دانشکده مهندسی کامپیوتر

عنوان درس:

ارزیابی کارایی سیستم‌های کامپیوتری

Performance Evaluation of Computer Systems (PECS)

جلسه ۷: مدلهای آماری

(Statistical Models)

مدرس:

محمد عبداللهی ازگمی

(Mohammad Abdollahi Azgomi)

azgomi@iust.ac.ir

جلسه ۷: مدلهای آماری

- مقدمه
- اصطلاحات و مفاهیم آمار و احتمالات
- مدلهای احتمالی سودمند
- توزیع‌های گسسته
- توزیع‌های پیوسته
- توزیع‌های تجربی
- جمع‌بندی

مقدمه

- دنیایی که مدل‌ساز با آن سر و کار دارد بیشتر احتمالی (probabilistic) است تا اینکه قطعی (deterministic) باشد.
  - از اینرو مدلهای احتمالی می‌توانند به‌خوبی تغییرات در چنین دنیایی را توصیف کنند.
- برای انتخاب یک مدل احتمالی مناسب می‌توان پدیده‌های مورد نظر را نمونه‌برداری (sampling) نموده و سپس:
  - یک توزیع شناخته شده را حدس زده و انتخاب کرد.
  - پارامتر(های) توزیع را تخمین زد.
  - آزمون برازندگی (goodness of fit) (نظیر K2 یا K-S) را جهت بررسی و حصول اطمینان از انتخاب صحیح توزیع و پارامترها انجام داد.
- برای این منظور در ادامه توزیع‌های احتمالی مهم مورد بررسی قرار گرفته و برخی از موارد کاربرد آنها ذکر می‌شود.

اصطلاحات و مفاهیم آمار و احتمالات

- بنا بر این در ادامه مفاهیم و تعاریف زیر ارائه می‌شوند:
  - متغیرهای تصادفی گسسته (discrete random variables)
  - متغیرهای تصادفی پیوسته (continuous random variables)
  - تابع توزیع تجمعی (cumulative distribution function)
  - امید ریاضی (expectation)

متغیرهای تصادفی گسسته

- به X یک متغیر تصادفی گسسته گفته می‌شود اگر تعداد مقادیر امکان‌پذیر X متناهی (finite) یا قابل شمارش (شمارا) (countably finite) باشد.
- برای مثال کارهای وارده به یک کارگاه را در نظر بگیرید:
  - تعداد کارهای وارده در یک هفته را با متغیر تصادفی X نشان می‌دهیم.
  - فضای برد (range space) مقادیر امکان‌پذیر X را با R_x نشان می‌دهیم: $R_x=\{0, 1, 2, \ldots\}$
  - احتمال اینکه مقدار متغیر برابر با x_i باشد را با $p(x_i)$ نشان می‌دهیم: $p(x_i) = P(X = x_i)$
- $p(x_i)$ برای $i = 1, 2, \ldots$ باید در شرایط زیر صدق کند:

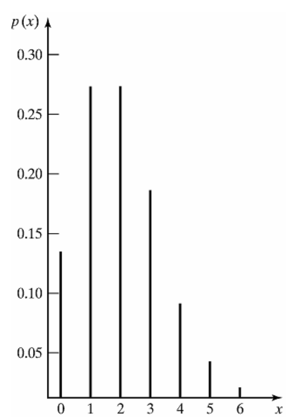
$$1.\ p(x_i) \geq 0, \text{ for all } i$$

$$2.\ \sum_{i=1}^{\infty} p(x_i) = 1$$

- به مجموعه زوجهای مرتب $[x_i, p(x_i)]$، $i=1, 2, \ldots$ توزیع احتمالی (probability distribution) و به $p(x_i)$ تابع جرم احتمالی (pmf: probability mass function) متغیر تصادفی X گفته می‌شود.

