آمار و احتمال مهندسی

توزیع‌های گسسته مهم

(Ross 4.7-4.8)

فرایند پواسون

- پیشامدهای نادری را در نظر بگیرید که در یک بازه زمانی خاص اتفاق می‌افتند:
  - زلزله، فروپاشی رادیواکتیو، تصادف، و …
  - بازه زمانی ثابت: یک سال، یک ثانیه، …
  - پیشامدها با نرخ ثابت اتفاق می‌افتند: λ پیشامد در بازه زمانی مفروض
- بازه زمانی را به $n \to \infty$ زیربازه زمانی کوچکتر تقسیم کنید:
  - به طوری که حداکثر یک پیشامد در هر زیربازه اتفاق بیافتد.
  - رخ دادن پیشامدها در زیربازه‌ها مستقل از هم باشد.
  - به دلیل وجود زیربازه‌های خیلی زیاد، احتمال وقوع پیشامد در یک زیربازه خاص داده شده بسیار کم است.
- در این حالت متغیر تصادفی $N(t)$ که نشان‌دهنده تعداد وقوع پیشامد در بازه زمانی اصلی است، دارای توزیع پواسون خواهد بود: $N(t)\sim\mathrm{Poi}(\lambda)$

مثال: بار سرور وب

- درخواست‌هایی که یک سرور وب در یک ثانیه دریافت می‌کند را در نظر بگیرید:
  - در گذشته مشاهده شده است که بار روی سرور به طور متوسط ۲ درخواست در ثانیه است.
  - متغیر تصادفی X را تعداد درخواست‌های دریافتی توسط سرور در یک ثانیه بگیرید
  - احتمال $P\{X=5\}$ چقدر است؟
- مدل:
  - فرض کنید سرور نمی‌تواند بیش از یک درخواست در یک میلی‌ثانیه را پاسخ دهد.
  - 1 sec = 1000 msec (بنابراین با n بزرگ مواجه هستیم)
  - احتمال وقوع درخواست در یک بازه مشخص به طول 1msec برابر است با 2/1000 که احتمال بسیار کوچکی است.
- بنابراین می‌توانیم فرض کنیم: $X \sim Poi(2)$

$$P\{X=5\} = e^{-2}\frac{2^5}{5!} \approx 0.0361$$

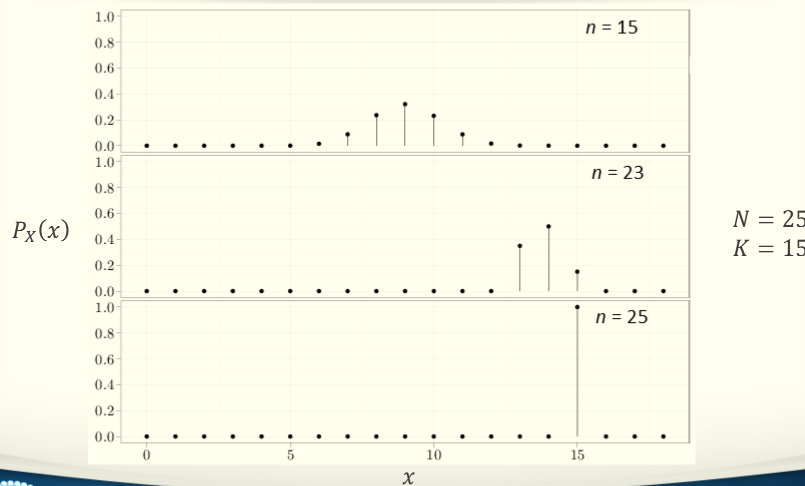
توزیع فوق هندسی (Hypergeometric)

- در مثال کارت حافظه‌های خراب اگر انتخاب بدون جایگزینی باشد، داریم:

$$P(x) = \frac{\text{تعداد انتخابهای مورد نظر}}{\text{تعداد کل}} = \begin{cases} \dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} & \max\{0, n-N+K\} \le x \le \min\{n, K\} \\ 0 & \text{otherwise} \end{cases}$$

- $\max\{0,n-N+K\}$ بدین خاطر است که تعداد کارت‌های سالم انتخاب‌شده $n-x$ بوده و نمی‌تواند از تعداد کل سالم‌ها یا $N-K$ بیشتر شود: $n-x \le N-K$
- $\min\{n,K\}$ نیز به این خاطر است که تعداد کارت‌های خراب انتخاب شده نمی‌تواند از تعداد کل کارت‌های خراب بیشتر شود.

