی بالایی $<$ آماره آزمون
مقدار بحرانی پایینی $>$ آماره آزمون

در شکل زیر آزمون دو دامنه با استفاده از توزیع نرمال نشان داده شده است. احتمال رد فرضیه H_0 و پذیرش فرضیه H_1 یعنی α را سطح معناداری می‌نامند.



- در سطح معناداری α آماره محاسبه شده آزمون با مقدار بحرانی جدول برای این سطح معناداری مقایسه می‌شود. برای مثال در سطح معناداری $\alpha = 0.05$ آماره محاسبه شده آزمون با مقدار بحرانی z یعنی ± 1.96 مقایسه می‌شود. مقادیر $\pm z_{\alpha/2} = \pm z_{0.025}$ متناظر با می‌باشند که دامنه‌ای از مقادیر z است که با احتمال ۹۵٪ در آن فاصله قرار می‌گیرد.

- اگر آماره محاسبه شده آزمون خارج از دامنه مقادیر بحرانی قرار گیرد (یعنی ۱/۹۶- > آماره آزمون یا ۱/۹۶ < آماره آزمون)، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که آماره نمونه به لحاظ آماری متفاوت از مقدار مورد فرض است.
- اگر آماره محاسبه شده آزمون در فاصله ± 1.96 قرار گیرد نتیجه می‌گیریم که آماره نمونه به لحاظ آماری با مقدار مورد فرض متفاوت نیست (برای مثال در مورد آزمون میانگین $\mu = \mu_0$) و نمی‌توان فرضیه صفر را رد کرد.
- بنابراین قاعده تصمیم‌گیری برای آزمون Z دو دامنه در $\alpha = 0.05$ را می‌توان اینگونه بیان کرد:

H_0 رد می‌شود اگر: آماره آزمون > -1.96 و آماره آزمون $< +1.96$

مثال ۱: در آزمون برای جامعه نرمال با انحراف معیار نامشخص و درجه آزادی کمتر از ۳۰ تابع نمونه‌ای آزمون (آماره آزمون) عبارت است از: (مدیریت ۷۲)

۱) $t = \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}}$ ۲) $Z = \frac{\overline{X} - \mu_0}{S_{\overline{X}}}$ ۳) $t = \frac{\overline{X} - \mu_0}{S_{\overline{X}}}$ ۴) $Z = \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}}$

جواب:

با توجه به اینکه جامعه نرمال و درجه آزادی (و بنابراین حجم نمونه) کمتر از ۳۰ است بنابراین گزینه ۳ صحیح است.

مثال ۲: محققی داده‌های بازده یک پورتفوی اختیارات خرید را در دوره ۲۵۰ روز اخیر را جمع‌آوری نموده است. بازده میانگین روزانه ۰/۱٪ و انحراف معیار نمونه ۰/۲۵٪ است. محقق معتقد است که بازده میانگین پورتفوی روزانه برابر با صفر نیست. آزمون فرضیه اعتقاد محقق را بسازید.

ابتدا فرضیه صفر و مقابل را تعیین می‌کنیم. فرضیه صفر آن چیزی است که محقق انتظار دارد رد شود. بنابراین

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

از آنجایی که فرضیه صفر برابری است بنابراین آزمون از نوع دو دامنه است. از آنجایی که حجم نمونه از ۳۰ بالاتر است بنابراین مقدار بحرانی Z در سطح معناداری ۵٪ که برابر با ± 1.96 است مورد استفاده قرار می‌گیرد. بنابراین قاعده تصمیم‌گیری به این شکل است:

H_0 رد می‌شود اگر: (یعنی ۱/۹۶- > آماره آزمون < ۱/۹۶)

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{0.0025}{\sqrt{250}} \rightarrow \frac{0.001}{\left(\frac{0.0025}{\sqrt{250}}\right)} = \frac{0.001}{0.000158} = 6.33$$

از آنجایی که $6.33 > 1.96$ فرضیه صفر مبنی بر برابر صفر بودن بازده میانگین اختیارات روزانه برابر با صفر است، رد می‌شود. باید توجه داشت که فرضیه صفر بر اساس انحراف معیار و حجم نمونه به نفع فرضیه مقابل رد می‌شود.