متغیرهای تصادفی گسسته (ادامه)

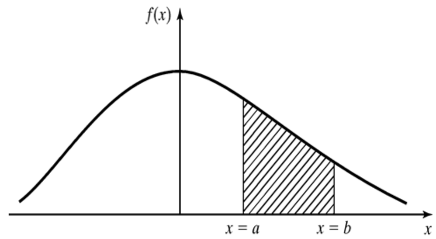
- مثال: تماسهای تلفنی درخواست سرویس به یک تعمیرکار کامپیوتر طبق یک توزیع گسسته پواسان است. تابع جرم احتمالی این توزیع در شکل زیر نشان داده شده است:



تعریف متغیرهای تصادفی پیوسته

- به X یک متغیر تصادفی پیوسته گفته می‌شود اگر فضای برد (R_x) آن یک بازه (پیوسته) یا مجموعه‌ای از بازه‌ها از اعداد حقیقی باشد.
- احتمال اینکه X در بازه [a, b] قرار گرفته باشد به‌صورت زیر بدست می‌آید:

$$P(a \le X \le b) = \int_a^b f(x)dx$$



- به f(x) تابع چگالی احتمالی (pdf: probability density function) گفته می‌شود.

تعریف متغیرهای تصادفی پیوسته

- f(x) باید در شرایط ذیل صدق نماید:

1. $f(x) \ge 0$, for all x in R_X
2. $\int_{R_X} f(x)dx = 1$
3. $f(x) = 0$, if x is not in R_X

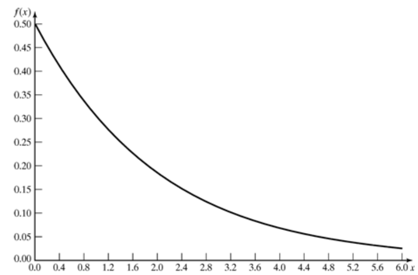
- خصوصیتها:

1. $P(X = x_0) = 0$, because $\int_{x_0}^{x_0} f(x)dx = 0$
2. $P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b)$

متغیرهای تصادفی پیوسته

- مثال: فرض کنید که X یک متغیر تصادفی پیوسته که نماینده طول عمر یک دستگاه بوده و دارای pdf زیر باشد:

$$f(x)=\begin{cases}\frac{1}{2}e^{-x/2}, & x\geq 0\\ 0, & \text{otherwise}\end{cases}$$



  - بعداً خواهیم دید که X دارای توزیع نمایی با میانگین ۲ سال است.
- احتمال اینکه طول عمر دستگاه بین ۲ تا ۳ سال باشد برابر خواهد بود با:

$$P(2\leq X\leq 3)=\int_2^3\frac{1}{2}e^{-x/2}dx=\frac{1}{2}\int_2^3 e^{-x/2}dx=0.145$$

تابع توزیع تجمعی

- تابع توزیع تجمعی (**cumulative distribution function (cdf)**) یک متغیر تصادفی X با F(X) یا F_x نشان داده شده و به‌صورت زیر تعریف می‌شود:
  - $F(x) = F_x = P(X\leq x)$
  - اگر X گسسته باشد: $F(x)=\sum_{\substack{\text{all}\\ x_i\leq x}} p(x_i)$
  - اگر X پیوسته باشد: $F(x)=\int_{-\infty}^{x} f(t)dt$
- خصوصیت‌ها:

1. F is nondecreasing function : If $a<b$, then $F(a)\leq F(b)$
2. $\lim_{x\to\infty} F(x)=1$
3. $\lim_{x\to-\infty} F(x)=0$

- همه پرسشهای احتمالی در باره X بر حسب cdf قابل پاسخگویی است. برای مثال:

$$P(a<X\leq b)=F(b)-F(a),\ \text{for all}\ \ a<b$$

تابع توزیع تجمعی

- برای مثال در مورد توزیع تجمعی دستگاه مورد بحث خواهیم داشت:

$$F(x)=\frac{1}{2}\int_0^x e^{-t/2}dt=1-e^{-x/2}$$

  - احتمال اینکه این دستگاه کمتر از دو سال عمر کند چقدر است؟

$$P(X\leq 2)=F(2)=1-e^{-1}=0.632$$

  - احتمال اینکه این دستگاه دو الی سه سال عمر کند چقدر است؟

$$P(2\leq X\leq 3)=F(3)-F(2)=(1-e^{-(3/2)})-(1-e^{-1})=0.145$$

امید ریاضی و واریانس

- امید ریاضی (expected value) یک متغیر تصادفی به‌صورت زیر تعریف می‌شود:
  - اگر X گسسته باشد: $E(x)=\sum_{\text{all } i} x_i p(x_i)$
  - اگر X پیوسته باشد: $E(x)=\int_{-\infty}^{\infty} xf(x)dx$
- واریانس یک متغیر تصادفی با σ^2 یا V(X) یا Var(X) نشان داده می‌شود و عبارت است از:
  - $V(X) = E[(X - E[X]^2)]$
  - یا $V(X) = E(X^2) - [E(x)]^2$
- به جذر واریانس، انحراف معیار (standard deviation) گفته می‌شود.

امید ریاضی و واریانس

- امید ریاضی دستگاه مورد بحث را بدست می‌آوریم:

$$E(X)=\frac{1}{2}\int_0^{\infty} xe^{-x/2}dx = -xe^{-x/2}\Big|_0^{\infty} + \int_0^{\infty} e^{-x/2}dx = 2$$

- سپس $E[X^2]$ را محاسبه می‌کنیم:

$$E(X^2)=\frac{1}{2}\int_0^{\infty} x^2e^{-x/2}dx = -x^2e^{-x/2}\Big|_0^{\infty} + \int_0^{\infty} e^{-x/2}dx = 8$$

- و حالا می‌توانیم واریانس و انحراف معیار را بدست آوریم:

$$V(X)=8-2^2=4$$

$$\sigma=\sqrt{V(X)}=2$$

مدلهای احتمالی سودمند

- در ادامه با برخی مدلهای احتمالی سودمند که در زمینه‌های زیر کاربرد دارند آشنا می‌شویم:
  - سیستم‌های صف (queueing systems)،
  - سیستم‌های موجودی و زنجیره تامین (inventory and supply-chain systems)
  - قابلیت اطمینان و قابلیت نگهداشت (reliability and maintainability)
  - مدلهای احتمالی داده‌های محدود (limited data):
    - موقعی که در مورد یک پدیده، داده‌های کافی در اختیار نداریم، استفاده می‌شوند.

مدلهای سودمند برای سیستم‌های صف

- در سیستم‌های صف الگوهای زمان بین ورود مشتریان و زمان سرویس می‌تواند احتمالی باشد.
- برای این منظور توزیع‌های احتمالی زیر قابل استفاده هستند:
  - اگر زمانهای بین ورود یا سرویس کاملاً تصادفی هستند: توزیع نمایی
  - اگر زمانهای بین ورود یا سرویس دارای مقادیر مثبت و منفی حول یک مقدار ثابت باشند: توزیع نرمال
  - در حالتهای عمومی‌تر توزیعهای گاما (gamma) و ویبول (Weibull) قابل استفاده هستند.

سیستم‌های موجودی و زنجیره تامین

- در سیستم‌های موجودی و زنجیره تامین، حداقل سه متغیر تصادفی وجود دارد:
  - تعداد واحدهای مورد تقاضا در هر سفارش،
  - زمان بین تقاضاها، و
  - زمان انجام سفارش تا رسیدن تقاضا (lead time).
- مدل آماری مناسب برای زمان انجام سفارش تا رسیدن تقاضا:
  - توزیع گاما
- تعداد واحدهای مورد تقاضا در هر سفارش:
  - توزیع پواسان،
  - توزیع دوجمله‌ای منفی، یا
  - توزیع هندسی.

قابلیت اطمینان و قابلیت نگهداشت

- توزیع مناسب برای زمان تا خرابی ((time to failure (TTF)):
  - اگر خرابیها تصادفی باشند: توزیع نمایی
  - برای حالتی که قطعات یدکی افزونه (standby redundancy) داریم و هر قطعه دارای TTF نمایی باشد، باشد در این صورت TTF کل سیستم: توزیع گاما
  - اگر خرابی به دلیل عیوب متعدد در قطعات سیستم باشد: توزیع ویبول
  - اگر خرابیها حاصل فرسودگی/پوشیدن باشد: توزیع نرمال

مدلهای احتمالی سودمند برای داده‌های محدود

- در صورتی که داده‌های محدودی در دسترس است توزیع‌های زیر قابل استفاده هستند:
  - یکنواخت،
  - مثلثی، و
  - بتا.
- سایر توزیع‌های قابل استفاده و مهم:
  - برنولی (Bernoulli)،
  - دوجمله‌ای (binomial)، و
  - فوق‌نمایی (hyperexponential).

