آمار و احتمالات

آنالیز ترکیبی

۱) اصل ضرب (اصل اساسی شمارش)

اگر دو آزمایش داشته باشیم و آزمایش اول دارای n نتیجه و برای هر نتیجه آن m نتیجه برای آزمایش دوم وجود داشته باشد آنگاه $m \times n$ نتیجه برای دو آزمایش وجود خواهد داشت.

نکته: اصل ضرب را می‌توان برای k آزمایش به صورت $m_1 \times m_2 \times \ldots \times m_k$ تعمیم داد که در این رابطه $m_1, \ldots, m_k$ تعداد نتایج k آزمایش هستند.

۲) جایگشت (ترتیب)

تعداد طرقی که می‌توان از n شیء متمایز k عدد انتخاب و آنها را مرتب نمود برابر است با: $p_n^k = (n)_k = \dfrac{n!}{(n-k)!}$

۳) ترکیب

تعداد طرقی که می‌توان از n شیء متمایز k عدد انتخاب نمود برابر است با: $C_n^k = \dbinom{n}{k} = \dfrac{n!}{k!(n-k)!}$

مثال: به چند طریق می‌توان n آنتن سالم و m آنتن خراب را در کنار هم قرار داد به طوری که دو آنتن خراب کنار هم قرار نگیرند؟

برای حل این مسأله کافی است ابتدا آنتن‌های سالم را چید و سپس محل‌های آنتن‌های خراب را در میان آنتن‌های سالم انتخاب نمود به طوری که هیچ دو آنتن خرابی کنار هم قرار نگیرند. با چیدن آنتن‌های سالم $n+1$ مکان برای آنتن‌های خراب معین می‌گردد، لذا پاسخ $\dbinom{n+1}{m}$ خواهد شد.

۴) تعداد طرقی که می‌توان از n شیء متمایز k شیء با جایگذاری انتخاب کرد برابر راست با: n^k

۵) تعداد طرقی که می‌توان n شیء متمایز را که n_1 تای آن مشابه هم، n_2 تای آن مشابه به هم و ..., n_k تای آن مشابه به هم هستند را مرتب نمود برابر است با: $C_n^{n_1, n_2, \ldots, n_k} = \dbinom{n}{n_1, \ldots, n_k} = \dfrac{n!}{n_1! n_2! \ldots n_k!} \quad , \quad \sum_{i=1}^{k} n_i = n$

مثال: به چند طریق می‌توان با حروف a,a,b,b,b,c,c کلمه ساخت؟ $\dfrac{7!}{2!3!2!}$

۶) تعداد طرقی که می‌توان n شیء متمایز را در k ظرف طوری مرتب نمود که ظرف اول n_1 تا، ظرف دوم n_2 تا و ... و ظرف kام n_k تا را شامل شود برابر است با: $C_n^{n_1, \ldots, n_k} = \dbinom{n}{n_1, \ldots, n_k} = \dfrac{n!}{n_1! n_2! \ldots n_k!} \quad , \quad \sum_{i=1}^{k} n_i = n$

مثال: به چند طریق می‌توان ۱۰ نفر را در ۵ اتاق مرتب نمود به طوری که در یک اتاق ٤ نفر و در یک اتاق ۳ نفر و در اتاق‌های دیگر هر کدام ۱ نفر قرار گیرند؟

$$\frac{10!}{4!3!1!1!1!} \times \underbrace{\binom{5}{1}}_{\text{انتخاب اتاق ٤ نفره}} \times \underbrace{\binom{4}{1}}_{\text{انتخاب اتاق ۳ نفره}} \times \underbrace{\binom{3}{3}}_{\text{انتخاب اتاق تک نفره}}$$

۷) تعداد طرقی که می‌توان n شیء متمایز را در k ظرف قرار داد به طوری که محدودیتی در قرارگیری اشیاء در ظرف‌ها نباشد، برابر است با:

$$r^n = \sum_{\substack{(n_1,n_2,\dots,n_k) \\ n_1+n_2+\dots+n_k=n}} \binom{n}{n_1,n_2,\dots,n_k}$$

۸) تعداد طرقی که می‌توان n شیء مشابه را در r ظرف قرار داد برابر است با:

$$\binom{n+r-1}{r-1}$$

مثال: به چند طریق می‌توان ۱۱ سیب را بین ٤ دختر بچه تقسیم کرد؟

$$\binom{11+4-1}{4-1} = \binom{15}{3}$$

۹) بسط دوجمله‌ای

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

تعمیم

$$(x_1+x_2+\dots+x_k)^n = \sum_{\substack{n_1,\dots,n_k \\ n_1+n_2+\dots+n_k=n}} \binom{n}{n_1,\dots,n_k} x_1^{n_1} x_2^{n_2},\dots,x_k^{n_k}$$

۱۰) تعداد طرقی که می‌توان n شیء مشابه را به k دستهٔ $x_1,x_2,\dots,x_k$ تقسیم کرد به نحوی که سهم هر دسته در شرایط $0 \le x_i \le m$ صدق کند برابر است با:

$$N = \sum_{s=0}^{k} (-1)^s C_k^s C_{n+k-(m+1)s-1}^{k-1}$$

تذکر: محاسبه جملات تا زمانی ادامه دارد که عبارت $C_{n+k-(m+1)S-1}^{k-1}$ معنی داشته باشد.

مثال: در بسط $(x_1+2x_2+3x_3+4x_4)^{10}$ ضریب جملهٔ $x_1^3x_2^3x_3^1x_4^3$ کدام است؟

حل: طبق بسط n جمله‌ای داریم: ضریب جمله مورد نظر: $\left(\frac{10!}{3!\times3!\times1!\times3!}\right)\times1^3\times2^3\times3^1\times4^3$

مقدمات احتمال

- **آزمایش:** فرآیندی است با نتایج محدود و غیر قابل پیش‌بینی مانند پرتاب تاس
- **فضای نمونه:** مجموعهٔ همهٔ نتایج ممکن حاصل از انجام یک آزمایش را فضای نمونه گویند. مثلاً مجموعه {۱و۲و...و٦} فضای نمونهٔ پرتاب یک تاس است.
- **پیشامد:** هر زیر مجموعه از فضای نمونه را یک پیشامد گویند. اگر این زیر مجموعه تنها یک عضو داشته باشد مانند {(شیر، شیر)} آن را پیشامد ساده و اگر بیش از یک عضو داشته باشد مانند {۱و۲و۳} آن را پیشامد مرکب می‌نامند.

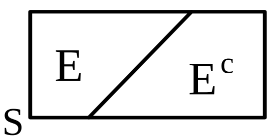
نکات مهم در جبر مجموعه‌ها

اگر پیشامدهای E و F را داشته باشیم:

– $E \cap F$: مجموعهٔ نتایج ممکن آزمایش که هم در E و هم در F هست.

– $E \cup F$: مجموعهٔ نتایج ممکن آزمایش که حداقل در یکی از E یا F هست.

– E^c متمم E است هرگاه در فضای نمونهٔ S داشته باشیم:



$E^c \cup E = S$ و $E^c \cap E = \phi$

– $E \subset F$ است هرگاه هر عضو E در F هم باشد، لذا داریم $E \cap F = E$ و $E \cup F = F$

- E و F ناسازگارند هرگاه $E \cap F = \phi$

قوانین جبر مجموعه‌ها

- خاصیت جابجایی $\quad F \cup E = E \cup F \qquad F \cap E = E \cap F$
- خاصیت شرکت‌پذیری $\quad E \cap (F \cap G) = (E \cap F) \cap G \qquad E \cup (F \cup G) = (E \cup F) \cup G$
- خاصیت توزیع‌پذیری $\quad E \cup (F \cap G) = (E \cup F) \cap (E \cup G) \qquad E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$
- **قوانین دمورگان**

$$\left(\bigcap_{i=1}^{n} E_i\right)^C = \bigcup_{i=1}^{n} E_i^C \qquad \left(\bigcup_{i=1}^{n} E_i\right)^C = \bigcap_{i=1}^{n} E_i^C$$

• **احتمال:** اگر همهٔ نقاط یا پیشامدهای سادهٔ فضای نمونه‌ای S هم‌شانس باشند آنگاه احتمال وقوع پیشامد E به صورت زیر تعریف می‌شود:

$$p(E) = \frac{n(E)}{n(S)}$$

که در آن n(E) و n(S) به ترتیب عبارتند از تعداد نقاط پیشامد E و فضای نمونه‌ای S، بدیهی است با چنین تعریفی خواص زیر برای تابع احتمال متصور است:

۱) $P(S) = 1$

۲) $0 \le P(E) \le 1$

۳) اگر Eها دوبه‌دو ناسازگار باشند یعنی داشته باشیم $\forall i \neq j \; E_i \cap E_j = \phi \;,\; \bigcup_{i=1}^{n} E_i = S$ آنگاه $\sum_{i=1}^{n} p(E_i) = 1$

مثال: احتمال اینکه از ۱۲ نفر هیچکدام در یک ماه مشترک به دنیا نیامده باشند چقدر است؟

حل در این مسأله چون شانس تولد افراد در ماه‌های مختلف هم‌شانس نمی‌باشد حل $\frac{12!}{12^{12}}$ غلط می‌باشد و لذا ما مسأله را در فضای هم‌شانس حل خواهیم نمود و به جای ماه‌های سال، روزهای سال را در نظر می‌گیریم:

$$p(A) = \frac{12!(31)^6 \times (30)^5 \times 29}{(365)^{12}}$$

مثال: حروف abcdefghizzzzzzz را به تصادف در یک دریف قرار می‌دهیم، احتمال اینکه هیچ دو z کنار هم نباشند چقدر است؟

حل

$$p(A) = \frac{9! \times \binom{10}{7}}{\binom{16}{1,1,\ldots,7}}$$

نکته:

$$p(E_1 \cup E_2 \cup \ldots \cup E_n) = \sum_{i=1}^{n} p(E_i) - \sum_{i_1 < i_2} p(E_{i_1} \cap E_{i_2}) + \ldots + (-1)^{n+1} \sum p(E_{i_1} \cap E_{i_2} \cap \ldots \cap E_{ik})$$

مثال: (**مسألهٔ انطباق**) فرض کنید n نفر کلاه خود را در کیسه‌ای قرار می‌دهند و آنگاه هر یک به تصادف کلاه خود را برمی‌دارند، احتمال اینکه هیچ‌کس کلاه خود را برندارد چقدر است؟

حل P (هیچکس کلاه خود را بر ندارد) = 1−p (حداقل یک نفر کلاه خود را بردارد)

A_i: پیشامد که i امین نفر کلاه خود را بر دارد

$$P\,(\text{حداقل یک نفر کلاه خود را بردارد}) = p(A_1 \cup A_2 \cup \ldots \cup A_n)$$

$$= \binom{n}{1} p(A_1) - \binom{n}{2} p(A_1 \cap A_2) \ldots + (-1)^n p(A_1 \cap A_2 \cap \ldots \cap A_n)$$

$$= \binom{n}{1} \times \frac{1}{n} - \binom{n}{2} \frac{1}{n(n-1)} + \ldots + (-1)^{n+1} \binom{n}{n} \frac{1}{n!} = 1 - \frac{1}{2!} + \ldots + (-1)^{n+1} \frac{1}{n!}$$

$$p\,(\text{هیچکس کلاه خود را بر ندارد}) = 1 - \left[1 - \frac{1}{2!} + \ldots + (-1)^{n+1} \frac{1}{n!}\right] = \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^n \frac{1}{n!}$$

مثال: در مثال قبل احتمال اینکه k نفر کلاه خود را بردارند چقدر است؟

حل: در این حالت اگر k نفر کلاه خود را بردارند (n-k) نفر کلاه خود را برنخواهند داشت، لذا با توجه به حل مثال قبل داریم:

$$p\,(\text{انطباق } k) = \frac{\binom{n}{k} \times (n-k)! \left[\frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^{n-k} \times \frac{1}{(n-k)!}\right]}{n!}$$

احتمال شرطی

احتمال وقوع پیشامد A وقتی بدانیم B رخ داده است را احتمال A به شرط B گویند و آن را به شکل P(A|B) نشان می‌دهند که

$$P(A \mid B) = \frac{P(A \cap B))}{P(B)}$$

مثال: اگر هیأت منصفهٔ یک دادگاه با احتمال ٪۹۵ در مورد یک متهم تصمیم درست بگیرد و ٪۷۰ از متهمان این دادگاه مجرم باشند، آنگاه احتمال اینکه فردی مجرم باشد و دادگاه نیز رأی به مجرم بودن او بدهد چقدر است؟

