# پردازش گفتار

## مروری بر آمار و احتمال


هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران – دانشکده علوم و فنون نوین

# فهرست

# معرفی

- یک کلمه خاص (مانند "گفتار") را ۱۰ بار مختلف ضبط کنید
  - در هیچ دو حالتی فایل های ضبط شده دیجیتالی آنها دقیقاً یکسان نیست!
  - فرآیند تولید گفتار دارای ذات تصادفی است
- نقش پررنگ تصادف و عدم قطعیت در پردازش زبان گفتاری
  - فرموله کردن مسائل پردازش گفتار در یک چارچوب احتمالاتی
  - رایج‌ترین روش‌ها و الگوریتم‌های پردازش گفتار آماری هستند

# احتمال . . .

- بیان میزان اطمینان از خروجی وقایعی (مشاهده‌هایی) که قطعی نیستند

- تعاریف
  - فضای نمونه (Sample Space): مجموعه‌ای از تمام خروجی‌های ممکن = S
  - رویداد (Event): زیرمجموعه‌ای از فضای نمونه =A
  - احتمال (Probability) یک رویداد: فراوانی نسبی رخداد آن رویداد با فرض تکرار این فرایند به تعداد دفعات زیاد تحت شرایط مشابه= P(A)

$$P(A) = \frac{N_A}{N_S}$$

تعداد مشاهده‌هایی که خروجی آن‌ها متعلق به رویداد A است

تعداد کل تمام مشاهده‌ها

$$0 \le P(A) \le 1 \text{ for all } A$$

# احتمال . . .

- افراز (Partition)
  - اگر تعداد n رویداد $A_1$، $A_2$، ....، $A_n$ داشته باشیم که $\bigcup_{i=1}^{n} A_i = S$
  - آنگاه داریم $P(A_1 \cup A_2 \cup \ldots A_n) = \sum_{i=1}^{n} P(A_i) = 1$

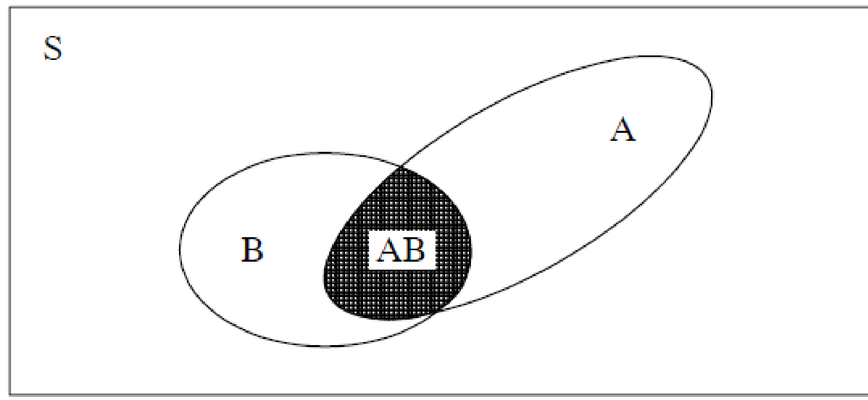
- احتمال توأم (Joint Probability)
  - برای دو رویداد A و B که به طور هم‌زمان اتفاق می‌افتند

$$P(AB) = \frac{N_{AB}}{N_S}$$

# احتمال . . .

- احتمال شرطی (Conditional Probability)
  - رخ دادن رویداد A با دانستن اینکه رویداد دیگری مانند Bرخ داده است

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{N_{AB}/N_S}{N_B/N_S}$$




- در بازشناسی گفتار

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{X}) = \arg\max_{\mathbf{W}} \frac{P(\mathbf{W})P(\mathbf{X}|\mathbf{W})}{P(\mathbf{X})}$$

کلمه بیان شده

سیگنال گفتاری

# احتمال . . .

## قاعده زنجیری (Chain Rule)

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{N_{AB}/N_S}{N_B/N_S} \Rightarrow P(AB)=P(A|B)P(B)=P(B|A)P(A)$$

- حالت ساده
- می‌تواند احتمال توأم چندین رویداد را بر حسب ضرب چند احتمال شرطی مشخص کند
- استفاده برای تجزیه یک مسئله احتمالاتی توأم پیچیده به زنجیره‌ای از احتمال‌های شرطی
- حالت کلی $P(A_1A_2\cdots A_n)=P(A_n|A_1\cdots A_{n-1})\cdots P(A_2|A_1)P(A_1)$

## مستقل (Independent) بودن

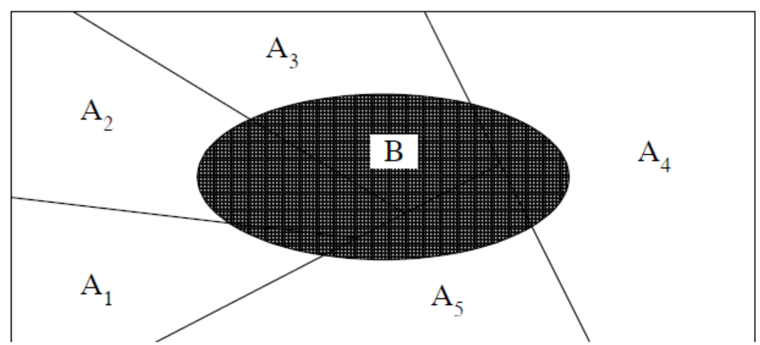
- رخ دادن یک رویداد هیچ ارتباط و تأثیری بر رخ دادن رویداد دیگر ندارد.
- احتمال شرطی: برابر با احتمال غیرشرطی است. $P(A|B)=P(A)$
- احتمال توأم : برابر با حاصل‌ضرب دو احتمال است. $P(AB)=P(A)\ P(B)$