توزیع‌های گسسته

- متغیرهای تصادفی گسسته برای توصیف یک پدیده تصادفی که فقط مقادیری از نوع اعداد صحیح می‌توانند رخ دهند، استفاده می‌شوند.
- در ادامه توزیع‌های زیر را معرفی می‌کنیم:
  - آزمونها و توزیع برنولی
  - توزیع دوجمله‌ای
  - توزیع هندسی و دوجمله‌ای منفی
  - توزیع پواسان

آزمونها و توزیع برنولی

- آزمونهای برنولی: یک آزمایش را که شامل n آزمون است که نتیجه هر کدام موفقیت یا شکست است را در نظر بگیرید:
  - $X_j = 1$ در صورتی که نتیجه jامین آزمایش موفقیت باشد.
  - $X_j = 0$ در صورتی که نتیجه jامین آزمایش شکست باشد.
- توزیع برنولی (یک آزمون) عبارت است از:
  - که در نتیجه خواهیم داشت:

$$p_j(x_j) = p(x_j) = \begin{cases} p, & x_j = 1, j = 1,2,\ldots,n \\ 1-p = q, & x_j = 0, j = 1,2,\ldots,n \\ 0, & \text{otherwise} \end{cases}$$

  - $E(X_j) = p$
  - $V(X_j) = p(1-p) = pq$
- فرآیند برنولی: شامل n آزمون مستقل برنولی است و به‌صورت زیر تعریف می‌شود:
  - $p(x_1,x_2,\ldots, x_n) = p_1(x_1)p_2(x_2) \ldots p_n(x_n)$

توزیع دوجمله‌ای

- متغیر تصادفی X که تعداد موفقیت‌ها در n آزمون برنولی را نشان می‌دهد یک توزیع دوجمله‌ای است:

$$p(x)=\begin{cases}\binom{n}{x}\underbrace{p^x q^{n-x}}, & x=0,1,2,\ldots,n\\ 0, & \text{otherwise}\end{cases}$$

تعداد برآمدهایی که دارای تعداد مورد نیاز موفقیت و شکست باشند.

احتمال اینکه x موفقیت و n-x شکست وجود داشته باشد

- امید ریاضی و واریانس:
  - E(x) = p + p + … + p = n*p
  - V(X) = pq + pq + … + pq = n*pq

توزیع هندسی و دوجمله‌ای منفی

- توزیع هندسی: نشان‌دهنده تعداد آزمونهای مورد نیاز (X) برای نیل به اولین موفقیت:

$$p(x)=\begin{cases}q^{x-1}p, & x=0,1,2,\ldots,n\\ 0, & \text{otherwise}\end{cases}$$

  - E(x) = 1/p و V(X) = q/p^2

- توزیع دوجمله‌ای منفی:
  - نشان‌دهنده تعداد آزمونهای برنولی (X) تا نیل به kامین موفقیت است.
  - اگر Y یک توزیع دوجمله‌ای منفی است با پارامتر p و k باشد، آنگاه:

$$p(x)=\begin{cases}\binom{y-1}{k-1}q^{y-k}p^k, & y=k,k+1,k+2,\ldots\\ 0, & \text{otherwise}\end{cases}$$