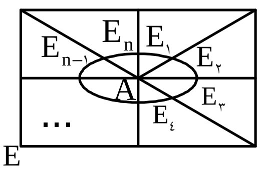
حل:

$$P\,(\text{دادگاه رای دهد و مجرم بودن}) = P\,(\text{مجرم}) \times P\,(\text{مجرم بودن} \mid \text{دادگاه رأی دهد}) = 0/7 \times 0/95$$

نکته:

$$P(E_1 \cap E_2 \cap \ldots \cap E_n) = P(E_1) P(E_2 \mid E_1) P(E_3 \mid E_1 \cap E_2) \ldots P(E_n \mid E_1 \cap \ldots E_{n-1})$$

قانون اول بیز



فرض کنید فضای نمونه E به پیشامدهای $E_1, E_2, \ldots, E_n$ افراز شده باشد یعنی:

$$\bigcup_{i=1}^{n} E_i = E \quad , \quad \forall i \neq j \quad E_i \cap E_j = \phi$$

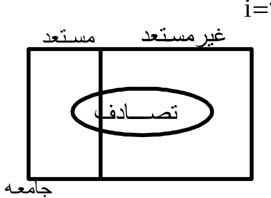
همچنین پیشامد A را در نظر بگیرید که با همه (یا برخی) از این افرازها دارای اشتراک است. محاسبهٔ احتمال پیشامد A داریم:

$$P(A) = \sum_{i-1}^{n} P(A_i) = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(A \mid E_i) P(E_i)$$

قانون دوم بیز

اگر بخواهیم احتمال وقوع هر یک از افرازهای E را به دست آوریم داریم:

$$P(E_k \mid A) = \frac{P(E_k \cap A)}{P(A)} = \frac{P(A \mid E_k)}{\sum_{i=1}^{n} P(A \mid E_i) P(E_i)}$$



مثال: در یک جامعه ٪۳۰ مستعد تصادف هستند، اگر احتمال تصادف برای افراد مستعد ۰/٤ و برای افراد غیر مستعد ۰/۲ باشد، چند درصد افراد جامعه تصادف می‌کنند.

حل

P (تصادف) =P (تصادف| مستعد بودن) P (مستعد بودن) +P (تصادف | غیر مستعد بودن) P (غیر مستعد بودن)

$$=0/4\times0/3+0/2\times0/7$$

مثال: فرض کنید هیأت منصفهٔ دادگاهی با احتمال ٪۹۵ در مورد هر متهم تصمیم درست بگیرد، اگر ٪۹۹ متهمین این دادگاه واقعاً گناهکار باشد مطلوب است احتمال اینکه یک شخص توسط دادگاه که گناهکار شناخته شده باشد، بی‌گناه باشد چقدر است؟

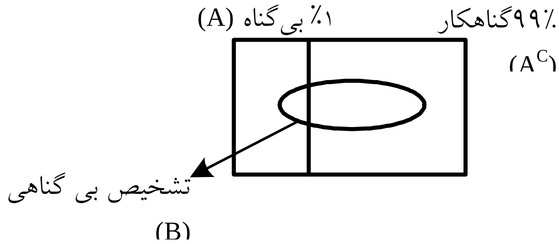
حل

P (تشخیص بی‌گناهی | گناهکاری)= ٪۵

P (تشخیص گناهکاری | گناهکاری) = ٪۹۵

P (تشخیص گناهکاری | بی‌گناهی)= ٪۵

P (تشخیص بی‌گناهی | بی‌گناهی)= ٪۹۵



$$P(\text{بی‌گناهی} \mid \text{تشخیص گناهکاری})=\frac{\overbrace{P(B|A)}^{0/05}\overbrace{P(A)}^{0/01}}{\underbrace{P(B|A)}_{0/05}\underbrace{P(A)}_{0/01}+\underbrace{P(A^C)}_{0/99}\underbrace{P(B|A^C)}_{0/05}}$$

نکته: $P(A|B)+P(A^C|B)=1$

استقلال دوپیشامد

اگر دانستن وقوع پیشامد E در احتمال وقوع پیشامد F تأثیری نداشته باشد، آنگاه گوییم F از E مستقل است.

$$P(F|E)=P(F)\Leftrightarrow P(E|F)=P(E)$$

نکته: اگر دو پیشامد ناتهی F و E ناسازگار باشند آنگاه آن دو حتماً وابسته هستند. $P(E|F)\neq P(E)$

نکته: اگر E و F مستقل باشند آنگاه F^C و E، E^C و F، E^C و F^C مستقل هستند.

متغیر های تصادفی

تعریف

متغیر تصادفی تابعی است که روی فضای نمونه بر روی زیرمجموعه‌ای ناتهی از اعداد حقیقی تعریف می‌شود.

انواع متغیر تصادفی

۱) متغیرهای تصادفی گسسته

متغیرهایی هستند که برد آنها مقادیر گسسته یا قابل شمارش می‌باشد. در این حالت از تابع احتمال برای محاسبهٔ احتمال فضای جدید استفاده می‌شود.

۲) متغیرهای تصادفی پیوسته

متغیرهایی هستند که برد آنها مقادیر پیوسته یا غیرقابل شمارش می‌باشد. در این حالت از تابع چگالی برای محاسبهٔ احتمال فضای جدید استفاده می‌شود.

۳) متغیرهای تصادفی آمیخته

متغیرهایی هستند که قسمتی از برد آنها گسسته و قسمتی دیگر پیوسته است.

تابع توزیع تجمعی

تابع توزیع یک متغیر تصادفی X که آن را با F نشان می‌دهیم برای همهٔ مقادیر حقیقی b به صورت زیر تعریف می‌شود:

$$F_X(b) = P(x \le b)$$

خصوصیات تابع توزیع تجمعی

۱) F تابعی غیر نزولی است یعنی اگر a<b باشد آنگاه $F(a) \le F(b)$ است.

۲) $\lim\limits_{b\to\infty} F(b) = ۱$

۳) $\lim\limits_{b\to-\infty} F(b) = ۰$

٤) F از سمت راست پیوسته است. یعنی برای هر b و هر دنباله صعودی b_n که به سمت b همگرا است، $\lim\limits_{n\to\infty} F(b_n) = F(b)$

توزیع یک متغیر تصادفی

توزیع یک متغیر تصادفی پخش شدن احتمال را بین متغیرهای تصادفی نشان می‌دهد.

تحلیل شکل توزیع هر متغیر تصادفی توسط معیارهای عددی و مشخص صورت می‌گیرد که این معیارها عبارتند از:

۱) شکل توزیع
- چولگی
- کشیدگی

۲) پراکندگی
- واریانس
- انحراف معیار
- ضریب تغییرات
- دامنه

۳) تمرکز توزیع
- امید ریاضی
- میانه
- مُد توزیع

امید ریاضی متغیرهای تصادفی گسسته

اگر x یک متغیر تصادفی گسسته با تابع احتمال P(x) باشد آنگاه امید ریاضی یا مقدار امید آن که با E(x) نمایش داده می‌شود، به صورت زیر تعریف می‌شود.

$$E(x) = \sum_{X:P(x)>۰} XP(x)$$

امید ریاضی تابعی از متغیر تصادفی گسسته

اگر g(x) یک تابع از متغیر تصادفی باشد برای محاسبهٔ امید ریاضی g(x) ، دو روش وجود دارد:

۱) محاسبهٔ تابع احتمالی g(x) و سپس محاسبهٔ امید ریاضی آن.

۲) محاسبهٔ ریاضی به طریق مستقیم از طریق رابطهٔ زیر:

$$E[g(x)] = \sum_i g(x_i) \times P(x_i)$$

که در آن $P(x_i)$ تابع احتمال متغیر تصادفی گسسته x است.

نکته: $E[aX+b] = aE(X)+b$

نکته: $E[h(n)+g(x)]=E[g(x)]+E[h(x)]$

مدتوزیع

نقطه‌ای است که بیشترین فراوانی را دارد. به عبارت دیگر متحمل‌ترین پیشامد است.

واریانس

اگر x یک متغیر تصادفی با امید ریاضی μ باشد، آنگاه واریانس x که آن را با $Var(X)$ نشان می‌دهیم، به صورت زیر تعریف می‌شود.

$$Var(X)=E\left[(X-\mu)^{۲}\right]=\sum_{x}(x-\mu)^{۲}p(x)$$

نکته: $Var(X)=E\left[x^{۲}\right]-[E(x)]^{۲}$

نکته: $Var(ax+b)=a^{۲}var(x)$

انحراف استاندارد

$$S.D(x)=\sqrt{var(x)}$$

نکته: $S.D(ax+b)=|a|S.D(x)$

ضریب تغییرات

اگر x یک متغیر تصادفی باشد آنگاه ضریب تغییرات که آن را با C.V نشان می‌دهیم، به صورت زیر تعریف می‌شود.

$$C.V(x)=\frac{S.D(x)}{E(x)}$$

انواع متغیر های تصادفی

متغیر تصادفی برنولی

آزمایشی را در نظر بگیرید که با احتمال p به موفقیت و با احتمال $۱-p$ به شکست منجر می‌شود، اگر موفقیت حاصل شود، متغیر x، عدد ۱ و اگر شکست حاصل شود عدد ۰ را می‌گیرد. در این صورت به این متغیر تصادفی برنولی گویند که تابع احتمال آن به صورت زیر است:

$$P(x=i)=\begin{cases}۱-p & i=۰\\ p & i=۱\end{cases} \quad \text{یا} \quad P(x=i)=P^{i}(۱-P)^{۱-i} \quad i=۰,۱$$

$$E(x)=P \qquad var(x)=P(۱-P)$$

نکته: اگر x دارای توزیع برنولی با احتمال موفقیت p باشد آنگاه متغیر تصادفی $y=x^{n}$ نیز دارای توزیع برنولی با احتمال موفقیت p^{n} است.

- **متغیر تصادفی دوجمله‌ای**

n آزمایش مستقل برنولی را آزمایش دوجمله‌ای گوییم.

$$P(x=i)=\binom{n}{i}P^{i}(۱-P)^{n-i} \qquad i=۰,۱,۲,\ldots,n$$

$$E(x)=nP \qquad var(x)=np(۱-p)$$

چند نکته:

- اگر $x_1, x_2, ..., x_n$ متغیرهای تصادفی مستقل برنولی با پارامتر p باشند آنگاه توزیع $y = x_1 + x_2 + ... + x_n$ دوجمله‌ای با پارامتر n و p است.
- اگر $x_1, x_2, ..., x_n$ متغیرهای تصادفی مستقل برنولی با پارامتر p باشند آنگاه توزیع $y = x_1 x_2 ... x_n$ برنولی با پارامتر p^n است.
- اگر $x_1, x_2, ..., x_n$ متغیرهای تصادفی مستقل برنولی با پارامتر p باشند آنگاه توزیع $y = \min\{x_1 x_2 ... x_n\}$ برنولی با پارامتر p^n است.
- اگر $x_1, x_2, ..., x_n$ متغیرهای تصادفی مستقل برنولی با پارامتر p باشند آنگاه توزیع $y = \max\{x_1, x_2, ..., x_n\}$ برنولی با پارامتر $1-(1-p)^n$ است.
- اگر $x_1, x_2, ..., x_n$ متغیرهای تصادفی مستقل برنولی با پارامتر p باشند آنگاه توزیع $y = x_1 x_2^2 ... x_n^n$ برنولی با پارامتر p^n است.
- اگر x_i یک متغیر تصادفی با توزیع دوجمله‌ای و پارامترهای n و p باشد آنگاه با افزایش i مقدار P(n=i) ابتدا افزایش و سپس کاهش می‌یابد. بنابراین مقدار مدتوزیع وقتی n زوج باشد $[(n+1)P]$ و وقتی فرد باشد دو نقطهٔ $(n+1)p$ و $(n+1)p-1$ است.

متغیر تصادفی پواسان

متغیر تصادفی x که یکی از مقادیر ...و۲و۱و۰ را اختیار می‌کند، یک متغیر تصادفی پواسان با پارامتر λ نامیده می‌شود، هرگاه برای $\lambda > 0$ تابع احتمال به صورت زیر باشد:

$$P(i) = P(x=i) = \frac{e^{-\lambda}\lambda^i}{i!} \qquad i = 0, 1, ...$$

$$E(x) = \lambda \qquad var(x) = \lambda$$

- اگر x یک متغیر تصادفی پواسان با پارمتر λ باشد، با افزایش مقدار i معمولاً مقدار P(x=1) ابتدا افزایش و سپس کاهش می‌یابد، لذا اگر λ یک عدد صحیح باشد، مد (محتمل‌ترین پیشامد) توزیع پواسان زمانی است که i مقدار λ یا $\lambda - 1$ را بگیرد و اگر λ عدد صحیح نباشد، مدتوزیع زمانی است که i مقدار $[\lambda]$ را بگیرد.