Hadi Veisi (h.veisi@ut.ac.ir)

# احتمال . . .

- افراز (Partition) رویداد B
  - اگر تعداد n رویداد $A_1$، $A_2$، ....، $A_n$ یک افراز از S باشد و B یک رویداد در S باشد
  - آنگاه رویدادهای $BA_1$، $BA_2$، ....، $BA_n$ یک افراز از B را شکل می‌دهد

$$B = A_1B \cup A_2B \cup \cdots \cup A_nB$$

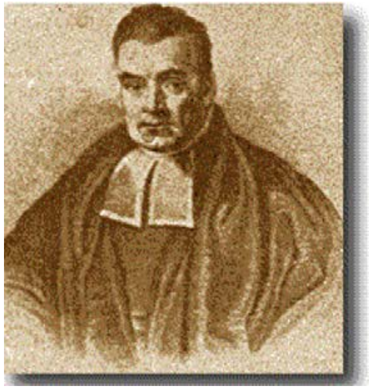



- چون رویدادهای $BA_1$، $BA_2$، ....، $BA_n$ مجزا هستند
  - احتمال رویداد B از حاصل جمع احتمال‌های توأم محاسبه می‌شود

$$P(B) = \sum_{k=1}^{n} P(A_kB)$$

احتمال حاشیه‌ای (Marginal Probability) رویداد B

# احتمال



Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370-418

## قانون بیز (Bayes´ Rule)

- با توجه به قاعده زنجیری

$$P(B) = \sum_{k=1}^{n} P(A_k)P(B \mid A_k)$$

$$\boxed{P(A_i \mid B) = \frac{P(A_i B)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k)P(A_k)}}$$

- این قانون مبنای بازشناسی الگو (مانند بازشناسی گفتار) است

مدل زبانی

مدل آوایی

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X}) = \arg\max_{\mathbf{W}} \frac{P(\mathbf{W})P(\mathbf{X} \mid \mathbf{W})}{P(\mathbf{X})}$$

کلمه بیان شده

سیگنال گفتاری

# متغیرهای تصادفی . . .

- متغیر تصادفی (Random Variable)
  - متغیر X که بیانگر یک کمیت عددی در یک فضای نمونه است
    - عناصر یک فضای نمونه را می‌توان شماره‌گذاری کرد و با آن شماره‌ها به آن‌ها ارجاع کرد.
  - تابعی که هر خروجی ممکن s در فضای نمونه S را به یک عدد حقیقی X(s) نگاشت می‌کند.
  - یک رویداد به‌صورت مجموعه‌ای از {s} نشان داده می‌شود که {s | X(s)=x}

- مثال: پرتاب سکه
  - فضای نمونه S = {شیر، خط}
  - متغیر تصادفی X

$$X(s) = \begin{cases} 1 & if \quad s = \text{شیر} \\ 0 & if \quad s = \text{خط} \end{cases}$$

- احتمال اینکه X=x باشد

$$P(X = x) = P(s \mid X(s) = x)$$

  - در مثال سکه

$$P(X = 1) = P(s|X(s) = 1) = P\left(s = \text{شیر}\right) = 0.5$$



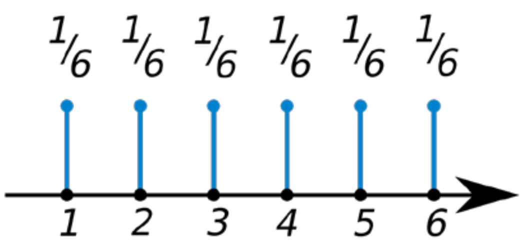
Hadi Veisi (h.veisi@ut.ac.ir)

# متغیرهای تصادفی . . .

- متغیر تصادفی گسسته (Discrete). . .
  - فقط تعداد متناهی n از مقادیر مختلف را می‌گیرد (دارای توزیع گسسته)
    - مثال: پرتاب تاس (فقط ۶ حالت دارد)
  - تابع احتمال (Probability Function)
    - یا تابع جرم احتمال (Probability Mass Function)
    - برای هر عدد حقیقی x، بیانگر میزان احتمال متغیر تصادفی گسسته است

$$p_X(x) = P(X = x)$$

- مثال: پرتاب تاس
  - متغیرهای تصادفی: 1 تا 6




- حاصل جمع جرم احتمال در تمام مقادیر متغیر تصادفی برابر با یک است

$$\sum_{k=1}^{n} p(x_i) = \sum_{k=1}^{n} P(X = x_i) = 1$$

# متغیرهای تصادفی . . .

- متغیر تصادفی گسسته (Discrete)

  - احتمال حاشیه‌ای $p_X(x_i) = P(X = x_i) = \sum_{k=1}^{m} P(X = x_i, Y = y_k) = \sum_{k=1}^{m} P(X = x_i \mid Y = y_k) P(Y = y_k)$

  - قانون زنجیری

$$P(X_1 = x_1, \cdots, X_n = x_n) = P(X_n = x_n \mid X_1 = x_1, \cdots, X_{n-1} = x_{n-1}) \cdots P(X_2 = x_2 \mid X_1 = x_1) P(X_1 = x_1)$$

