

# چکیده‌ی احتمال و آمار

علی فهیم

## فهرست مطالب

۱	احتمال	۱
۲	آزمایش تصادفی، فضای نمونه و پیشامد	۱.۱
۲	عملهای جبری بر پیشامدها	۲.۱
۳	تعریف احتمال و ویژگیهای آن	۳.۱
۳	روش‌های اساسی شمارش	۴.۱
۴	احتمال شرطی	۵.۱
۴	قضیه بیز	۶.۱
۴	متغیرهای تصادفی گسسته	۲
۵	امید ریاضی و واریانس	۱.۲
۵	گشتاورها و تابع مولد گشتاور	۲.۲
۶	توزیع یکنواخت گسسته	۳.۲
۶	توزیع برنولی و دوجمله‌ای	۴.۲
۶	توزیع هندسی و دوجمله‌ای منفی	۵.۲
۷	توزیع فوق‌هندسی	۶.۲
۷	فرآیند و توزیع پواسون	۷.۲
۷	قانون توانی و توزیع زیف و زتا	۸.۲
۸	متغیرهای تصادفی پیوسته	۳
۱۱	توزیع یکنواخت در بازه‌ی $(a, b)$	۱.۳
۱۱	توزیع نمایی	۲.۳
۱۱	توزیع گاما	۳.۳
۱۱	ویژگی بی‌حافظگی یا مارکفی	۴.۳
۱۲	توزیع وایبول (Weibull)	۵.۳
۱۲	توزیع نرمال یا بهنجار	۶.۳
۱۲	توزیع نرمال استاندارد	۷.۳
۱۳	توزیع لانگرمال	۸.۳
۱۴	توزیع بتا	۹.۳
۱۴	توزیع کوشی	۱۰.۳
۱۴	توزیع پارتو	۱۱.۳
۱۴	متغیرهای تصادفی باهم	۴
۲۰	توابع عمومی از متغیرهای تصادفی	۵
۲۴	نمونه‌گیری	۶

۳۱	برآورد نقطه‌ای	۷
۳۲	روش گشتاوری . . . . .	۱.۷
۳۲	روش بیشینه درست‌نمایی . . . . .	۲.۷
۳۳	ارزیابی برآوردها - ناریبی . . . . .	۳.۷
۳۴	ارزیابی برآوردها - کارابی . . . . .	۴.۷
۳۵	روش بوت استرپ (خودگردان) . . . . .	۵.۷
۳۶	برآورد بازه‌ای یا بازه‌ی اطمینان	۸
۳۶	بازه‌ی اطمینان برای میانگین جامعه . . . . .	۱.۸
۲۸	بازه اطمینان برای نسبت در جامعه . . . . .	۲.۸
۲۸	بازه اطمینان برای انحراف معیار جامعه . . . . .	۳.۸
۲۸	پیش‌بینی بازه‌ای . . . . .	۴.۸
۳۹	پیش‌بینی بازه‌ای (Bootstrap) روش خودگردان . . . . .	۵.۸
۳۹	آزمون فرضیه - ۱	۹
۴۴	آزمون فرضیه - ۲	۱۰
۴۴	۱۱ آزمون‌های نیکویی برآش و استقلال	
۴۴	۱.۱۱ آزمون نیکویی برآش کای دو	
۴۴	۲.۱۱ آزمون استقلال . . . . .	
۴۶	۱۲ مدل سازی‌آماری: رگرسیون	
۴۶	۱.۱۲ مدل رگرسیون خطی . . . . .	
۴۸	۲.۱۲ تحلیل آماری مدل . . . . .	
۴۹	۳.۱۲ ملاحظات مدل رگرسیونی . . . . .	
۴۹	۴.۱۲ همبستگی و تحلیل آن . . . . .	

## ۱ احتمال

### ۱.۱ آزمایش تصادفی، فضای نمونه و پیشامد

آزمایش عملی است که نتیجه‌ی اجرای آن جمع‌آوری اطلاعات است. یک آزمایش با حرف  $E$  نمایش داده می‌شود. آزمایش تصادفی به آزمایشی گفته می‌شود که گرچه همه نتایج ممکن آن مشخص است، اما نتیجه دقیق آن قبل از آزمایش معلوم نیست. هر نتیجه آزمایش را یک برآمد می‌نامیم. مجموعه برآمدهای هر آزمایش تصادفی را فضای نمونه آزمایش می‌نامیم و با  $S$  نشان می‌دهیم. هر زیرمجموعه از یک فضای نمونه‌ای  $S$  را یک پیشامد می‌نامیم و معمولاً با حرف بزرگ نمایش می‌دهیم ( $A, B, \dots$ ). یک پیشامد را می‌گوییم رخداده است اگر برآمد یک آزمایش تصادفی متعلق به آن پیشامد باشد. فضای نمونه‌ای  $S$  را پیشامد حتمی و مجموعه تهی را پیشامد نامحتمل می‌نامیم. پیشامدهای تک عضوی را پیشامدهای ساده می‌نامیم و با حرف کوچک نشان می‌دهیم ( $a, b, \dots$ ).

### ۲.۱ عمل‌های جبری بر پیشامدها

اجتماع دو پیشامد: یکی و یا هر دو رخداند  $A \cup B$ . اشتراک دو پیشامد: هر دو پیشامد رخداند  $A \cap B$  و اگر  $A \cap B = \emptyset$  دو پیشامد ناسازگار نامیده می‌شوند. متمم پیشامد  $A$ : که با  $\bar{A}$  نشان داده می‌شود یعنی نتیجه آزمایش در  $A$  نباشد. تفاضل پیشامدها:  $A - B = A \cap \bar{B}$  یعنی  $A$  رخداند اما  $B$  رخداند. زیرپیشامد:  $A \subseteq B$  یعنی اگر  $A$  رخداند نتیجه می‌گیریم پیشامد  $B$  نیز رخداند.

## ۳.۱ تعریف احتمال و ویژگیهای آن

الف - کلاسیک: مبتنی بر هم‌شانسی پیشامدهای ساده می‌باشد. نسبت تعداد اعضای پیشامد  $A$  به تعداد اعضای فضای نمونه‌ای

$$P(A) = \frac{n(A)}{n(S)}$$

ب - فراوانی نسبی: تکرار زیاد  $n$  آزمایش با شرایط یکسان که به مقدار ثابتی میل می‌کند.

$$P(A) = \frac{n(A)}{n}$$

پ - درجه باور: ...  
ت - اصل موضوع:

- $P(A) \geq 0$

- $P(S) = 1$

- اصل جمع‌پذیری برای برای پیشامدهای ناسازگار

$$A_i \bigcap_{i \neq j} A_j = \emptyset \Rightarrow P\left(\bigcup_{i=1}^m A_i\right) = \sum_{i=1}^m P(A_i)$$

قضایا:

- $P(\emptyset) = 0$

- $P(\bar{A}) = 1 - P(A)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- $P(A - B) = P(A) - P(A \cap B)$

- $B \subseteq A \Rightarrow P(A - B) = P(A) - P(B)$

- $B \subseteq A \Rightarrow P(B) \leq P(A)$

## ۴.۱ روش‌های اساسی شمارش

اگر کار  $A$  با  $m$  روش و کار  $B$  با  $n$  روش قابل انجام باشند آنگاه انجام متوالی این دو کار با  $m \times n$  روش امکان‌پذیر است.

جایگشت: هر آرایش مرتب از کنار هم قرار گرفتن  $n$  شی را جایگشتی از آن اشیاء می‌نامیم. طبق اصل شمارش  $n$  شی متمایز را با  $n!$  روش می‌توانیم کنار هم قرار دهیم.

جایگشت با عناصر همانند: تعداد  $n$  شی که  $n_1$  تای آنها مانند هم و ... و  $n_k$  تای آنها نیز مانند هم باشند با  $\frac{n!}{n_1! \dots n_k!}$  تعداد روش می‌توان کنار هم قرار داد که در آن  $n = n_1 + \dots + n_k$  است.

ترتیب: یعنی از  $n$  شی متمایز  $r$  تا را با رعایت ترتیب انتخاب کنیم.

ترکیب: یعنی از  $n$  شی متمایز  $r$  تا را بدون رعایت ترتیب انتخاب کنیم.

روش خط و نقطه: اگر  $n$  شی یکسان را در  $m$  مکان بخواهیم قرار دهیم،  $n$  نقطه و  $m - 1$  خط در نظر می‌گیریم و عمل جایگشت با عناصر همانند را روی آنها اعمال می‌کنیم. تعداد حالت‌های ممکن می‌شود  $\frac{(n+m-1)!}{n!(m-1)!}$ . مثلا برای قرار دادن ۵ شی یکسان در ۳ مکان، یکی از حالت‌ها می‌شود ۰|00|00 و یک حالت دیگر می‌شود 0|0|000 و تعداد حالت‌های کل  $= 21 = \frac{7!}{(5!2!)}$  می‌شود.

## ۵.۱ احتمال شرطی

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0$$

با توجه به اصل موضوعه رابطه‌ی بالا احتمال است زیرا:

$$\begin{cases} P(A|B) \geq 0 \\ P(S|B) = 1 \\ A_i \bigcap_{i \neq j} A_j = \emptyset \Rightarrow P(\bigcup_{i=1}^m A_i|B) = \sum_{i=1}^m P(A_i|B) \end{cases}$$

$$P(A \cap B \cap C \cap \dots) = P(A) P(B|A) P(C|A \cap B) P(\dots|A \cap B \cap C) \dots$$

استقلال پیشامدها: با ۲ تعریف هم ارز

$$\begin{aligned} P(A|B) &= P(A) \\ P(A \cap B) &= P(A)P(B) \end{aligned}$$

استقلال با ناسارگاری متفاوت است. در ناسارگاری دو پیشامد  $P(A \cap B) = 0$  است.

## ۶.۱ قضیه بیز

اگر فضای نمونه‌ای به  $k$  پیشامد  $B$  افزار شود یعنی:

$$B_i \bigcap_{i \neq j} B_j = \emptyset \quad \text{and} \quad \bigcup_{i=1}^k B_i = S$$

آنگاه

$$A = A \cap S = A \cap \left( \bigcup_{i=1}^k B_i \right) = (A \cap B_1) \cup \dots \cup (A \cap B_k)$$

$$P(A) = \sum_{i=1}^k P(A \cap B_i) = \sum_{i=1}^k P(B_i)P(A|B_i)$$

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^k P(B_j)P(A|B_j)}$$

که در آن  $P(B_i|A)$  احتمال پسین و  $P(A|B_i)$  احتمال پیشین می‌باشد. همچنین (Likelihood) می‌نامیم.

## ۲ متغیرهای تصادفی گسته

متغیر تصادفی  $X$  تابعی از فضای نمونه‌ای آزمایش ( $S$ ) به اعداد حقیقی می‌باشد. برد متغیر را با  $S_X$  نمایش می‌دهیم.

$$X : S \rightarrow S_X \subseteq R$$

هرگاه  $S_X$  شمارا باشد متغیر گسته و هرگاه پیوسته باشد، متغیر پیوسته می‌باشد. تابع جرم احتمال: تابعی از  $R$  به بازه‌ی  $[0, 1]$  می‌باشد که احتمال برآمد مقدار  $x$  را معلوم می‌کند.

$$\begin{aligned} P(X = x) &= f_X(x), \quad x \in R \\ f_X(x) &\geq 0 \\ \sum_{x'} f_X(x') &= 1 \end{aligned}$$

و احتمال هر پیشامد مانند  $A$  از رابطه‌ی محاسبه می‌شود.  
تابع توزیع تجمعی: تابعی از  $R$  به بازه‌ی  $[0, 1]$  می‌باشد که احتمال برآمد همه مقادیر کمتر و برابر  $x$  را معلوم می‌کند.

$$\begin{aligned} F : R &\rightarrow [0, 1] \\ P(X \leq x) &= F_X(x) = \sum_{x' \leq x} f_X(x') \end{aligned}$$

نکته:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

## ۱.۲ امید ریاضی و واریانس

امید ریاضی:

$$\mu_X = E(X) = \sum_{x'} x' f_X(x')$$

$$E(g(x)) = \sum_{x'} g(x') f_X(x')$$

$$E(ag_1(x) + bg_2(x)) = aE(g_1(x)) + bE(g_2(x))$$

واریانس:

$$\sigma_X^2 = V(X) = E((X - \mu_X)^2) = \sum_{x'} (x' - \mu_X)^2 f_X(x') = E(X^2) - E(X)^2$$

انحراف معیار:

$$\sigma_x = \sqrt{\sigma_X^2}$$

$$\sigma_{aX+b}^2 = V(aX + b) = a^2 V(X) + a^2 \sigma_X^2$$

## ۲.۱ گشتاورها و تابع مولد گشتاور

گشتاور مرتبه  $k$ :

$$m_k = E(X^k) = \sum_{x'} x'^k f_X(x')$$

گشتاور مرکزی  $m!k$ :

$$\mu_k = E((X - \mu_X)^k) = \sum_{x'} (x' - \mu_X)^k f_X(x')$$

تابع مولد گشتاور: به تابع

$$M_X(t) = E(e^{tX}) = \sum_{x'} e^{tx'} f_X(x')$$

تابع مولد گشتاور گفته می‌شود، به شرط آنکه عدد مثبت  $h$  وجود داشته و تابع در فاصله‌ی  $(-h, h)$  همگرا باشد، در صورت وجود  $M_X(t)$ ، با مشتقات آن نسبت به  $t$  در نقطه‌ی  $t = 0$  گشتاورهای متغیر تصادفی  $X$  محاسبه می‌شود.  
قضیه یکتایی: اگر تابع مولد دو متغیر تصادفی حول نقطه‌ی صفر با هم برابر باشند، دو متغیر دارای تابع جرم احتمال یکسانی هستند.

## ۳.۲ توزیع یکنواخت گسسته

$$P(X = x) = f_x(x) = \frac{1}{N}, \quad x = x_1, \dots, x_N$$

## ۴.۲ توزیع برنولی و دوجمله‌ای

آزمایش برنولی: به آزمایشی گفته می‌شود که نتیجه‌ی آن تنها ۲ حالت دارد، موفقیت  $(x = 1)$  یا شکست  $(x = 0)$ . اگر احتمال موفقیت  $p$  باشد آنگاه

$$P(X = x) = f_X(x) = p^x(1 - p)^{(1-x)}, \quad x = 0, 1$$

$$E(X) = p, \quad V(X) = p(1 - p)$$

اگر متغیر تصادفی، تعداد موفقیت‌ها در  $n$  آزمایش برنولی با شرایط یکسان باشد، تابع جرم احتمال آن از توزیع دوجمله‌ای پیروی می‌کند.

$$X \sim b(n, p)$$

$$P(X = x) = f_X(x) = \frac{n!}{x!(n-x)!} p^x(1 - p)^{(n-x)}, \quad x = 0, 1, \dots, n$$

$$E(X) = np, \quad V(X) = np(1 - p)$$

## ۵.۲ توزیع هندسی و دوجمله‌ای منفی

اگر متغیر تصادفی، تعداد آزمایش‌های برنولی با شرایط یکسان، تا رسیدن به اولین موفقیت باشد، از توزیع هندسی پیروی می‌کند.

$$X \sim Ge(p)$$

$$P(X = x) = f_X(x) = p(1 - p)^{(x-1)}, \quad x = 1, 2, \dots$$

$$E(X) = 1/p, \quad V(X) = \frac{1-p}{p^2}$$

اگر متغیر تصادفی تعداد آزمایش‌های برنولی با شرایط یکسان، تا رسیدن به  $r$  امین موفقیت باشد، متغیر از توزیع دوجمله‌ای منفی پیروی می‌کند.

$$X \sim NB(r, p)$$

$$P(X = x) = f_X(x) = \frac{(x-1)!}{(r-1)!(x-r)!} p^r(1 - p)^{(x-r)}, \quad x = r, r+1, \dots$$

$$E(X) = r/p, \quad V(X) = \frac{r(1-p)}{p^2}$$

## ۶.۲ توزیع فوق‌هندسی

مانند توزیع دوجمله‌ای، متغیر تصادفی تعداد نمونه‌های موفق در  $n$  نمونه‌گیری (آزمایش برنولی) است با این تفاوت که نیازی به جایگذاری نمونه‌ها برای حفظ شرایط پیکسان نیست. لذا اگر متغیر تصادفی تعداد نمونه‌های موفق در نمونه‌گیری بدون جایگذاری باشد، از توزیع فوق‌هندسی پیروی می‌کند. بطور مثال اگر یک نمونه‌ی  $n$  قابی از  $M$  مهره که تنها آنها سفید است را به تصادف و بدون جایگذاری انتخاب کنیم، احتمال آنکه تعداد مهره‌های سفید در  $n$  مهره‌ی انتخاب شده  $x$  باشد از رابطه‌ی زیر بدست می‌آید.

$$X \sim HG(N, M; n)$$

$$P(X = x) = f_X(x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}},$$

$$\max(0, n - (N - M)) \leq x \leq \min(n, M)$$

## ۷.۲ فرآیند و توزیع پواسون

یک فرآیند را پواسون می‌نامیم اگر تعداد اتفاقاتی که در یک بازه‌ی پیوسته‌ی رخ می‌دهد در شرایط زیر صدق کند.

- $P((t, t+h] \cap \text{اتفاق}) \simeq \lambda h$
- $P((t, t+h] \cap \text{اتفاق}) \simeq 0$
- رخ دادن اتفاقات در بازه‌های جدا، از هم مستقل باشند.

فرض کنید متغیر تصادفی، تعداد اتفاقات در بازه‌ی  $0$  تا  $T$  باشد. این بازه را به تعداد (زیاد)  $n$  قسمت مساوی تقسیم می‌کنیم. پس پهنانی هر قسمت  $h = T/n$  می‌شود و احتمال رخ دادن یک اتفاق در هر قسمت  $p = \lambda T/n$  است. پس احتمال رخدادن  $k$  اتفاق در بازه‌ی  $0$  تا  $T$  از رابطه‌ی زیر محاسبه می‌شود.

$$\begin{aligned} P(N(T) = k) &= f_N(k) = \lim_{n \rightarrow \infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda T}{n}\right)^k \left(1 - \frac{\lambda T}{n}\right)^{n-k} \\ &= \frac{e^{-\lambda T} (\lambda T)^k}{k!}, \quad k = 0, 1, \dots \end{aligned}$$

که برای اثبات آن از رابطه‌ی زیر استفاده شده است.

$$\lim_{n \rightarrow \infty} \left(1 - \frac{x}{n}\right)^n = e^{-x}$$

معمولًا در یک فرآیند پواسون، طول بازه را  $T = 1$  در نظر می‌گیرند و توزیع فقط پارامتر  $\lambda$  دارد.

$$X \sim P(\lambda), \quad P(X = x) = f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \dots$$

$$E(X) = \lambda, \quad V(X) = \lambda$$

## ۸.۲ قانون توانی و توزیع زیف و زتا

قانون توانی: اگر احتمال آنکه متغیر تصادفی، مقدار خاصی شود، متناسب با توانی از معکوس آن مقدار باشد، گفته می‌شود که متغیر از قانون توانی پیروی می‌کند.

$$P(X = x) \propto \frac{1}{x^\alpha}, \quad x > 0, \quad \alpha \geq 0$$

توزیع زیف: توزیعی که از قانون توانی پیروی می‌کند.

$$X \sim Zipf(N, \alpha),$$

$$P(X = x) = f_X(x) = \frac{x^{-\alpha}}{H(N, \alpha)}, \quad \alpha \geq 0, \quad x = 1, 2, \dots, N$$

$$H(N, \alpha) = \sum_{x=1}^N x^{-\alpha}, \quad \text{عدد هارمونیک}$$

متغیر تصادفی مربوطه گستته و مقادیر محدود تا  $N$  می‌گیرد.