- اگر تعداد دفعات رخ دادن پیشامدی در واحد زمان از توزیع پواسان با پارامتر λ پیروی کند، تعداد دفعات رخ دادن آن پیشامد در بازهٔ زمانی t از توزیع پواسان با پارامتر $t\lambda$ پیروی می‌کند.

- اگر تعداد دفعات رخ دادن پیشامدی در واحد مکان از توزیع پواسان با پارامتر λ پیروی کند، تعداد دفعات رخ دادن آن پیشامد در بازهٔ مکانی k از توزیع پواسان با پارامتر $k\lambda$ پیروی می‌کند.

- اگر تعداد دفعات رخ داد پیشامدی در واحد زمان/مکان از توزیع پواسان با پارامتر λ پیروی کند، و هر پیشامدی از این توزیع پواسان را بتوان به دو واقعهٔ ناسازگار و متمم E و E^C با احتمال‌های p و $1-p$ تفکیک نمود، آنگاه تعداد دفعات رخ دادن پیشامد E در واحد زمان/مکان از توزیع پواسان با پارامتر $P\lambda$ پیروی می‌کند.

آمار و احتمالات

متغیرهای تصادفی پیوسته

فرض کنید یک بازهٔ یک بعدی A و تابع f(x) را بگونه‌ای داشته باشیم که:

الف) $\forall x \in A \Rightarrow f(x) \geq 0$

ب) در هر زیر مجموعهٔ متناهی از A، تعداد نقاط ناپیوستگی f(x) متناهی باشد.

ج) $\int_A f(x)dx = 1$

در این صورت x یک متغیر تصادفی پیوسته با تابع چگالی احتمال پیوسته f(x) است.

برای یک متغیر تصادفی پیوسته داریم: $\int_{-\infty}^{x} f_X(t)dt = F(x)$

که در آن f(x) تابع چگالی احتمال X و F(x) تابع توزیع آن است. بدیهی است که مانند آنچه دربارهٔ تابع توزیع یک متغیر تصادفی گسسته گفته شد برای یک تابع توزیع متغیر تصادفی پیوسته نیز داریم: $F(-\infty) = 0 \quad , \quad F(+\infty) = 1$

علاوه بر آن اگر f(x) پیوسته باشد، F(x) نیز در تمام نقاط پیوسته است. هم چنین داریم: $f(x) = \frac{dF(x)}{dx}$

نکته: احتمال این که متغیر تصادفی X در بازه [a, b) باشد برابر است با $p(a < x \leq b) = \int_a^b f(x)dx = F(b) - F(a)$

لذا احتمال این که متغیر تصادفی X دارای یک مقدار مشخص مانند α باشد صفر است

$$p(X = \alpha) = \lim_{\varepsilon \to 0} \int_{x-\varepsilon}^{\alpha+\varepsilon} f(x)dx = \lim_{\varepsilon \to 0} [F(\alpha+\varepsilon) - F(\alpha-\varepsilon)] = 0$$

نکته: امید ریاضی تابعی از یک متغیر تصادفی پیوسته و نیز واریانس آن همانند متغیر تصادفی گسسته قابل محاسبه است.

$$E[g(X)] = \int_x g(x)f(x)dx$$

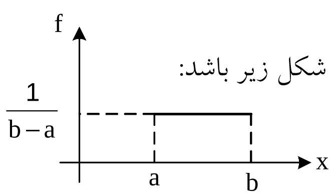
نامساوی چبیشف

اگر X یک متغیر تصادفی و g(X) یک تابع حقیقی غیرمنفی از X باشد، داریم: $p[g(X) \geq k] \leq \frac{1}{k} E[g(X)] \quad , \quad k > 0$

در حالت خاص با انتخاب $g(X) = (X-2)^2$ و $k = k^2\sigma^2$ داریم: $p[|x - \mu| \geq k\sigma] \leq \frac{1}{k^2}$

متغیر تصادفی یکنواخت

متغیر تصادفی پیوستهٔ X را یکنواخت می‌نامیم هر گاه در فاصلهٔ [a, b] دارای تابع چگالی احتمال ثابت و به شکل زیر باشد:



$$f(x) = \frac{1}{b-a} \qquad a \leq x \leq b$$

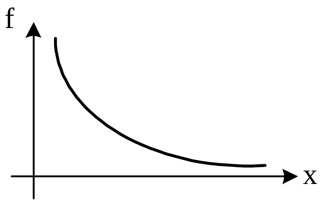
امید ریاضی و واریانس این متغیر تصادفی عبارتند از: $E(X) = \frac{a+b}{2} \qquad var(X) = \frac{(b-a)^2}{12}$

نکته: در صورتی که بخواهیم احتمال یک بازه را بیابیم داریم:

$$p(\alpha \le X \le \beta) = \int_{\alpha}^{\beta} \frac{۱}{b-a} dx = \frac{\beta-\alpha}{b-a}$$

همان طور که از رابطه مشخص است، احتمال قرار گرفتن متغیر تصادفی یکنواخت در یک بازه به طول $(\beta-\alpha)$ تنها به طول آن بازه بستگی دارد نه به مقادیر β,α.

متغیر تصادفی نمایی



متغیر تصادفی X دارای توزیع نمایی است اگر تابع چگالی احتمال آن به صورت زیر باشد:

$$f(x) = \lambda e^{-\lambda x} \qquad x \ge ۰$$

امید ریاضی و واریانس این متغیر تصادفی عبارتند از:

$$E(X) = \frac{۱}{\lambda} \qquad var(X) = \frac{۱}{\lambda^{۲}}$$

نکته: توزیع نمایی بی‌حافظه است، یعنی داریم:

$$p(x>s+t|x>t)= p(x>s)$$

مثال: اگر مدت زمان کارکرد یک دستگاه تا زمان خرابی دارای توزیع نمایی با میانگین ۵ ساعت باشد و بدانیم ایـن دسـتگاه در یـک ساعت اول خراب نشده است، احتمال این که تا ساعت سوم هم خراب نشود چقدر است؟

$$p(x>۳|x>۱) = p(x>۲) = e^{-۲\times\frac{۱}{۵}}$$

نکته: اگر تعداد دفعات وقوع پیشامدی دارای توزیع پواسان با پارامتر λ باشد، فاصلهٔ زمانی میان هر دو پیشامد متوالی دارای توزیع نمـایی بـا پارامتر λ است.

متغیر تصادفی گاما

متغیر تصادفی X دارای توزیع گاما است اگر دارای تابع چگالی احتمال به فرم زیر باشد:

$$f(x) = \frac{(\lambda x)^{n-۱} \lambda e^{-\lambda x}}{\Gamma(n)} \qquad x \ge ۰$$

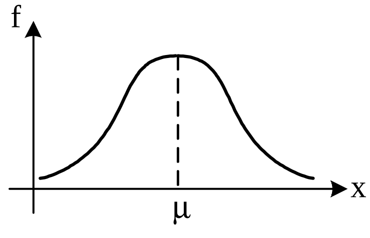
که در آن $\Gamma(n)$ تابع گاما بوده و داریم:

$$r(n+۱) = nr(n) \Rightarrow n \in N : \Gamma(n) = (n-۱)! \;, \Gamma(n+۱) = n\Gamma(n)$$

امید ریاضی و واریانس متغیر تصادفی گاما عبارتند از:

$$E(X) = \frac{n}{\lambda} \qquad var(X) = \frac{n}{\lambda^{۲}}$$

متغیر تصادفی نرمال



متغیر تصادفی X دارای توزیع نرمال است هرگاه تابع چگالی احتمال آن به فرم زیر باشد:

$$f(x) = \frac{۱}{\sigma\sqrt{۲\pi}} e^{-\frac{(x-\mu)^{۲}}{۲\sigma^{۲}}} \qquad -\infty < x < +\infty$$

امید ریاضی و واریانس متغیر تصادفی نرمال عبارتند از:

$$E(X) = \mu \qquad var(X) = \sigma^{۲}$$

هم چنین احتمال متغیر تصادفی X در بازهٔ [a, b] برابر است با

$$p(a < x \le b) = \int_{a}^{b} \frac{۱}{\sigma\sqrt{۲\pi}} e^{-\frac{(x-\mu)^{۳}}{۲\sigma^{۲}}}$$

نکته: به جهت سهولت در انجام محاسبات برای این توزیع متغیر تصادفی نرمال Z به شکل زیر تعریف می‌شود

$$Z=\frac{X-\mu}{\sigma} \qquad var(Z)=1 \qquad E(Z)=0$$

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \qquad -\infty<z<+\infty$$

$$F(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=\phi(z)$$

معمولاً تابع توزیع متغیر تصادفی نرمال استاندارد (متغیر تصادفی نرمال با واریانس ۱ و میانگین ۰) را با $\phi(z)$ نمایش می‌دهیم که دارای خواص روبرو است:

$$\phi(-\infty)=0 \qquad \phi(+\infty)=1 \qquad \phi(0)=\frac{1}{2} \qquad \phi(-a)=1-\phi(a)$$

نکته: جدول توزیع نرمال در واقع جدول محاسبهٔ مقادیر تابع $\phi(z)$ برحسب مقادیر مختلف Z است و بدین ترتیب داریم:

$$p(a<x<b)=p\left(\frac{a-\mu}{\sigma}<z<\frac{b-\mu}{\sigma}\right)=p\left(z<\frac{b-\mu}{\sigma}\right)-p\left(z<\frac{a-\mu}{\sigma}\right)=\phi\left(\frac{b-\mu}{\sigma}\right)-\phi\left(\frac{a-\mu}{\sigma}\right)$$

تقریب توزیع دوجمله‌ای بوسیلهٔ توزیع نرمال

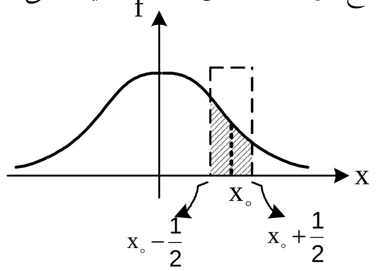
اگر در توزیع دو جمله‌ای $n\to\infty$ و $p\approx\frac{1}{2}$ آنگاه توزیع دو جمله‌ای را می‌توان بوسیلهٔ توزیع نرمال تقریب زد که برای این توزیع نرمال داریم:

$$\mu=np \qquad \sigma^2=np(1-p)$$

در این حالت برای محاسبهٔ $p(X=x_0)$ با استفاده از توزیع دو جمله‌ای داریم:

$$p(X=x_0)=C_n^{x_0}p^{x_0}(1-p)^{n-x_0}$$

که برابر طول میلهٔ فراوانی این نقطه در شکل است. برای محاسبهٔ این احتمال با استفاده از توزیع نرمال، همانگونه که در شکل می‌بینید مساحت تقریبی مستطیل در اطراف نقطهٔ x_0 برابر است با مساحت ذوزنقه نشان داده شده است و برابر است با حاصلضرب ارتفاع در میانگین دو قاعده، که ارتفاع این ذوزنقه برابر $x_0+\frac{1}{2}-\left(x_0-\frac{1}{2}\right)=1$ و میانگین دو قاعده تقریباً همان احتمال x_0 بر مبنای توزیع دو جمله‌ای است. بدین ترتیب مساحت این ذوزنقه تقریباً برابر همان $F(x_0)$ در توزیع دو جمله‌ای است، یعنی:



$$p(X=x_0)=\phi\left(\frac{x_0+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)-\phi\left(\frac{x_0-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

به این ترتیب مساحت ناحیهٔ ذوزنقه حول نقطهٔ x_0 و قبل از آن برابر مجموع چگالی‌های دو جمله‌ای به ازای xهای کوچکتر یا مساوی x_0 است. یعنی با استفاده از تقریب نرمال خواهیم داشت:

$$p(X\le x_0)=\phi\left(\frac{x_0+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) \quad , \quad p(a<x\le b)=\phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)-\phi\left(\frac{a+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

تقریب توزیع پواسان بوسیلهٔ توزیع نرمال

یک توزیع پواسان با λ بزرگ (تقریباً بزرگ‌تر یا مساوی ۵) را می‌توان به یک توزیع نرمال با میانگین و واریانس $\mu=\lambda$ و $\sigma^2=\lambda$ تقریب زد، بدین ترتیب روابط $p(X=x_0), p(X\le x_0)$ و $p(a<X\le b)$ مشابه روابط ذکر شده در تقریب توزیع دو جمله‌ای نرمال است با این تفاوت که به جای مقادیر np و np(1-p) مقدار λ قرار می‌گیرد.