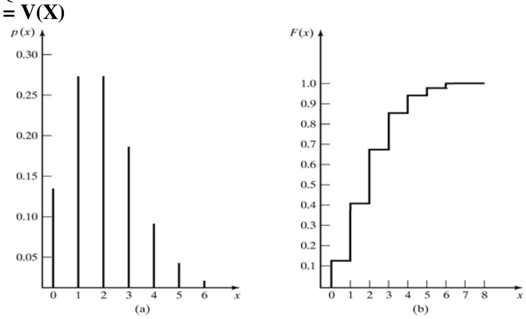
  - E(Y) = k/p و V(X) = kq/p^2

توزیع پواسان

- توزیع پواسان فرآیندهای تصادفی متعددی را به‌خوبی توصیف می‌کند و از نظر ریاضی کاملاً ساده است.
برای $\alpha > 0$، **pdf** و **cdf** این توزیع به شکل زیر است:

$$p(x) = \begin{cases} \dfrac{e^{-\alpha}\alpha^x}{x!}, & x = 0,1,\ldots \\ 0, & \text{otherwise} \end{cases} \qquad F(x) = \sum_{i=0}^{x} \frac{e^{-\alpha}\alpha^i}{i!}$$

- **E(X) = α = V(X)**



توزیع پواسان

- **مثال: تماسهای تلفنی درخواست سرویس به یک تعمیرکار کامپیوتر طبق یک توزیع پواسان با نرخ $\alpha = 2$ در هر ساعت است.**
  - احتمال سه درخواست در یک ساعت آینده چقدر است؟

p(3) = $e^{-2}2^3/3!$ = 0.18

همچنین به طور مشابه و با توجه به نمودار F(x) داریم:

p(3) = F(3) − F(2) = 0.857-0.677 = 0.18

  - احتمال دو یا بیشتر درخواست در مدت یک ساعت چقدر است؟

p(2 or more) = 1 − p(0) − p(1)
= 1 − F(1)
= 0.594

توزیع‌های پیوسته

- متغیرهای تصادفی پیوسته می‌توانند برای توصیف پدیده‌های تصادفی استفاده شوند که در آنها متغیر می‌تواند هر مقداری را در یک بازه داشته باشد (مقادیری از مجموعه اعداد حقیقی).
- در ادامه توزیع‌های پیوسته زیر معرفی می‌شوند:
  - یکنواخت (uniform)
  - نمایی (exponential)
  - نرمال (normal)
  - ویبول (Weibull)
  - لاگ‌نرمال (lognormal)

توزیع یکنواخت

- یک متغیر تصادفی X در بازه (a, b) دارای توزیع یکنواخت است و با U(a, b) نشان داده می‌شود اگر pdf و cdf آن به‌صورت زیر باشد:

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b\\ 0, & \text{otherwise}\end{cases}\qquad F(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\le x<b\\ 1, & x\ge b\end{cases}$$

- خصوصیت‌ها:
  - $P(x_1 < X < x_2)$ به نسبت طول بازه $[F(x_2) - F(x_1) = (x_2-x_1)/(b-a)]$ است.
  - امید ریاضی: $E(X) = (a+b)/2$ و واریانس: $V(X) = (b-a)^2/12$
- $U(0, 1)$ وسیله‌ای برای تولید اعداد تصادفی (random numbers) است که از این اعداد تصادفی می‌توان سایر متغیرهای تصادفی را تولید نمود.

توزیع نمایی

- یک متغیر تصادفی پیوسته X دارای توزیع نمایی با پارامتر λ که λ>0 است، اگر تابع چگالی احتمالی آن به‌صورت زیر باشد:

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0\\ 0, & x<0\end{cases}$$

  - یا اگر CDF آن به‌صورت زیر باشد:

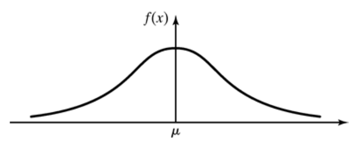
$$F(x)=\int_0^x \lambda e^{-\lambda x}dx=1-e^{-\lambda x}, x\ge 0$$

- توزیع نمایی دارای خصوصیت‌های جالبی است که در جلسه ۸ به بررسی دقیق‌تر آنها خواهیم پرداخت.