توزیع زتا: مانند توزیع زیف می‌باشد اما متغیر تصادفی که از این توزیع پیروی می‌کند همه مقادیر گستته بیشتر از صفر را می‌تواند قبول کند.

$$X \sim Zeta(\alpha),$$

$$P(X = x) = f_X(x) = \frac{x^{-\alpha}}{\zeta(\alpha)}, \quad \alpha > 1, \quad x = 1, 2, \dots$$

$$\zeta(\alpha) = \sum_{x=1}^{\infty} x^{-\alpha}, \quad \text{تابع زتا ریمان}$$

### ۳ متغیرهای تصادفی پیوسته

اگر مجموعه مقادیری که متغیر تصادفی  $X$  می‌گیرد یعنی  $S_X$  پیوسته باشد، متغیر تصادفی  $X$  متغیر تصادفی پیوسته گفته می‌شود.

$$P(X \in A) = \int_A f_X(x) dx$$

که  $f_X(x)$  چگالی احتمال است.

اگر مجموعه  $A$  مجموعه تمام مقادیری باشد که متغیر تصادفی  $X$  می‌گیرد یعنی  $A = S_X$  خواهیم داشت.

$$P(X \in S_X) = \int_{S_X} f_X(x) dx = 1 \Rightarrow \int_{-\infty}^{\infty} f_X(x) dx = 1$$

شرط چگالی بودن یک تابع

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

احتمال اینکه متغیر تصادفی پیوسته مقدار خاصی بگیرد صفر است.

$$\int_a^a f_X(x) dx = 0$$

تابع توزیع تجمعی: احتمال آنکه متغیر تصادفی  $X$  مقدار  $x$  یا کوچکتر بگیرد.

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt \quad x \in R$$

درصورت مشتق‌پذیری تابع توزیع تجمعی

$$f_X(x) = \frac{d}{dx} F_X(x)$$

امید ریاضی یا میانگین متغیر تصادفی  $X$

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f_X(x) dx$$

و برای تابعی از متغیر تصادفی می‌شود

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

واریانس متغیر تصادفی پیوسته: معیاری برای بررسی پراکندگی مقادیر  $X$  نسبت به میانگین

$$V(X) = \sigma_X^2 = E((X - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

و انحراف معیار می‌شود

$$\sigma_X = \sqrt{V(X)}$$

$$V(X) = E(X^2) - E(X)^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx - \mu_X^2$$

$$E(aX + b) = aE(X) + b$$

$$V(aX + b) = a^2 V(X)$$

تعريف میانه  $m$  و چارک اول  $Q_1$  و سوم  $Q_3$  به ترتیب برابر است:

$$P(X \leq m) = 0.5 \quad P(X \leq Q_1) = 0.25 \quad P(X \leq Q_2) = 0.75$$

مد یا نما برای هر توزیع یا متغیر تصادفی (در صورت وجود) نقطه‌ای است که مقدار تابع جرمی احتمال (چگالی احتمال) بیشینه باشد.

گشتاور مرتبه  $k$ : ام  $k$

$$m_k = E(X^k) = \int x^k f_X(x) dx$$

گشتاور مرکزی مرتبه  $k$ : ام  $k$

$$\mu_k = E((X - \mu_X)^k) = \int (x - \mu_X)^k f_X(x) dx$$

تابع مولد گشتاور (تبدیل لابلس):

$$M_X(x) = E(e^X) = \int_{-\infty}^{\infty} e^x f_X(x) dx$$

شرط وجود و یکتاپی تابع مولد مشابه گستته است.  
ضریب چولگی (Skewness) معیاری از تقارن توزیع

$$\gamma_X = Sk(X) = E\left(\left(\frac{X - \mu_X}{\sigma}\right)^3\right) = \frac{\mu_3}{\sigma^3}$$

در یک توزیع متقاضی این ضریب صفر و در توزیع متمایل به راست مثبت و متمایل به چپ منفی است.  
ضریب کشیدگی (Kurtosis)

$$Kurt(X) = E\left(\left(\frac{X - \mu_X}{\sigma}\right)^4\right) - 3 = \frac{\mu_4}{\sigma^4} - 3$$

در یک توزیع نرمال ضریب کشیدگی صفر و اگر توزیع کشیده‌تر و پخته باشد این ضریب مثبت‌تر و اگر جمع‌تر و تیزتر باشد منفی‌تر است.

نامساوی یا کران مارکف: اگر  $X$  متغیر تصادفی نامنفی با میانگین متناهی  $\mu$  باشد، بازء هر  $a > 0$  داریم:

$$\begin{aligned} E(X) &= \int_0^a t f_X(t) dt + \int_a^\infty t f_X(t) dt \geq \int_a^\infty t f_X(t) dt \geq \int_a^\infty a f_X(t) dt = a P(X \geq a) \\ &\Rightarrow P(X \geq a) \leq \frac{\mu}{a} \end{aligned}$$

نامساوی یا کران چبیشف: اگر  $X$  متغیر تصادفی نامنفی با میانگین  $\mu$  و واریانس متناهی  $\sigma^2$  باشد، بازء هر  $k > 0$  داریم:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

اثبات - چون  $(X - \mu)^2$  یک متغیر تصادفی نامنفی است، با به کار بردن نامساوی مارکف (با انتخاب  $a = k^2$ ) به دست می آوریم:

$$\begin{aligned} P(|X - \mu| \geq k) &= P((X - \mu)^2 \geq k^2) \leq \frac{E((X - \mu)^2)}{k^2} \\ &\Rightarrow P(|X - \mu| \geq k) \leq \frac{E((X - \mu)^2)}{k^2} = \frac{\sigma^2}{k^2} \end{aligned}$$

نامساوی یا کران چرنف: اگر  $X$  متغیر تصادفی نامنفی با تابع مولد گشتاور متناهی  $M_X(t)$  باشد، بازء هر  $t > 0$  داریم:

$$P(X \geq a) \leq e^{-ta} M_X(t)$$

تابع مشخصه: برخلاف تابع مولد برای همه توزیع‌ها وجود دارد و تعریف شده است.

$$\Phi(t) = E(e^{itX}) = M(it) \quad i = \sqrt{-1}$$

برای توزیع‌های پیوسته تابع مشخصه همان تبدیل فوریه است.

## مثال

فرض کنید بدانیم که تعداد کالاهای تولید شده در یک کارخانه در طول یک هفته متغیری تصادفی با میانگین 50 است.

۱. در مورد این احتمال که تولیدات این هفته از 75 تجاوز کند چه می‌توان گفت؟
۲. اگر واریانس تولیدات هفتگی مساوی 25 باشد آنگاه در مورد احتمال این که تولیدات این هفته بین 40 و 60 است چه می‌توان گفت؟

حل: فرض کنید  $X$  تعداد کالاهایی است که در یک هفته تولید می‌شود:

۱. با استفاده از نامساوی مارکف:

$$P\{X > 75\} \leq \frac{E[X]}{75} = \frac{50}{75} = \frac{2}{3}$$

۲. با استفاده از نامساوی چبیشف:

$$P\{|X - 50| \geq 10\} \leq \frac{\sigma^2}{10^2} = \frac{1}{4}$$

$$P\{|X - 50| < 10\} \geq 1 - \frac{1}{4} = \frac{3}{4}$$

بنابراین احتمال این که تولیدات این هفته بین 40 و 60 باشد حداقل 75% است.

## ۱.۳ توزیع یکنواخت در بازه‌ی $(a, b)$

$$X \sim U(a, b)$$

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{Others} \end{cases}$$

$$E(X) = \frac{a+b}{2} \quad V(X) = \frac{(b-a)^2}{12}$$

## ۲.۳ توزیع نمایی

$$X \sim E(\lambda)$$

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad x > 0, \quad \lambda > 0$$

$$E(X) = \frac{1}{\lambda} \quad V(X) = \frac{1}{\lambda^2} \quad M(t) = \frac{\lambda}{\lambda-t}$$

## ۳.۳ توزیع گاما

$$X \sim \Gamma(\alpha, \beta) : \quad f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\beta x)^{\alpha-1} e^{-\beta x} \quad x > 0, \quad \alpha, \beta > 0$$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad \alpha > 0$$

$$E(X) = \frac{\alpha}{\beta} \quad V(X) = \frac{\alpha}{\beta^2} \quad M(t) = \left( \frac{\lambda}{\lambda-t} \right)^\alpha \quad t < \lambda$$

توزیع گاما بازه  $1 = \alpha$  منجر به توزیع نمایی می‌شود.

اگر  $n, X_n, ..., X_1$  متغیرهای تصادفی مستقل و هر یک با توزیع نمایی با پارامتر  $\lambda$  باشند، آنگاه  $S_n = \sum_{i=1}^n X_i$  توزیع گاما با پارامترهای  $\lambda$  می‌باشد.

## ۴.۳ ویژگی حافظگی یا مارکفی

$$P(X > s+t | X > t) = P(X > s), \quad \forall t, s \geq 0$$

اگر پیشامدی تا زمان  $t$  رخ نداده است احتمال آنکه تا زمان  $s+t$  نیز رخ ندهد برابر با آن است که ابتدا تا زمان  $s$  رخ ندهد. به عبارتی مستقل از زمان، احتمال آنکه تا  $s$  ثانیه بعد پیشامدی رخ ندهد همواره ثابت است و تنها به  $s$  بستگی دارد نه به  $t$ .

$$\text{e.g. } X \sim E(\lambda) : \quad P(X > s+t | X > t) = \frac{P(X > s+t)}{P(X > t)} \\ = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s)$$

توزیع هندسی تنها توزیع گستته و توزیع نمایی تنها توزیع پیوسته بدون حافظه هستند. رابطه بین توزیع پواسون و نمایی: اگر  $T$  فاصله‌ی زمانی بین دو پیشامد در یک فرآیند پواسون با پارامتر  $\lambda$  باشد، پس در تمام زمان‌های  $t < T$  هیچ پیشامدی رخ نمی‌دهد.  $T$  متغیر تصادفی است.

$$P(T > t) = P(N(t) = 0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}$$

يعنى متغير تصادفي فاصله زمانی بین دو رویداد پواسون از توزيع نمایی پیروی می کند. برای توزيع تجمعی آن داریم:

$$t \geq 0 \quad F_T(t) = P(T \leq t) = 1 - e^{-\lambda t}$$

تعداد رویدادها در واحد زمان توزيع پواسون با میانگین  $\lambda$  معادل فاصله زمانی بین هر دو رویداد پی در پی توزيع نمایی با میانگین  $1/\lambda$  است. زمان انتظار تا  $n$  امین رویداد یا به عبارتی فاصله زمانی بین هر  $n+1$  رویداد متواالی در فرآیند پواسون توزيع گاما با پارامترهای  $n$  و  $\lambda$  است. گاما متغير تصادفي زمان انتظار پیوسته برای  $n$  امین رویداد و توزيع های هندسی و دوجمله ای زمان انتظار گستته می باشند.تابع قابلیت اعتماد (بقاء، ماندگاری) در نقطه  $t$  یعنی تا زمان  $t$  رویدادی مثل خرابی رخ ندهد، به شکل زیر تعریف می شود:

$$R(t) = P(T > t) = 1 - F_T(t) \quad t \geq 0$$

### 5.3 توزيع وایبول (Weibull)

$$X \sim W(\alpha, \beta) : \quad f(x; \alpha, \beta) = \alpha \beta (\beta x)^{\alpha-1} e^{-(\beta x)^\alpha} \quad x > 0, \quad \alpha, \beta > 0$$

در این توزيع  $\alpha$  پارامتر شکل و  $\beta$  پارامتر مقیاس می باشد. در حالت خاص که  $\alpha = 1, \beta = \lambda$  همان توزيع نمایی می شود.

$$F_X(x) = 1 - e^{-(\beta x)^\alpha} \quad x > 0$$

$$E(X) = \frac{1}{\beta} \Gamma(1 + \frac{1}{\alpha}) \quad V(X) = \frac{1}{\beta^2} \left\{ \Gamma(1 + \frac{2}{\alpha}) - \left( \Gamma(1 + \frac{1}{\alpha}) \right)^2 \right\}$$

### 6.3 توزيع نرمال یا بهنجار

$$X \sim N(\mu, \sigma^2) : \quad f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \quad x, \mu \in R \quad \sigma > 0$$

$$E(X) = \mu \quad V(X) = \sigma^2$$

ضریب تغییرات:

$$CV = \frac{\sigma}{\mu}$$

مجموع متغيرهای تصادفي: ثابت می شود مجموع تعداد زیاد از متغيرهای تصادفي مستقل هم توزيع (از هر توزیعی) تقریبا توزيع نرمال می شود که به قضیه حد مرکزی مشهور است.

### 7.3 توزيع نرمال استاندارد

$$Z = \frac{X - \mu}{\sigma} \quad \text{where} \quad X \sim N(\mu, \sigma)$$

$$Z \sim N(0, 1) : \quad f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < +\infty$$

$$\Phi(z) = P(Z \leq z)$$

با استفاده از توزيع نرمال استاندارد می توان احتمال های مربوط به توزيع نرمال را محاسبه کرد.

$$\text{e.g. } P(X < 7) = P\left(\frac{X - \mu}{\sigma} < \frac{7 - \mu}{\sigma}\right) = P(Z < \frac{7 - \mu}{\sigma})$$

چندک  $\alpha$  توزیع نرمال استاندارد:

$$P(Z \leq z_\alpha) = \alpha$$

مثال: فرض کنید در یک کانال ارتباطی دیجیتال، تعداد بیت‌های دریافتی به اشتباہ را می‌توان با یک متغیر تصادفی دو جمله‌ای مدل‌سازی کرد، و فرض کنید که احتمال اینکه یک بیت به اشتباہ دریافت شود  $10^{-5}$  است. اگر ۱۶ میلیون بیت ارسال شود، احتمال وقوع ۱۵۰ یا کمتر خطأ چقدر است؟

تقریب نرمال برای توزیع دو جمله‌ای:

$$X \sim b(n, p) \quad \text{if } np > 5 \quad \text{approximately then } X \sim N(np, np(1-p))$$

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

و  $Z$  تقریباً توزیع نرمال استاندارد دارد.  
اگر  $X$  متغیر تصادفی با توزیع پواسون و پارامترهای  $\lambda$  و  $E(X) = \lambda$  باشد، متغیر

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

تقریباً توزیع نرمال استاندارد دارد. این تقریب برای  $\lambda > 5$  خوب است.

### ۸.۳ توزیع لاغرمال

$$X \sim \ln N(\theta, \delta^2) : \quad f_X(x; \theta, \delta^2) = \frac{1}{x\sqrt{2\pi\delta}} e^{-\frac{1}{2}\frac{(\ln x - \theta)^2}{\delta^2}} \quad x, \theta, \delta > 0$$

$$\mu = E(X) = e^{\theta + (\delta^2/2)} \quad \sigma^2 = V(X) = e^{2\theta + 2\delta^2} - e^{2\theta + \delta^2}$$

$$\theta = \ln \mu - \frac{\delta^2}{2} \quad \delta^2 = \ln\left(1 + \frac{\sigma^2}{\mu^2}\right)$$

$$Y = \ln X \sim N(\theta, \delta^2)$$

$$P(X \leq x) = P(\ln X \leq \ln x) = P(Y \leq \ln x) = P\left(\frac{Y - \theta}{\delta} \leq \frac{\ln x - \theta}{\delta}\right) = \Phi\left(\frac{\ln x - \theta}{\delta}\right)$$

که در آن  $\Phi$  تابع توزیع تجمعی نرمال استاندارد می‌باشد.

شرایط اعتبار توزیع لاغرمال: اگر متغیر تصادفی  $X$  نتیجه تعداد زیادی علتهای مستقل باشد که اثرهای آن‌ها مثبت و علتها به صورت ضربی ترکیب شوند و اثر هر یک در مقابل اثر کلی ناچیز باشد آنگاه  $X$  توزیع لاغرمال دارد.

ضریب تغییرات: اگر ضریب تغییرات چندان بزرگ نباشد یعنی  $CV = \frac{\sigma}{\mu} \leq 0.3$  آنگاه

$$\delta^2 = \ln\left(1 + \frac{\sigma^2}{\mu^2}\right) \simeq \frac{\sigma^2}{\mu^2}$$

یعنی پارامتر  $\delta$  همان ضریب تغییرات توزیع است.

$$m = e^\theta \quad \mu = m\sqrt{1 + CV^2}$$

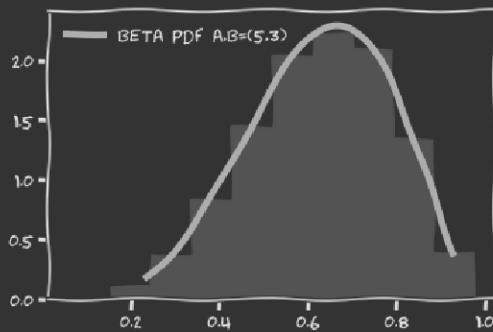
که در آن  $m$  میانه و  $\mu$  میانگین توزیع می‌باشد.

$$X \sim Beta(\alpha, \beta) : f_X(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \\ 0 < x < 1, \quad \alpha, \beta > 0$$

تعریف تابع بتا:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$



### ۱۰.۳ توزیع کوشی

$$X \sim C(\theta, \gamma) : f_X(x; \theta, \gamma) = \frac{1}{\pi} \frac{1}{\gamma(1+(\frac{x-\theta}{\gamma})^2)} \\ -\infty < x < \infty \quad \theta \in R \quad \gamma > 0$$

برای این توزیع میانگین، واریانس و تابع مولد وجود ندارد اما گشتاورهای مراتب کوچکتر از ۱ وجود دارد. شکل توزیع کوشی مانند شکل توزیع نرمال است البته با دم‌های سنگین‌تر. در این توزیع  $\theta$  میانه توزیع است و  $\gamma$  پارامتر مقیاس (میزان پراکندگی). مقدار  $\gamma$  برابر نصف دامنه‌ی میان چارکی (IQR) است.