متغیرهای تصادفی دو بعدی (و چند بعدی)

تابع چگالی احتمال توأم متغیرهای تصادفی در حالت دو بعدی برای متغیرهای تصادفی گسسته و پیوسته به صورت زیر تعریف می‌شود.

$$\begin{cases} ۰ \le f(x,y) \le ۱ & \forall x,y \\ \sum\limits_{y}\sum\limits_{x} f(x,y) = ۱ \end{cases}$$ (برای متغیرهای گسسته)

$$\begin{cases} f(x,y) \ge ۰ & \forall x,y \\ \int\limits_{y}\int\limits_{x} f(x,y)dxdy = ۱ \end{cases}$$ (برای متغیرهای پیوسته)

همچنین امید ریاضی تابعی از متغیرهای تصادفی Y, X به صورت زیر تعریف می‌شود:

$$E[g(X,Y)] = \sum_{y}\sum_{x} g(x,y)f(x,y)\,dx\,dy$$ (برای متغیرهای گسسته)

$$E[g(X,Y)] = \int_{y}\int_{x} g(x,y)f(x,y)\,dx\,dy$$ (برای متغیر پیوسته)

توابع چگالی احتمال حاشیه‌ای Y, X نیز عبارتند از:

$f(x) = \sum\limits_{y} f(x,y)$ (برای متغیرهای گسسته) — $f(x) = \int\limits_{y} f(x,y)dy$ (برای متغیرهای پیوسته)

$f(y) = \sum\limits_{x} f(x,y)$ (برای متغیرهای گسسته) — $f(y) = \int\limits_{x} f(x,y)dx$ (برای متغیرهای پیوسته)

تابع توزیع توام Y, X عبارتست از:

$$F_{X,Y}(x,y) = p(X \le x, Y \le y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(t_۱,t_۲)\,dt_۱dt_۲$$

به این ترتیب احتمال این که متغیرهای تصادفی X, Y در ناحیهٔ دو بعدی A قرار گیرند، برابر انتگرال دو گانه تابع چگالی احتمال توأم آنها روی بازهٔ A خواهد بود.

$$p\{(x,y) \in A\} = \iint_{A} f(x,y)dxdy$$

ناهمبستگی، استقلال

$$cov(X, Y)= E(X Y)- E(X) E(Y)$$

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x\sigma_y}$$ ضریب همبستگی Y, X

در صورتی که ضریب همبستگی Y, X و در نتیجه cov(X, Y) صفر باشد، Y, X ناهمبسته هستند، بدین ترتیب شرط لازم و کافی برای ناهمبستگی متغیرهای تصادفی Y, X عبارتست از: $E[X Y]= E[X]\, E[Y]$

هم چنین اگر حاصلضرب توأم چگالی احتمال حاشیه‌ای Y, X برابر تابع چگالی توأم Y, X باشد یعنی: $f_X(x) \times f_Y(y) = f_{X,Y}(x,y)$

در این صورت متغیرهای تصادفی Y, X مستقل هستند. هم چنین در چنین حالتی $F_X(x)F_y(y) = F_{x,y}(x,y)$

نکته: در صورتی که دو متغیر تصادفی Y, X مستقل باشند، ناهمبسته نیز هستند، یعنی استقلال دو متغیر تصادفی متضمن ناهمبستگی آنها است ولی عکس این مطلب همواره صحیح نیست.

تابع چگالی احتمال شرطی

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \qquad f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$E[g(X,Y)|x] = \int_{y} g(x,y)f(y|x)dy \qquad E[g(X,Y)|y] = \int_{x} g(x,y)f(x|y)dx$$

$$E[Y|x] = \int_{y} yf(y|x)dy \qquad var[Y|x] = E\left[Y^۲|x\right] - \left[E[Y|x]\right]^۲$$

تابع چگالی احتمال تابعی از متغیرهای تصادفی

متغیر تصادفی X را با تابع چگالی احتمال $f_X(x)$ در نظر بگیرید. هم چنین فرض کنید که تابع $Y=g(X)$ موجود باشد. هدف بدست آوردن تابع چگالی احتمال $F_Y(y)$ است. برای این کار می‌بایست ابتدا $F_Y(y)$ را یافت و سپس از آن مشتق گرفت:

$$F_Y(y)=p(Y\le y)=p(g(x)\le y)=p(x\le y^{-1}(y))=\int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx$$

تحقق تساوی فوق منوط به شرایطی است: ۱) یک به یک بودن تابع $Y=g(X)$ و ۲) صعودی بودن تابع تا جهت نامساوی حفظ شود (و گرنه جهت آن عکس خواهد شد)

$$f_Y(y)=\left|\frac{dg^{-1}(y)}{dy}\right| f_X\left(g^{-1}(y)\right)$$

نکته: در صورتی که تابع $Y=g(X)$ یک به یک نباشد. با تقسیم بازهٔ تعریف آن به زیر بازه‌های یک به یک و محاسبهٔ $F_Y(y)$ در هر یک از این بازه‌ها، از رابطهٔ زیر استفاده می‌کنیم:

$$f_Y(y)=\sum_i \left|\frac{dg^{-1}(y)}{dy}\right| f_X(g^{-1}(y))$$

نکته: در حالت دو و چند بعدی نیز روند محاسبات مشابه حالت یک بعدی است. اگر W, Z توابعی از X و Y به شکل زیر باشند

$$W=g(X,Y)\qquad,\qquad Z=h(X,Y)$$

در این صورت توابع وارون W, Z عبارتند از:

$$X=g_1^{-1}(Z,W)\ ,\quad Y=h_1^{-1}(Z,W)\quad ,\quad F_{Z,W}(z,w)=|j|\,F_{X,Y}(g^{-1}(z,w)\,,\,h^{-1}(z,w))$$

که در آن $|j|$، قدر مطلق ژاکوبین Y, X نسبت به W, Z است:

$$j=\frac{\partial(x,y)}{\partial(z,w)}=\begin{vmatrix}\frac{\partial x}{\partial z} & \frac{\partial y}{\partial z}\\ \frac{\partial x}{\partial w} & \frac{\partial y}{\partial w}\end{vmatrix}$$

تابع چگالی احتمال ماکزیمم و مینیمم متغیرهای تصادفی

اگر متغیرهای تصادفی Y, X دارای تابع چگالی توأم $F(x,y)$ باشند و توابع W, Z برابر باشند با $Z=\max(X,Y)$ و $W=\min(X,Y)$.
برای یافتن چگالی احتمال W, Z ابتدا توابع توزیع آنها را یافته و سپس با مشتق‌گیری نسبت به W, Z توابع چگالی آنها پیدا می‌کنیم.

$$F_Z(z)=p(Z\le z)=p\left(\max(X,Y)\le z\right)=\int_{-\infty}^{z}\int_{-\infty}^{z} f(x,y)\,dxdy$$

$$F_W(w)=p(W\le w)=p\left(\min(X,Y)\le w\right)=1-p(\min(X,Y)\ge w)=1-\int_w^{\infty}\int_w^{\infty} f(x,y)\,dxdy$$

مثال: اگر Y, X دارای توزیع نمایی با پارامتر λ,μ باشند، توزیع $W=\min(X,Y)$ چگونه است؟

$$f(x,y)=\lambda\mu e^{-\lambda x-\mu y}\qquad x,y>0\qquad W=\min(X,Y)$$

$$F_W(w)=p(W\le w)=1-\int_w^{\infty}\int_w^{\infty}\lambda\mu e^{-\lambda x-\mu y}\,dxdy=1-e^{-(\lambda+\mu)w}\Rightarrow f(w)=(\lambda+\mu)e^{-(\lambda+\mu)w}\qquad w\ge 0$$

تابع مولد گشتاور

امید ریاضی تابع e^{tx} را طبق تعریف تابع مولد گشتاور متغیر تصادفی X می‌نامند.

$$m_X(t)=E\left[e^{tx}\right]=\sum_{i=1}^{n} e^{tx_i} f(x_i)$$

تابع مولد گشتاور در واقع گشتاورهای متغیر تصادفی X حول مبدأ، را تولید می‌کند.

قضیه: مشتق مرتبهٔ k ام تابع مولد گشتاور نسبت به t و در نقطهٔ t= 0، برابر گشتاور مرتبهٔ k ام متغیر تصادفی X حول نقطهٔ صفر است. یعنی

$$\frac{\partial^k}{\partial t^k} m_X(t)\Big|_{t=۰} = E\left[X^k\right]$$

قضیه: ضریب $\frac{t^k}{k!}$ در بسط تیلور تابع مولد گشتاور X حول نقطهٔ صفر، برابر گشتاور مرتبهٔ k ام X حول نقطهٔ صفر است، یعنی:

$$m_X(t) = \sum_{k=۰}^{\infty} E\left[X^k\right]\frac{t^k}{k!}$$

مثال: تابع مولد گشتاور متغیر تصادفی X با تابع چگالی $F(x) = ۱ - |x|$ $(-۱ \le x \le ۱)$ را محاسبه کنید:

$$m_X(t) = E\left[e^{tx}\right] = \int_{-۱}^{۱} e^{tx}(۱-|x|)dx = \int_{-۱}^{۰} e^{tx}(۱+x)dx + \int_{۰}^{۱} e^{tx}(۱-x)dx = \frac{۲(\cosh t - ۱)}{t^۲}$$

تابع مولد گشتاور برخی از توزیعهای خاص در جدول زیرآورده شده است:

توزیعای گسسته	تابع مولد گشتاور
برنولی با پارامتر p:	$(pe^t + ۱ - p)$
دوجمله ای با پارامتر (n,p):	$(pe^t + ۱ - p)^n$
هندسی با پارامتر p:	$\frac{pe^t}{۱-(۱-p)e^t}$
دوجمله ای منفی با پارامتر (r,p):	$\left[\frac{pe^t}{۱-(۱-p)e^t}\right]^r$
پواسون با پارامتر λ:	$e^{\lambda(e^t - ۱)}$

توزیعهای پیوسته	تابع مولد گشتاور
یکنواخت در بازه (a,b):	$\frac{e^{tb} - e^{ta}}{t(b-a)}$
نمایی با پارامتر λ:	$\frac{\lambda}{\lambda - t}$
گاما با پارامترهای (λ, α):	$\left[\frac{\lambda}{\lambda - t}\right]^{\alpha}$
نرمال با پارامتر های (μ, σ):	$e^{\mu t + \frac{\sigma^۲ t^۲}{۲}}$

آمار و احتمالات

آمار توصیفی

آمار توصیفی اولین بخش علم آمار را تشکیل می‌دهد. در آمار توصیفی اطلاعات جمع‌آوری شده درباره اعضای جامعه به صورت منظم گردآوری می‌شود تا بتوان درک شهودی از صفت یا صفات مورد نظر افراد جامعه به دست آورد.

در ادامه برخی از مشخصات اولیه جامعه آماری که در آمار توصیفی کاربرد دارد توضیح داده خواهد شد.

۱) میانگین حسابی

$$\bar{X} = \bar{X}_A = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

که در آن f_i ها فراوانی هر یک از دسته‌های x_i هستند.

۲) میانگین هندسی

$$\bar{X}_G = \left(\prod_{i=1}^{n} x_i^{f_i}\right)^{\frac{1}{n}} \qquad n = \sum_{i=1}^{n} f_i$$

۳) میانگین همنوا یا هارمونیک

$$\bar{X}_H = \frac{n}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

نکته: همواره $\bar{x}_H \le \bar{x}_G \le \bar{x}_A$

۴) مد: از میان X_iها دسته‌ای که دارای بیشترین فراوانی (f_i) باشد مد نامیده می‌شود.

۵) میانه: کوچکترین مقدار x که فراوانی تجمعی آن بیشتر یا مساوی $\frac{N}{۲}$ باشد میانه است.