  - قانون بیز

$$P(X = x_i \mid Y = y) = \frac{P(X = x_i, Y = y)}{P(Y = y)} = \frac{P(Y = y \mid X = x_i) P(X = x_i)}{\sum_{k=1}^{n} P(Y = y \mid X = x_k) P(X = x_k)}$$

  - اگر متغیرهای تصادفی X و Y به لحاظ آماری از هم مستقل باشند

$$P(X = x_i, Y = y_j) = P(X = x_i) P(Y = y_j) = p_X(x_i) p_Y(y_j) \ \forall \text{ all } i \text{ and } j$$

# متغیرهای تصادفی . . .

- متغیر تصادفی پیوسته (Continuous) . . .
  - دارای مقادیر پیوسته (و در نتیجه توزیع پیوسته) است
    - مثال: قد افراد یک کشور، مقدار دامنه سیگنال گفتار

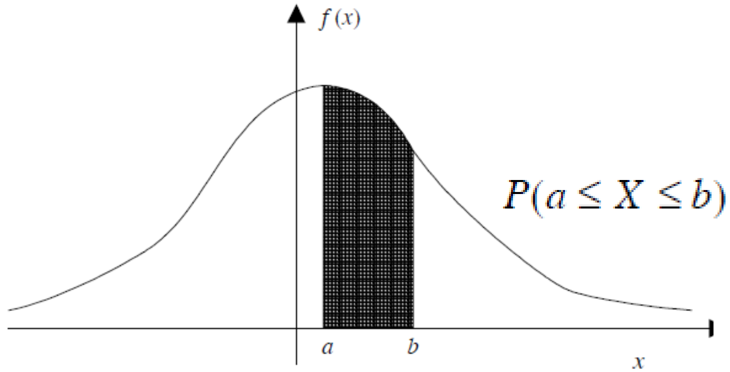
- اگر تابع غیرمنفی f وجود داشته باشد که روی مقادیر حقیقی تعریف شده و برای بازه A

$$f(x) \geq 0 \text{ for } -\infty \leq x \leq \infty$$

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

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

- $f_X$: تابع توزیع احتمال (pdf: Probability Density Function)
  - احتمال در یک بازه معنی دارد
  - احتمال در یک نقطه برابر با صفر است



# متغیرهای تصادفی . . .

- متغیر تصادفی پیوسته (Continuous) . . .
  - شرایط تابع pdf

$$f(x) \geq 0 \text{ for } -\infty \leq x \leq \infty$$

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

  - احتمال حاشیه‌ای

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)dy$$

  - قانون زنجیری

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n) = f_{X_n|X_1,\cdots,X_{n-1}}(x_n \mid x_1,\cdots,x_{n-1}) \cdots f_{X_2|X_1}(x_2 \mid x_1) f_{X_1}(x_1)$$

  - قانون بیز

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y \mid x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y \mid x) f_X(x)dx}$$

# متغیرهای تصادفی . . .

- تابع توزیع (Distribution Function)
  - یا تابع توزیع تجمعی (CDF: Cumulative Distribution Function)
  - بیانگر [جمع] احتمال‌های مقادیر کوچک‌تر از x
  - برای متغیر تصادفی گسسته یا پیوسته

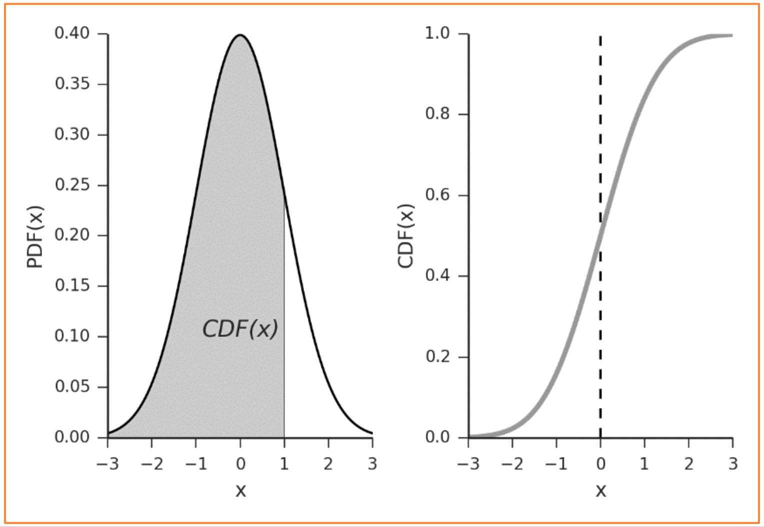
$$F(x) = P(X \le x) \quad \text{for } -\infty \le x \le \infty$$