### ۱۱.۳ توزیع پارتولو

$$X \sim Pa(m, \alpha) : f_X(x; m, \alpha) = \frac{\alpha m}{x^{\alpha+1}} \quad m \leq x \quad m, \alpha > 0$$

مانند کوشی دم‌سنگین است اما برخلاف آن چولگی دارد. تابع مولد ندارد اما گشتاورهای آن بازه برشی مقادیر  $\alpha$  وجود دارد.

$$E(X) = \frac{m\alpha}{\alpha - 1} \quad \alpha > 1, \quad V(X) = \frac{m^2\alpha}{(\alpha - 1)^2(\alpha - 2)} \quad \alpha > 2$$

## ۴ متغیرهای تصادفی باهم

تابع جرم احتمال باهم:

$$f_{XY}(x, y) \geq 0$$

$$\sum_{X,Y} f_{XY}(x, y) = 1$$

$$f_{XY}(x, y) = P(X = x, Y = y)$$

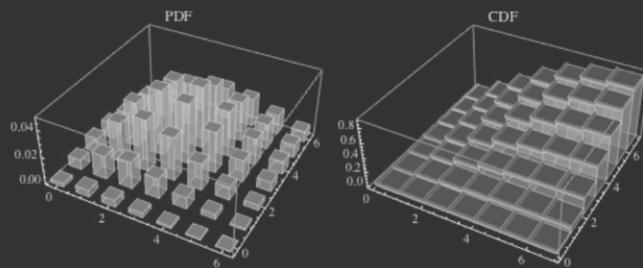
	$Y_1$	$\dots$	$Y_n$	Marginal Probabilities of X
$X_1$	$p_{11}$	$\dots$	$p_{1n}$	$P_1$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X_m$	$p_{m1}$	$\dots$	$p_{mn}$	$P_m$
Marginal Probabilities of Y	$P'_1$	$\dots$	$P'_n$	1

تابع چگالی احتمال باهم:

$$f_{XY}(x, y) \geq 0$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1$$

$$\int \int_R f_{XY}(x, y) dx dy = P((X, Y) \in R)$$



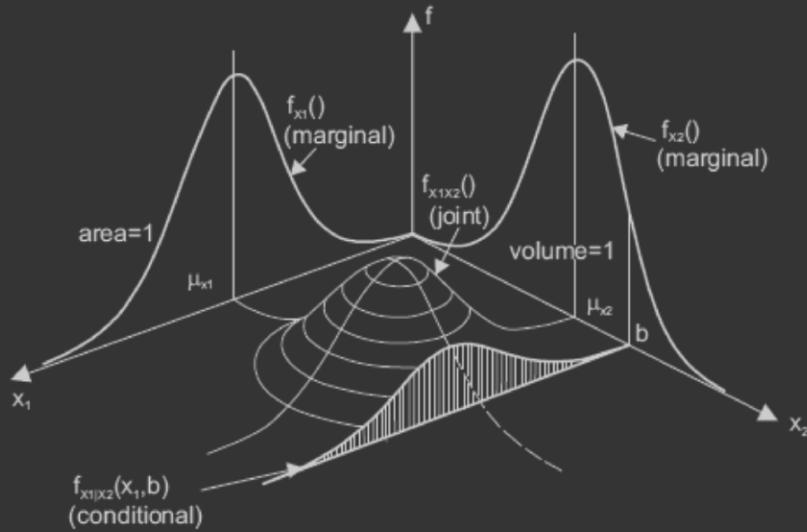
توزیع احتمال حاشیه‌ای:

$$f_X(x) = \int f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int f_{X,Y}(x, y) dx$$

توزیع احتمال شرطی: تابع چگالی احتمال با هم  $f_{X,Y}(x, y)$  را در نظر بگیرید. تابع چگالی احتمال رویداد  $Y$  به شرطی که  $X = x$  از رابطه زیر محاسبه می‌شود.

$$f_{Y|x}(y) = \frac{f_{X,Y}(x, y)}{f_X(x)} \quad \text{for } f_X(x) > 0$$



$$f_{Y|x}(y) \geq 0$$

$$\int f_{Y|x}(y) dy = 1$$

$$P(Y \in B | X = x) = \int_B f_{Y|x}(y) dy$$

$$\mu_{Y|x} = E(Y|x) = \int_Y y f_{Y|x}(y) dy$$

$$\sigma_{Y|x}^2 = V(Y|x) = \int_Y (y - \mu_{Y|x})^2 f_{Y|x}(y) dy = \int_Y y^2 f_{Y|x}(y) dy - \mu_{Y|x}^2$$

استقلال:

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$$

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

$$\begin{aligned} f_{Y|x}(y) &= f_Y(y) \\ f_{X|y}(x) &= f_X(x) \end{aligned}$$

تعتميم به  $p$  متغير با هم:

$$f_{X_1, X_2, \dots, X_p}(x_1, x_2, \dots, x_p) \geq 0$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1, X_2, \dots, X_p}(x_1, x_2, \dots, x_p) dx_1 dx_2 \dots dx_p = 1$$

$$P((X_1, X_2, \dots, X_p) \in B) = \int \int \cdots \int_B f_{X_1, X_2, \dots, X_p}(x_1, x_2, \dots, x_p) dx_1 dx_2 \dots dx_p$$

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1, \dots, X_p}(x_1, \dots, x_p) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_p$$

$$E(X_i) = \int_{-\infty}^{\infty} x_i f_{X_i}(x_i) dx_i$$

$$V(X_i) = \int_{-\infty}^{\infty} (x_i - \mu_{X_i})^2 f_{X_i}(x_i) dx_i$$

$$f_{X_1 X_2 X_3 | x_4 x_5}(x_1, x_2, x_3) = \frac{f_{X_1, X_2, X_3, X_4, X_5}(x_1, x_2, x_3, x_4, x_5)}{f_{X_4, X_5}(x_4, x_5)} \text{ for } f_{X_4, X_5}(x_4, x_5) > 0$$

independent are  $X_1, X_2, \dots, X_p$  if

$$f_{X_1, X_2, \dots, X_p}(x_1, x_2, \dots, x_p) = f_{X_1}(x_1) f_{X_2}(x_2) \dots f_{X_p}(x_p) \quad \text{all for } x_1, x_2, \dots, x_p$$

$$P(X_1 \in A_1, X_2 \in A_2, \dots, X_p \in A_p) = P(X_1 \in A_1) P(X_2 \in A_2) \dots P(X_p \in A_p)$$

مقدار چشمداشتی یک تابع از ۲ متغیر تصادفی:

$$E(h(X, Y)) = \begin{cases} \sum \sum h(X, Y) f_{XY}(x, y) & X, Y \text{ discrete} \\ \int \int h(X, Y) f_{XY}(x, y) dx dy & X, Y \text{ continuous} \end{cases}$$

کواریانس بین دو متغیر تصادفی  $X$  و  $Y$ :

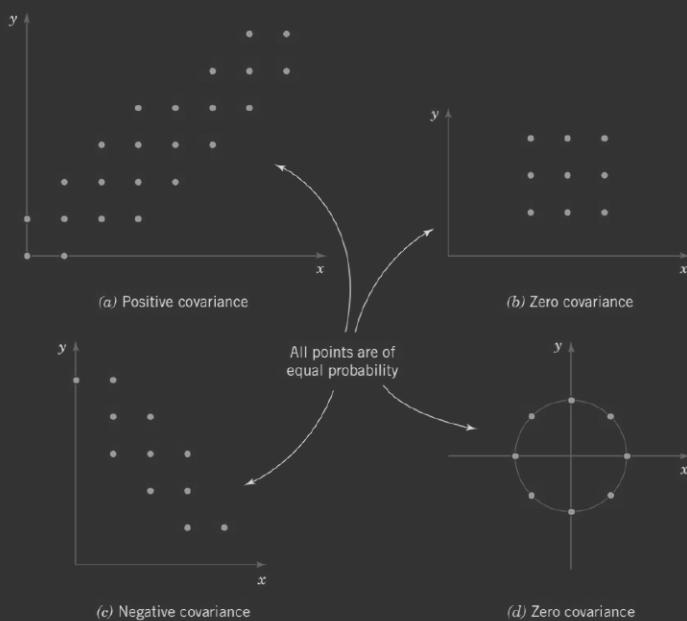
$$\sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - \mu_X \mu_Y$$

همبستگی بین دو متغیر تصادفی  $X$  و  $Y$ :

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

independent: are  $Y$  and  $X$  If

$$\sigma_{XY} = \rho_{XY} = 0$$



توزیع‌های چندمتغیر با هم معمول:

توزیع احتمال چندجمله‌ای:

فرض کنید که یک آزمایش تصادفی از تکرار  $n$  آزمایش ساده تشکیل شده است. فرض کنید که

- ۱ - نتیجه هر تکرار در یک از کلاس های  $k$  طبقه بندی می شود.
  - ۲ - احتمال رخداد هر رویداد در هر یک از کلاس های  $1, 2, \dots, k$  به ترتیب برابر  $p_1, p_2, \dots, p_k$  می باشد.
  - ۳ - تکرارها از هم مستقل هستند.
- متغیرهای تصادفی  $X_1, X_2, \dots, X_k$  تعداد نتایج رخداد در هر یک از کلاس ها می باشند.

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

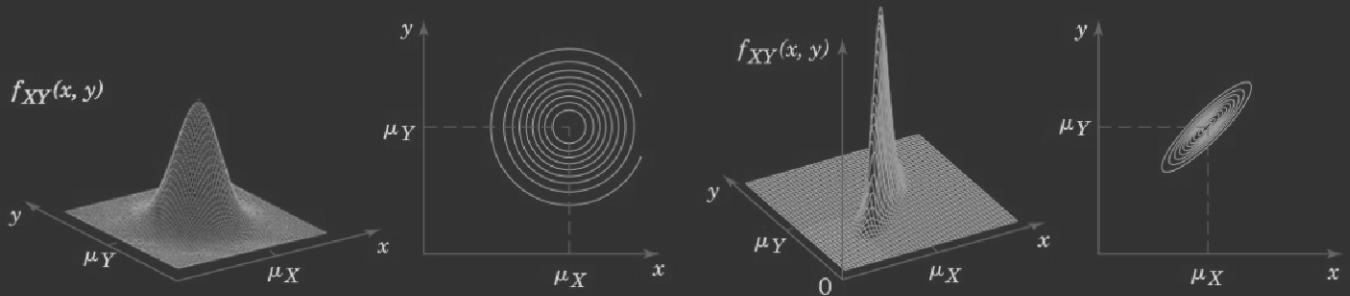
$$x_1 + \dots + x_k = n \text{ and } p_1 + \dots + p_k = 1$$

$$E(X_i) = np_i \quad V(X_i) = np_i(1-p_i)$$

توزیع نرمال دومتغیره:

$$f_{XY}(x, y; \sigma_X, \sigma_Y, \mu_X, \mu_Y, \rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)} \left( \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right)\right)$$

$$-\infty < x, y, \mu_X, \mu_Y < \infty \quad 0 < \sigma_X, \sigma_Y \quad -1 < \rho < 1$$



اگر متغیرهای تصادفی  $X$  و  $Y$  از توزیع نرمال دومتغیره پیروی کنند. و چگالی احتمال باهم آنها برابر ( $f_{XY}(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ ) باشد. توزیع احتمال حاشیه‌ای آنها برای  $X$  و برای  $Y$  توزیع نرمال با میانگین و انحراف معیار به ترتیب  $\mu_X, \mu_Y, \sigma_X, \sigma_Y$  می باشد. توزیع احتمال شرطی  $Y$  برای  $X = x$  توزیع نرمال با میانگین

$$\mu_{Y|x} = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$$

و واریانس

$$\sigma_{Y|x}^2 = \sigma_Y^2 (1 - \rho^2)$$

می باشد.

ضریب همبستگی بین  $X$  و  $Y$  برابر  $\rho$  است. و اگر  $\rho = 0$  باشد، دو متغیر از هم مستقل هستند. توابع خطی از متغیرهای تصادفی:

$$Y = c_1 X_1 + \dots + c_p X_p$$

$$E(Y) = c_1 E(X_1) + \dots + c_p E(X_p)$$

$$V(Y) = c_1^2 V(X_1) + \dots + c_p^2 V(X_p) + 2 \sum_{i < j} c_i c_j \text{cov}(X_i, X_j)$$

بطور مثال اگر متغیرها از هم مستقل و توزیع نرمال یکسانی با پارامترهای  $\mu$  و  $\sigma^2$  داشته باشند.

$$E(\bar{X}) = E\left(\frac{X_1 + \dots + X_p}{p}\right) = \mu$$

$$V(\bar{X}) = V\left(\frac{X_1 + \dots + X_p}{p}\right) = \frac{\sigma^2}{p}$$

این همان مفهوم انتشار خطا است و نشان می‌دهد با تکرار آزمایش‌های یکسان مقدار میانگین تغییر نمی‌کند اما خطای آماری کاهش می‌یابد.

### مثال

فرض کنید تابع چگالی توانم به صورت زیر باشد.

$$f(x, y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{در غیر این صورت} \end{cases}$$

مطلوب است محاسبه:

$$P(X < a), \quad P(X < Y), \quad P(X > 1, Y < 1)$$

حل:

$$P\{X > 1, Y < 1\} = \int_0^1 \int_1^\infty 2e^{-x}e^{-2y} dx dy = \int_0^1 2e^{-2y} (-e^{-x}) \Big|_1^\infty dy = e^{-1} \int_0^1 2e^{-2y} dy = e^{-1} (1 - e^{-2})$$

$$P\{X < Y\} = \iiint_{(x,y):x < y} 2e^{-x}e^{-2y} dx dy = \int_0^\infty \int_0^y 2e^{-x}e^{-2y} dx dy = \int_0^\infty 2e^{-2y}(1 - e^{-y}) dy$$

$$= \int_0^\infty 2e^{-2y} dy - \int_0^\infty 2e^{-3y} dy = 1 - \frac{2}{3} = \frac{1}{3}$$

$$P\{X < a\} = \int_0^a \int_0^\infty 2e^{-2y}e^{-x} dy dx = \int_0^a e^{-x} dx = 1 - e^{-a}$$

### مثال

فرض کنید  $X$  و  $Y$  متغیرهای تصادفی مستقل باشند با تابع چگالی احتمال مشترک

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{جاهای دیگر} \end{cases}$$

مطلوب است تابع چگالی احتمال متغیر تصادفی  $\frac{X}{Y}$  را به دست آوریم.

حل: ابتدا تابع توزیع تجمعی  $\frac{X}{Y}$  را به دست می‌آوریم

$$F_{X/Y}(a) = P\left(\frac{X}{Y} \leq a\right) = \iint_{\frac{x}{y} \leq a} f(x, y) dx dy = \iint_{\frac{x}{y} \leq a} e^{-x}e^{-y} dx dy = \int_0^\infty \int_0^{ay} e^{-x}e^{-y} dx dy$$

$$= \int_0^\infty (1 - e^{-ay}) e^{-y} dy = \int_0^\infty \left( -e^{-y} + \frac{e^{-(a+1)y}}{a+1} \right) dy = \left[ -e^{-y} + \frac{e^{-(a+1)y}}{a+1} \right]_0^\infty = 1 - \frac{1}{a+1}$$

با مشتق‌گیری از این مقدار، تابع چگالی  $\frac{X}{Y}$  به صورت زیر به دست می‌آید:

$$f_{X/Y}(a) = \frac{1}{(a+1)^2}, \quad 0 < a < \infty.$$

## ۵ توابع عمومی از متغیرهای تصادفی

فرض کنید  $X$  متغیر تصادفی گسسته با توزیع احتمال  $f_X(x)$  باشد. تابع  $Y = h(X)$  تبدیل یک به یک بین متغیرهای  $X$  و  $Y$  به نحوی تعریف می‌کند که همواره معادله‌ی  $y = h(x)$  دارای پاسخ یکتا برای  $x$  برحسب  $y$  داشته باشد که آن را  $u(y)$  بگیرید. حال توزیع احتمال متغیر تصادفی  $Y$  از رابطه‌ی زیر بدست می‌آید.

$$f_Y(y) = f_X(u(y))$$

این رابطه به دو متغیر تصادفی قابل تعمیم می‌باشد.

$$f_{Y_1 Y_2}(y_1, y_2) = f_{X_1 X_2}(u_1(x_1, x_2), u_2(x_1, x_2))$$

و برای توزیع احتمال حاشیه‌ای آن می‌توان رابطه‌ی زیر را نوشت.

$$f_{Y_1}(y_1) = \sum_{y_2} f_{Y_1 Y_2}(y_1, y_2)$$

حال اگر فرض کنیم که  $X$  متغیر تصادفی پیوسته است. توزیع چگالی احتمال از رابطه‌ی زیر بدست می‌آید.

$$f_Y(y) = f_X(u(y)) |J| \quad J = u'(y)$$

تعمیم این رابطه به دو متغیر پیوسته به شکل زیر تبدیل می‌شود:

$$f_{Y_1 Y_2}(y_1, y_2) = f_{X_1 X_2}(u_1(y_1, y_2), u_2(y_1, y_2)) |J|$$

که در اینجا از قدرمطلق ژاکوبی استفاده شده است. ژاکوبی برای دو متغیر دترمینان زیر می‌باشد.

$$J = \begin{vmatrix} \partial_{y_1} x_1 & \partial_{y_2} x_1 \\ \partial_{y_1} x_2 & \partial_{y_2} x_2 \end{vmatrix}$$

مجدد می‌توان توزیع حاشیه‌ای آن را نیز محاسبه نمود.

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1 Y_2}(y_1, y_2) dy_2$$

اگر تبدیل  $(X) = h(Y)$  یک به یک نباشد می‌توان بازه‌ی  $X$  را به یک سری زیر بازه‌ها تقسیم کرد که در هر یک از این زیر بازه‌ها تبدیل یک به یک باشد و چگالی احتمال را از رابطه‌ی زیر محاسبه نمود.

$$f_Y(y) = \sum_{i=1}^m f_X(u_i(y)) |J_i| \quad J_i = u'_i(y), \quad i = 1, \dots, m$$

گشتاور ام ۲ برای متغیر تصادفی  $X$  از رابطه‌ی زیر محاسبه می‌شود.

$$\mu_r = E(X^r) = \begin{cases} \sum x^r f_X(x), & X \text{ discrete} \\ \int_{-\infty}^{\infty} x^r f_X(x) dx, & X \text{ continuous} \end{cases}$$

تابع مولد گشتاور برای متغیر تصادفی  $X$  نیز از رابطه‌ی زیر حاصل می‌شود.

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum e^{tx} f_X(x), & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx, & X \text{ continuous} \end{cases}$$

$$\mu_r = \left. \frac{d^r M_X(t)}{dt^r} \right|_{t=0}$$


---

برای محاسبه‌آمار روی توابع می‌توان از بسط تیلور توابع نیز استفاده نمود. برای یادآوری بسط تیلور چند متغیره یک تابع به شکل زیر تعریف می‌شود.

$$\mathbf{x} = (x_1, \dots, x_d)^T$$

$$f_X(\mathbf{x}) = f(\mu) + \sum_{j=1}^d \frac{\partial f}{\partial x_j} \Big|_{\mu_j} (x_j - \mu_j) + \frac{1}{2} \sum_{j=1}^d \sum_{k=1}^d \frac{\partial^2 f}{\partial x_j \partial x_k} \Big|_{\mu_j, \mu_k} (x_j - \mu_j)(x_k - \mu_k)$$

$$+ \frac{1}{6} \sum_{j=1}^d \sum_{k=1}^d \sum_{\ell=1}^d \frac{\partial^3 f}{\partial x_j \partial x_k \partial x_\ell} \Big|_{\mu_j, \mu_k, \mu_\ell} (x_j - \mu_j)(x_k - \mu_k)(x_\ell - \mu_\ell),$$

بطور مثال وقتی تابع مدنظر از دو متغیر تصادفی  $X$  و  $Y$  باشد بطور خلاصه تا تقریب مرتبه ۲ بسط به شکل زیر خواهد شد.

$$f(X, Y) \approx f(\mu_X, \mu_Y) + \frac{\partial f}{\partial X} \Big|_{\mu_X, \mu_Y} (X - \mu_X) + \frac{\partial f}{\partial Y} \Big|_{\mu_X, \mu_Y} (Y - \mu_Y)$$

$$+ \frac{1}{2} \frac{\partial^2 f}{\partial X^2} \Big|_{\mu_X, \mu_Y} (X - \mu_X)^2 + \frac{1}{2} \frac{\partial^2 f}{\partial Y^2} \Big|_{\mu_X, \mu_Y} (Y - \mu_Y)^2 + \frac{\partial^2 f}{\partial X \partial Y} \Big|_{\mu_X, \mu_Y} (X - \mu_X)(Y - \mu_Y)$$

برای محاسبه میانگین این تابع با تقریب مرتبه اول:

$$E(f_{XY}(x, y)) = E(f(\mu_X, \mu_Y)) + \frac{\partial f}{\partial x} E(x - \mu_X) + \frac{\partial f}{\partial y} E(y - \mu_Y) = f(\mu_X, \mu_Y)$$

و یا برای محاسبه واریانس با تقریب مرتبه اول داریم:

$$V(f_{XY}(x, y)) = V(f(\mu_X, \mu_Y)) + (\frac{\partial f}{\partial x})^2 V(x - \mu_X) + (\frac{\partial f}{\partial y})^2 V(y - \mu_Y)$$

$$= 0 + (\frac{\partial f}{\partial x})^2 V(X) + (\frac{\partial f}{\partial y})^2 V(Y)$$

### مثال

فرض کنید تابعی به صورت زیر داده شده است:

$$f(X, Y) = X^2 + XY + Y^2$$

که  $X$  و  $Y$  دو متغیر تصادفی مستقل با میانگین‌های  $\mu_X$  و  $\mu_Y$  و واریانس‌های  $\sigma_X^2$  و  $\sigma_Y^2$  هستند.  
الف) از بسط تیلور مرتبه دوم برای تقریب تابع  $f(X, Y)$  حول نقطه  $(\mu_X, \mu_Y)$  استفاده کنید و بیان کنید که تابع تقریب‌زده شده چگونه خواهد بود.