۶) چولگی: میزان تمایل یک منحنی به سمت راست یا چپ چولگی نامیده می‌شود. محاسبه میزان چولگی هر منحنی با استفاده از فرمول زیر صورت می‌گیرد.

$$\text{kewness} = \frac{\frac{\sum (x_i - \bar{x})^3}{n}}{S^3}$$

۷) کشیدگی: میزان کشیدگی منحنی هر منحنی با استفاده از فرمول زیر قابل محاسبه است.

$$\text{Cartusis} = \frac{\frac{\sum (\bar{x}_i - \bar{x})^3}{n}}{S^4}$$

۸) واریانس: میانگین مربع انحرافات داده‌ها از میانگین، واریانس نامیده می‌شود که یک میزان برای سنجش میزان پراکندگی داده‌ها نیز هست. محاسبه واریانس از طریق رابطه زیر صورت می‌گیرد.

$$S^2 = \frac{1}{n-۱}\sum f_i (x_i - \bar{x})^2 = \frac{1}{n-۱}\sum f_i x_i^2 - n\bar{x}^2$$

آمار استنتاجی

در آمار استنتاجی برخلاف آمار توصیفی می‌توان با کمک متغیرها در مورد کل جامعه تصمیم گرفت. برخی از تعاریف که در آمار استنتاجی کاربرد دارند عبارتند از:

پارامتر: پارامتر اندازه‌ای است که خصوصیت خاصی را در جامعه بیان می‌کند، به عنوان مثال میانگین جامعه (μ)، واریانس جامعه $(\sigma^۲)$ و غیره.

نمونه تصادفی: در علم آمار همواره با دو نوع نمونه‌گیری مواجه هستیم: ۱) نمونه‌گیری از جامعه محدود و ۲) نمونه‌گیری از جامعه نامحدود. در صورتی که نمونه‌گیری از جامعه محدود (نمونه‌گیری بدون جایگذاری) انجام شود، نمونه‌گیری تصادفی نخواهد بود اما در صورتی که نمونه‌گیری ازجامعه نامحدود (نمونه‌گیری با جایگذاری) صورت گیرد، نمونه‌گیری به صورت تصادفی خواهد بود.

بر این اساس نمونه تصادفی، نمونه‌ای است که عناصر آن متغیرهای تصادفی (متغیرهای هم توزیع و مستقل) باشند.

می توان نشان داد از آنجا که اعضای نمونه تصادفی مستقل و هم توزیع هستند، تابع چگالی توأم آنها حاصل‌ضرب تک‌تک توابع چگالی آنها است.
$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n} f(x_i)$$

آماره: جهت بدست آوردن تخمینی از پارامترهای جامعه لازم است تا تابعی بر روی نمونه تصادفی تعریف شود، به این تابع آماره گویند، به عنوان مثال $\overline{X}$, $S^۲$ دو آماره اصلی هستند که بر روی نمونه تصادفی تعریف می شوند.

$$\overline{X}=\frac{\sum_{i=۱}^{n} X_i}{n} \qquad ,S^۲=\frac{\sum_{i=۱}^{n}\left(X_i-\overline{X}\right)^۲}{n-۱}$$

توزیع‌های نمونه‌ای

از آنجا که آماره‌ها به صورت تابعی روی متغیرهای تصادفی تعریف می‌شوند، خود نیز متغیر تصادفی هستند و دارای توزیع می‌باشند، به توزیع آماره‌ها، توزیع‌های نمونه‌ای گفته می‌شود.

انواع توزیع‌های نمونه‌ای معروف عبارتند از:

- توزیع کامی اسکور $(X^۲)$
- توزیع t
- توزیع F

نکته: در صورتی که X_iها یک نمونه تصادفی از جامعه‌ای با میانگین μ واریانس σ^2 باشند آنگاه داریم:

$$E(\overline{x})=\mu \quad , \quad var(\overline{x})=\frac{\sigma^2}{n} \quad , \quad E(s^2)=\sigma^2 \quad , \quad var(s^2)=\frac{1}{n}\left(\mu_4-\frac{n-3}{n-1}\sigma^4\right)$$

که μ_4، گشتاور مرکزی مرتبه چهارم جامعه است.

نکته: در صورتی که نمونه‌گیری از یک جامعه محدود (یا نمونه‌گیری بدون جایگذاری) صورت گیرد برای محاسبه واریانس میانگین نمونه از رابطه زیر استفاده می‌کنیم:
$$var(\overline{x})=\frac{\sigma^2}{n}\times\frac{N-n}{N-1}$$

که N حجم جامعه و n حجم نمونه است.

قضیه حد مرکزی:

اگر $X_1,\dots,X_n$ نمونه‌ای تصادفی از یک جامعه نامحدود با میانگین μ و واریانس σ^2 باشند، آنگاه اگر تعداد اعضای نمونه بالا باشد متغیر تصادفی $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ دارای توزیع نرمال استاندارد است. به عبارت دیگر براساس قضیه حد مرکزی صرف نظر از توزیع X_iها، میانگین نمونه‌های

مختلف بیشتر به سمت میانگین جامعه گرایش دارند و هر چه در هر یک از دو طرف از میانگین جامعه دور شویم، احتمال اینکه میانگین نمونه‌ای که از جامعه می‌گیریم فاصله زیادی تا میانگین داشته باشد، کاهش می‌یابد.

نکته: منظور از اندازه نمونه بالا برای استفاده از قضیه حد مرکزی $n>30$ است. که این مقدار بستگی به نوع توزیع X_iها نیز دارد، مثلا اگر X_i دارای توزیع نرمال باشند هیچ قیدی برای n وجود ندارد.

توزیع کای اسکور $\left(X^2\right)$

تابع چگالی متغیر تصادفی X^2 با k درجه آزادی برابر است با:

$$f_x(x)=\frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}x^{\left(\frac{k}{2}-1\right)}e^{-\frac{1}{2}x}}{\Gamma\left(k/2\right)}$$

بنابراین توزیع X^2، توزیع گاما دارد با پارامترهای $\left(\lambda=\frac{1}{2},\alpha=\frac{k}{2}\right)$

قضیه: اگر $X_1,X_2,\ldots,X_n$ متغیرهای تصادفی مستقل نرمال استاندارد باشند، در این صورت متغیر تصادفی $y=x_1^2+x_2^2+\cdots+x_n^2$ دارای توزیع X^2 با n درجه آزادی است.

نکته: با توجه به توزیع X^2 با n درجه آزادی، امید ریاضی واریانس آن برابر است با: $E(x)=n \qquad var(x)=2n$

نکته: اگر x یک متغیر تصادفی نرمال استاندارد باشد توزیع متغیر تصادفی $Y=x^2$، $X_{(1)}^2$ (کای اسکور با ۱ درجه آزادی) است.

قضیه: اگر $X_1,X_2,\ldots,X_n$ متغیرهای تصادفی مستقل دارای توزیع کای اسکور، هر کدام با $n_1,n_2,\ldots,n_n$ درجه آزادی باشند توزیع $Y=\sum_{i=1}^{n}X_i$، کای اسکور است با $k=\sum_{i=1}^{n}n_i$ درجه آزادی.

نکته: طبق قضیه حد مرکزی وقتی متغیر تصادفی x توزیع $X_{(n)}^2$، درجه آزادی $n>30$ داشته باشد، توزیع آن به نرمال با $\left(\mu=n \quad , \quad \sigma=\sqrt{2n}\right)$ میل می‌کند.

نکته: در صورتی که در توزیع $X_{(n)}^2$، n زوج باشد، X^2 دارای توزیع ارلنگ خواهد بود.

نکته: در توزیع متغیر تصادفی $X_{(n)}^2$، اگر $n=2$ باشد، توزیع نمایی با پارامتر $\lambda=\frac{1}{2}$ خواهد شد.

نکته: اگر متغیر تصادفی X_1 دارای توزیع کای اسکور با n_1 درجه آزادی باشد، متغیر تصادفی $Y=X_1+X_2$ دارای توزیع کای اسکور با n_1+n_2 درجه آزادی باشد، توزیع X_2 حتما کای اسکور با n_2 درجه آزادی بوده است.

توزیع t

اگر x یک متغیر تصادفی با توزیع t و n درجه آزادی باشد آنگاه تابع چگالی احتمال آن برابر است با:

$$f_x(x)=\frac{1}{\sqrt{\pi n}}\times\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\qquad -\infty<x<\infty$$

$$E(x)=0\quad,\quad var(x)=\frac{n}{n-2}\quad n>2$$

قضیه: اگر Z,Y متغیرهای تصادفی مستقل با توزیع‌های نرمال استاندارد و کای اسکور با n درجه آزادی باشند آنگاه توزیع متغیر تصادفی t با n درجه آزادی به صورت زیر تعریف خواهد شد:

$$t_{(n)}=\frac{Y}{\sqrt{\frac{Z}{n}}}$$

نکته: طبق قضیه حد مرکزی توزیع $t_{(n)}$ برای $n>۳۰$ به توزیع نرمال استاندارد میل می‌کند.

نکته: Z_2, Z_1 متغیرهای تصادفی مستقل با توزیع نرمال استاندارد باشند آنگاه متغیرهای تصادفی $X=\frac{|z_1|}{z_2}\ ,\ W=\frac{z_1}{|z_2|}$ دارای توزیع $t_{(1)}$ هستند.

نکته: توزیع t یک توزیع متقارن است لذا داریم $t_{(1-\alpha),n}=-t_{\alpha,n}$ ($t_{\alpha,n}$، نقطه از توزیع $t_{(n)}$ است که احتمال سمت راست آن α باشد)

نکته: تابع چگالی متغیر تصادفی X با توزیع t_1 از رابطه زیر به دست می‌آید.

$$f_x(x)=\frac{1}{\pi}\times\frac{1}{1+x^2}$$

توزیع F

اگر x یک متغیر تصادفی با توزیع $F_{m,n}$، (F با m و n درجه آزادی) باشد آنگاه تابع چگالی احتمال آن برابر است با:

$$f_x(x)=\frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{2}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\times x^{\left(\frac{m}{2}-1\right)}\times\left(1+\frac{m}{n}x\right)^{\left(-\frac{m+n}{2}\right)}\quad x>0\qquad,\qquad E(x)=\frac{n}{n-2}\quad n>2\qquad var(x)=2\left(\frac{1}{n}+\frac{1}{m}\right)$$

قضیه: اگر V,U متغیرهای تصادفی مستقل توزیع $X^۲$ به ترتیب با n,m درجه آزادی باشند آنگاه متغیر تصادفی Y به صورت $Y=\frac{U/m}{V/n}$ دارای توزیع $F_{m,n}$ است.

چند نکته:

– اگر X یک متغیر تصادفی با توزیع $t_{(n)}$ باشد آنگاه توزیع $Y=X^2$ عبارتست از: $F_{1,n}$

– اگر نقطه‌ای در توزیع $F_{m,n}$ که احتمال راست آن α باشد را $F_{\alpha,m,n}$ بنامیم داریم: $F_{1-\alpha,n,m}=\frac{1}{F_{\alpha,m,n}}$

– توزیع F چوله به راست است لذا داریم: $mod<\tilde{x}<\bar{x}$

– میانه توزیع $F_{n,n}$ برابر با ۱ است.

مثال: اگر x متغیر تصادفی دارای توزیع $F_{10,10}$ باشد احتمال اینکه x مقادیر کوچکتر از یک را بگیرد چقدر است؟

حل: با توجه به اینکه میانه توزیع $F_{n,n}$، ۱ است، داریم $P(x<1)=\frac{1}{2}$

چند نکته:

– اگر S^2 واریانس یک نمونه تصادفی nتایی از یک جامعه با توزیع نرمال با میانگین μ و واریانس σ^2 باشد آنگاه داریم:

$$\frac{(n-1)s^2}{\sigma^2}\sim X^2_{(n-1)}$$

همچنین متغیر تصادفی S^2 توزیع گاما با پارامترهای $\left(\alpha=\frac{n-1}{2},\lambda=\frac{n-1}{2\sigma^2}\right)$ خواهد داشت.

– تحت شرایط بالا $\text{var}\left(s^2\right)=\frac{2\sigma^4}{n-1}$, $E=\left(s^2\right)=\sigma^2$ (لازم به ذکر است صرف نظر از نوع توزیع جامعه همواره $E\left(s^2\right)=\sigma^2$ است)

– همچنین اگر $x_1,x_2,\dots,x_n$ اعضای نمونه تصادفی باشند متغیر تصادفی $\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{\sigma^2}$ ($\bar{x}$ میانگین نمونه) دارای توزیع $X^2_{(n-1)}$ است. و متغیر تصادفی $\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}$ دارای توزیع $X^2_{(n)}$ است.

– متغیر تصادفی $\frac{\bar{x}-\mu}{S/\sqrt{n}}$ دارای توزیع $t_{(n-1)}$ است.

– اگر S_1^2,S_2^2 واریانس‌های دو نمونه تصادفی به حجم n_1 و n_2 باشند آنگاه متغیر تصادفی $F=\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}$ دارای توزیع F با n_1-1 و n_2-1 درجه آزادی خواهد بود.