آزمون یک دامنه میانگین جامعه را می‌توان به صورت زیر ساخت:

- آزمون دامنه بالا یا به سمت راست

$$\begin{cases} H_0 : \mu \le 0 \\ H_1 : \mu > 0 \end{cases}$$

- آزمون دامنه به سمت پایین یا چپ

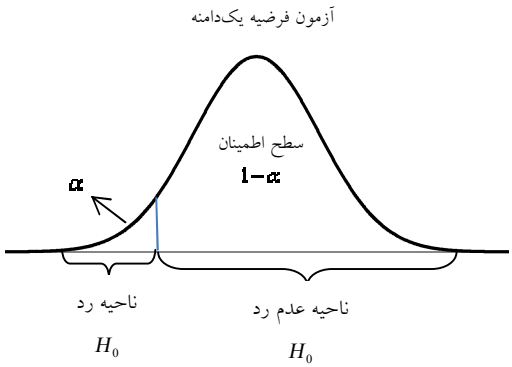
$$\begin{cases} H_0 : \mu \ge 0 \\ H_1 : \mu < 0 \end{cases}$$

فرضیه مناسب با توجه به انتظار محقق در مورد اینکه میانگین جامعه کمتر (دنباله پایین) یا بیشتر (دنباله بالا) از مقدار مورد فرض هست یا نه تعیین می‌شود. در آزمون یک دامنه میانگین جامعه مقدار بحرانی آزمون z برای سطح اطمینان α برابر خواهدبود با Z_α. مثلاً برای آزمون فرضیه به سمت بالا و سطح اطمینان ۵٪ داریم $Z_{0.05} = 1.645$ و برای آزمون فرضیه به سمت پایین و سطح اطمینان ۵٪ داریم $-Z_{0.05} = -1.645$.

قاعده عمومی تصمیم‌گیری برای آزمون یک دامنه به سمت بالای میانگین جامعه عبارت است از:

- اگر آماره محاسبه شده آزمون بزرگتر از Z_α باشد نتیجه می‌گیریم که آماره نمونه به لحاظ آماری بزرگتر از مقدار مورد فرض است یعنی فرضیه صفر را رد می‌کنیم.
- اگر آماره محاسبه شده آزمون کمتر از Z_α باشد نتیجه می‌گیریم که آماره نمونه به لحاظ آماری با مقدار مورد فرض تفاوت ندارد یعنی نمی‌توان فرضیه صفر را رد کرد.

در شکل زیر آزمون دو دامنه با استفاده از توزیع نرمال نشان داده شده است.



مثال ۳: با استفاده از اطلاعات داده‌های پورتفوی اختیارات در مثال ۲ این اعتقاد که بازده اختیارات مثبت است را آزمون نمایید:

جواب:

در این مورد از آزمون یک دامنه استفاده می‌شود که بر اساس آن آزمون فرضیه به این شکل خواهد بود:

$$\begin{cases} H_0 : \mu \le 0 \\ H_1 : \mu > 0 \end{cases}$$

بنابراین قاعده تصمیم‌گیری برای آزمون Z یک دامنه در سطح معناداری ۵٪ عبارت است از:

فرضیه H_0 رد می‌شود اگر: آماره آزمود > ۱/۶۴۵

نحوه محاسبه آماره آزمون فارغ از اینکه آزمون یک دامنه باشد یا دو دامنه به یک روش محاسبه می‌شود و تنها تفاوت در نحوه محاسبه مقدار بحرانی است که در این مورد $Z_\alpha = 1.645$ است. بنابراین با توجه به مقدار آماره آزمون که در مثال ۲ محاسبه شد $6.33>1.96$ و فرضیه صفر رد شده و نتیجه می‌گیریم که بازده میانگین به لحاظ آماری در سطح معناداری ۵٪ بزرگتر از صفر است.