ب) با استفاده از تقریب مرتبه اول بسط تیلور، میانگین  $E(f(X, Y))$  و واریانس  $V(f(X, Y))$  را محاسبه کنید.  
راهنمایی: برای حل این سوال، لازم است مشتقات جزئی تابع  $f(X, Y)$  را نسبت به  $X$  و  $Y$  محاسبه کنید و سپس میانگین و واریانس را با استفاده از بسط تیلور و فرض استقلال  $X$  و  $Y$  به دست آورید.

الف) تابع  $f(X, Y) = X^2 + XY + Y^2$  حول نقطه  $(\mu_X, \mu_Y)$  با استفاده از بسط تیلور مرتبه دوم تقریب می‌زنیم.

$$\frac{\partial f}{\partial X} = 2X + Y, \quad \frac{\partial f}{\partial Y} = 2Y + X$$

$$\frac{\partial^2 f}{\partial X^2} = 2, \quad \frac{\partial^2 f}{\partial Y^2} = 2, \quad \frac{\partial^2 f}{\partial X \partial Y} = 1$$

پس بسط تیلور مرتبه دوم تابع مدنظر به صورت زیر خواهد بود:

$$f(X, Y) \approx f(\mu_X, \mu_Y) + (2\mu_X + \mu_Y)(X - \mu_X) + (2\mu_Y + \mu_X)(Y - \mu_Y) \\ + \frac{1}{2} \cdot 2(X - \mu_X)^2 + \frac{1}{2} \cdot 2(Y - \mu_Y)^2 + 1 \cdot (X - \mu_X)(Y - \mu_Y)$$

ساده‌سازی این عبارت به صورت زیر خواهد بود:

$$f(X, Y) \approx \mu_X^2 + \mu_X \mu_Y + \mu_Y^2 + (2\mu_X + \mu_Y)(X - \mu_X) + (2\mu_Y + \mu_X)(Y - \mu_Y) \\ + (X - \mu_X)^2 + (Y - \mu_Y)^2 + (X - \mu_X)(Y - \mu_Y)$$

ب) برای محاسبه میانگین  $E(f(X, Y))$  با استفاده از تقریب مرتبه اول، فقط بخش ثابت و بخش‌های خطی در  $X - \mu_X$  و  $Y - \mu_Y$  را در نظر می‌گیریم، زیرا امید ریاضی  $X - \mu_X$  و  $Y - \mu_Y$  برابر صفر هستند. بنابراین، داریم:

$$E(f(X, Y)) \approx f(\mu_X, \mu_Y) = \mu_X^2 + \mu_X \mu_Y + \mu_Y^2$$

برای محاسبه واریانس  $f(X, Y)$  از تقریب مرتبه اول، باید واریانس بخش‌های درجه دوم را محاسبه کنیم.

$$V(f(X, Y)) \approx \text{Var}((X - \mu_X)^2 + (Y - \mu_Y)^2 + (X - \mu_X)(Y - \mu_Y))$$

با استفاده از استقلال  $X$  و  $Y$

$$V(f(X, Y)) \approx \sigma_X^2 + \sigma_Y^2$$

در نتیجه واریانس تقریبی تابع به شکل بالا به دست می‌آید.

## مثال

سرعت  $V$  یک مولکول در یک گاز در تعادل دارای تابع چگالی احتمال (pdf) زیر است:

$$f(v) = av^2 e^{-bv^2}, v > 0; f(v) = 0, \text{ otherwise.}$$

در اینجا  $b = m/2kT$  است، که در آن  $m$  جرم مولکول،  $T$  دمای مطلق و  $k$  ثابت بولتزمن است.

(الف) ثابت  $a$  را بر حسب  $b$  محاسبه کنید.

(ب) در نظر بگیرید که  $Y = (m/2)V^2$  ارزی جنبشی مولکول باشد. تابع توزیع تجمعی  $G(y) = P(Y \leq y)$  را برای  $Y$  محاسبه کنید.

(ج) تابع چگالی احتمال  $g(y) = G'(y)$  را محاسبه کنید.

(الف) محاسبه ثابت  $a$  بر حسب  $b$ :  
چون  $f(v)$  یک تابع چگالی احتمال است، باید انتگرال آن در کل دامنه برابر با 1 باشد:

$$\int_0^\infty f(v) dv = 1. \quad \rightarrow \quad \int_0^\infty av^2 e^{-bv^2} dv = 1.$$

برای حل این انتگرال، از تغییر متغیر  $dv$  استفاده می‌کنیم، که نتیجه می‌دهد:

$$v dv = \frac{du}{2b} \quad \rightarrow \quad \int_0^\infty v^2 e^{-bv^2} dv = \int_0^\infty \frac{u^{1/2} e^{-u}}{2(b)^{3/2}} du.$$

این انتگرال با توجه به خواص تابع گاما حل می‌شود. با توجه به اینکه  $\Gamma(3/2) = \frac{\sqrt{\pi}}{2}$  و  $\int_0^\infty u^{n-1} e^{-u} du = \Gamma(n)$  داریم:

$$\int_0^\infty v^2 e^{-bv^2} dv = \frac{\sqrt{\pi}}{4b^{3/2}}. \quad \rightarrow \quad a \cdot \frac{\sqrt{\pi}}{4b^{3/2}} = 1. \quad \rightarrow \quad a = \frac{4b^{3/2}}{\sqrt{\pi}}.$$

(ب) محاسبه تابع توزیع تجمعی  $G(y) = P(Y \leq y)$  بر حسب  $V^2 = \frac{2Y}{m}$  تعیین کرد. نتیجه می‌دهد.

$$G(y) = P(Y \leq y) = P\left(V^2 \leq \frac{2y}{m}\right) = P\left(V \leq \sqrt{\frac{2y}{m}}\right).$$

بنابراین:

$$G(y) = \int_0^{\sqrt{2y/m}} f(v) dv. \quad \rightarrow \quad G(y) = \int_0^{\sqrt{2y/m}} \frac{4b^{3/2}}{\sqrt{\pi}} v^2 e^{-bv^2} dv.$$

برای محاسبه این انتگرال از تغییر متغیر  $dv = \frac{du}{2b}$  استفاده می‌کنیم. بنابراین،

$$v^2 e^{-bv^2} dv = \frac{u}{b} e^{-u} \cdot \frac{du}{2b} = \frac{u}{2b^2} e^{-u} du.$$

$$\rightarrow G(y) = \int_0^{2by/m} \frac{4b^{3/2}}{\sqrt{\pi}} \cdot \frac{u}{2b^2} e^{-u} du. = \frac{2\sqrt{b}}{\sqrt{\pi}} \int_0^{2by/m} ue^{-u} du.$$

انتگرال  $\int ue^{-u} du$  را می‌توان با روش جزء به جزء حل کرد:

$$\int ue^{-u} du = -ue^{-u} - \int -e^{-u} du = -ue^{-u} + e^{-u} = (1-u)e^{-u}.$$

$$\rightarrow G(y) = \frac{2\sqrt{b}}{\sqrt{\pi}} \left[ -(1-u)e^{-u} \right]_0^{2by/m}.$$

$$G(y) = \frac{2\sqrt{b}}{\sqrt{\pi}} \left( 1 - \frac{2by}{m} \right) e^{-\frac{2by}{m}} - \frac{2\sqrt{b}}{\sqrt{\pi}}.$$

(ج) محاسبه تابع چگالی احتمال  $g(y) = G'(y)$

$$G(y) = \int_0^{\sqrt{2y/m}} \frac{4b^{3/2}}{\sqrt{\pi}} v^2 e^{-bv^2} dv.$$

برای مشتق‌گیری از  $G(y)$  نسبت به  $y$ ، از قاعده زنجیره‌ای استفاده می‌کنیم. با تعریف  $u = \sqrt{2y/m}$  به عنوان کران بالای انتگرال، داریم:

$$G'(y) = \frac{d}{dy} \int_0^u \frac{4b^{3/2}}{\sqrt{\pi}} v^2 e^{-bv^2} dv.$$

طبق قاعده اساسی حساب دیفرانسیل و انتگرال:

$$G'(y) = \frac{4b^{3/2}}{\sqrt{\pi}} \left( u^2 e^{-bu^2} \right) \cdot \frac{du}{dy}.$$

اکنون  $u = \sqrt{2y/m}$  را نسبت به  $y$  مشتق می‌گیریم:

$$\frac{du}{dy} = \frac{1}{\sqrt{2y/m}} \cdot \frac{1}{\sqrt{m/2}} = \frac{1}{\sqrt{2ym}}.$$

پس:

$$G'(y) = \frac{4b^{3/2}}{\sqrt{\pi}} \left( \frac{2y}{m} \right) e^{-b\frac{2y}{m}} \cdot \frac{1}{\sqrt{2ym}}.$$

با ساده‌سازی، نتیجه نهایی به صورت زیر خواهد بود:

$$g(y) = G'(y) = \frac{4b^{3/2}}{\sqrt{\pi}} \cdot \frac{2y}{m} \cdot e^{-b\frac{2y}{m}} \cdot \frac{1}{\sqrt{2ym}} = \frac{2b^{3/2}}{\sqrt{\pi y}} e^{-\frac{2by}{m}}.$$

بنابراین، تابع چگالی احتمال  $Y$  به صورت زیر است:

$$g(y) = \frac{2b^{3/2}}{\sqrt{\pi y}} e^{-\frac{2by}{m}}.$$

## ۶ نمونه‌گیری

مجموعه هدف برای بررسی ویژگی‌های آن را جامعه آماری می‌گوییم. نمونه تصادفی با اندازه  $n$  از یک جامعه آماری عبارت از  $n$  متغیر تصادفی  $X_1, X_2, \dots, X_n$  مستقل و همه با توزیع یکسان  $f_X(x)$  می‌باشد. بنابر فرض استقلال و هم توزیع بودن، توزیع باهم این نمونه عبارت است:

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i)$$

به محض نمونه‌گیری، کمیتهای مشاهده‌پذیر  $X_1, X_2, \dots, X_n$  (متغیرهای تصادفی) مقادیر تحقق یافته (مشاهده شده)  $x_1, x_2, \dots, x_n$  را می‌گیرند. مستقل و هم توزیع  $f_{X_1, \dots, X_n}(x_1, \dots, x_n)$  distributed identically and (independent) نشان داده می‌شود. بطور مثال:

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$$

هر تابع از مشاهدات نمونه‌ای را آماره (Statistics) می‌نامند. آماره‌ها تابع متغیرهای تصادفی هستند و در نتیجه خود متغیر تصادفی هستند و دارای توزیع احتمال می‌باشند که توزیع نمونه‌ای آماره نامیده می‌شود. از مشهورترین آماره‌ها میانگین و واریانس نمونه‌ای هستند.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

شاخص‌های جامعه را پارامترهای جامعه می‌نامند. بطور مثال میانگین و واریانس جامعه  $\mu$  و  $\sigma^2$  می‌باشد. آماره‌ها متغیرهای تصادفی و دارای تابع توزیع احتمال هستند و بعد از اندازه‌گیری مقادیر آنها مشخص می‌شود در حالیکه پارامترها گرچه مجھول باشند اما اعداد ثابت و غیر تصادفی هستند.

**قضیه:** اگر  $X_1, \dots, X_n$  نمونه‌ای تصادفی از جامعه‌ای با میانگین  $E(X) = \mu$  و واریانس  $V(X) = \sigma^2$  باشد آنگاه:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu \quad V(\bar{X}) = \sigma_{\bar{X}}^2 = \sigma^2/n$$

**قضیه:**

$$\begin{aligned} X_1, \dots, X_n &\stackrel{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n) \\ \Rightarrow S_n = X_1 + \dots + X_n &= N(n\mu, n\sigma^2) \end{aligned}$$

**قضیه:** اگر  $X_i$ ‌ها نمونه تصادفی از جامعه با پارامترهای  $\mu$  و  $\sigma^2$  باشند، با استفاده از قضیه چبیشف داریم:

$$P(|\bar{X} - \mu| \geq \epsilon) \leq \sigma^2/n\epsilon^2 \Rightarrow P(|\bar{X} - \mu| < \epsilon) \geq 1 - \sigma^2/n\epsilon^2$$

**قضیه حد مرکزی:** اگر  $X_i$ ‌ها نمونه تصادفی از جامعه‌ای با میانگین  $\mu$  و واریانس  $\sigma^2$  باشد، هرگاه  $n$  بزرگ باشد توزیع  $\bar{X}$  تقریباً نرمال با میانگین  $\mu$  و واریانس  $\sigma^2/n$  است. هر قدر توزیع جامعه از حالت نرمال و متقاضی دور باشد، مقادیر بزرگتری لازم است تا تقریب خوبی داشته باشیم. (برای یک تقریب خوب  $n \geq 25$ )

تعیین قضیه حد مرکزی: حتی اگر متغیرهای تصادفی هم توزیع نباشند، به شرطی که متغیرهای تصادفی نسبت به مجموع کل تاثیر کمی داشته باشند، همچنان قضیه حد مرکزی معتبر می‌باشد.

$$\mu \rightarrow \frac{1}{n} \sum_{i=1}^n \mu_i \quad \sigma^2 \rightarrow \frac{1}{n} \sum_{i=1}^n \sigma_i^2$$

### مثال

فرض کنید یک دستگاه بطری‌پرکن داریم، به گونه‌ای تنظیم شده که به طور متوسط مقدار  $\mu$  گرم مایع در هر بطری پر می‌کند. می‌دانیم که میزان مایع پر شده در هر بطری توسط این دستگاه از توزیع نرمال پیروی می‌کند و انحراف معیار آن برابر  $10.0 \text{ gr} = \sigma$  است. از خروجی دستگاه، یک دسته نمونه تصادفی با اندازه  $n = 9$  بطری پر شده انتخاب و میزان مایع در هر یک از این بطری‌ها را اندازه‌گیری می‌کنیم. هدف ما این است که احتمال اینکه میانگین نمونه‌ها در فاصله ۳ گرمی از میانگین واقعی  $\mu$  قرار گیرد، را محاسبه کنیم.

$$P(|\bar{X} - \mu| \leq 3) = ?$$

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

$$\begin{aligned} P(|\bar{X} - \mu| \leq 3) &= P(-3 \leq (\bar{X} - \mu) \leq 3) \\ &= P\left(\frac{-3}{\sigma/\sqrt{n}} \leq \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \leq \frac{3}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{-3}{10/\sqrt{9}} \leq Z \leq \frac{3}{10/\sqrt{9}}\right) \\ &= P(-0.9 \leq Z \leq 0.9) \end{aligned}$$

$$P(|\bar{X} - \mu| \leq 3) = 0.6318$$

### مثال

در مثال قبل، دسته نمونه‌ی انتخاب شده از خروجی دستگاه با چه اندازه‌ای باشد که با احتمال 0.95 میانگین آنها در فاصله‌ی 3 گرمی از میانگین واقعی  $\mu$  بدست آید.

$$n = ? \quad \rightarrow \quad P(|\bar{X} - \mu| \leq 3) = 0.95$$

از توزیع نرمال استاندارد  $Z$  می‌دانیم که:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

می‌خواهیم:

$$P(|\bar{X} - \mu| \leq 3) = P\left(\frac{-3}{10/\sqrt{n}} \leq Z \leq \frac{3}{10/\sqrt{n}}\right) = 0.95$$

پس:

$$0.3/\sqrt{n} = 1.96 \quad \rightarrow \quad n = \left(\frac{1.96}{0.3}\right)^2 = 42.68 \quad \rightarrow \quad n = 43$$

### قضیه

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2) \quad \rightarrow \quad \begin{cases} \frac{nS_\mu^2}{\sigma^2} = \frac{\sum(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2 \\ \frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2 \\ \bar{X} \perp\!\!\!\perp S^2 \end{cases}$$

$$P(X \leq \chi_{n,\alpha}^2) = \alpha$$

که در آن  $\chi_n^2$  توزیع کای دو (خی دو) با درجه آزادی  $n$  می‌باشد. این توزیع به شکل زیر تعریف می‌شود.

$$X \sim \chi_n^2 : \quad f_X(x) = \frac{1}{\Gamma(n/2)2^{n/2}} e^{-x/2} x^{(n/2-1)} \quad x > 0, n > 0$$

## مثال

فرض کنید دستگاهی داریم که موقعیت یک شیء واقع در مختصات  $(a, b)$  را ثبت می‌کند. موقعیت ثبت شده توسط دستگاه در مختصات  $(X, Y)$  قرار دارد، به طوری که:

$$(X, Y) = (a + X_1, b + X_2)$$

در اینجا، متغیرهای تصادفی  $X_1$  و  $X_2$  دارای توزیع نرمال مستقل و یکسان هستند، یعنی:

$$X_i \sim N(0, \sigma^2) \quad \text{برای } i = 1, 2$$

این یعنی مختصات ثبت شده توسط دستگاه دچار خطای تصادفی با توزیع نرمال هستند. حالا متغیر تصادفی  $R$  به شکل زیر تعریف شده است:

$$R = \sqrt{(X - a)^2 + (Y - b)^2} = \sqrt{X_1^2 + X_2^2}$$

این متغیر  $R$  فاصله بین موقعیت واقعی شیء و موقعیت ثبت شده توسط دستگاه را نشان می‌دهد. نشان دهید که:

$$P(R \leq r) = P\left(\chi_2^2 \leq \left(\frac{r}{\sigma}\right)^2\right)$$

از آنجا که  $X_1$  و  $X_2$  هر دو دارای توزیع نرمال  $N(0, \sigma^2)$  هستند، می‌توانیم آنها را به صورت زیر بنویسیم:

$$X_i \sim N(0, \sigma^2) \Rightarrow \frac{X_i}{\sigma} \sim N(0, 1)$$

بنابراین، اگر  $Z_i = \frac{X_i}{\sigma}$  باشد، آنگاه  $Z_1$  و  $Z_2$  دارای توزیع نرمال استاندارد  $N(0, 1)$  هستند. با استفاده از  $Z_1$  و  $Z_2$ ، می‌توان  $R$  را به صورت زیر بازنویسی کرد

$$R^2 = X_1^2 + X_2^2 = \sigma Z_1^2 + (\sigma Z_2)^2 = \sigma^2(Z_1^2 + Z_2^2)$$

از آنجا که  $Z_1$  و  $Z_2$  هر دو دارای توزیع نرمال استاندارد مستقل  $N(0, 1)$  هستند، مجموع مربعات آنها، یعنی  $Z_1^2 + Z_2^2$ ، از توزیع کای دو با ۲ درجه آزادی پیروی می‌کند:

$$Z_1^2 + Z_2^2 \sim \chi_2^2$$

اکنون می‌خواهیم احتمال  $P(R \leq r)$  را به دست آوریم. با توجه به بازنویسی  $R$  داریم:

$$P(R \leq r) = P(R^2 \leq r^2) = P\left(\frac{R^2}{\sigma^2} \leq \frac{r^2}{\sigma^2}\right) = P\left(Z_1^2 + Z_2^2 \leq \left(\frac{r}{\sigma}\right)^2\right) = P\left(\chi_2^2 \leq \left(\frac{r}{\sigma}\right)^2\right)$$

## مثال

یک گلوله توپ به سمت هدف شلیک می‌شود. فاصله (به متر) از نقطه برخورد تا هدف با متغیر تصادفی  $\chi_2^2$  ۷.۲۲ مدل‌سازی شده است. اگر گلوله در فاصله‌ای کمتر از ۱۰ متر از هدف به هدف برخورد کند، هدف نابود می‌شود.