مقدار حقیقی

  - برای حالت پیوسته داریم

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

$$f_X(x) = \frac{dF(x)}{dx}$$



# میانگین و واریانس . . .

- امید ریاضی (Expectation) یا میانگین (Mean)
  - برای متغیر تصادفی گسسته X

$$E(X)=\sum_x xf(x)$$

  - برای متغیر تصادفی پیوسته X

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

  - برای تابعی از متغیر X

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

    - مرکز جرم توزیع احتمال

  - امید ریاضی یک عملگر خطی است؛ دارای ویژگی‌های جمع‌پذیری و همگنی
    - حتی در صورت مستقل نبودن $X_i$ها

$$E(a_1X_1+\cdots+a_nX_n+b)=a_1E(X_1)+\cdots+a_nE(X_n)+b$$

مقدار ثابت

# میانگین و واریانس . . .

- امید ریاضی (Expectation) یا میانگین (Mean)
  - مثال ۱: انداختن یک تاس
    - متغیر تصادفی با شش مقدار $1, 2, \ldots, 6$ با احتمال برابر ۱/۶ برای هر کدام

$$E(X) = \sum_x xf(x)$$

$$= \tfrac{1}{6} \times 1 + \tfrac{1}{6} \times 2 + \tfrac{1}{6} \times 3 + \tfrac{1}{6} \times 4 + \tfrac{1}{6} \times 5 + \tfrac{1}{6} \times 6 = 3.5$$

  - مثال ۲: محاسبه معدل درسی یک دانشجو
    - متغیر تصادفی (x)؟
      - نمره درس
    - احتمال (تابع توزیع)؟
      - متناسب با تعداد واحد درس‌ها (تعداد واحد درس تقسیم بر کل واحد

| درس | نمره | تعداد واحد |
|---|---|---|
| آواشناسی | ۱۸ | ۲ |
| برنامه نویسی | ۱۲ | ۴ |
| ریاضیات | ۱۵ | ۳ |
| روش تحقیق | ۲۰ | ۱ |

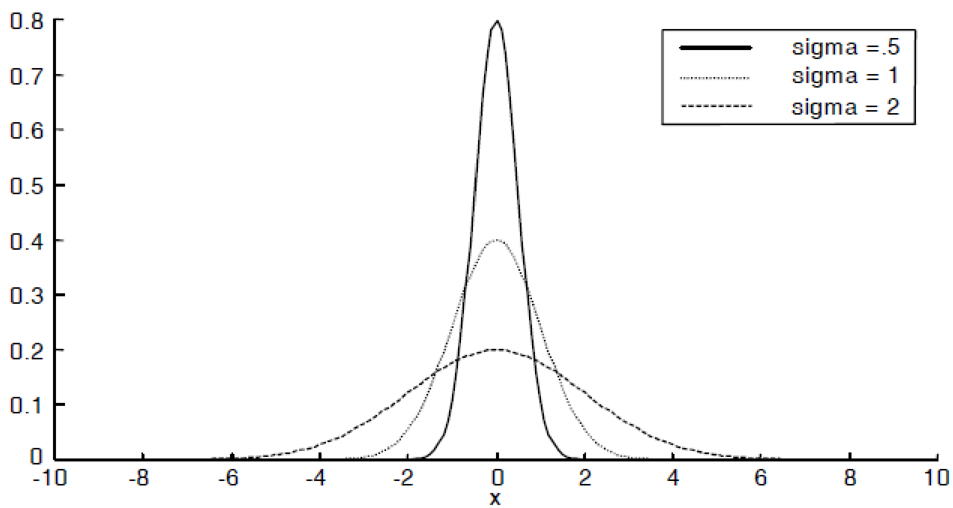
$$E(x) = \frac{2}{10} \times 18 + \frac{4}{10} \times 12 + \frac{3}{10} \times 15 + \frac{1}{10} \times 20 = 14.9$$

# میانگین و واریانس . . .

- واریانس (Variance) . . .
  - میانگین متغیر X: μ=E(X)

$$Var(X) = \sigma^2 = E\left[(X-\mu)^2\right]$$

  - مجذور غیرمنفی واریانس = انحراف معیار (Standard Deviation) = σ
  - بیانگر میزان پراکندگی یا انتشار توزیع در اطراف میانگین
    - مقدار کوچک واریانس = توزیع فشرده احتمال در اطراف میانگین
    - مقدار بزرگ واریانس = توزیع احتمال در اطراف میانگین دارای پراکندگی بیشتری است



# میانگین و واریانس . . .

## واریانس (Variance)

- ممان (گشتاور) kام X = امید ریاضی $E(X^k)$
  - برای هر متغیر تصادفی X و هر عدد صحیح مثبت k
- برای واریانس داریم

$$Var(X) = \sigma^2 = E\left[(X-\mu)^2\right] = E(X^2) - \left[E(X)\right]^2$$

- واریانس اختلاف بین ممان دوم و مربع ممان اول است
- ویژگی‌های واریانس
  - جمع‌پذیری: در صورت مستقل بودن متغیر تصادفی X و Y $\quad Var(X+Y) = Var(X) + Var(Y)$
  - همگنی: برقرار نیست $\quad Var(aX) = a^2 Var(X)$
  - واریانس مقدار ثابت صفر است

$$Var(a_1X_1 + \cdots + a_nX_n + b) = a_1^2 Var(X_1) + \cdots + a_n^2 Var(X_n)$$

# میانگین و واریانس . . .

## امید ریاضی شرطی (Conditional Expectation)

- برای متغیرهای گسسته X و Y
  - امید ریاضی شرطی Y: تابعی از X

$$E_{Y|X}(Y \mid X = x) = \sum_y y f_{Y|X}(y \mid x)$$

- برای متغیرهای پیوسته X و Y

$$E_{Y|X}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy$$

- خود E(Y | X) یک متغیر تصادفی است
  - تابعی از متغیر تصادفی X است

$$E_X\left[E_{Y|X}(Y \mid X)\right] = E_{X,Y}(Y)$$

- فرض کنید X و Y یک توزیع توأم پیوسته دارند و g(X,Y) تابعی از  Xو Y است

$$E_{Y|X}\left[g(X,Y) \mid X = x\right] = \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y \mid x) dy$$

$$E_X\left\{E_{Y|X}\left[g(X,Y) \mid X\right]\right\} = E_{X,Y}\left[g(X,Y)\right]$$