تعیین آماره مناسب برای آزمون فرضیه میانگین

آزمون فرضیه شامل دو آماره است: آماره محاسبه شده آزمون از داده‌های نمونه و مقدار بحرانی آماره آزمون. مقدار آماره محاسبه شده آزمون نسبت به مقدار بحرانی گام کلیدی در ارزیابی درستی فرضیه است. آماره محاسبه شده آزمون یک متغیر تصادفی است که بسته به ویژگی‌های نمونه و جامعه می‌تواند توزیع‌های مختلفی داشته باشد. مقدار بحرانی برای آماره آزمون یعنی مقداری که آماره محاسبه شده آزمون با آن مقایسه می‌شود، به توزیع آن آماره بستگی دارد.

در مورد آزمون فرضیه میانگین جامعه استفاده از مقدار بحرانی بر اساس توزیع t یا توزیع Z به حجم نمونه، توزیع جامعه و اینکه واریانس جامعه مشخص باشد یا نباشد بستگی دارد. همانند بحث مطرح شده در مورد ساختن فاصله اطمینان برای میانگین جامعه در مورد آزمون فرضیه نیز داریم:

- اگر واریانس جامعه نامشخص باشد و یکی از شرایط زیر وجود داشته باشد از توزیع t برای آزمون فرضیه میانگین جامعه استفاده می‌شود:
  ۱) حجم نمونه بزرگ باشد $(n \ge 30)$
  ۲) حجم نمونه کوچک باشد (کمتر از ۳۰)، اما توزیع جامعه نرمال یا تقریباً نرمال باشد.

اگر نمونه کوچک و توزیع غیرنرمال باشد، آزمون آماری قابل اتکا وجود ندارد. در این حالت آماره t با n-1 درجه آزادی برای آزمون فرضیه بر اساس رابطه زیر محاسبه می‌شود:

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

این آماره با مقدار بحرانی که از جدول توزیع t برای درجه آزادی n-1 و سطح اطمینان مشخص بدست آمده مقایسه می‌شود. از آنجایی که در دنیای واقعی معمولاً واریانس جامعه مشخص نیست توزیع t از کاربرد بسیاری برخوردار است.

- زمانی که واریانس جامعه مشخص و توزیع آن نرمال است از آزمون Z برای انجام آزمون فرضیه میانگین جامعه استفاده می‌شود. آماره محاسبه شده آزمون با استفاده از آزمون Z آماره Z نامیده می‌شود. آماره Z برای آزمون فرضیه صفر میانگین جامعه به صورت زیر محاسبه می‌شود:

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

برای آزمون یک فرضیه آماره Z با مقدار بحرانی Z متناظر با سطح اطمینان آزمون مقایسه می‌شود. مقدار بحرانی Z برای سطوح اطمینان معمول در جدول زیر ارائه شده است. بهتر است این مقادیر را به ذهن بسپارید.

سطح اطمینان	آزمون دو دامنه	آزمون یک دامنه
٪۱۰=۰/۱	± ۱/۶۵	-۱/۲۸ یا +۱/۲۸
٪۵=۰/۰۵	± ۱/۹۶	-۱/۶۵ یا +۱/۶۵
٪۱=۰/۰۱	± ۲/۵۸	-۲/۳۳ یا +۲/۳۳

- زمانی که حجم نمونه بزرگ و واریانس جامعه نامشخص است، آماره Z عبارت است از:

$$z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

برای مثال در مورد مثال ۲ و ۳ به دلیل اینکه واریانس جامعه مشخص نیست آزمون مناسب آزمون t است اما با ۲۵۰ مشاهده نمونه بزرگ تلقی می‌شود بنابراین آزمون Z نیز مناسب است. بنابراین به دلیل بزرگ بودن حجم نمونه مقادیر بحرانی Z و t تقریباً مشابه هم هستند. بنابراین در این حالت تفاوتی در احتمال رد فرضیه صفر بر اساس دو توزیع وجود ندارد.