(الف) احتمال اینکه هدف توسط گلوله نابود شود را بیابید.

(ب) اگر گلوله‌ها به صورت متوالی شلیک شوند، چند شلیک لازم است تا با احتمال ۹۰٪ هدف نابود شود؟ توجه: تصمیم‌گیری در مورد تعداد گلوله‌های شلیک شده باید از پیش انجام شود، زیرا نمی‌توانیم مشاهده کنیم که آیا به هدف برخورد کرده‌ایم یا خیر.

می‌دانیم  $X = 7.22\chi_2^2$  و می‌خواهیم  $P(X \leq 10)$  را بیابیم. چون  $\chi_2^2$  دارای توزیع کای-دو با 2 درجه آزادی است، می‌نویسیم:

$$P(X \leq 10) = P\left(\chi_2^2 \leq \frac{10}{7.22}\right) = P\left(\chi_2^2 \leq 1.385\right) \approx 0.75$$

برای قسمت (ب)، می‌خواهیم  $n$  را طوری پیدا کنیم که احتمال نابودی حداقل 90% باشد. داریم:

$$1 - (1 - 0.75)^n \geq 0.9 \quad \rightarrow \quad (1 - 0.75)^n \leq 0.1 \quad \rightarrow \quad 0.25^n \leq 0.1$$

با حل این رابطه، به  $n \approx 6$  می‌رسیم. پس حداقل 6 شلیک لازم است.

### مثال

نوسان (Volatility) یک سهام به انحراف معیار بازدههای آن گفته می‌شود. فرض کنید بازدههای هفتگی یک سهام به طور نرمال با انحراف معیار  $s = 0.042$  توزیع شده‌اند. را به عنوان انحراف معیار نمونه‌ای بازده‌ها در نظر بگیرید که بر اساس یک نمونه شامل 13 بازده هفتگی به دست آمده است (این به معنای ثبت بازده‌ها برای یک‌چهارم سال است). مطلوب است:

$$\begin{aligned} \text{(الف)} & P(s < 0.0522) \\ \text{(ب)} & P(s > 0.0556) \end{aligned}$$

برای حل این مسئله، توجه کنید که  $s$ ، انحراف معیار نمونه‌ای بازده‌ها است و چون توزیع بازده‌ها نرمال است، توزیع انحراف معیار نمونه‌ای  $s$  از یک توزیع کای-دو ( $\chi^2$ ) پیروی می‌کند. با استفاده از توزیع کای-دو، می‌توانیم احتمالات خواسته شده را محاسبه کنیم.

- تعداد مشاهدات  $n = 13$ .
- انحراف معیار جمعیت  $\sigma = 0.042$ .
- توزیع  $s$  به صورت زیر توزیع می‌شود:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2 \quad \rightarrow \quad \frac{12 \cdot s^2}{0.042^2} \sim \chi_{12}^2.$$

(الف) محاسبه  $P(s < 0.0522)$

$$P(s < 0.0522) = P\left(\frac{12 \cdot s^2}{0.042^2} < \frac{12 \cdot (0.0522)^2}{0.042^2} \approx 18.52\right) = P\left(\chi_{12}^2 < 18.52\right).$$

با استفاده از جدول توزیع کای-دو یا محاسبه عددی مقدار تقریبی این احتمال به دست می‌آید:

$$P(s < 0.0522) \approx 0.85.$$

(ب) محاسبه  $P(s > 0.0556)$

$$P(s > 0.0556) = P\left(\frac{12 \cdot s^2}{0.042^2} > \frac{12 \cdot (0.0556)^2}{0.042^2} \approx 21.07\right) = P\left(\chi_{12}^2 > 21.07\right).$$

با استفاده از جدول توزیع کای-دو یا محاسبه عددی، این مقدار به صورت تقریبی به دست می‌آید:

$$P(s > 0.0556) \approx 0.20.$$

قضیه:

$$\left. \begin{array}{l} Z \sim N(0, 1) \\ U \sim \chi_n^2 \\ U \perp\!\!\!\perp Z \end{array} \right\} \Rightarrow \frac{Z}{\sqrt{U/n}} \sim T_n$$

$$\left. \begin{array}{l} X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2) \\ Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \\ U = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2 \end{array} \right\} \Rightarrow \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$$

که در آن توزیع  $t$  (ستیودنت) با درجه آزادی  $n$  می‌باشد.

$$T_n \sim t_n : \quad f_T(t) = \frac{1}{\sqrt{n}\pi} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad -\infty < t < \infty, n > 0$$

$$E(T_n) = 0 \quad n \geq 2, \quad V(T_n) = \frac{n}{n-2} \quad n \geq 3$$

### مثال

مقاومت کششی نوعی سیم دارای توزیع نرمال با میانگین نامعلوم  $\mu$  و واریانس نامعلوم  $\sigma^2$  است. شش قطعه سیم به طور تصادفی از یک رول بزرگ انتخاب شده‌اند؛ مقاومت کششی هر قطعه  $i$  اندازه‌گیری شده است، که  $i = 1, 2, \dots, 6$  می‌باشد. میانگین جامعه  $\mu$  و واریانس جامعه  $\sigma^2$  را می‌توان به ترتیب توسط میانگین نمونه‌ای  $\bar{Y}$  و واریانس نمونه‌ای  $S^2$  تخمین زد. از آنجا که  $\text{Var}(\bar{Y}) = \sigma^2/n$  ، نتیجه می‌شود که  $\text{Var}(\bar{Y})$  را می‌توان با  $S^2/n$  تقریب زد. احتمال اینکه  $\bar{Y}$  در فاصله‌ی  $2S/\sqrt{n}$  از میانگین واقعی جامعه  $\mu$  باشد را پیدا کنید.

$$\begin{aligned} P\left(-\frac{2S}{\sqrt{n}} \leq (\bar{Y} - \mu) \leq \frac{2S}{\sqrt{n}}\right) &= P\left(-2 \leq \sqrt{n}\left(\frac{\bar{Y} - \mu}{S}\right) \leq 2\right) \\ &= P(-2 \leq T_{n-1} \leq 2) \\ &= 0.8981 \simeq 0.90 \end{aligned}$$

### قضیه:

$$\left. \begin{array}{l} U \sim \chi_{n_1}^2 \\ V \sim \chi_{n_2}^2 \\ U \perp\!\!\!\perp V \end{array} \right\} \Rightarrow \frac{U/n_1}{V/n_2} \sim F_{n_1, n_2}$$

$$\left. \begin{array}{l} X \sim N(\mu_1, \sigma_1^2) \rightarrow U = \frac{\sum(X_i - \bar{X})^2}{\sigma_1^2} \sim \chi_{n_1-1}^2 \\ Y \sim N(\mu_2, \sigma_2^2) \rightarrow V = \frac{\sum(Y_i - \bar{Y})^2}{\sigma_2^2} \sim \chi_{n_2-1}^2 \end{array} \right\} \Rightarrow \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$$

$$\sigma_1 = \sigma_2 = \sigma \rightarrow \frac{S_1^2}{S_2^2} = \frac{\sum_{n_1=1}^{n_1-1} (X_i - \bar{X})^2}{\sum_{n_2=1}^{n_2-1} (Y_i - \bar{Y})^2} = F_{n_1, n_2}$$

که در آن توزیع  $F$  با درجات آزادی صورت  $n_1$  و مخرج  $n_2$  می‌باشد. این توزیع به شکل زیر تعریف می‌شود.

$$X \sim F_{n_1, n_2} : f_X(x) = \frac{\Gamma(\frac{n_1+n_2}{2}) n_1^{(\frac{n_1}{2})} n_2^{(\frac{n_2}{2})}}{\Gamma(\frac{n_1}{2}) \Gamma(\frac{n_2}{2})} \frac{x^{\frac{n_1}{2}-1}}{(n_1 + n_2 x)^{\frac{n_1+n_2}{2}}} \quad x > 0$$

$$E(X) = \frac{n_2}{n_2 - 2} \quad n_2 > 2, \quad V(X) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)(n_2 - 4)} \quad n > 4$$

$$\Gamma(n) = (n - 1)!$$

### مثال

فرض کنید یک مدیر کنترل کیفیت می‌خواهد نسبت واریانس‌های وزن محصولات تولید شده توسط دو دستگاه مختلف، یعنی دستگاه A و B، را برآورد کند. اطلاعات زیر برای این منظور جمع‌آوری شده است:

- دستگاه A با  $n_1 = 16$  نمونه و واریانس نمونه‌ای 4.0
- دستگاه B با  $n_2 = 21$  نمونه و واریانس نمونه‌ای 2.5

مدیر می‌خواهد یک بازه اطمینان 95% برای نسبت واریانس‌های جامعه، یعنی  $\sigma_1^2 / \sigma_2^2$ ، محاسبه کند.

با توجه به قضیه بالا

$$S_{12} = (S_1 / S_2)^2 = 4 / 2.5 = 1.6, \quad \sigma_{12} = (\sigma_1 / \sigma_2)^2, \quad S_{12} / \sigma_{12} \sim F_{\nu_1, \nu_2}$$

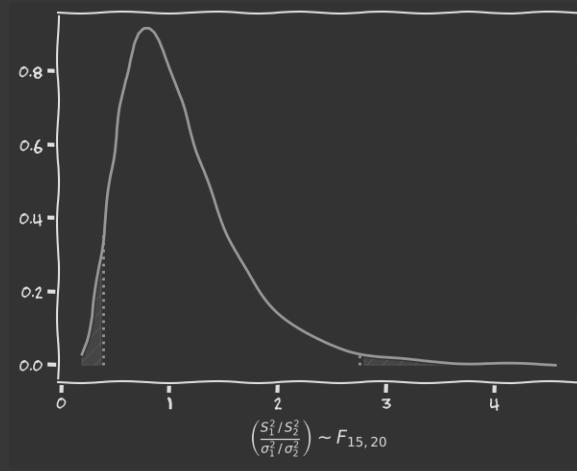
از آنجایی که سطح اطمینان 95% است،  $\alpha = 0.05$  و نیز درجه آزادی برای دستگاه A و B به ترتیب 15 و  $\nu_2 = n_2 - 1 = 21 - 1 = 20$  می‌باشد. با جایگذاری این اعداد و استفاده از جداول توزیع F برای چندک‌های بالا و پایین بدست می‌آوریم:

$$F_{\alpha/2, \nu_1, \nu_2} = F_{0.025, 15, 20} \approx 0.36, \quad F_{1-\alpha/2, \nu_1, \nu_2} = F_{0.975, 15, 20} \approx 2.57$$

پس برای بازه اطمینان 95% داریم:

$$P(F_{\alpha/2, \nu_1, \nu_2} \leq S_{12} / \sigma_{12} \leq F_{1-\alpha/2, \nu_1, \nu_2}) = 1 - \alpha$$

$$P(0.36 \leq 1.6 / \sigma_{12} \leq 2.57) = 0.95$$



$$P(1.6/2.57 \leq \sigma_{12} \leq 1.6/0.36) = 0.95$$

$$P\left(0.6218 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 4.4094\right) = 0.95$$

## ۷ برآورد نقطه‌ای

برآورده: به تابعی از نمونه که برای برآورد پارامتری از جامعه به کار می‌رود برآورده‌گر آن پارامتر گوییم. مقدار عددی حاصل برای برآورده، برآورده نامیده می‌شود. برآورده‌گر مقدار تصادفی است اما مرسوم است برای هر دو از یک نماد استفاده شود. مثلا:

$$\hat{\mu} = \bar{X}, \quad \hat{\mu} = \bar{x}$$

مثال: تعداد حوادث در یک بازه‌ی زمانی از توزیع پواسون پیروی می‌کند.

$$P(X = x) = \frac{e^{-\theta} \theta^x}{x!}$$

که در آن  $\theta$  پارامتر توزیع است. فرض کنید هدف برآورد تعداد حوادث در یک هفته‌ی خاص است.

$$g(\theta) = P(X \leq 1) = e^{-\theta} + \theta e^{-\theta} = (1 + \theta)e^{-\theta}$$

به این منظور باید برآورده‌گر مناسبی پیشنهاد داد. بطور مثال:

$$\widehat{g(\theta)} = (1 + \hat{\theta})e^{-\hat{\theta}}$$

که در آن باید برآورد  $\theta$  را محاسبه نمود که از یکی از دو روش زیر بطور مثال می‌تواند محاسبه شود.

$$\hat{\theta}_1 = \bar{X}, \quad \hat{\theta}_2 = median(X_1, X_2, \dots, X_n)$$

در روش دوم می‌توان از برآورده‌گر زیر استفاده نمود:

$$\widehat{g(\theta)} = \widehat{P}(X \leq 1) = \frac{\text{week per observed event one or zero of number}}{\text{week per observation event total of number}}$$

فرض کنید در ۲۰ هفته مشاهدات تعداد حوادث در هر هفته به شرح زیر باشد:

$$N = \{2, 3, 2, 0, 0, 1, 2, 3, 2, 1, 1, 1, 0, 2, 5, 1, 1, 0, 2, 1\}$$

$$\hat{g}_1 = 0.56, \quad \hat{g}_2 = 0.74, \quad \hat{g}_3 = 0.55$$

پس علاوه بر پیدا کردن برآوردهای مختلف، مساله بعد ارزیابی و انتخاب بهترین آنها است.

## ۱.۷ روش گشتاوری

در این روش پارامتر را برحسب گشتاور اول توزیع بیان می‌کنیم. اگر دو پارامتر داشته باشیم از گشتاور دوم نیز استفاده می‌کنیم و به همین ترتیب اگر تعداد پارامترها افزایش یابند از گشتاورهای بالاتر نیز استفاده خواهد شد.  
بطور مثال فرض کنید  $X_1, X_2, \dots, X_n$  نمونه‌ای تصادفی از جامعه‌ای با توزیع نرمال  $N(\mu, \sigma^2)$  باشد. برآوردهای گشتاوری  $\mu$  و  $\sigma$  را بیابید.

$$E(X) = \mu, \quad V(X) = E(X^2) - E(X)^2 = \sigma^2$$

پس برآوردهای مدنظر را به شکل زیر می‌توان نوشت.

$$\begin{aligned} \hat{\mu} &= \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \\ \hat{\sigma} &= \sqrt{\bar{X}^2 - (\bar{X})^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

## ۲.۷ روش بیشینه درستنمایی

متغیرهای تصادفی  $X_1, X_2, \dots, X_n$  با توزیع جرمی (چگالی)  $f(x_1, \dots, x_n; \theta)$  را در نظر بگیرید. تابع  $L(\theta; \vec{x}) = f(x_1, \dots, x_n; \theta)$  که تابع پارامتر توزیع است را تابع درستنمایی (likelihood) می‌نامند. بردار  $\vec{x}$  بیانگر مقادیر مشاهده شده نمونه  $(x_1, \dots, x_n)$  می‌باشد. با فرض هم‌توزیع و استقلال متغیرهای تصادفی  $X_i$  تابع جرم (چگالی) مشترک برابر حاصلضرب توابع توزیع متغیرهای تنها است.

$$L(\theta; \vec{x}) = \prod_{i=1}^n f(x_i; \theta)$$

هرچه  $L(\theta; \vec{x})$  بزرگ‌تر باشد، شناس مشاهده‌ی مقدار  $(x_1, \dots, x_n) = \vec{x}$  بیشتر است و البته مقدار  $L(\theta; \vec{x})$  به پارامتر مجھول  $\theta$  بستگی دارد. پس کاملاً منطقی است که برآورد خوب  $\theta$ ، مربوط به مقداری از فضای پارامترها باشد که تابع درستنمایی را بیشینه کند. به عبارتی برآورده مشخص می‌کند که بازای چه مقدار  $\theta$  آنچه رخداده است محتمل‌ترین حالت بوده است.  
بر اساس نمونه تصادفی تایی  $n$  از جامعه‌ای با توزیع نمایی برآورد  $\theta$  را بیابید.

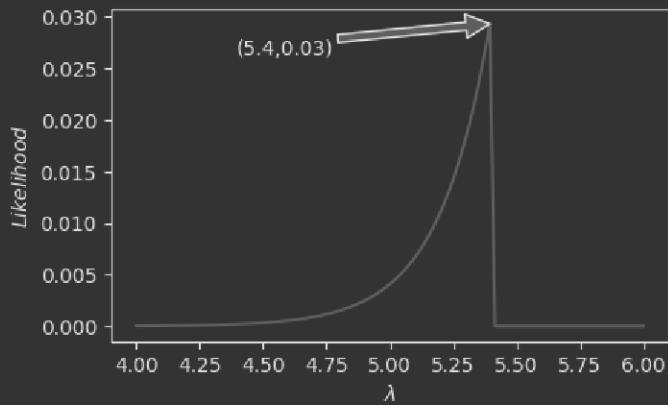
$$\begin{aligned} L(\theta; \vec{x}) &= \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i}, \quad x_i > 0 \\ \ln L(\theta; \vec{x}) &= n \ln \theta - \theta \sum_{i=1}^n x_i \\ \frac{\partial \ln L(\theta; \vec{x})}{\partial \theta} &= 0 \quad \rightarrow \quad \hat{\theta} = \frac{1}{\bar{x}} \end{aligned}$$

مثال: از جامعه با توزیع نمایی قطع شده زیر، نمونه  $n = 5$  تایی به تصادف گرفته شده است. پارامتر  $\lambda$  را برآورد کنید.

$$\begin{aligned} f(x; \lambda) &= e^{-(x-\lambda)}, \quad x > \lambda \\ x &= \{5.5, 6.7, 7.2, 5.7, 5.4\} \end{aligned}$$

$$\begin{aligned} E(X) &= \int_{\lambda}^{\infty} x f(x; \lambda) dx = \int_{\lambda}^{\infty} x e^{-(x-\lambda)} dx = \lambda + 1 \\ \lambda &= E(X) - 1, \quad \rightarrow \quad \hat{\lambda} = \bar{x} - 1 = 6.1 - 1 = 5.1 \end{aligned}$$

$$L(\lambda; \vec{x}) = \begin{cases} e^{-\sum_{i=1}^n x_i + n\lambda} & \lambda \leq x_m \\ 0 & \lambda > x_m \end{cases} \quad x_m = \min(x_1, \dots, x_n), \quad \hat{\lambda} = 5.4$$



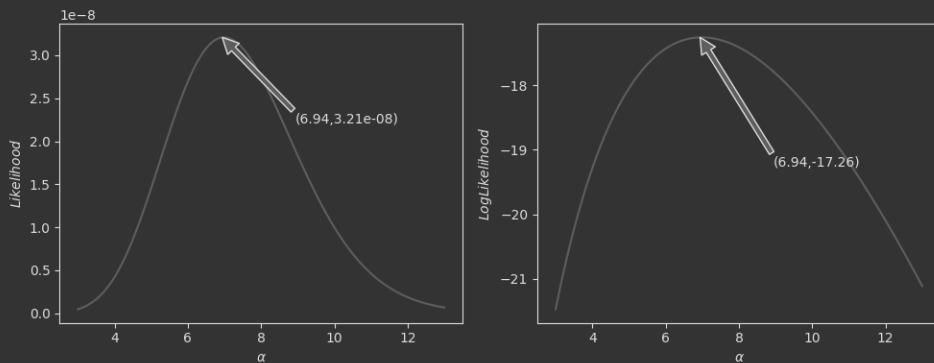
مثال: جامعه‌ای از مدل وایبل  $W(\alpha, \beta)$  با مقدار پارامتر  $\beta = 0.15$  پیروی می‌کند. برآورده  $\hat{\alpha}$  را با روش بیشینه‌ی درستنمایی بیابید. سپس با استفاده از نمونه‌های زیر مقدار درست  $\alpha$  را برآورد کنید.

$$x = \{4.1, 5.5, 5.7, 4.8, 4.9, 6.0, 6.3, 5.3, 5.9, 6.7, 5.7, 6.8\}$$

$$L(\alpha; \vec{x}) = \prod_{i=1}^{12} 0.15\alpha(0.15x_i)^{\alpha-1}e^{-(0.15x_i)^\alpha}$$

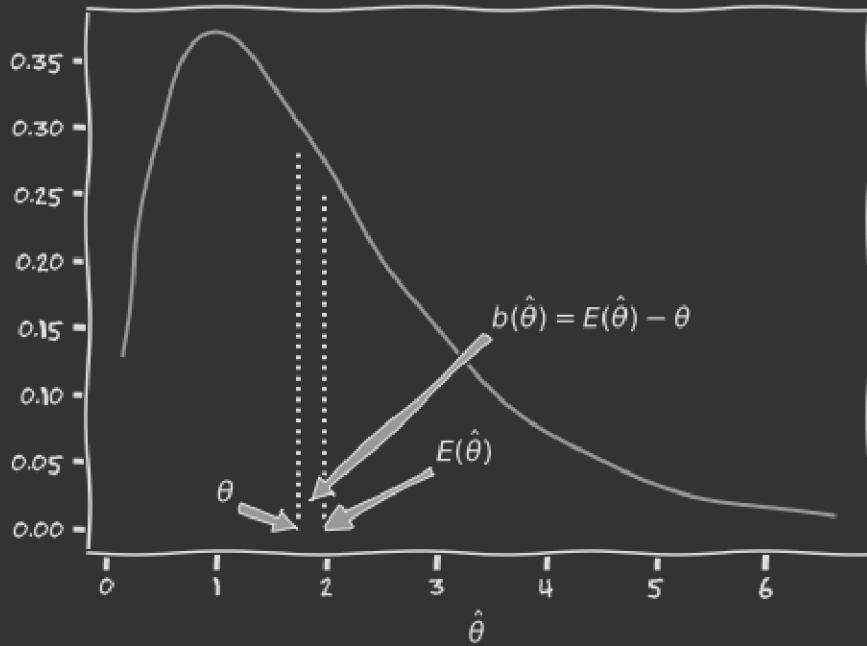
$$\ln L(\alpha; \vec{x}) = 12 \ln(0.15\alpha) + (\alpha - 1) \sum_{i=1}^{12} \ln(0.15x_i) - \sum_{i=1}^{12} (0.15x_i)^\alpha$$

برای پیدا کردن بیشینه، مشتق گرفتن کار ساده‌ای اینجا نیست. بهتر است با رسم منحنی بیشینه را پیدا کنیم.