******

برآورد

اگر جامعه‌ای متشکل از عناصر متعدد در اختیار داشته باشیم و بخواهیم در مورد صفت خاصی از جامعه تحقیق کنیم، بدیهی است دستیابی به تمام عناصر جامعه، اندازه و صفت مورد نظر آنها غیر ممکن است. بنابراین جهت این کار باید توزیع نمونه تصادفی (یا به عبارت دیگر توزیع تعدادی از افراد جامعه) را به دست آورد و از روی آن در مورد پارامترهای جامعه تصمیم‌گیری کرد یا پارامترهای جامعه را برآورد نمود.

پارامترهای نمونه (مانند $\bar{X},S^2$) را می‌توان به دو صورت زیر برای برآورد پارامترهای جامعه به کار برد.

۱) **برآورد نقطه‌ای:** مقدار به دست آمده برای یک پارامتر نمونه را مساوی مقدار پارامتر جامعه قرار دهیم. یعنی از پارامتر معلوم نمونه را برابر پارامتر نامعلوم جامعه قرار دهیم.

۲) **برآورد فاصله‌ای:** در این روش حول پارامتر معلوم نمونه یک فاصله تعیین می‌کنیم و ادعا می‌کنیم که با احتمالی مشخص پارامتر جامعه در این فاصله قرار دارد.

نکته: هر دو برآورد دارای میزان خطایی خواهند بود.

خواص برآورد کننده

اگر برای برآورد پارامتر θ از جامعه‌ای، برآورد کننده $\hat{\theta}$ را که با استفاده از نمونه تصادفی به دست امده، به کار بریم، برآورد کننده θ می‌تواند دارای خصوصیات زیر باشد:

۱) **نااریبی**: با توجه به اینکه مقدار $\hat{\theta}$ دقیقا با θ برابر نیست، محاسبه $\hat{\theta}$ باید به‌گونه‌ای باشد که با محاسبه $\hat{\theta}$های مختلف از روی نمونه‌های مختلف و میانگین گرفتن از آنها مقدار θ به دست آید به بیان دیگر باید امید ریاضی متغیر تصادفی $\hat{\theta}$ با θ برابر باشد. این خاصیت نااریبی گفته می‌شود $\left(E\left(\hat{\theta}\right)=\theta\right)$

بنابراین برای پارامترهای $\bar{X},S^2$ که روی نمونه تصادفی تعریف می‌شوند همواره داریم: $E(\bar{x})=\mu\ ,\ E\left(s^2\right)=\sigma^2$

۲) سازگاری: $\hat{\theta}$ را یک برآورد کننده سازگار θ گوییم هرگاه داشته باشیم: $E(\hat{\theta})=\theta$, 2) $\lim var(\hat{\theta})=0$

برای مثال چون $E(\bar{x})=\mu$, $\lim\limits_{x\to\infty} var(\bar{x})=\lim\limits_{x\to\infty} \sigma^2/n=0$ پس $\bar{X}$ یک برآورد کننده سازگار μ است.

۳) **کفایت**: $\hat{\theta}$ را یک برآورد کننده کافی θ می‌نامیم، اگر و فقط اگر تابع چگالی احتمال توأم نمونه‌ها را بتوان به صورت زیر تجزیه نمود.

$$f(x_1,x_2,\cdots,x_n,\theta)=g(\hat{\theta},\theta)\times h(x_1,x_2,\cdots,x_n)$$

٤) **کارایی**: $\hat{\theta}$ را یک برآورد کننده کارا θ می‌نامیم هرگاه واریانس کوچکتری نسبت به سایر برآورد کننده‌ها داشته باشد.

تعریف: برای دو برآورد کننده $\hat{\theta}_2,\hat{\theta}_1$ معیار کارایی نسبی برآورد کننده $\hat{\theta}_2$ نسبت به $\hat{\theta}_1$ به صورت $\dfrac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$ تعریف می‌شود.که در آن

MSE میانگین مربع خطا برای تخمین زننده است و به صوزت زیر تعریف می شود: $MSE(\hat{\theta})=E\left[(\hat{\theta}-\theta)^2\right]=var(\hat{\theta})+\left[E(\hat{\theta})-\theta\right]^2$

مثال: اگر یک از جامعه‌ای، دو نمونه تصادفی یکی به حجم ۱۰ و میانگین ۱۲ و انحراف استاندارد ۲ و دیگری به حجم ۳۰، میانگین ۱۰ و انحراف استاندارد ۳ گرفته باشیم کدام نمونه تصادفی برآورد کاراتری از میانگین جامعه به ما می‌دهد؟

$$\bar{x}_1 \text{ به } \bar{x}_2 \text{ کارایی } =\frac{var(\bar{x}_1)}{var(\bar{x}_2)}=\frac{\frac{\sigma_1^2}{n_1}}{\frac{\sigma_2^2}{n_2}}=\frac{\frac{4}{10}}{\frac{9}{3}}=\frac{4}{3}$$

بنابراین نمونه دوم از نمونه اول کاراتر است.

نکته: اگر $\hat{\theta}$ یک برآورد کننده نا اریب باشدو از میان کلیه برآورد کننده دارای واریانس کمتری باشد (کارایی بیشتری داشته) آنگاه $\hat{\theta}$ بهترین برآورد کننده θ خواهد بود.

برآورد نقطه‌ای

برای به دست آوردن یک برآورد نقطه‌ای از پارامترهای دلخواه جامعه دو روش اصلی وجود دارد:

۱) **برآورد به روش گشتاورها**: اگر $x_1,x_2,\ldots,x_n$ نمونه‌ای تصادفی از یک جامعه باشند آنگاه k امین گشتاور نمونه‌ای این مشاهدات میانگین توان kام آنها است یعنی داریم:

$$(m'_k) \text{ امین گشتاور نمونه } k=\frac{1}{n}\sum_{i=1}^{n}x_i^k$$

اگر تابع چگالی احتمال متغیر X در جامعه دارای پارامترهای محصول $\theta_p,\cdots,\theta_2,\theta_1$ و به شکل $f(x,\theta_1,\cdots,\theta_p)$ باشد.

گشتاورهای مرتبه kام جامعه عبارتند از: $\mu'_k=E(x^k)=\int_x x^k f(\theta_1,\theta_2,\cdots,\theta_p)dx=g_k(\theta_1,\cdots,\theta_p)$

بدین ترتیب با حل دستگاه معادلات زیر می‌توان مقدار $\hat{\theta}_i$ را به عنوان برآورد نقطه $\theta_p,\cdots,\theta_1$ به دست آورد.

$$m'_k=\mu'_k=g_k(\theta_1,\cdots,\theta_p) \qquad k=1,\ldots P$$

مثال: نمونه تصادفی ۰/۵، ۰/۳، ۰/٤، ۰/۸ را از توزیع $f(x)=mx^{m-1}$, $0<x<1$ مشاهده کرده‌ایم، برآورد گشتاوری m را به دست آورید.

$$E(x)=\mu=\frac{\sum xi}{4}=\int_0^1 xmx^{m-1}dx \Rightarrow \frac{0/8+0/4+0/3+0/5}{4}=\frac{m}{m+1}x^{m+1}\Big|_0^1 \Rightarrow m=1$$

۲) برآورد به روش درستنمایی ماکزیمم: اگر نمونه تصادفی $x_1,...,x_n$ از یک جامعه‌ای استخراج شده باشند، برای به دست آوردن تابع چگالی احتمال توأم $f(x,\theta_1,\cdots,\theta_p)$ می‌توان مقادیری از θ_iها را انتخاب نمود که به ازای آنها مقدار تابع چگالی توأم x_iها ماکزیمم شود. در واقع می‌توان نمونه‌های به دست آمده را محتمل‌ترین پیشامد استخراج n نمونه دانست.

در این حالت تابع درستنمایی این نمونه‌ها عبارتست از: $L(\theta_1,\cdots\theta_p)=f(x_1,x_2,\cdots,x_n,\theta_1,\cdots,\theta_p)$

حال با مشتق‌گیری از تابع L برحسب θ_iها مساوی صفر قرار دادن آنها و حل معادلات به دست آمده $\hat{\theta}_i$ را محاسبه می‌نماییم (می‌توان به جای مشتق‌گیری از تابع L، از $\ln(L)$ نیز مشتق گرفت)

قضیه: اگر $g(\theta)$ تابعی یک به یک از θ باشد، برآورد درستنمایی ماکزیمم برای $g(\theta)$ برابر $g(\hat{\theta})$ خواهد بود.

مثال: متغیر تصادفی x با توزیع احتمال $f(x)=\theta(1-\theta)^x, x=0,1,2,\ldots$ مفروض است $(0<\theta<1)$. اگر یک نمونه تصادفی n تایی از این جامعه انتخاب کنیم در این صورت برآورد کننده ماکزیمم درستنمایی (MLE) برای θ و $E(X)$ را به دست آورید.

$$L(\theta)=f(x_1,x_2,\cdots,x_n,\theta)=\theta^n x(1-\theta)^{\sum xi} \quad , \quad \ln[L(\theta)]=n\times\ln(\theta)+\sum x_i\times\ln(1-\theta)$$

$$\frac{\sigma\ln(\theta)}{\sigma\theta}=\frac{n}{\theta}-\frac{\sum x_i}{1-\theta}=0\Rightarrow\frac{n}{\theta}=\frac{\sum x_i}{1-\theta}\Rightarrow\theta=\frac{1}{1+\overline{X}}$$

$$\overline{X}=\frac{1-\theta}{\theta}$$

برآورد فاصله‌ای

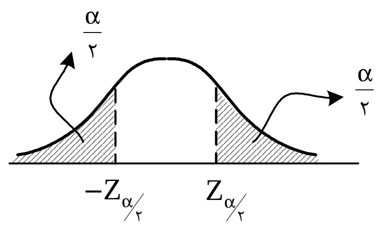
برآورد فاصله‌ای پارمتری مانند θ از جامعه عبارتست از فاصله‌ای حول پارامتر $\hat{\theta}$ که به احتمالی معین شامل θ است.

$$P(\hat{\theta}-A<\theta<\hat{\theta}+A)=1-\alpha$$

بدیهی است که مقدار A تابعی از α است و هر چه α کوچکتر باشد، A بزرگتر خواهد بود. فاصله اطمینان همواره به صورت $[\hat{\theta}-A,\hat{\theta}+A]$ نیست. ساختن فاصله اطمینان به توزیع احتمالی $\hat{\theta}$ بستگی دارد.

فاصله اطمینان برای میانگین جامعه نرمال با واریانس معلوم

اگر متغیر تصادفی x دارای توزیع $N(\mu,\sigma)^2$ با μ نامعلوم و σ^2 معلوم باشد، متغیر تصادفی $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ دارای توزیع نرمال $N(0,1)$ است و داریم:

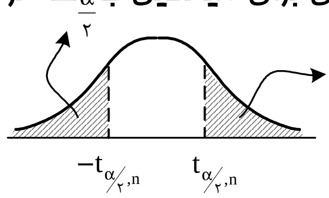


$$p\left[-Z_{\alpha/2}\le\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\le Z_{\alpha/2}\right]=1-\alpha\Rightarrow P\left[\overline{x}-Z_{\alpha/2}\times\frac{\sigma}{\sqrt{n}}\le\mu\le\overline{x}+Z_{\alpha/2}\times\frac{\sigma}{\sqrt{n}}\right]=1-\alpha$$

لازم به ذکر است در صورتی که جامعه دارای توزیع نرمال نباشد ولی اندازه نمونه (n) بالا باشد، نیز می‌توان رابطه فوق را استفاده کرد. به طور کلی اگر دو شرط از سه شرط، (معلوم بودن واریانس جامعه)، (نرمال بودن جامعه)، (بزرگ بودن اندازه نمونه) برقرار باشد آنگاه می‌توان از رابطه فوق استفاده کرد.

فاصله اطمینان برای میانگین جامعه نرمال با واریانس مجهول با تعداد نمونه کم

$$P\left[\overline{x}-t_{\alpha/2,(n-1)}\times\frac{S}{\sqrt{n}}\le\mu\le\overline{x}+t_{\alpha/2,(n-1)}\times\frac{S}{\sqrt{n}}\right]=1-\alpha$$



فاصله اطمینان برای تفاضل میانگین‌های دو جامعه نرمال با واریانس‌های معلوم

$$P\left[(\bar{x}_2-\bar{x}_2)-Z_{\alpha/2}\left[\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right]^{\frac{1}{2}}\le\mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+Z_{\alpha/2}\left[\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right]^{\frac{1}{2}}\right]=1-\alpha$$

که در آن $\bar{x}_2,\bar{x}_1$ میانگین نمونه‌های دو جامعه، n_2,n_1 اندازه نمونه دو جامعه و σ_2^2,σ_1^2 واریانس‌های دو جامعه هستند.