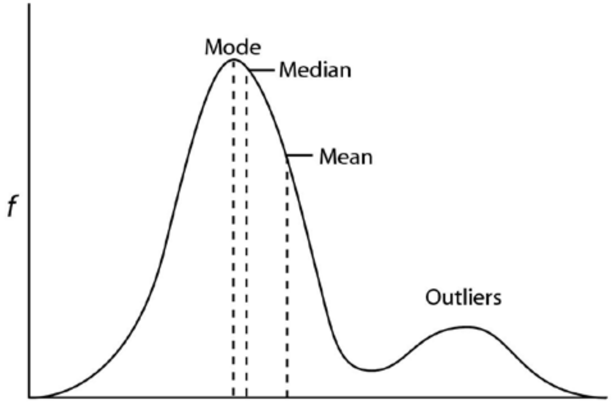
# میانگین و واریانس

- میانه (Median)
  - احتمال کل را به دو قسمت مساوی تقسیم می‌کند
  - میانه توزیع متغیر X نقطه‌ای مانند m است

$$P(X \le m) \ge 1/2 \text{ and } P(X \ge m) \ge 1/2$$

  - احتمال سمت چپ m و احتمال سمت راست m دقیقاً 0.5 است

- نما (Mode)
  - جایی که تابع توزیع دارای بیشترین مقدار خود است
  - یک توزیع می‌تواند بیش از یک میانه داشته باشد



# قانون اعداد بزرگ . . .

- قانون اعداد بزرگ (Law of Large Numbers)
  - میانگین نمونه (Sample Mean) و واریانس نمونه (Sample Variance)
    - مقدار میانگین و واریانس تعدادی از نمونه‌های حاصل از یک آزمایش آماری است (آنچه ما در عمل محاسبه می‌کنیم)
- فرض کنید یک توزیع با میانگین $\mu$ و واریانس $\sigma^2$ داریم
    - متغیرهای تصادفی $x_1, x_2, \ldots, x_n$ از این توریع تولید می‌شوند
  - متغیرهای تصادفی i.i.d: - Independent Identically Distributed
    - مستقل و با توزیع یکسان
    - هر یک دارای $\mu$ و واریانس $\sigma^2$
  - میانگین حسابی n نمونه
    - همان میانگین نمونه

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

خودش متغیری تصادفی است

# قانون اعداد بزرگ

- داریم
  - میانگین ”میانگین نمونه“ $E(\bar{X}_n) = \mu$
    - میانگین ”میانگین نمونه“ برابر با میانگین توزیع است
  - واریانس ”میانگین نمونه“ $\mathrm{Var}(\bar{X}_n) = \dfrac{\sigma^2}{n}$
    - واریانس ”میانگین نمونه“برابر با 1/n واریانس توزیع است
    - توزیع ”میانگین نمونه“به نسبت توزیع اصلی در اطراف میانگین متمرکزتر است
- قانون اعداد بزرگ
  - بیان می‌کند ”میانگین نمونه“ به میانگین توزیع نزدیک می‌شود
    - وقتی اندازه نمونه (n) بزرگ باشد

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| < \varepsilon\right) = 1 \quad \text{for any given number } \varepsilon > 0$$

# کوواریانس و همبستگی . . .

- کواریانس متغیرهای تصادفی X و Y

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = Cov(Y,X)$$

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

- ضریب همبستگی (Correlation Coefficient)

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

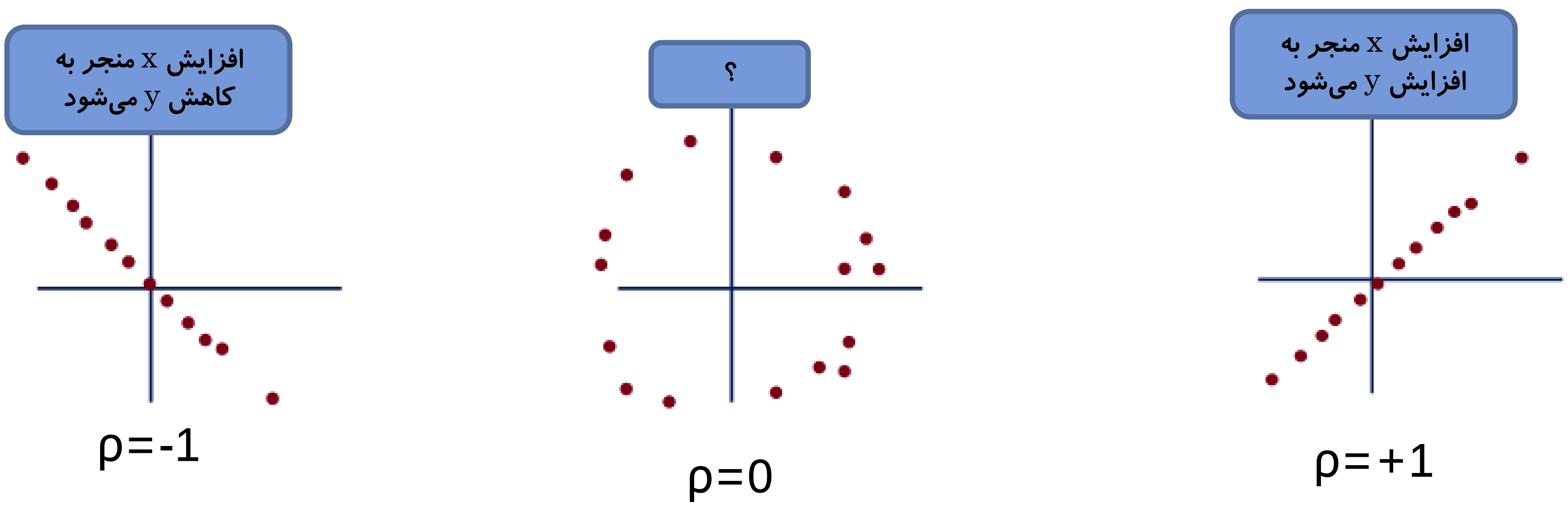
  - مقدار در بازه ۱ و ۱- $-1 \le \rho(X,Y) \le 1$
  - بیانگر همبستگی خطی (linear dependency) بین دو متغیر X و Y
  - دو متغیر تصادفی X و Y متعامد (Orthogonal) هستند اگر $E(XY) = 0$