### ۳.۷ ارزیابی برآوردها - نااریب

مقدار اریب برآورده  $\hat{\theta}$  به شکل  $b(\hat{\theta}) = E(\hat{\theta}) - \theta$  تعریف می‌کنیم و اگر این مقدار صفر باشد، برآورده را نااریب می‌نامیم.



مثال: نمونه‌هایی از جامعه پواسونی در نظر بگیرید و برآورده نااریب برای  $\lambda^2$  بیابید. می‌دانیم که  $T = \bar{X}$  برآورده نااریب برای  $\lambda$  در توزیع پواسون است.

$$\begin{aligned} E(T^2) &= E(\bar{X}^2) = V(\bar{X}) + E(\bar{X}^2) = \frac{\lambda}{n} + \lambda^2, \\ &\rightarrow E(\bar{X}^2) - \frac{\lambda}{n} = \lambda^2 \\ E(\bar{X}^2) - \frac{E(\bar{X})}{n} &= E(\bar{X}^2 - \frac{\bar{X}}{n}) = \lambda^2, \quad \hat{\lambda}^2 = \bar{X}^2 - \frac{\bar{X}}{n} \end{aligned}$$

مثال: نمونه تصادفی از جامعه‌ای دلخواه با میانگین  $\mu$  و واریانس  $\sigma^2$  را در نظر بگیرید و برآورده نااریب برای  $\sigma^2$  بیابید.

$$E(X_i) = \mu, \quad V(X_i) = \sigma^2, \quad E(X_i^2) = \mu^2 + \sigma^2$$

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \sigma^2/n, \quad E(\bar{X}^2) = \mu^2 + \sigma^2/n, \quad E(\bar{X}^2) = \mu^2 + \sigma^2$$

$$E(\bar{X}^2 - \bar{X}^2) = \sigma^2 - \sigma^2/n = \frac{n-1}{n}\sigma^2$$

$$\begin{aligned} &\rightarrow \frac{n}{n-1}E(\bar{X}^2 - \bar{X}^2) = \frac{n}{n-1}E\left(\frac{\sum(X_i - \bar{X})^2}{n}\right) = \sigma^2, \\ &\rightarrow \hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i^2 - \bar{X})^2 \end{aligned}$$

## ۴.۷ ارزیابی برآوردها - کارایی

میانگین مربع خطای (Mean Square Error - MSE) برآورده  $T$  برای پارامتر  $\theta$  به شکل زیر تعریف می‌شود:

$$\text{MSE}[T] = \mathbb{E}[(T - \theta)^2] = \text{Var}[T] + b[T]^2$$

برآورده  $MSE$  کمتری دارد، کاراتر می‌باشد.

$$\text{MSE}(T_1) \leq \text{MSE}(T_2), \quad \theta \in \Theta$$

دست کم برای یک  $\theta$  اکیدا کوچک باشد.  
برآورده  $T$  برای پارامتر  $\theta$  را در نظر بگیرید. میانگین مربع خطای MSE برای آن به این شکل تعریف می‌شود که:

$$\text{MSE}(T) = \mathbb{E}[(T - \theta)^2]$$

و نیز با مراحل زیر آن را به دو مولفه‌ی واریانس و اریبی تجزیه کرد.

$$\begin{aligned} (T - \theta)^2 &= (T - \mathbb{E}[T] + \mathbb{E}[T] - \theta)^2 \\ &= (T - \mathbb{E}[T])^2 + 2(T - \mathbb{E}[T])(\mathbb{E}[T] - \theta) + (\mathbb{E}[T] - \theta)^2 \\ \mathbb{E}[(T - \theta)^2] &= \underbrace{\mathbb{E}[(T - \mathbb{E}[T])^2]}_{\text{Var}[T]} + \underbrace{2\mathbb{E}[(T - \mathbb{E}[T])(\mathbb{E}[T] - \theta)]}_{2(\mathbb{E}[T] - \theta)\mathbb{E}[T - \mathbb{E}[T]]} + \underbrace{\mathbb{E}[(\mathbb{E}[T] - \theta)^2]}_{\text{b}[T]^2} \end{aligned}$$

## ۵.۷ روش بوتاسترپ (خودگردان)

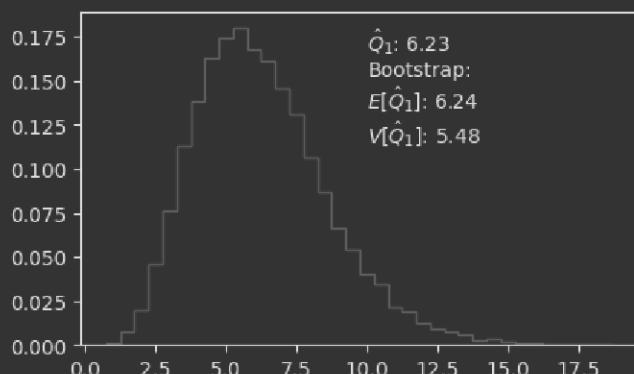
در ارزیابی برآورده‌ها باید برآوری از واریانس برآورده داشته باشیم. محاسبه واریانس برآورده مستلزم دانستن توزیع برآورده است که در بسیاری از موارد نامشخص است و این موارد عبارتند از  
 ۱ - حالت‌هایی که اساساً توزیع جامعه مجھول و یافتن توزیع برآورده ممکن نیست (حالت ناپارامتری)  
 ۲ - و یا توزیع جامعه معلوم ولی توزیع برآورده بسیار پیچیده است و به روش تحلیلی قابل محاسبه نیست (حالت پارامتری)  
 در حالت ناپارامتری از نمونه‌های محدود به روش جایگشتی همراه با جایگذاری نمونه‌های زیاد درست و سپس واریانس برآورده را محاسبه می‌کنیم و در حالت پارامتری که توزیع جامعه را داریم با استفاده از توزیع تعداد زیاد داده‌ی مصنوعی تولید و با آن‌ها واریانس برآورده را محاسبه می‌نماییم.

تعداد ۷ نمونه از یک جامعه به تصادف گرفته شده است و پارامتر  $t$  در این نمونه‌ها اندازه‌گیری شده است. چارک اول این پارامتر در جامعه را برآورد کنید و سپس با فرض این که (الف) بدانید جامعه از توزیع نمایی پیروی می‌کند (ب) توزیع جامعه نامشخص باشد، اریبی و میانگین مربع خطای برآورده خود را محاسبه نمایید.

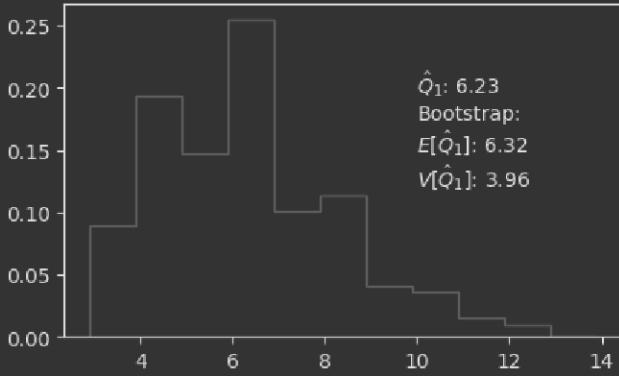
$$x = \{15.63, 63.09, 25.14, 9.17, 17.10, 11.14, 10.28\}$$

(الف)

$$\begin{aligned} f(x; \lambda) &= (1/\lambda)e^{-x/\lambda} \quad \rightarrow \quad \text{ML}(\lambda) : \quad \hat{\lambda} = \bar{x} \\ \int_0^{Q_1} f(x; \lambda) dx &= 0.25 \quad \rightarrow \quad 1 - e^{-Q_1/\lambda} = 0.25 \quad \rightarrow \quad Q_1 = -\ln(0.75)\lambda \\ \hat{Q}_1 &= -\ln(0.75)\hat{\lambda} \end{aligned}$$



(ب) توزیع را نمی‌شناسیم اما به طریق برآورده را داریم. حالا تعداد زیاد جایگشت همراه با جایگذاری انتخاب می‌کنیم (choice)

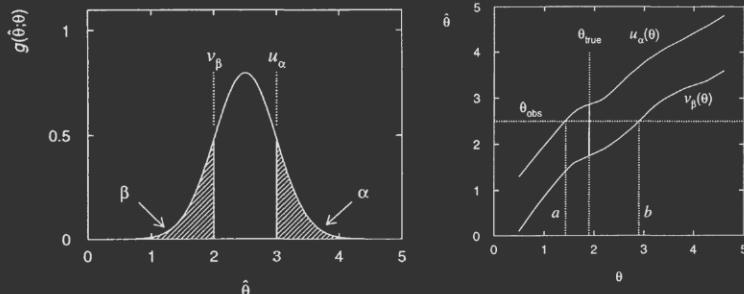


## ۸ براورد بازه‌ای یا بازه‌ی اطمینان

فرض کنید  $X_i$  نمونه‌ای تصادفی از جامعه‌ای با تابع جرمی یا چگالی احتمال  $f(x; \theta)$  باشد. گوییم بازه‌ی  $(L(X), U(X))$  یک بازه‌ی اطمینان در سطح  $\alpha - 1$  برای پارامتر  $\theta$  است اگر

$$P(L(X) < \theta < U(X)) = 1 - \alpha$$

که در آن  $L$  و  $U$  دو متغیر تصادفی هستند پس بازه‌ی اطمینان نیز یک بازه‌ی تصادفی است و برسی نمونه تغییر می‌کند. بعد از نمونه‌گیری و محاسبه  $L(x)$  و  $U(x)$  می‌توان ادعا کرد که با اطمینان  $\alpha - 1$  بازه‌ی  $(L, U)$  شامل مقدار واقعی  $\theta$  است. به این معنی است که اگر به دفعات زیاد نمونه‌های تصادفی تایب  $n$  از جامعه تحت بررسی مشاهده کنیم و هر بار بازه اطمینان  $(L(X), U(X))$  را بسازیم، آنگاه در  $100\%(\alpha - 1)$  دفعات این بازه‌ها در برگیرنده  $\theta$  می‌باشند.



تا زمانی که نمونه‌گیری انجام نشده است از کلمه احتمال می‌توان استفاده کرد و گفت  $\theta$  با احتمال  $\alpha - 1$  در بازه‌ی اطمینان قرار دارد اما بعد از نمونه‌گیری بازه دیگر متغیر تصادفی نیست و بنابر این احتمال برای آن درست نیست لذا از کلمه اطمینان استفاده می‌کنیم و می‌گوییم اندازه  $\alpha - 1$  اطمینان داریم که این بازه مشخص شده شامل  $\theta$  می‌شود. دقیت کنید که ما فقط یک بازه محاسبه کردیم اما اگر این محاسبه را می‌توانستیم تکرار کنیم در  $100\%(\alpha - 1)$  دفعات  $\theta$  واقعی در این بازه محاسبه شده قرار می‌گرفت پس اکنون هم به این بازه‌ای که محاسبه کردیم به همین میزان اطمینان داریم. معمولاً  $\alpha$  یکی از مقادیر  $0.01, 0.05, 0.1$  در نظر گرفته می‌شود.

### ۱.۸ بازه‌ی اطمینان برای میانگین جامعه

(الف) فرض می‌کنیم توزیع جامعه نرمال و  $\sigma$  (انحراف معیار جامعه) معلوم است. بازه‌ی اطمینان در سطح  $\alpha - 1$  برای  $\mu$  بازه‌ی زیر است.

$$\left( \bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

$$\rightarrow P \left( \bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

(ب) همچنان فرض می‌کنیم توزیع جامعه نرمال و  $n$  نمونه تصادفی انتخاب می‌کنیم. اما  $\sigma$  (انحراف معیار جامعه) و  $\mu$  (میانگین جامعه) مجھول باشند. آنگاه بازه‌ی اطمینان در سطح  $\alpha - 1$  برای  $\mu$  از رابطه‌ی زیر بدست می‌آید.

$$\left(\bar{X} - t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}\right)$$

$$\rightarrow P\left(\bar{X} - t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

که در آن  $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$  انحراف معیار نمونه است.

مثال: تعداد ۱۰ نمونه از یک جامعه با توزیع نرمال گرفته و در زیر آورده شده است. بازه‌ی اطمینان در سطح ۹۰ درصد برای میانگین این جامعه را محاسبه نمایید.

$$X = \{3.74, 3.16, 2.18, 4.47, 3.92, 2.94, 3.6, 4.97, 3.41, 1.7\}$$

برای تعیین بازه اطمینان یک پارامتر جمعیت، نیاز داریم تا توزیع برآورده‌گری از جمعیت را برحسب پارامتر مدنظر خود بدانیم. در این مساله از آنجاییکه خود جمعیت توزیع نرمال دارد، با استفاده از قضیه ۶.۶ می‌توانیم برآورده‌گر جدید  $\hat{\theta}$  را به شکل زیر تعریف کنیم که وابسته به پارامتر مد نظرمان یعنی  $\mu$  و همچنین برآورده‌گرهای میانگین نمونه  $\bar{X}$  و واریانس نمونه  $S^2$  می‌باشد و مهم آنکه می‌دانیم این برآورده‌گر جدید از توزیع  $t$  با درجه آزادی  $1 - n$  پیروی می‌کند.

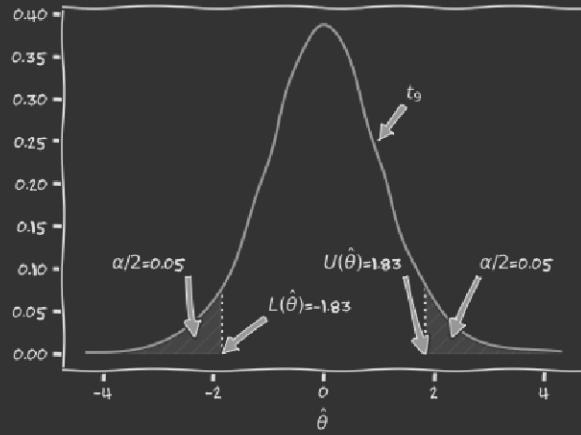
$$\hat{\theta} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$\bar{X} = 3.41, \quad S = 0.93,$$

$$P(-1.83 < \hat{\theta} < 1.83) = 0.90, \quad P(-1.83 < \frac{3.41 - \mu}{0.93/\sqrt{10}} < 1.83)$$

$$= P(2.87 < \mu < 3.95) = 0.90$$

یعنی اگر ۱۰۰ دفعه این آزمایش را تکرار کنیم و نمونه‌های ۱۰۰ اتایی از این جمعیت بگیریم و هر دفعه حد بالا و پایین را برای میانگین جمعیت محاسبه کنیم، در ۹۰ دفعه میانگین واقعی جمعیت در بازه‌ی پیشنهادی قرار خواهد داشت و تنها در ۱۰ دفعه بازه‌ی اشتباه پیشنهاد خواهیم کرد. پس می‌توانیم بگوییم که با ۹۰ درصد اطمینان میانگین جمعیت بین ۲.۸۷ و ۳.۹۵ می‌باشد.