توزیع فوق هندسی



ویژگی‌های توزیع فوق هندسی

- ویژگی‌های یک متغیر تصادفی با توزیع فوق هندسی:
  - آزمایش‌های تصادفی بر روی یک جمعیت Nتایی که Kتا از آن‌ها ویژگی خاصی دارند (موفقیت آزمایش) انجام می‌گیرد.
  - هر آزمایش دو خروجی ممکن دارد: شکست و موفقیت
  - تعداد آزمایش‌های تصادفی ثابت است: n
  - آزمایش‌های تصادفی مستقل از هم نیستند.
  - احتمال موفقیت در هر آزمایش متفاوت از آزمایش قبل است.

$$\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

- کاربردهای توزیع فوق هندسی :
  - در آزمایش‌هایی که با نمونه‌برداری بدون جایگزینی سر و کار داریم.
  - کنترل کیفیت
  - بازی پوکر

میانگین و واریانس توزیع فوق هندسی

$$\mathrm{P}_X(k) = \mathrm{P}\{X = k\} = \frac{\binom{Np}{k}\binom{N-Np}{n-k}}{\binom{N}{n}}$$

$$\max(0, n - N + K) \le k \le \min(n, K) \quad , \qquad p = \frac{K}{N}$$

- می‌توان نشان داد که:

$$\mathrm{E}(X) = n\frac{K}{N} = np$$

$$\mathrm{var}(X) = \underbrace{\frac{nK(N-K)}{N^2}}_{npq} \times \frac{N-n}{N-1}$$

توزیع هندسی (Geometric)

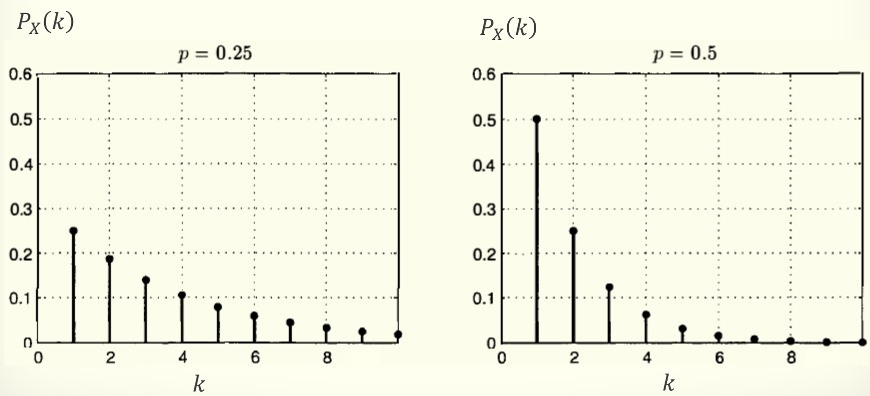
- توزیع متغیر تصادفی تعداد دفعات لازم برای تکرار یک آزمایش برنولی تا رسیدن به موفقیت را توزیع هندسی می‌گوییم:

$$X \sim \mathrm{Geo}(p): \quad P\{X = k\} = p(1-p)^{k-1} \;:\quad k = 1,2, \dots$$

- این توزیع یک دنباله هندسی است و داریم:

$$\sum_{k=1}^{\infty} p(1-p)^{k-1} = \sum_{m=0}^{\infty} p(1-p)^m = \frac{p}{1-(1-p)} = 1$$

توزیع هندسی



ویژگی‌های توزیع هندسی

- ویژگی‌های یک متغیر تصادفی با توزیع هندسی:
  - آزمایش‌های تصادفی مستقل از هم و در شرایط یکسان انجام می‌گیرند.
  - هر آزمایش دو خروجی ممکن دارد: شکست و موفقیت
  - احتمال موفقیت در همه آزمایش‌ها یکسان است: p
  - تعداد آزمایش‌های تصادفی ثابت نیست و تا زمانی که یک موفقیت مشاهده نشود ادامه می‌یابند.
- کاربردهای توزیع هندسی:
  - تعداد آزمایش‌های برنولی لازم برای این که یک فرایند گسسته تغییر حالت دهد.
  - تعداد تماس‌های لازم برای برقراری تماس موفق با یک خط تلفن که ۸۰٪ اوقات اشغال است
  - تعداد چاه‌هایی که در یک منطقه باید حفر شوند تا به یک چاه آب برسیم.
  - تعداد دفعاتی که یک تاس باید پرتاب شود تا عدد ۶ ظاهر شود.