فاصله اطمینان برای تفاضل میانگین‌های دو جامعه نرمال با واریانس‌های مجهول و برابر $\left(\sigma_1^2=\sigma_2^2\right)$

$$P\left[(\bar{x}_1-\bar{x}_2)-t_{\alpha/2,n+m-2}\times S_P\left[\frac{1}{n}+\frac{1}{m}\right]^{\frac{1}{2}}\le\mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)+t_{\alpha/2,n+m-2}\times S_P\left[\frac{1}{n}+\frac{1}{m}\right]^{\frac{1}{2}}\right]=1-\alpha$$

که در آن m,n اندازه نمونه‌های دو جامعه و S_p^2 واریانس‌های مشترک دو جامعه به صورت زیر است:

$$S_P^2=\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$$

فاصله اطمینان برای نسبت P با تعداد نمونه زیاد

اگر بخواهیم نسبت افراد به خصوصی را در کل جامعه محاسبه کنیم. اگر از جامعه نمونه n تایی بگیریم که x تای آن دارای چنین خصوصیتی باشند نسبت نمونه مورد نظر برابر $\hat{P}=\frac{x}{n}$ خواهد شد.

$$P\left[\hat{P}-z_{\alpha/2}\left[\frac{\hat{P}(1-\hat{P})}{n}\right]^{-\frac{1}{2}}\le P\le\hat{P}+z_{\alpha/2}\left[\frac{\hat{P}(1-\hat{P})}{n}\right]^{-\frac{1}{2}}\right]=1-\alpha$$

فاصله اطمینان برای تفاضل نسبت‌های دو جامعه با تعداد نمونه‌های زیاد

$$P\left[\hat{P}_1-\hat{P}_2-Z_{\alpha/2}\left(\frac{\hat{P}_1(1-\hat{P})}{n_1}+\frac{\hat{P}_2(1-\hat{P}_2)}{n_2}\right)^{\frac{1}{2}}\le P_1-P_2\le\hat{P}_1-\hat{P}_2+Z_{\alpha/2}\left(\frac{\hat{P}_1(1-\hat{P})}{n_1}+\frac{\hat{P}_2(1-\hat{P}_2)}{n_2}\right)^{\frac{1}{2}}\right]=1-\alpha$$

فاصله اطمینان برای واریانس جامعه نرمال

۱) در صورتی که میانگین جامعه نامعلوم باشد:

$$P\left[\frac{(n-1)S^2}{X^2_{\alpha/2,(n-1)}}\le\sigma^2\le\frac{(n-1)S^2}{X^2_{1-\alpha/2,(n-1)}}\right]=1-\alpha$$

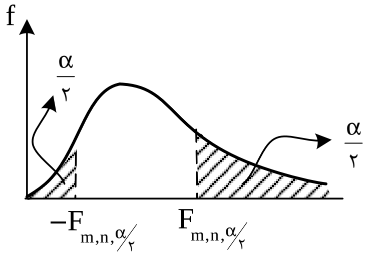
۲)در صورتی که میانگین جامعه معلوم باشد:

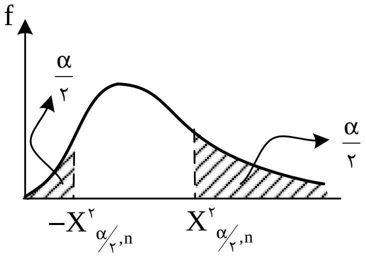
$$P\left[\frac{nS^2}{X^2_{\alpha/2,n}}\le\sigma^2\le\frac{nS^2}{X^2_{1-\alpha/2,n}}\right]=1-\alpha$$

فاصله اطمینان برای نسبت واریانس‌های دو جامعه نرمال

$$P\left[\frac{S_1^2}{S_2^2}\times F_{n-1,m-1,1-\alpha/2}\le\frac{\sigma_1^2}{\sigma_2^2}\le\frac{S_1^2}{S_2^2}\times F_{n-1,m-1,\alpha/2}\right]=1-\alpha$$

که در آن S_1^2,S_2^2 واریانسهای نمونه های تصادفی مستقل به اندازه n, m از دو جامعه نرمال هستند.





آمار و احتمالات

آزمون فرض

فرضیه آماری عبارتست از فرض یا حکمی که درباره یک یا چند پارامتر نامعلوم از یک جامعه و یا، یک گروه از جوامع، مطرح می‌شود که بر اساس آن می‌توان مقایسه‌هایی در مورد پارامتر مورد نظر صورت داد.

هر فرضیه آماری ممکن است درست یا غلط باشد لذا دو فرض در آزمون فرض همواره مقابل هم قرار می‌گیرند. معمولا فرضیه‌ای که در انتظار رد آن هستیم «فرض صفر» یا «فرض خنثی» نام دارد که با H_0 آنرا نمایش می‌دهند و در مقابل فرضیه دیگری نیز وجود دارد که در صورت رد H_0 آن را می‌پذیریم، که به آن «فرض مقابل» گفته می‌شود و آنرا با H_1 نمایش می‌دهیم.

هیچ الزامی وجود ندارد که رد فرض صفر به معنای نادرستی آن باشد لذا هنگامی که دلایل کافی و لازم برای پذیرش فرض صفر وجود ندارد ما آن را رد می‌کنیم و بر عکس.

انواع فرضیه‌های آماری

فرض ساده: اگر فرضیه‌ای به طور کامل توزیع یک جامعه را مشخص کند به آن فرض ساده گوییم، مانند $H_0 : p = ۰/۵$ در مثال قبل پارامتر توزیع برنولی به طور کامل معین است لذا، توزیع متغیر تصادفی انتخاب شده از جامعه کاملا مشخص است.

فرض مرکب: اگر فرضیه‌ای به طور کامل توزیع یک جامعه را مشخص نکند به آن فرض مرکب گوییم، مانند $H_0 : p \neq ۰/۵$

انواع خطا در آزمون فرض

به جهت اینکه در آزمون فرض ما براساس نمونه‌ای از یک جامعه در مورد پارامتری از آن جامعه (یا به طور کلی در مورد توزیع جامعه) تصمیم‌گیری می‌کنیم، لذا همواره مقدار خطایی در انجام آزمون وجود خواهد داشت.

تصمیم‌گیری در مورد جامعه بر مبنای نمونه تصادفی به چهار نتیجه منتهی می‌شود که در جدول زیر مشخص شده است:

وضعیت پارامتردر حالت واقعی / نتیجه آزمون	H_0 واقعا صحیح است	H_1 واقعا صحیح است
رد H_0	α	$۱-\beta$
پذیرش H_0	$۱-\alpha$	β

براساس جدول فوق همواره با دو نوع خطا مواجه هستیم:

خطای نوع اول: حالتی که فرض صفر صحیح باشد ولی ما آنرا بپذیریم.

$\alpha = $ احتمال خطای نوع I $= P($رد H_0 | H_0 درست است$)$

خطای نوع دوم: حالتی که فرض صفر صحیح نباشد (فرض مقابل صحیح باشد) و ما آنرا نپذیریم.

$\beta = $ احتمال خطای نوع II $= P($پذیرش H_0 | H_0 صحیح نیست$)$

نکته: به خطای نوع اول اصطلاحا «ریسک تولید کننده» و به خطای نوع دوم «ریسک مصرف کننده» نیز گویند.

مراحل انجام آزمون فرض

۱) مشخص نمودن فرضیه‌ها

۲) تعریف یک آماره روی نمونه مشاهده شده.

تصمیم در مورد رد یا پذیرش یک فرضیه بستگی به آماره تعریف شده براساس آن فرضیه دارد. آماره هر آزمون فرض معمولاً به صورت زیر تعریف می‌شود.

$$\text{آماره آزمون} = \frac{\text{بر آورده کننده پارامتر}\ \theta - \text{مقدار پارامتر براساس فرض صفر}}{\text{خطای معیار}\ \theta}$$

نکته: انتخاب آماره آزمون فرض براساس فرض صفر شکل می‌گیرد.

۳) مشخص نمودن ناحیه پذیرش (یا رد) براساس احتمال مورد نظر α

ناحیه پذیرش مجموعه‌ای از مقادیر آماره است که به ازای آن H_0 باید قبول شود.

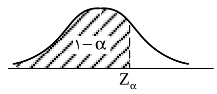
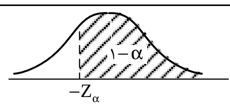
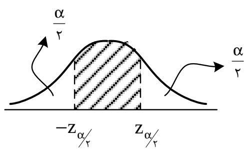
٤) محاسبه مقدار آماره آزمون

۵) نتیجه‌گیری

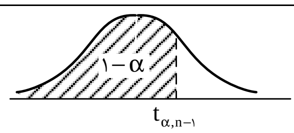
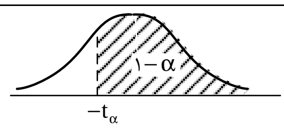
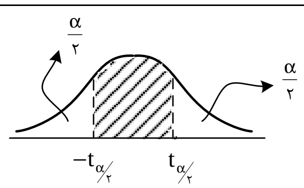
اگر مقدار محاسبه شده آماره آزمون در ناحیه رد قرار گیرد فرض H_0 را به نفع فرض H_1 رد می‌کنیم.

آزمون‌های میانگین

۱) حالتی که σ معلوم باشد:

ناحیه پذیرش	ناحیه رد	آماره		
	$Z > Z_\alpha$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	آزمون یکطرفه
	$Z < -Z_\alpha$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	
	$Z < -Z_{\alpha/2}$ و $Z > Z_{\alpha/2}$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	آزمون دو طرفه

۲) حالتی که σ نامعلوم باشد:

ناحیه پذیرش	ناحیه رد	آماره		
	$t > t_\alpha$	$t_{(n-1)} = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	آزمون یکطرفه
	$t < -t_\alpha$	$t_{(n-1)} = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	
	$t < -t_{\alpha/2}$ و $t > t_{\alpha/2}$	$t_{(n-1)} = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	آزمون دو طرفه

آزمون مربوط به اختلاف میانگین دو جامعه

۱) حالتی که واریانس جوامع معلوم است:

ناحیه رد	آماره		
$Z > Z_\alpha$	$z=\dfrac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1>\mu_2$	آزمون یکطرفه
$Z < -Z_\alpha$	$z=\dfrac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1<\mu_2$	آزمون یکطرفه
$Z > Z_{\alpha/2}$ و $Z < -Z_{\alpha/2}$	$z=\dfrac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1\neq\mu_2$	آزمون دو طرفه

۲) حالتی که واریانس جوامع نامعلوم است:

الف) حالتی که واریانس جوامع نامعلوم ولی مساوی باشد.

$$S_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

ناحیه رد	آماره		
$t_{(n_1+n_2-2)} > t_\alpha$	$t=\dfrac{\bar{x}-\bar{Y}}{sp\sqrt{1/n_1+1/n_2}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1>\mu_2$	آزمون یکطرفه
$t < -t_\alpha$	$t=\dfrac{\bar{x}-\bar{Y}}{sp\sqrt{1/n_1+1/n_2}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1<\mu_2$	آزمون یکطرفه
$t > t_{\alpha/2}$ و $t < -t_{\alpha/2}$	$t=\dfrac{\bar{x}-\bar{Y}}{sp\sqrt{1/n_1+1/n_2}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1\neq\mu_2$	آزمون دو طرفه

ب) حالتی که واریانس جوامع نامعلوم باشد و لزوما هم واریانس‌ها با هم برابر نباشند.

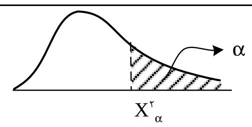
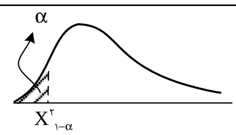
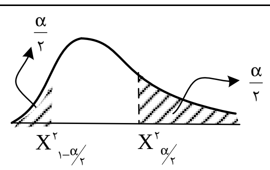
ناحیه رد	آماره		
$t'_{(v)} > t_\alpha$	$t'=\dfrac{\bar{x}-\bar{Y}}{\sqrt{\frac{s_1^2}{n_2}+\frac{s_2^2}{n_2}}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1>\mu_2$	آزمون یکطرفه
$t'_{(v)} < -t_\alpha$	$t'=\dfrac{\bar{x}-\bar{Y}}{\sqrt{\frac{s_1^2}{n_2}+\frac{s_2^2}{n_2}}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1<\mu_2$	آزمون یکطرفه
$t'_{(v)} > t_\alpha$ و $t'_{(v)} < -t_\alpha$	$t'=\dfrac{\bar{x}-\bar{Y}}{\sqrt{\frac{s_1^2}{n_2}+\frac{s_2^2}{n_2}}}$	$H_0:\mu_1=\mu_2$ $H_1:\mu_1\neq\mu_2$	آزمون دو طرفه

آزمون اختلاف واریانس‌های دو جامعه

این آزمون جهت تست برابری واریانس دو جامعه به کار می‌رود و در انتخاب فرمول t یا فرمول t' جهت آزمون برابری میانگین‌های دو جامعه مفید واقع می‌شود.