(پ) توزیع جامعه نیز نامعلوم باشد. در این حالت اگر اندازه‌ی نمونه تصادف به اندازه کافی بزرگ باشد، متغیر تصادف  $\frac{\bar{X} - \mu}{S/\sqrt{n}}$  تقریباً از توزیع نرمال استاندارد پیروی می‌کند.

$$\lim_{n \rightarrow \infty} \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1)$$

پس برای بازه‌ی اطمینان می‌توان نوشت:

$$\left(\bar{X} - z_{1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{S}{\sqrt{n}}\right)$$

$$\rightarrow P\left(\bar{X} - z_{1-\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

اگر واریانس جامعه آزمایش‌های قبلی برآورده شده باشد  $\hat{\sigma}^2$  و بخواهیم حداقل خطا  $h$  باشد تعداد نمونه از رابطه‌ی زیر بدست می‌آید.

$$n = \left( \frac{z_{1-\alpha/2} \hat{\sigma}}{h} \right)^2$$

## ۲.۸ بازه اطمینان برای نسبت در جامعه

نمونه‌های تصادفی به اندازه کافی بزرگ از جامعه‌ی صفر و یک‌ها در نظر بگیرید. این جامعه از توزیع برنولی پیروی می‌کند ( $b(p)$ ). از آنجاییکه تعداد نمونه‌ها بسیار زیاد است، بنابر قضیه حد مرکزی، میانگین نمونه‌ها از توزیع نرمال پیروی می‌کند ( $\bar{X} \sim N(p, p(1-p))$ ). بازه‌ی اطمینان تقریبی در سطح  $\alpha - 1$  برای  $p$  نسبت یک‌ها (موفق‌ها) در جامعه از رابطه زیر بدست می‌آید.

$$\hat{p} = \bar{X}$$

$$\left( \hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

## ۳.۸ بازه اطمینان برای انحراف معیار جامعه

اگر نمونه‌های  $X$  از جامعه‌ای با توزیع نرمال  $N(\mu, \sigma^2)$  باشند، متغیر تصادفی  $S^2/\sigma^2$  از توزیع مرربع کای با درجه‌ی آزادی  $1 - n$  پیروی می‌کند. پس بازه‌ی اطمینان برای واریانس جامعه در سطح  $\alpha - 1$  از رابطه‌ی زیر بدست می‌آید.

$$X \sim N(\mu, \sigma^2) \rightarrow \frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\rightarrow P\left(\chi^2_{n-1, \alpha/2} < \frac{(N-1)S^2}{\sigma^2} < \chi^2_{n-1, 1-\alpha/2}\right) = 1 - \alpha$$

$$\frac{(N-1)S^2}{\chi^2_{n-1, 1-\alpha/2}} < \sigma^2 < \frac{(N-1)S^2}{\chi^2_{n-1, \alpha/2}}$$

## ۴.۸ پیش‌بینی بازه‌ای

بر اساس نمونه‌ای از مشاهدات مقدار مشاهده‌ای جدید را برآورد کنیم. محاسبات را محدود به توزیع نرمال می‌کنیم.  
(الف) هر دو مقدار  $\mu$  و  $\sigma^2$  معلوم باشند. یک بازه‌ی پیش‌بینی با اطمینان  $\alpha - 1$  برای  $X$  به شکل زیر محاسبه می‌شود.

$$X \sim N(\mu, \sigma^2), \quad Z = \frac{X - \mu}{\sigma} \sim N(0, 1),$$

$$P\left(\mu - z_{1-\alpha/2}\sigma < X < \mu + z_{1-\alpha/2}\sigma\right) = 1 - \alpha$$

(ب) مقدار  $\sigma^2$  معلوم باشد اما  $\mu$  مجھول باشد.

$$E(X - \bar{X}) = \mu - \mu = 0, \quad V(X - \bar{X}) = V(X) + V(\bar{X}) = \sigma^2 + \sigma^2/n = \sigma^2(1 + 1/n)$$

ترکیب خطی از متغیرهای نرمال مستقل توزیع نرمال دارند.

$$(X - \bar{X}) \sim N(0, \sigma^2(1 + 1/n)), \quad \frac{X - \bar{X}}{\sqrt{\sigma^2(1 + 1/n)}} \sim N(0, 1),$$

$$P\left(\bar{X} - z_{1-\alpha/2}\sqrt{\sigma^2(1 + 1/n)} < X < \bar{X} + z_{1-\alpha/2}\sqrt{\sigma^2(1 + 1/n)}\right) = 1 - \alpha$$

(پ) هر دو مقدار  $\mu$  و  $\sigma^2$  مجھول باشند.

$$\frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1},$$

$$\frac{X - \bar{X}}{\sqrt{S^2(1 + 1/n)}} \sim t_{n-1}$$

$$P\left(\bar{X} - t_{n-1, 1-\alpha/2}\sqrt{S^2(1 + 1/n)} < X < \bar{X} + t_{n-1, 1+\alpha/2}\sqrt{S^2(1 + 1/n)}\right) = 1 - \alpha$$

که در آن  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  واریانس نمونه و مستقل از میانگین نمونه  $\bar{X}$  است.

---

## ۵.۸ روش خودگردان (Bootstrap)

ابتدا با استفاده از توزیع جامعه و یا جایگذاری پیاپی  $B$  دفعه نمونه‌ی  $n$  تابی تولید می‌کنیم و در هر دفعه پارامتر مدنظر  $\theta$  را برآورده می‌کنیم.

$$\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$$

که در هر یک برآورده مربوطه از روی نمونه‌های تولید شده محاسبه شده‌اند.

$$\hat{\theta}^* = T(x_1^*, x_2^*, \dots, x_n^*)$$

برای محاسبه‌ی بازه اطمینان می‌توانیم از صدک‌های مربوطه استفاده کنیم. برای این منظور برآورده‌گرهای مربوطه را از کوچک به بزرگ مرتب می‌کنیم. سپس حد بالا و پایین برابر خواهد بود.

$$\hat{\theta}_L = \hat{\theta}_{(\alpha/2B^{th})}^*, \quad \hat{\theta}_U = \hat{\theta}_{((1-\alpha/2)B^{th})}^*$$

بطور مثال اگر 1000 نمونه‌ی  $n$  تابی تولید کرده باشیم و  $1 - \alpha = 0.95$  باشد. بعد از مرتب کردن 1000 تا  $\hat{\theta}^*$  از کوچک به بزرگ، بازه‌ی اطمینان 0.95 بین  $\hat{\theta}_{25}^*$  و  $\hat{\theta}_{75}^*$  می‌باشد.

## ۹ آزمون فرضیه - ۱

فرضیه آماری ادعایی است درباره توزیع احتمال یک و یا چند متغیر تصادفی. (عبارتی است درباره پارامترهای یک یا چند جامعه). فرضیه‌ی اصلی را فرضیه صفر  $H_0$  می‌نامیم و فرضیه‌های جایگزین (alternative) را با  $H_a$  یا گاهی  $H_1$  نمایش می‌دهیم.

آزمون فرضیه آماری شیوه‌ای که بر پایه مقادیری از متغیرهای تصادفی تصمیم به پذیرش یا رد فرضیه‌ی  $H_0$  می‌گیریم. این تصمیم با تعریف یک آماره‌ی آزمون (Test Statistic) که تابعی از متغیرهای تصادفی می‌باشد صورت می‌گیرد. یادآوری می‌کنیم که آماره‌ها دارای توزیع بوده و از مشاهدات نمونه‌ای مقدار آن‌ها مشخص می‌شود.

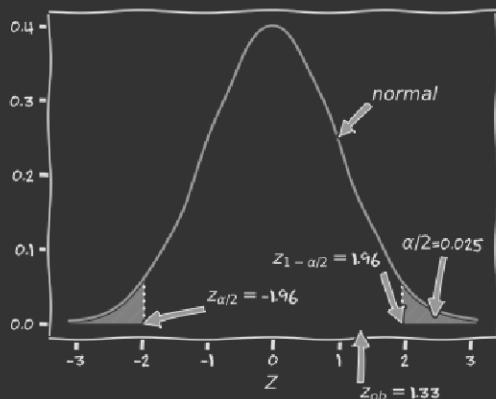
مثال: میدانیم که ۱۰ نمونه‌ی زیر از یک جامعه با توزیع نرمال که انحراف معیار  $2.0 = \sigma$  دارد، بطور تصادفی انتخاب شده‌اند. ادعا می‌شود که میانگین این جامعه  $3.0 = \mu$  می‌باشد.

$$X = \{5.35, 2.96, 3.58, 4.24, 2.18, 3.2, 2.63, 4.94, 4.9, 4.42\}$$

ابتدا باید آماره‌ی مناسبی برای بررسی صحت ادعا طراحی کنیم. این آماره باید ۲ شرط اساسی داشته باشد. ۱ - شامل پارامتر مورد ادعا باشد و ۲ - توزیع آن برای ما مشخص باشد.

بطور مثال متغیر تصادفی  $Z$  با تعریف زیر تابعی از پارامتر مورد ادعا  $\mu$  می‌باشد و با توجه به قضایای فصل ۶ می‌دانیم توزیع نرمال استاندارد دارد.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{2.0/\sqrt{10}} = 1.58(\bar{X} - \mu)$$



مقدار  $Z$  کاملاً تصادف و بطور مثال با احتمال ۹۵٪ می‌تواند عددی بین  $-1.96$  تا  $+1.96$  مشاهده شود. پس اگر آماره  $Z$  مشاهده شده را محاسبه کنیم و در این بازه قرار گیرد کاملاً نتایج طبیعی و نرمال می‌باشد. اما اگر آماره مشاهده شده خارج این بازه باشد، می‌توانیم بگوییم که با اطمینان ۹۵٪ این فرضیه برای میانگین جامعه  $\mu = 3$  رد می‌شود و درست نمی‌باشد. پس ناحیه‌ای از مقادیر ممکن آماره را به عنوان ناحیه پذیرش و ناحیه‌ای را ناحیه رد یا بحرانی انتخاب می‌کنیم که اگر مقدار مشاهده شده آماره در هر ناحیه قرار بگیرد تصمیم متناسب آن ناحیه را می‌گیریم یعنی فرضیه را قبول و یا رد می‌کنیم. در این مثال ناحیه  $[+1.96, -1.96]$  به عنوان ناحیه پذیرش و خارج آن، ناحیه رد انتخاب می‌شود. و از آنجاییکه مقدار مشاهده شده

$$z_{ob.} = 1.58(3.84 - 3.0) = 1.33$$

است، پس فرضیه  $\mu = 3.0$  از این آزمون موفق خارج می‌شود.

باید دقت کنیم که آماره طراحی شده یک متغیر کاملاً تصادف است و مقدار مشاهده شده می‌توانست در ناحیه پذیرش قرار گیرد در حالی که فرضیه درست نباشد و بالعکس مقدار مشاهده شده در ناحیه رد (بحرانی) بدست آید در حالیکه فرضیه درست باشد. پس در آزمون فرضیه‌ها با توجه به ماهیت تصادفی آماره، همواره می‌توانیم مرتكب خطأ شویم که این خطأ از دو نوع می‌تواند باشد.

خطای نوع اول: رد فرضیه صفر  $H_0$  زمانی که درست است.

خطای نوع دوم: عدم رد فرضیه صفر  $H_0$  زمانی که نادرست است.

خطای نوع اول معمولاً کم اثیرتر می‌باشد. لذا همیشه سعی می‌کنیم فرضیه  $H_0$  را به نحوی در نظر بگیریم که خطای نوع اول خطای جدی‌تری باشد تا اثر کمتری از اشتباہ خود متحمل شویم.

نکته: در آمار رد کردن یک فرضیه با قاطعیت است اما در پذیرفتن آن قاطعیت وجود ندارد.

Thinking of the rejection of  $H_0$  is a strong conclusion, and its acceptance is a weak one.

احتمال رخ دادن خطای نوع اول را سطح معنی‌دار (significance level) می‌نامیم و با  $\alpha$  نشان می‌دهیم..

$$\alpha(\theta) = P(\text{Type I error}) = P(\text{Reject } H_0 \text{ when } H_0 \text{ is true})$$

احتمال رخ دادن خطای نوع دوم را با  $\beta$  نمایش می‌دهیم.

$$\beta(\theta) = P(\text{Type II error}) = P(\text{Fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

خطاهای تابعی از پارامترهای مورد ادعای فرضیه می‌باشند.

تابع توان: نشان دهنده‌ی توانایی آزمون در اثبات غلط بودن فرضیه  $H_0$  است، وقتی واقعاً  $H_0$  غلط باشد.

$$\pi(\theta) = P_\theta(\text{Rej. } H_0) = \begin{cases} \alpha(\theta) & \theta \in \Theta_0 \\ 1 - \beta(\theta) & \theta \in \Theta_a \end{cases}$$

در حالت ایده‌آل وقتی فرضیه درست است باید احتمال رد کردن آن صفر باشد و وقتی نادرست است باید احتمال رد کردن آن ۱ باشد.

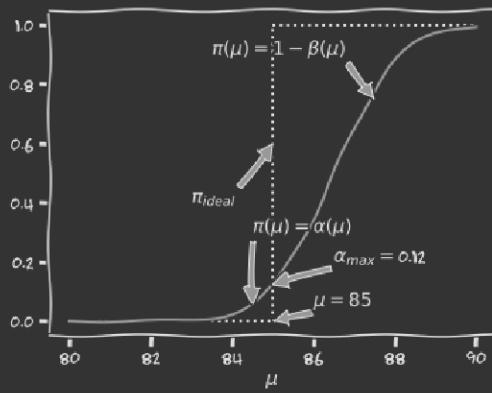
$$\pi_{\text{ideal}}(\theta) = \begin{cases} 0 & \theta \in \Theta_0 \\ 1 & \theta \in \Theta_a \end{cases}$$

تابع مشخصه عملکرد هر آزمون متمم تابع توان آن آزمون می‌باشد.

$$OC(\theta) = P_\theta(\text{Acc. } H_0) = 1 - \pi(\theta)$$

مثال: فرض کنید تعداد  $n = 15$  نمونه از یک جامعه نرمال با انحراف معیار  $\sigma = 5$  انتخاب می‌کنیم. فرضیه صفر ادعا می‌کند میانگین جامعه از 85 بیشتر نیست. برای آزمون این فرضیه از آماره‌ی میانگین نمونه‌ها استفاده می‌کنیم و ناحیه رد فرضیه را  $\bar{X} > 86.5$  انتخاب می‌کنیم. پس برای تابع توان محاسبه می‌کنیم

$$\left\{ \begin{array}{ll} H_0 : & \mu \leq 85 \\ H_1 : & \mu > 85 \end{array} \right. \rightarrow \pi = P(\bar{X} > 86.5) = 1 - P(Z \leq \frac{86.5 - \mu}{5\sqrt{15}})$$



اندازه‌ی آزمون: حداقل احتمال ارتکاب خطای نوع اول اندازه‌ی آزمون نامیده می‌شود.

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(\text{Rej. } H_0 | \text{True } H_0) = \max_{\theta \in \Theta_0} \pi(\theta)$$

در مثال بالا همانطور که در شکل دیده می‌شود، اندازه‌ی آزمون  $\alpha_{max} = 0.12$  است. مقدار-p: کوچکترین مقداری از اندازه‌ی آزمون  $\alpha$  که می‌توان فرضیه  $H_0$  را در آن سطح رد نمود. به عبارتی دیگر احتمال بدست آوردن مشاهداتی معادل یا غریب‌تر از آنچه مشاهده شده است، می‌باشد.

پاسخ: از آنجاییکه تعداد نمونه‌ها زیاد است، می‌توان از آماره‌ی  $Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$  که توزیع نرمال استاندارد دارد استفاده کرد.

$$\left\{ \begin{array}{ll} H_0 : & \mu \leq 2100 \\ H_1 : & \mu > 2100 \end{array} \right.$$

روش مقدار-p:

$$\begin{aligned} p-value &= P_{\mu=2100}(\bar{X} \geq 2139) = P(Z \geq \frac{2139 - 2100}{152/\sqrt{30}}) = P(Z \geq 1.41) \\ &= 1 - P(Z < 1.41) = 1 - 0.9207 = 0.0793 \end{aligned}$$

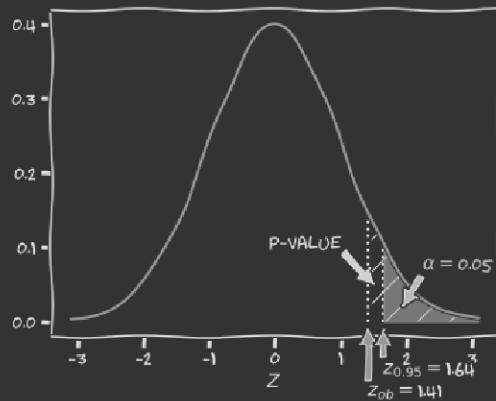
$$p-value > \alpha \rightarrow \text{Retain } H_0$$

روش اندازه‌ی آزمون:

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \rightarrow z_{ob} = \frac{2139 - 2100}{152/\sqrt{30}} = 1.41$$

$$z_{1-\alpha} = z_{0.95} = 1.64$$

$z_{ob} < z_{1-\alpha} \rightarrow \text{Retain } H_0$



مثال: تعداد  $n = 30$  نمونه از یک جامعه به تصادف انتخاب شده است. میانگین و انحراف معیار نمونه‌ها به ترتیب  $\bar{x} = 6.2149$  و  $s = 0.05075$  می‌باشد. ادعا می‌شود میانگین جامعه  $\mu = 6.20$  می‌باشد. این فرضیه را در سطح  $\alpha = 0.05$  بیازمایید.

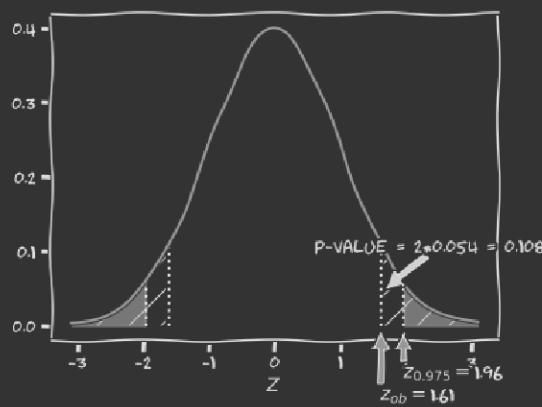
پاسخ: فرضیه‌ها عبارت است

$$\begin{cases} H_0 : \mu = 6.20 \\ H_1 : \mu \neq 6.20 \end{cases}$$

با توجه به اینکه تعداد نمونه‌ها زیاد است، آماره‌ی مطلوب از تبدیل نرمال استاندارد بدست می‌آید.

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \rightarrow z_{ob} = \frac{6.2149 - 6.20}{0.05075\sqrt{30}} = 1.61 \\ \rightarrow p-value = 2(0.054) = 0.108 > \alpha = 0.05 \rightarrow \text{Retain } H_0$$

$$z_{1-\alpha/2} = z_{0.975} = 1.96 > z_{ob} = 1.61 \rightarrow \text{Retain } H_0$$



محاسبه‌ی اندازه‌ی نمونه: همواره با افزایش حجم نمونه، میزان خطا کاهش می‌یابد. البته افزایش تعداد نمونه با افزایش هزینه همراه است و نیز از آنجاییکه خطاهای نوع اول و دوم به هم ارتباط دارند نمی‌توان هر دو را با هم کمینه نمود. اما در صورتیکه اندازه نمونه‌ها مشخص نباشد می‌توان آن را به نحوی انتخاب کرد که خطای نوع اول و دوم با هم کمترین مقدار خود را بگیرند.

$$n = \frac{\sigma^2(|Z_0| + |Z_1|)^2}{(\mu_0 - \mu_1)^2}$$

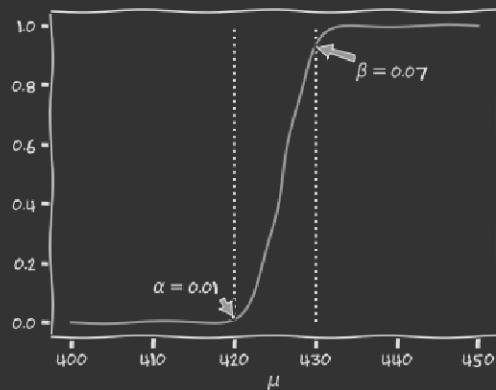
آزمون	$Z_0$	$Z_1$
یک طرفه راست	$Z_{1-\alpha}$	$Z_\beta$
یک طرفه چپ	$Z_\alpha$	$Z_{1-\beta}$
دو طرفه	$Z_{1-\alpha/2}$	$Z_\beta$

پاسخ:

$$\begin{cases} H_0 : \mu \leq 420 \\ H_1 : \mu > 430 \end{cases}$$

$$n = \left( \frac{18(2.33 + 1.28)}{(420 - 430)} \right)^2 \approx 43$$

$$\begin{cases} \frac{\bar{X}-420}{18/\sqrt{n}} = z_{1-\alpha} = z_{0.99} = 2.33 \\ \frac{\bar{X}-430}{18/\sqrt{n}} = z_\beta = z_{0.1} = -1.28 \end{cases} \rightarrow \begin{cases} n = 42.22 \rightarrow n = 43 \\ \bar{X} = 426.45 \end{cases}$$



مثال: تعداد  $n = 10$  نمونه از جامعه‌ای نرمال با انحراف معیار  $\sigma = 0.005$  انتخاب نموده‌ایم. میانگین نمونه‌ها  $\bar{x} = 0.152$  است. در سطح  $\alpha = 5\%$  بررسی کنید که آیا میانگین جامعه  $\mu_0 = 0.15$  می‌باشد یا نه؟ پاسخ: آزمون دو طرفه می‌باشد.

$$\begin{cases} H_0 : \mu = 0.15 \\ H_1 : \mu \neq 0.15 \end{cases}$$

$$\alpha = P(|\bar{x} - \mu_0| > C | H_0 \text{ is True}) = P(|\bar{x} - \mu_0| > C | \mu = \mu_0) = P(|Z| > \frac{\sqrt{n}C}{\sigma})$$

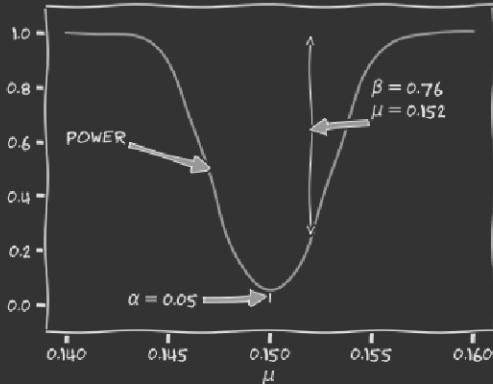
$$\rightarrow \frac{\sqrt{n}C}{\sigma} = z_{1-\alpha/2} = 1.96 \rightarrow C = \frac{1.96(0.005)}{\sqrt{10}} = 0.003$$

$$|\bar{x} - \mu_0| = |0.152 - 0.150| = 0.002 < C = 0.003 \rightarrow \text{Retain } H_0$$

تابع توان:

$$\pi(\mu) = P(\text{Rej } H_0 | \mu \neq \mu_0) = P(\bar{x} < \mu_0 - C) + P(\bar{x} > \mu_0 + C)$$

$$= \Phi\left(\frac{\mu_0 - C - \mu}{\sigma/\sqrt{n}}\right) + 1 - \Phi\left(\frac{\mu_0 + C - \mu}{\sigma/\sqrt{n}}\right) = 1 + \Phi\left(\frac{0.147 - \mu}{0.005/\sqrt{10}}\right) - \Phi\left(\frac{0.153 - \mu}{0.005/\sqrt{10}}\right)$$



از طریق بازه‌ی اطمینان نیز می‌توان مساله را حل نمود. بازه‌ی اطمینان 95% برای میانگین جامعه از رابطه‌ی زیر محاسبه می‌شود.

$$\bar{x} \pm c = 0.152 \pm 0.003 = [0.149, 0.155]$$

از آنجاییکه میانگین جامعه 15 در بازه‌ی اطمینان 95% قرار دارد فرضیه  $H_0$  ابقا می‌شود و نمی‌توانیم آن را رد کنیم.