توزیع نرمال

- متغیر تصادفی X دارای توزیع نرمال است اگر pdf آن به صورت زیر باشد:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], -\infty<x<\infty$$



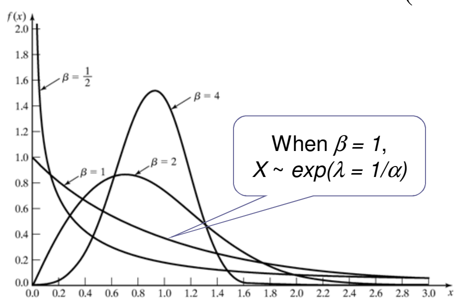
  - میانگین آن: $-\infty<\mu<\infty$
  - و واریانس آن: $\sigma^2>0$
  - و به شکل $X \sim N(\mu,\sigma^2)$ نشان داده می‌شود.
- خصوصیت‌های جالب توزیع نرمال:
  - خصوصیت‌های $f(x)$: $\lim_{x\to-\infty} f(x)=0$, and $\lim_{x\to\infty} f(x)=0$
  - pdf دارای تقارن حول μ است: $f(\mu-x)=f(\mu+x)$
  - حداکثر مقدار pdf در x = μ رخ می‌دهد.

توزیع ویبول

- یک متغیر تصادفی X دارای توزیع ویبول است اگر pdf آن به شکل زیر باشد:

$$f(x)=\begin{cases}\dfrac{\beta}{\alpha}\left(\dfrac{x-\nu}{\alpha}\right)^{\beta-1}\exp\left[-\left(\dfrac{x-\nu}{\alpha}\right)^{\beta}\right], & x\ge \nu\\ 0, & \text{otherwise}\end{cases}$$

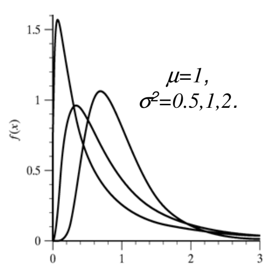
- این توزیع دارای سه پارامتر است:
  - پارامتر مکان (location): υ $(-\infty<\nu<\infty)$
  - پارامتر مقیاس (scale): β $(\beta>0)$
  - پارامتر شکل (shape): α $(\alpha>0)$
- مثال: $\upsilon = 0$ و $\alpha = 1$



توزیع لاگ‌نرمال

- یک متغیر تصادفی X که دارای توزیع لاگ‌نرمال (lognormal) است دارای pdf به شکل زیر است:

$$f(x)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\dfrac{(\ln x-\mu)^2}{2\sigma^2}\right], & x>0\\ 0, & \text{otherwise}\end{cases}$$



  - میانگین: $E(X) = e^{\mu+\sigma^2/2}$
  - واریانس: $V(X) = e^{2\mu+\sigma^2/2}(e^{\sigma^2} - 1)$
- رابطه توزیع لاگ‌نرمال با توزیع نرمال به این صورت است که:
  - اگر $Y \sim N(\mu, \sigma^2)$ آنگاه: $X = e^Y \sim \text{lognormal}(\mu, \sigma^2)$ (یا $Y = \ln X$)
  - توجه کنید که پارامترهای μ و σ² میانگین و واریانس لاگ‌نرمال نیستند، بلکه مربوط به نرمال هستند.

توزیع‌های تجربی

- به توزیعی که پارامترهایش مقادیر مشاهده شده (observed values) حاصل از داده‌های نمونه‌برداری شده باشد، توزیع تجربی (empirical distribution) گفته می‌شود.
  - این نوع توزیع‌ها در مواقعی که داشتن یک متغیر تصادفی ممکن یا ضروری نباشد، مورد استفاده قرار می‌گیرند.
  - برای این منظور داده‌های نمونه برداری شده تبدیل به جدولها یا نمودارهایی می‌شوند که با روشهای درون‌یابی (interpolation) یا برون‌یابی (extrapolation) می‌توان داده‌های ورودی بیشتری را از آن تولید نمود.
  - مزیت: هیچگونه فرضی خارج از مقادیر مشاهده شده در داده‌های نمونه لازم نیست انجام شود.
  - عیب: نمونه‌ها ممکن است که کل محدوده مقادیر امکان‌پذیر را پوشش ندهند.

جمع‌بندی

- در این جلسه انواع مدلهای احتمالی پیوسته و گسسته و خصوصیت‌ها و کاربردهایشان را به اختصار معرفی نمودیم.
- برای اطلاعات بیشتر به فصل مدلسازی ورودی (input modeling) مراجعه شود.
- در جلسه بعد این بحث را با فرآیندهای تصادفی پی می‌گیریم.