میانگین توزیع هندسی

- از فرم بسته سری هندسی داریم:

$$|x| < 1 \; : \; \sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$$

- با مشتق‌گیری از این رابطه داریم:

$$|x| < 1 \; : \; \sum_{i=1}^{\infty} i.x^{i-1} = \frac{1}{(1-x)^2}$$

- میانگین توزیع هندسی برابر است با:

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = \frac{p}{\left(1-(1-p)\right)^2} = \frac{p}{p^2} = \frac{1}{p}$$

واریانس توزیع هندسی

$$\sum_{k=1}^{\infty} k\, x^{k-1} = \frac{1}{(1-x)^2} \xrightarrow{\text{ضرب در } x} \sum_{k=1}^{\infty} k\, x^{k} = \frac{x}{(1-x)^2}$$

با مشتق‌گیری از طرفین داریم:

$$\sum_{k=1}^{\infty} k^2 x^{k-1} = \frac{1+x}{(1-x)^3}$$

$$E[X^2] = \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = p\sum_{k=1}^{\infty} k^2(1-p)^{k-1}$$

$$= \frac{p(1+1-p)}{\left(1-(1-p)\right)^3} = \frac{p(2-p)}{p^3} = \frac{2-p}{p^2}$$

واریانس توزیع هندسی

$$E[X^2] = \frac{2-p}{p^2}$$

$$\Rightarrow \; Var(X) = E[X^2] - (E[X])^2 = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

$$Var(X) = \frac{1-p}{p^2} = \frac{q}{p^2}$$

بی‌حافظگی توزیع هندسی

- توزیع متغیر تصادفی X را بی‌حافظه (memoryless) گویند، اگر برای هر t و s، داشته باشیم:

$$P\{X \geq s+t \;|\, X \geq t\} = P\{X \geq s\}$$

- توزیع هندسی تنها توزیع احتمال گسسته دارای خاصیت بی‌حافظگی است.

- بی‌حافظه بودن توزیع هندسی به این معنی است که اگر ما t شکست متوالی را در انجام یک آزمایش تصادفی مشاهده کرده‌ایم، احتمال مشاهده s شکست دیگر قبل از رسیدن به موفقیت (در کل مشاهده $s+t$ شکست)، برابر با احتمال این پیشامد است که هنوز هیچ آزمایشی انجام نداده‌ایم و s شکست تا اولین موفقیت فاصله داریم.

اثبات بی‌حافظگی توزیع هندسی

$$P\{X \geq t\} = \sum_{i=t}^{+\infty} pq^i = pq^t \sum_{i=t}^{+\infty} q^{i-t} = \frac{pq^t}{1-q} = \frac{pq^t}{p} = q^t$$

بنابراین:

$$P\{X \geq s+t \,|X \geq t\} =$$

$$\frac{P\{(X \geq s+t) \cap (X \geq t)\}}{P\{X \geq t\}} =$$

$$\frac{P\{X \geq s+t\}}{P\{X \geq t\}} =$$

$$\frac{q^{s+t}}{q^t} = q^s = P\{X \geq s\}$$

مثال: جمع‌آوری کوپن (Coupon Collector)

- یک جعبه که داخل آن n کوپن مختلف قرار دارد، در اختیار داریم. در هر بار آزمایش یک کوپن از جعبه خارج کرده و بعد از مشاهده به درون جعبه باز می‌گردانیم. به طور متوسط چند کوپن باید از جعبه خارج کنیم تا هر کوپن حداقل یک بار مشاهده شده باشد؟

T = مدت زمان لازم (تعداد دفعات انجام آزمایش) برای جمع‌آوری هر n کوپن

T_i = زمانی که بعد از مشاهده $(i-1)$ کوپن مختلف طول می‌کشد تا کوپن متفاوت i-ام را مشاهده کنیم.