Hadi Veisi (h.veisi@ut.ac.ir)

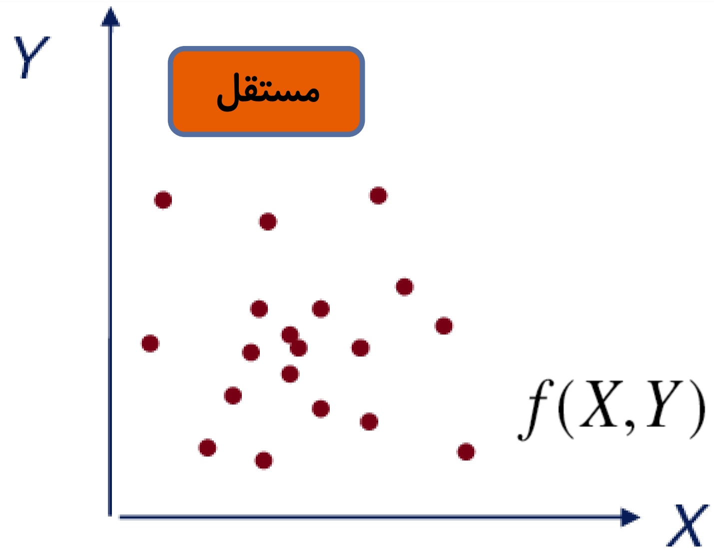
# کوواریانس و همبستگی . . .

- ضریب همبستگی (Correlation Coefficient)
  - اگر $\rho>0$ باشد، X و Y همبستگی مثبت دارند
  - اگر $\rho<0$ باشد، همبستگی منفی دارند
  - اگر $\rho=0$ باشد، همبسته (Correlated) نیستند



# کوواریانس و همبستگی . . .

- متغیرهای تصادفی مستقل (Independent)
  - اگر داشته باشیم $Cov(X,Y) = \rho_{XY} = 0$
  - آیا ناهمبسته بودن (uncorrelated)، استقلال (independence) را نتیجه می‌دهد؟
    - مستقل بودن، ناهمبسته بودن را نیز نتیجه می‌دهد، اما برعکس آن درست نیست (به غیر از توزیع نرمال)




$$f(X,Y) = f(X)\,f(Y)$$

# کوواریانس و همبستگی

- چند قضیه
  - اگر رابطه متغیرهای X و Y به صورت Y=aX+b باشد، آنگاه
    - اگر a>0 باشد، $\rho_{XY} = +1$
    - اگر a<0 باشد، $\rho_{XY} = -1$
  - برای هر دو متغیر X و Y داریم

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

  - اگر n متغیر تصادفی $x_1, x_2, \ldots, x_n$ داشته باشیم، آنگاه

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} Cov(X_i, X_j)$$

# بردارهای تصادفی . . .

## بردار تصادفی

- وقتی یک متغیر تصادفی یک بردار باشد و نه یک عدد

$$\mathbf{X} = (X_1, \cdots, X_n)$$

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

برداری با n مولفه

- بردار میانگین
  - یک بردار n بعدی که مؤلفه‌های آن امید ریاضی‌های تک تک مؤلفه‌های $\mathbf{X}$ است

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix}$$

- ماتریس کوواریانس
  - مؤلفه‌های قطر اصلی ماتریس کوواریانس = واریانس‌های هر کدام از $X_i$ها
  - کوواریانس متقارن است $Cov(X_i, X_j) = Cov(X_j, X_i)$

$$Cov(\mathbf{X}) = E\left[[X - E(X)][X - E(X)]^t\right] = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & & \vdots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{bmatrix}$$

# بردارهای تصادفی

- رابطه خطی بردارهای تصادفی
  - بردار n بعدی X
  - بردار m بعدی Y

یک ماتریس m×n

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{B}$$

یک بردار m بعدی

- میانگین و کوواریانس Y بر حسب میانگین و کوواریانس X
  - کاربرد در تبدیل ویژگی‌ها (کاهش بعد)
  - کاربرد در تطبیق (adapt) پارامترهای مدل‌های آوایی و زبانی