		آماره	ناحیه رد
آزمون یکطرفه	$H_0: \sigma_1 = \sigma_2$ $H_1: \sigma_1 > \sigma_2$	$F_{n_1-1, n_2-1} = \dfrac{S_1^2}{S_2^2}$	$F > F_{n_1-1, n_2-1, \alpha}$
	$H_0: \sigma_1 = \sigma_2$ $H_1: \sigma_1 < \sigma_2$	$F_{n_2-1, n_1-1} = \dfrac{S_2^2}{S_1^2}$	$F < F_{n_2-1, n_1-1, \alpha}$
آزمون دو طرفه	$H_0: \sigma_1 = \sigma_2$ $H_1: \sigma_1 \neq \sigma_2$	$F_{n_M-1, n_m-1} = \dfrac{S_M^2}{S_m^2}$ $S_M^2 > S_m^2$	$F < F_{n_M-1, n_m-1, \frac{\alpha}{2}}$

آزمون واریانس

		آماره	ناحیه رد	
آزمون یکطرفه	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$\chi^2_{(n-1)} = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 > \chi^2_{(n-1), \alpha}$	
	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$\chi^2_{(n-1)} = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 < \chi^2_{n-1, 1-\alpha}$	
آزمون دو طرفه	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$\chi^2_{(n-1)} = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 > \chi^2_{n-1, \alpha/2}$ $\chi^2 < \chi^2_{n-1, 1-\alpha/2}$	

آزمون اختلاف میانگین‌های دو جامعه وابسته (آزمون زوجی)

اگر n مشاهده از یک جامعه استخراج شده (x_i) و پس از اعمال تغییرات یا اعمال فرآیندی بر روی جامعه مجددا همان نمونه استخراج شود و مشاهدات y_i ثبت شود جهت آزمون تأثیر فرآیند روی صفت اندازه‌گیری شده زیر وجود دارد: $d_i = x_i - y_i$

		آماره	ناحیه رد
آزمون یکطرفه	$H_0: \mu_d = 0$ $H_1: \mu_d > 0$	$t_{n-1} = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	$t > t_\alpha$
	$H_0: \mu_d = 0$ $H_1: \mu_d < 0$	$t_{n-1} = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	$t < -t_\alpha$
آزمون دو طرفه	$H_0: \mu_d = 0$ $H_1: \mu_d \neq 0$	$t_{n-1} = \dfrac{\bar{d}}{s_d/\sqrt{n}}$	$t > t_{\alpha/2}$ $t < -t_{\alpha/2}$

آزمون برآورد نسبت یک جامعه

نکته: استفاده از آماره Z جهت آزمون فرض منوط به شرط بزرگی n یا اندازه نمونه است، معمولا برای n>50 آماره فوق توزیع Z دارد.

$$\hat{P} = \frac{x}{n}$$

		آماره	ناحیه رد
آزمون یکطرفه	$H_0: P = P_0$ $H_1: P > P_0$	$Z = \dfrac{\hat{P} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$	$Z > Z_\alpha$
	$H_0: P = P_0$ $H_1: P < P_0$	$Z = \dfrac{\hat{P} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$	$Z < -Z_\alpha$
آزمون دو طرفه	$H_0: P = P_0$ $H_1: P \neq P_0$	$Z = \dfrac{\hat{P} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$	$Z > Z_{\alpha/2}$ $Z < -Z_{\alpha/2}$

آزمون برآورد اختلاف نسبت دو جامعه

$$\hat{P}_1 = \frac{X_1}{n_1}, \ \hat{P}_2 = \frac{X_2}{n_2}$$

		آماره	
آزمون یکطرفه	$H_0: P_1 = P_2$ $H_1: P_1 > P_2$	$Z = \dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{P_1(1-P_1)}{n_1} + \dfrac{P_2(1-P_1)}{n_2}}}$	$Z > Z_\alpha$
	$H_0: P_1 = P_2$ $H_1: P_1 < P_2$	$Z = \dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{P_1(1-P_1)}{n_1} + \dfrac{P_2(1-P_1)}{n_2}}}$	$Z < -Z_\alpha$
آزمون دو طرفه	$H_0: P_1 = P_2$ $H_1: P_1 \neq P_2$	$Z = \dfrac{\hat{P}_1 - \hat{P}_2}{\sqrt{\dfrac{P_1(1-P_1)}{n_1} + \dfrac{P_2(1-P_1)}{n_2}}}$	$Z > Z_{\alpha/2}$ $Z < -Z_{\alpha/2}$

نکته: در حالتی که نسبت مشترک دو جامعه (P_1,P_2) مشخص نباشد که معمولا چنین است به جای P_1 و P_2 در محاسبه آماره Z از P' استفاده می‌کنیم که طبق رابطه زیر به دست می‌آید.

$$P' = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

آزمون‌های مربع‌کای

۱) **آزمون X^2 برای نیکویی برازش:** فرض کنید مجموعه‌ای از فراوانی‌ها در یک نمونه‌گیری از جامعه به دست آمده باشد. می‌خواهیم آزمون کنیم که آیا این نمونه‌ها می‌توانند از یک توزیع نظری خاص استخراج شده باشند یا خیر.

لذا فرض صفر آزمون به این صورت تعریف می‌شود: «هیچ تفاوت معناداری میان مقادیر مشاهده شده و مقادیر نظری وجود ندارد. آماره

آزمون به صورت زیر تعریف می‌شود:

$$\chi^2_{(n-1)} = \sum_{i=1}^{n} \frac{(O_i + E_i)^2}{E_i} = \sum_{i=1}^{n} \frac{O_i^2}{E_i} - N$$

آماره فوق دارای توزیع کای دو با n-1 درجه آزادی است که در آن O_i مقادیر مشاهده شده و E_i مقادیر نظری متناظر با آنها و N مجموع کل فراوانی‌ها است.

نکته: در صورتی که پارامترهای توزیع نظری معلوم نباشند، بایستی آنها را با توجه به نمونه‌ها برآورد نمود. برای برآورد هر یک از پارامترهای توزیع نظری، یک درجه از آزادی کاسته می‌شود. همچنین اگر فراوانی دسته‌ای از ۵ کمتر باشد بایستی با دسته قبل در هم ادغام شوند که در این صورت هم با هر ادغام یک درجه آزادی کم می‌شود.

نکته: توزیع داده‌های اصلی را نباید با فراوانی نسبی یا متناسب با آن جانشین کرد، بلکه اطلاعات اصلی باید به کار برده شود.

۲) آزمون X^2 برای نابستگی: جدول فراوانی دوبعدی زیر را که براساس یک نمونه N تایی در بررسی توأم دو صفت x و y به دست آمده است را در نظر بگیرید:

X | Y	y_1	y_2	…	y_i	…	y_m	f_x
x_1	f_{11}	f_{12}		…		f_{1n}	f_{10}
x_2	f_{21}	f_{22}		…		f_{2n}	f_{20}
⋮	⋮	⋮				⋮	⋮
x_i	f_{i1}	f_{i2}	…	f_{ij}		f_{in}	f_{i0}
⋮	⋮	⋮		⋮		⋮	⋮
x_m	f_{m1}	f_{m2}	…	f_{mi}	…	f_{mn}	f_{m0}
f_y	f_{01}	f_{02}		…		f_{0n}	N

اگر x و y مستقل باشند، برای مقدار فراوانی مورد انتظار توأم x_i و y_i

$$E_{ij} = NP\left(X = x_i, Y = y_j\right) = N \times P\left(x = x_i\right)P\left(Y = y_j\right) = N \times \frac{f_{0i}}{N} \times \frac{f_{j0}}{N} = \frac{f_{0i} \times f_{j0}}{N}$$

در این حالت آماره مورد استفاده برای آزمون استقلال دو پارامتر x و y به صورت زیر تعریف می‌گردد.

$$\chi^2_{(n-1)(m-1)} = \sum_i \sum_j \frac{\left(f_{ij} - E_{ij}\right)^2}{E_{ij}}$$

محدوده پذیرش فرض صفر (استقلال دو پارامتر) در بازه $\left(0, \chi^2_\alpha\right)$ است.

۳) آزمون X^2 برای واریانس جامعه نرمال: اگر بخواهیم آزمون کنیم که یک توزیع نرمال با میانگین مجهول آیا دارای واریانس خاصی است $\left(H_0 : \sigma^2 = \sigma_0^2\right)$، در صورتی که $x_1,\dots,x_n$ یک نمونه تصادفی nتایی از جامعه نرمال مورد نظر باشد و S^2 واریانس نمونه مورد نظر باشد، جهت آزمون فرض صفر آماره زیر را تعریف می‌کنیم:

$$\chi^2_{(n-1)} = \frac{(n-1)S^2}{\sigma_0^2}$$

ناحیه پذیرش فرض صفر برای این آماره همانند قسمت آزمون واریانس جامعه خواهد بود.

نکته: در صورتی که میانگین جامعه معلوم باشد درجه آزادی توزیع X^2 یک واحد اضافه خواهد شد، یعنی به n خواهد رسید.

$$\chi^2_{(n)} = \frac{ns^2}{\sigma_0^2} \; , \; s^2 = \frac{\sum (x_i - \mu)^2}{n}$$

آنالیز واریانس

آنالیز واریانس یا تجزیه و تحلیل تغییرات نسبت به میانگین، در واقع تقسیم تغییرات موجود در کل داده‌ها بین گونه‌های مختلفی از نمونه‌هاست. به عنوان مثال، فرض کنید صفت به خصوصی برای افراد یا نمونه‌های k جامعه مختلف اندازه‌گیری شده باشد. تعداد نمونه‌ها در جامعه jام، برابر n_j بوده و $\sum_{j=1}^{k} n_j = N$ می‌باشد، مقدار صفت اندازه‌گیری شده برای نمونه iام در جامعه یا گونه jام را با X_{ij}، میانگین کل نمونه‌ها را با $\bar{x}$ و میانگین نمونه‌های گونه jام را با $\bar{X}_j$ نشان می‌دهیم. هدف مقایسه میانگین صفت مورد نظر در جامعه یا گونه‌های مختلف است، داریم:

$$\left(X_{ij} - \bar{X}\right) = \left(\bar{X}_j - \bar{X}\right) + \left(X_{ij} - \bar{X}_j\right)$$

یا به بیان دیگر:

انحراف نمونه‌های هر گونه از میانگین خود گونه + انحراف میانگین گونه‌ها از میانگین کل = انحراف مشاهدات از میانگین کل

می‌توان نشان داد:

$$\underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}\right)^2}_{SST \text{ یا } SS_{\text{کل}}} = \underbrace{\sum_{j=1}^{k} n_j \left(\bar{X}_j - \bar{X}\right)^2}_{SS_{tt} \text{ یا } SSA \text{ یا } SS_{\text{گونه}}} + \underbrace{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}\right)^2}_{SSE \text{ یا } SS_{\text{خطا}} \text{ یا } SS_{\text{مانده}}}$$

آزمون برابری میانگین‌های چند جامعه

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

H_1 : حداقل یکی از μ_i با سایرین متفاوت باشد

در صورتی که بخواهیم برابری میانگین‌های چند جامعه را با هم آزمون کنیم بهتر است به جای مقایسه زوجی μ_i از آزمون واریانس استفاده کنیم زیرا مقایسه زوجی خطای نوع I آزمون را بسیار بالا خواهد برد. در این حالت آماره آزمون جهت فرض صفر به صورت زیر تعریف می‌شود.

$$F_{k-1, N-k} = \frac{SSA \text{ میانگین}}{SSE \text{ میانگین}} = \frac{\frac{SSA}{k-1}}{\frac{SSE}{N-k}}$$

همان‌طور که اشاره گردیده است آماره فوق دارای توزیع F با k-1 و N-k درجه آزادی است. در این حالت ناحیه پذیرش به صورت $\left(0, F_{k-1, N-k, \alpha}\right)$ تعریف می‌شود.