واضح است که: $T = T_1 + T_2 + \cdots + T_n$

احتمال مشاهده کوپن جدید بعد از مشاهده $(i-1)$ کوپن مختلف برابر است با:

$$p_i = \frac{n-(i-1)}{n}$$

مثال

- بنابراین T_i یک متغیر تصادفی هندسی با احتمال موفقیت p_i است: $T_i \sim Geo(p_i)$

$$E[T_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

- با استفاده از خاصیت خطی بودن امید ریاضی داریم:

$$E[T] = E[T_1] + E[T_2] + \cdots + E[T_n]$$
$$= \frac{n}{n} + \frac{n}{n-1} + \ldots + \frac{n}{1}$$
$$= n\left(\frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{n}\right)$$
$$\approx n \log(n)$$

توزیع دوجمله‌ای منفی (توزیع پاسکال)

- مساله: سکه‌ای را آنقدر پرتاب می‌کنیم که r بار شیر بیاید. اگر تعداد کل پرتاب‌ها را X بنامیم، احتمال $X = n$ را به دست آورید. $(n = r, r+1, \ldots)$
- در حالت کلی اگر یک آزمایش برنولی را آنقدر تکرار کنیم تا پیشامد A که $P(A) = p$، r بار اتفاق افتد و متغیر تصادفی X به صورت زیر تعریف شود:

X: تعداد کل آزمایش‌های انجام شده تا r بار وقوع A

احتمال r-امین موفقیت که حتماً باید در n-امین آزمایش باشد.

$$P\{X = n\} = \overset{\uparrow}{p}\underbrace{\binom{n-1}{r-1} p^{r-1} q^{(n-1)-(r-1)}}_{\text{احتمال } r-1 \text{ موفقیت در میان بقیه } (n-1) \text{ آزمایش}} = \binom{n-1}{r-1} p^r q^{n-r}$$

توزیع دوجمله‌ای منفی

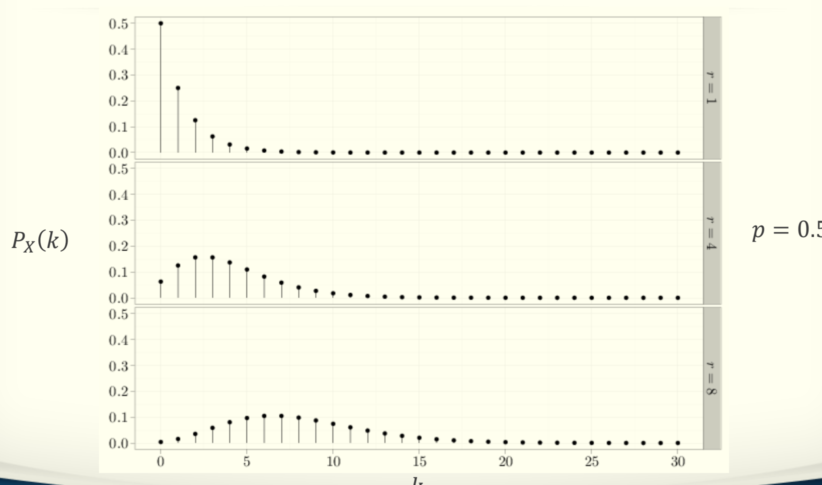
- این توزیع را دوجمله‌ای منفی می‌نامند زیرا بر خلاف توزیع دوجمله‌ای در اینجا تعداد موفقیت‌ها ثابت است و تعداد کل آزمایش‌های انجام شده متغیر.
- پارامترهای توزیع دوجمله‌ای منفی تعداد موفقیت‌ها (r) و احتمال موفقیت (p) هستند:

$$X\sim\mathrm{NegBin}(r,p):\ \ P(X=n)=\binom{n-1}{r-1}p^r(1-p)^{n-r}\,,\qquad n\geq r$$