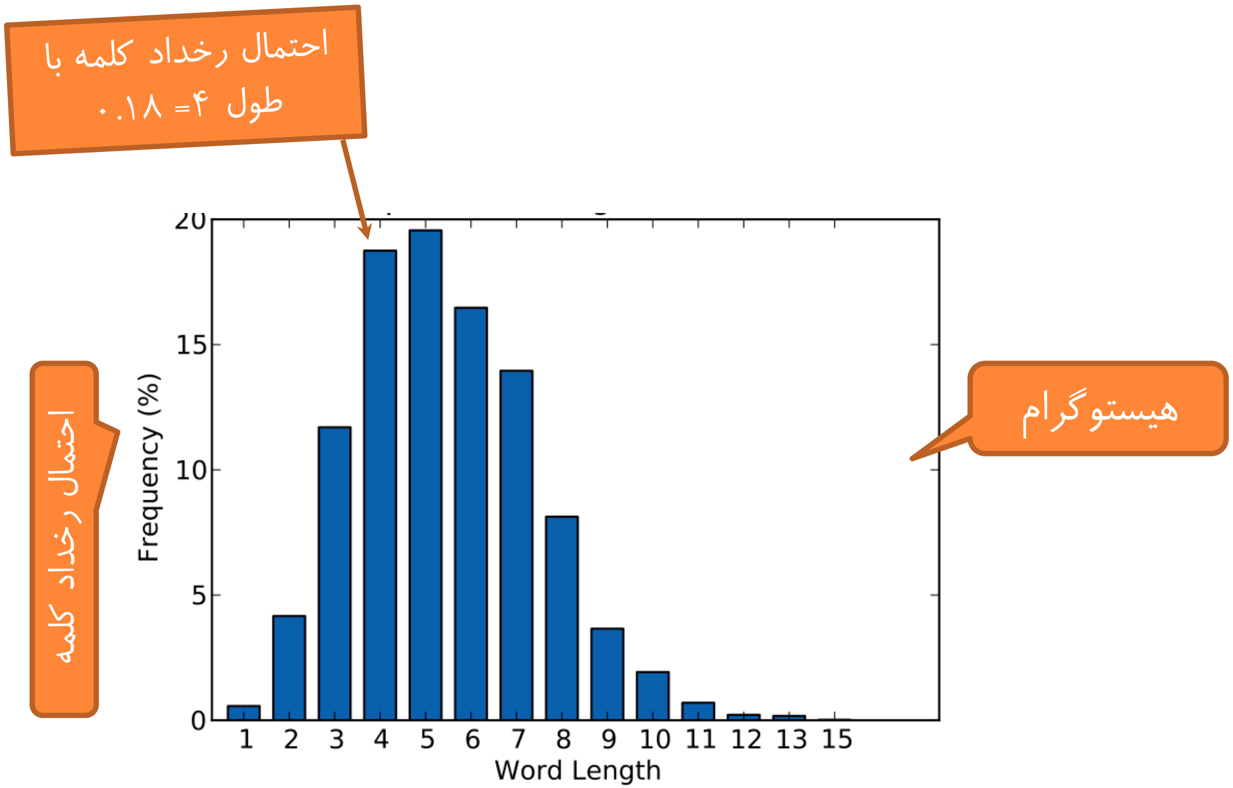
$$E(\mathbf{Y}) = \mathbf{A}E(\mathbf{X}) + \mathbf{B}$$

$$Cov(\mathbf{Y}) = \mathbf{A}Cov(\mathbf{X})\mathbf{A}^t$$

ترانهاده

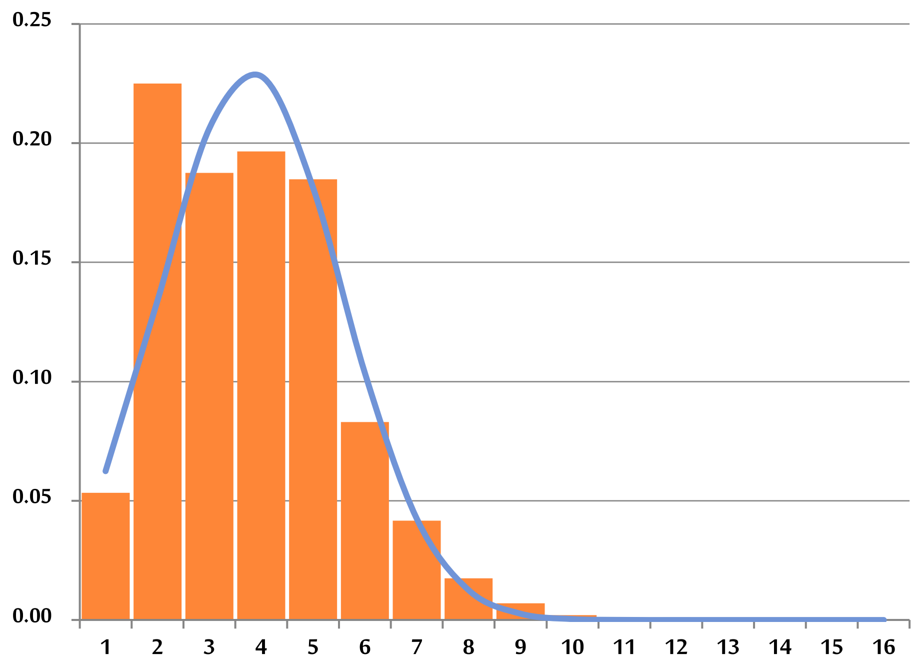
# توابع توزیع . . .