تا ابتدای خطای نوع اول و نوع دوم

نمونه تست آزمون:

۱) شش نفر کارشناس مدیریت را به چند طریق می‌توان به ۳ شهر اعزام کرد به طوری که تعداد افراد اعزامی به دو شهر برابر نباشند؟

۱- ۶۰ ۲- ۱۸۰ ۳- ۲۴۰ ۴- ۳۶۰

۲) در یک مسابقه دوچرخه‌سواری ۴۳ دوچرخه‌سوار قرار است در یک جاده کمربندی دور شهری مسابقه دهند. در چند حالت دوچرخه‌سواران می‌توانند مقام اول، دوم و سوم را کسب نمایند؟

۱- ۱۲۹ ۲- ۲۵۲۰ ۳- ۲۶۳۰ ۴- ۵۰۴۰

۳) اگر $A\cup B$ برابر فضای نمونه، $P(A)=0.7$ و $P(B)=0.6$ باشد، مقدار $P(B-A')$ چقدر است؟

۱- ۰/۱۸ ۲- ۰/۳ ۳- ۰/۴۲ ۴- ۰/۷

۴) ارقام ۳، ۳، ۲، ۲، ۱ و ۱ به تصادف کنار هم قرار می‌گیرند، با کدام احتمال بین هر دو رقم یکسان دو رقم متمایز قرار می‌گیرند؟

۱- $\frac{1}{10}$ ۲- $\frac{1}{12}$ ۳- $\frac{1}{15}$ ۴- $\frac{1}{18}$

۵) در یک رمز عبور شش رقمی بدون صفر با کدام احتمال دقیقاً سه رقم مضرب ۳ و یک رقم مضرب ۴ می‌باشد؟

۱- $\frac{320}{81\times81}$ ۲- $\frac{640}{81\times81}$ ۳- $\frac{80}{27\times27}$ ۴- $\frac{160}{27\times27}$

۶) از ۱۰ پست در یک اداره می‌خواهند ۳ پست را به علت کمی مراجعه حذف کنند. احتمال اینکه پست بخصوصی حذف نشود چقدر است؟

۱- $\frac{7}{9}$ ۲- $\frac{7}{10}$ ۳- $\frac{3}{9}$ ۴- $\frac{3}{10}$

۷) در ظرف اول ۱ مهره سفید و ۴ مهره سیاه و در ظرف دوم ۳ مهره سفید و ۲ مهره سیاه وجود دارد. به تصادف یک مهره از ظرف اول برداشته و بدون رویت در ظرف دوم قرار می‌دهیم. سپس از ظرف دوم دو مهره با هم خارج می‌کنیم. با کدام احتمال هر دو مهره خارج شده سفید است؟

۱- ۰/۱۲ ۲- ۰/۱۸ ۳- ۰/۲۴ ۴- ۰/۳۶

۸) احتمال وجود سفره زیرزمینی نفتی در مناطق مختلف یک استان ۰/۴ است. احتمال برخورد چاه حفر شده به نفت حتی در حالت وجود سفره نفتی تنها ۰/۳ است. اگر یک چاه در این استان به تصادف حفر شود، احتمال عدم برخورد آن به نفت چقدر است؟

۱- ۰/۱۲ ۲- ۰/۲۸ ۳- ۰/۳۰ ۴- ۰/۸۸

۹) تعداد دانشجویان کلاس الف دو برابر دانشجویان کلاس ب است و نسبت دختران در این دو کلاس به ترتیب ۰/۴ و ۰/۶ است. اگر دختری به تصادف از این کلاس انتخاب شود، احتمال اینکه متعلق به کلاس الف باشد چقدر است؟

۱- ۰/۴ ۲- ۰/۶ ۳- ۰/۵۷ ۴- ۰/۸۶

۱۰) اگر $P(A_1)=0.4$ ، $P(A_2)=0.6$ ، $P(B|A_1)=0.2$ و $P(B|A_2)=0.05$ باشد احتمال $P(A_1|B)$ عبارت است از؟

۱- ۰/۷۲ ۲- ۰/۲۷ ۳- ۰/۱۱ ۴- ۰/۰۳