- توزیع هندسی حالت خاص توزیع دوجمله‌ای منفی برای $r=1$ است.
- متغیر با توزیع دوجمله‌ای منفی را می‌توان به صورت مجموع r متغیر تصادفی هندسی نوشت:

$$X=X_1+X_2+\cdots+X_r\quad:\quad X_i\sim\mathrm{Geo}(p)$$

توزیع دوجمله‌ای منفی



ویژگی‌های توزیع دوجمله‌ای منفی

- ویژگی‌های یک متغیر تصادفی با توزیع دوجمله‌ای منفی:
  - آزمایش‌های تصادفی مستقل از هم و در شرایط یکسان انجام می‌گیرند.
  - هر آزمایش دو خروجی ممکن دارد: شکست و موفقیت
  - احتمال موفقیت در همه آزمایش‌ها یکسان است: p
  - تعداد آزمایش‌های تصادفی ثابت نیست و تا زمانی که r موفقیت مشاهده نشود ادامه می‌یابند.
- کاربردهای توزیع دوجمله‌ای منفی:
  - مدت زمان بستری شدن در بیمارستان
  - توزیع بسیاری از فرایندهای الکتروشیمیایی
  - در صنعت بیمه و بازاریابی
  - در همه مسائلی که نیاز به تعداد ثابتی آزمایش موفق برای پایان دادن به آزمایش‌های تصادفی داریم.

میانگین و واریانس توزیع دوجمله‌ای منفی

- دیدیم که متغیر تصادفی $X\sim\text{NegBin}(r,p)$ را می‌توان به صورت مجموع r متغیر تصادفی هندسی نوشت:

$$X = X_1 + X_2 + \cdots + X_r \;:\; X_i\sim\text{Geo}(p)$$

- با استفاده از خاصیت خطی بودن امید ریاضی داریم:

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_r] = \frac{1}{p} + \frac{1}{p} + \cdots + \frac{1}{p} = \frac{r}{p}$$

- همچنین می‌توان نشان داد که:

$$\text{Var}(X) = \frac{r(1-p)}{p^2} = \frac{rq}{p^2}$$

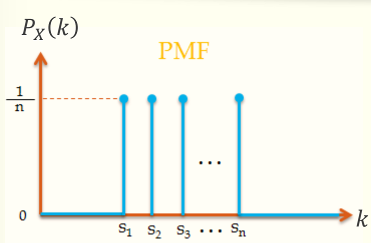
مثال

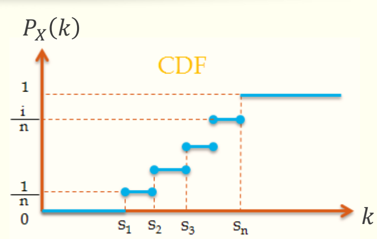
```
function findTeen(L, N)
    found = 0; i = 0;
    location = (0, 0);
    while ((found < 2) and (i < N))
        if (teen (L[i]))
            location[found] = i;
            found = found + 1;
        i = i + 1;
    return location;
```

- لیست مرتب‌نشده N نفره‌ای در اختیار داریم که ۲۰٪ آن‌ها نوجوان هستند که به طور یکنواخت در لیست پراکنده هستند. تابع findTeen به عنوان ورودی لیست L و اندازه آن (N) را دریافت می‌کند و مکان دو نوجوان اول در لیست را باز می‌گرداند.
- متوسط تعداد دفعاتی که حلقه while تکرار می‌شود، چقدر است؟
- هر درایه لیست با احتمال $p = 0.2$ شامل یک نوجوان است. احتمال موفقیت ثابت و محتوای درایه‌ها مستقل از هم هستند.
- تعداد دفعات تکرار حلقه برای رسیدن به دومین موفقیت دارای توزیع دوجمله‌ای منفی با پارامترهای $r = 2$ و $p = 0.2$ است: $X \sim \text{NegBin}(2, 0.2)$

$$E[X] = \frac{r}{p} = \frac{2}{0.2} = 10$$

توزیع یکنواخت (Uniform) گسسته





$$P\{X = s_i\} = \frac{1}{n}$$
$$i = 1, 2, \ldots, n$$

- کاربردهای توزیع یکنواخت گسسته:
  - پرتاب تاس یا سکه سالم
  - الگوریتم مولد اعداد تصادفی گسسته
  - متغیر تصادفی گسسته‌ای که از توزیع آن اطلاعی نداریم