## ۱۰ آزمون فرضیه - ۲

### ۱۱ آزمون‌های نیکویی برازش و استقلال

#### ۱۱.۱ آزمون نیکویی برازش کای دو

در بحث آزمون فرضیه‌ها، همواره توزیع حاکم بر جامعه مشخص است و تنها فرضیه‌هایی در مورد پارامترهای موجود در توزیع‌ها آزموده می‌شوند. اما در آزمون نیکویی برازش، دیگر توزیع حاکم بر جامعه مشخص نمی‌باشد و اساساً این‌که جامعه از چه توزیعی پیروی می‌کند مورد آزمون قرار می‌گیرد.

$$\left[ \begin{array}{l} H_0: \text{جامعه توزیع} \\ H_1: \text{پیروی نمی‌کند} \end{array} \right. \quad \begin{array}{l} F_0 \text{ دارد} \\ F_0 \text{ جامعه از توزیع} \end{array}$$

مراحل آزمون:

۱ - اگر در توزیع پیشنهادی فرضیه صفر یعنی  $F_0$  پارامتری وجود داشته باشد از طریق بیشینه‌ی درستنایی برآورده شرکت می‌شود.

۲ - اگر داده‌ها پیوسته باشند، آن‌ها را در  $k$  رده تنظیم می‌کنیم. در رده‌بندی داده‌ها دقت می‌کنیم که بنا بر فرض توزیع پیشنهادی فرضیه  $H_0$  فراوانی انتظاری در رده‌ها با هم یکسان باشند و هیچ یک کمتر از 5 نباشند ( $e_i \geq 5$ ). درصورتیکه داده‌ها گسته و شمارشی باشند، فقط نیازی به رده‌بندی نیست.

۳ - فراوانی مشاهدات نمونه‌ای ( $f_i$ ) و نیز فراوانی مورد انتظار ( $e_i$ ) تحت توزیع  $F_0$  را در هر رده بدست می‌آوریم. دقت می‌کنیم که باید:

$$\sum_{i=1}^k f_i = n, \quad \sum_{i=1}^k e_i = n$$

۴ - آماره‌ی این آزمون  $\chi^2$  با تعریف زیر است.

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$$

همانطور که دیده می‌شود هر چه فراوانی مشاهدات با فراوانی مورد انتظار تحت  $H_0$  به هم نزدیک‌تر باشند، شواهد قوی‌تری به نفع می‌باشد. اثبات می‌شود اگر تعداد نمونه‌ها بزرگ باشد، آماره‌ی زیر تقریباً از توزیع کای-دو با درجه آزادی  $1 - r - k = r$  پیروی می‌کند که  $r$  تعداد پارامترهایی است که در ابتدا با روش بیشینه درستنایی درون نمونه‌ها محاسبه کردیم. مقدار مشاهده شده‌ی آماره آزمون  $\chi^2_{ob}$  را در نمونه‌ها محاسبه می‌کنیم.

۵ - با توجه به ناحیه بحرانی برای رد فرضیه  $H_0$  در اندازه‌ی خطای  $\alpha$ :

$$\chi_{ob}^2 > \chi_{\nu, 1-\alpha}^2$$

در مورد رد یا ابقاء فرضیه  $H_0$  تصمیم می‌گیریم.

مثال: در اندازه‌ی خطای نوع اول  $\alpha = 0.05$  بررسی نمایید که آیا داده‌های زیر از توزیع پواسون پیروی می‌کنند؟

$$x = \{0, 1, 0, 0, 2, 3, 1, 0, 2, 1, 0, 1, 1, 1, 5, 2, 1, 1, 0, 4, 0, 2, 1, 3, \\ 0, 0, 2, 0, 1, 1, 2, 1, 3, 0, 4, 0, 2, 1, 1, 2, 1, 3, 0, 0, 1, 2, 1, 0, 2, 1, 2, 1, \\ 3, 1, 2, 7, 1, 4, 0, 0, 1, 2, 1, 0, 2, 2, 3, 1, 6, 1, 0, 5, 2, 5, 1, 2, 0, 0, 0, 3, \\ 2, 1, 0, 4, 1, 3, 4, 0, 1, 1, 2, 1, 0, 0, 1, 4, 0\}$$

پاسخ:

$$f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad \hat{\lambda} = MLE(\lambda) = \bar{x} = 1.5208, \\ \rightarrow e_i = n f(x_i) = 96 \frac{e^{-1.52} 1.52^{x_i}}{x_i!}$$

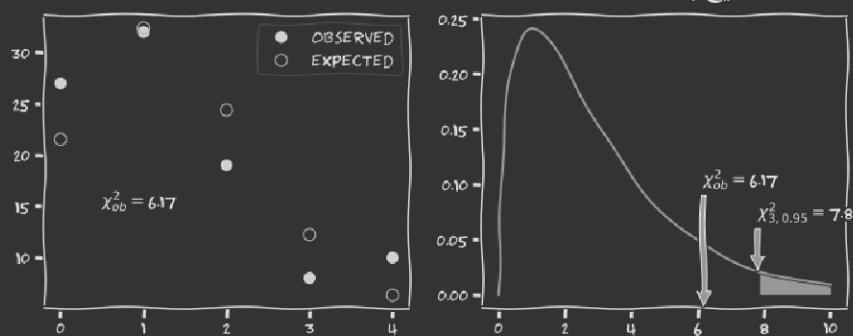
$x_i$	تعداد رویدادها	$f_i$	فرابوی انتظاری
0		26	20.98
1		32	31.90
2		19	24.26
3		8	12.30
4		6	4.68
5		3	1.42
6		1	0.35
7		1	0.08

از آنجاییکه در هر رده نباید فرابوی کمتر از ۵ داشته باشیم، رده‌ها ۴ تا ۷ را با هم ادغام می‌کنیم.

$$\chi_{ob}^2 = \frac{(26-20.98)^2}{20.98} + \frac{(32-31.90)^2}{31.90} + \frac{(19-24.26)^2}{24.26} + \frac{(8-12.30)^2}{12.30} + \frac{(11-6.53)^2}{6.53} = 6.90$$

$$\chi_{ob}^2 < \chi_{(5-1-1), 0.95}^2 = 7.815, \quad \rightarrow \text{Retain } H_0$$

پس این امکان وجود دارد که داده‌ها با توزیع پواسون توصیف شوند.



## ۲.۱۱ آزمون استقلال

در واقع کاملاً مانند آزمون برآنش می‌باشد که برای توزیع دو متغیره بکار گرفته شود. از آنجاییکه احتمال رخ دادن دو متغیر تصادفی با هم، در صورت استقلال از هم، برابر حاصلضرب احتمال رخ دادن تک تک آن دو می‌باشد، لذا مقدار مشاهده شده در هر رده را با مقدار انتظاری ناشی از حاصلضرب ضرب احتمال رخ داد تک تک مقایسه می‌کنیم.

	$B_1$	...	$B_c$	جمع سطرها
$A_1$	$n_{11}$	...	$n_{1c}$	$n_1$
...	...	...	...	...
$A_r$	$n_{1r}$	...	$n_{rc}$	$n_r$
جمع ستونها	$n_1$	...	$n_c$	$n$

فرضیه استقلال:

$$\left[ \begin{array}{ll} H_0 : p_{ij} = p_i p_j & \forall i, j \\ H_1 : p_{ij} \neq p_i p_j & \exists i, j \end{array} \right] \quad \text{where } p_{ij} = n_{ij}/n \quad \text{and} \quad p_i = n_i/n$$

آماره‌ی آزمون:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad \text{where } e_{ij} = n p_i p_j$$

ناحیه‌ی رد یا بحرانی:

$$\chi^2_{ob} > \chi^2_{(r-1)(c-1), 1-\alpha}$$

مثال: در سطح خطای نوع اول  $\alpha = 0.05$  استقلال دو متغیر تصادفی  $A$  و  $B$  را با توجه به مشاهدات داده شده بررسی نمایید.

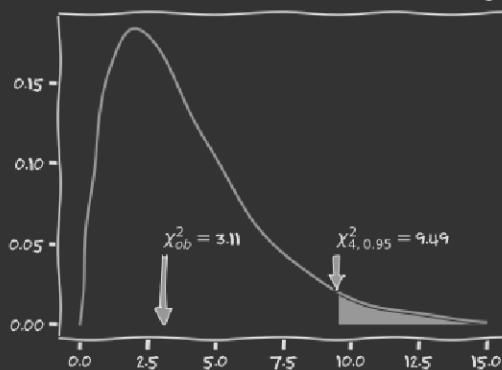
	$A_1$	$A_2$	$A_3$	sum
$B_1$	13	12	9	34
$B_2$	9	20	12	41
$B_3$	14	26	18	58
sum	36	58	39	n=133

پاسخ:

$$e_{ij} = \frac{n_i n_j}{n}$$

$$\begin{aligned} \chi^2_{ob} &= \sum_{i=1}^3 \sum_{j=1}^3 \frac{(n_{ij} - n_i n_j / n)^2}{n_i n_j / n} = \frac{(13-34 \times 36/133)^2}{34 \times 36/133} + \dots \\ &= 3.11 < \chi^2_{(3-1) \times (3-1), (1-0.05)} = \chi^2_{4, 0.95} = 9.49 \end{aligned}$$

پس بین متغیرهای  $A$  و  $B$  می‌تواند استقلال وجود داشته باشد.



## ۱۲ مدل سازی آماری: رگرسیون

### ۱.۱۲ مدل رگرسیون خطی

مسئله‌ی رگرسیون عبارت است از

$$E(Y|X = x) = \int y f(y|x) dy$$

اساسی‌ترین مدل رگرسیون، رگرسیون خطی ساده است.

$$E(Y|x) = \alpha + \beta x$$

مدل برآورده شده:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

مقدار خطای SSE باید کمینه شود:

$$SSE = \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

$$\frac{\partial SSE}{\partial \beta} = 0, \quad \rightarrow \hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$

$$\frac{\partial SSE}{\partial \alpha} = 0, \quad \rightarrow \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

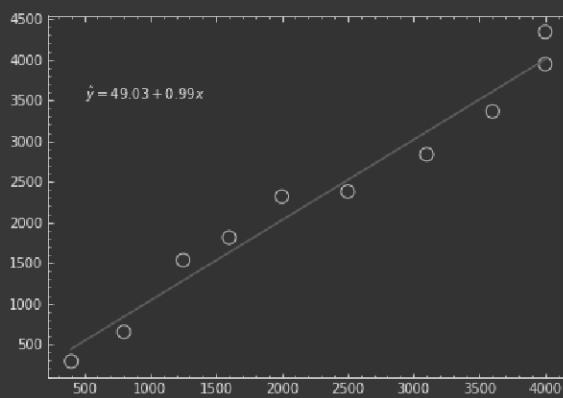
### مثال ۱

با توجه به جدول زیر مقدار  $y$  در نقطه‌ی  $x = 2400$  را پیش‌بینی کنید.

X	400	800	1250	1600	2000	2500	3100	3600	4000	4000
Y	291	654	1535	1813	2318	2378	2835	3365	4341	3945

$$\bar{x} = 2325.0 \quad \bar{y} = 2347.5 \quad S_{xy} = 15507275 \quad S_x^2 = 15686250$$

$$\hat{\beta} = 0.99 \quad \hat{\alpha} = 49.0 \quad Y(2400) = 2421.6$$



$$\begin{cases} Y_i = \alpha + \beta x_i + e_i \\ e_1, e_2, \dots, e_n \stackrel{iid}{\sim} N(0, \sigma^2) \end{cases} \rightarrow Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 0, 1, \dots, n$$

برآوردهای ناریب تحت مفروضات رگرسیونی نرمال برای  $\alpha$  و  $\beta$

$$\begin{cases} E(\hat{\alpha}) = \alpha \quad Var(\hat{\alpha}) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_x^2} \right) \\ E(\hat{\beta}) = \beta \quad Var(\hat{\beta}) = \frac{\sigma^2}{S_x^2} \end{cases} \rightarrow \begin{cases} \hat{\alpha} \sim N \left( \alpha, \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_x^2} \right) \right) \\ \hat{\beta} \sim N \left( \beta, \frac{\sigma^2}{S_x^2} \right) \end{cases}$$

$$\rightarrow \begin{cases} \frac{\hat{\beta} - \beta}{\sigma/S_x} \sim z \\ \frac{\hat{\alpha} - \alpha}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}}} \sim z \end{cases}$$

برآوردهای ناریب برای  $\sigma^2$  تحت مفروضات رگرسیونی

$$\widehat{\sigma^2} = S^2 = \frac{1}{n-2} SSE = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}, \quad \chi^2_{n-2} \perp \hat{\alpha}, \hat{\beta}$$

واریانس مولفه‌های خطای  $\sigma^2$  نامشخص می‌باشند. پس توزیع‌ها از نرمال به t تغییر می‌کنند.

$$\rightarrow \begin{cases} \frac{\beta - \hat{\beta}}{S/S_x} \sim t_{n-2} \\ \frac{\alpha - \hat{\alpha}}{S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}}} \sim t_{n-2} \end{cases}$$

## مثال ۲

با توجه به جدول زیر و پذیرش مفروضات رگرسیونی نرمال یک بازه‌ی ۹۵ درصدی اطمینان برای  $\alpha$  و  $\beta$  بیابید.

X	400	800	1250	1600	2000	2500	3100	3600	4000	4000
Y	291	654	1535	1813	2318	2378	2835	3365	4341	3945

با توجه به جدول زیر و پذیرش مفروضات رگرسیونی نرمال در سطح خطای (درجه اهمیت) ۵ درصد، بررسی کنید که آیا اصولاً  $Y$  به  $X$  بستگی خطی دارد؟

X	400	800	1250	1600	2000	2500	3100	3600	4000	4000
Y	291	654	1535	1813	2318	2378	2835	3365	4341	3945

$$H_0 : \beta = 0, \quad H_a : \beta \neq 0$$


---

بازه‌ی اطمینان برای میانگین پاسخ ( $E(Y|x^*)$ )

$$\begin{cases} E(\hat{\alpha} + \hat{\beta}x^*) = \alpha + \beta x^* \\ Var(\hat{\alpha} + \hat{\beta}x^*) = \sigma^2 \left( \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x^2} \right) \end{cases} \rightarrow \frac{(\hat{\alpha} + \hat{\beta}x^*) - (\alpha + \beta x^*)}{S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x^2}}} \sim t_{n-2}$$

بازه‌ی اطمینان برای پاسخ  $y^*$

$$\begin{cases} E(\hat{\alpha} + \hat{\beta}x^*) = \alpha + \beta x^* \\ Var(\hat{\alpha} + \hat{\beta}x^*) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x^2} \right) \end{cases} \rightarrow \frac{(\hat{\alpha} + \hat{\beta}x^*) - (\alpha + \beta x^*)}{S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_x^2}}} \sim t_{n-2}$$

با توجه به جدول زیر و پذیرش مفروضات رگرسیونی نرمال بازه‌ی اطمینان ۹۵ درصدی برای  $y$  و میانگین آن در  $x = 2400$  را بیابید.

X	400	800	1250	1600	2000	2500	3100	3600	4000	4000
Y	291	654	1535	1813	2318	2378	2835	3365	4341	3945

## ۳.۱۲ ملاحظات مدل رگرسیونی

- \* درونیابی و برونیابی: پیش‌بینی مدل‌های رگرسیونی خارج از دامنه‌ی داده‌های استفاده شده در محاسبه‌ی ضرایب قابل اعتماد نیستند.
- \* داده‌های پرت: روش کمترین مربعات خطای نسبت به داده‌هایی که از الگوی اصلی پیروی نمی‌کنند بسیار حساس هستند.
- \* فرضیات رگرسیونی نرمال - خطی بودن رابطه‌ی زیربنایی - استقلال خطاهای - ثابت بودن واریانس خطاهای - نرمال بودن توزیع خطاهای
- \* ثابتیت متغیر مستقل

## ۴.۱۲ همبستگی و تحلیل آن

ضریب همبستگی پیرسون

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{S_{xy}}{S_x S_y} = \frac{\hat{\beta} S_x}{S_y}$$

ضریب تعیین:  $r^2$

$$\begin{aligned} SSE(\text{Error}) &= \sum_{i=1}^n \left( y_i - \hat{\alpha} - \hat{\beta} x_i \right)^2 = S_y^2 - \hat{\beta}^2 S_x^2 \\ SST(\text{Total}) &= S_y^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \\ SSR(\text{Regression}) &= \hat{\beta}^2 S_x^2 \end{aligned}$$

$$SST = SSE + SSR$$

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

تقریب Z فیشر: استنباط از R به  $\rho$  برای نمونه‌ی تصادفی n تایی

$$Z^* = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

$$Z = \frac{Z^* - \mu_{Z^*}}{\sigma_{Z^*}} = N(0, 1), \quad \rightarrow \rho = \left( \tanh\left(Z^* - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right), \tanh\left(Z^* + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right) \right)$$

$$\tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$$

### مثال ۵

ضریب همبستگی و ضریب تعیین را محاسبه کنید.

X	8.8	9.3	10.4	11.2	12.3	13.	13.8	14.4	15.8	16.7	17.3
Y	11.8	11.9	11.4	11.6	12.3	12.7	13.3	13.7	13.8	14.4	14.5