- مثال: پردازش متن (طول کلمات) . . .
  - یک لغت‌نامه داریم
  - می‌خواهیم تعداد کلمات با طول ۱ حرف، با طول ۲ حرف، ...، طول ۲۰ حرف را بشماریم



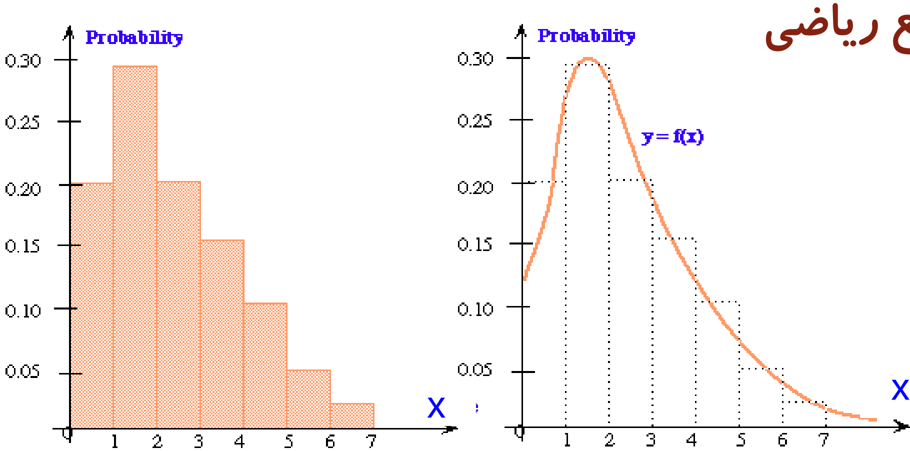
# توابع توزیع . . .

- مثال: پردازش متن (طول کلمات)– برای فارسی
  - روی پیکره کوچک
    - متوسط طول کلمات: ۳.۸

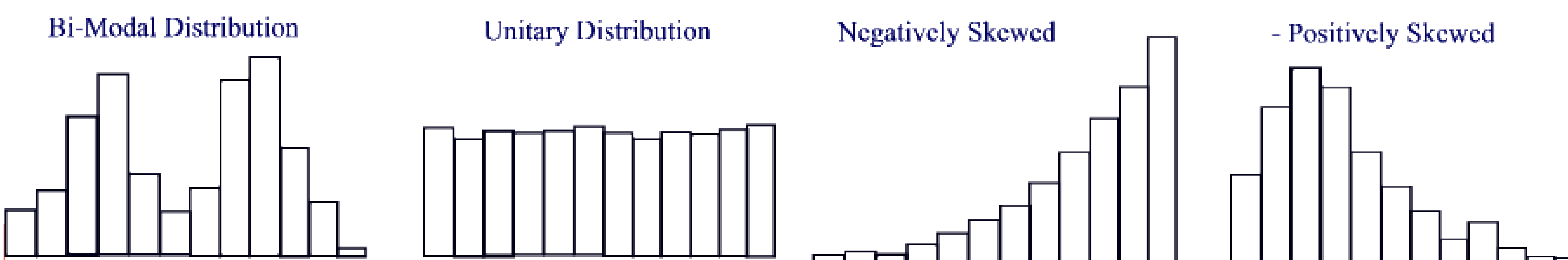


# توابع توزیع . . .

- از هیستوگرام به تابع توزیع
  - نمایش شکل توزیع احتمال‌ها با یک تابع ریاضی




  - می‌تواند شکل‌های مختلفی داشته باشد



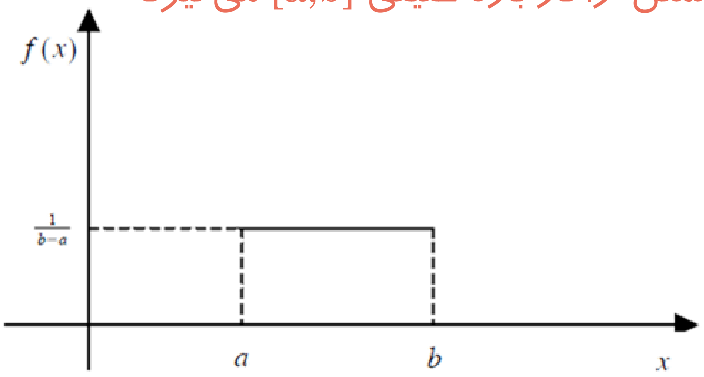
# توابع توزیع . . .

- توزیع یکنواخت (Uniform Distribution)
  - تابع احتمال یا تابع توزیع احتمال، یک تابع ثابت است
  - احتمال وقوع همه مقادیر یکسان است
    - مثال: احتمال انتخاب هر نقطه در یک بازه عددی مشخص
  - برای متغیر گسسته
    - X فقط مقادیر ممکن از ۱ تا n را می‌گیرد

$$P(X = x_i) = \frac{1}{n} \quad 1 \le i \le n$$

  - برای متغیر پیوسته
    - X فقط مقادیر ممکن را در بازه حقیقی [a,b] می‌گیرد

$$f(x) = \frac{1}{b-a} \quad a \le x \le b$$



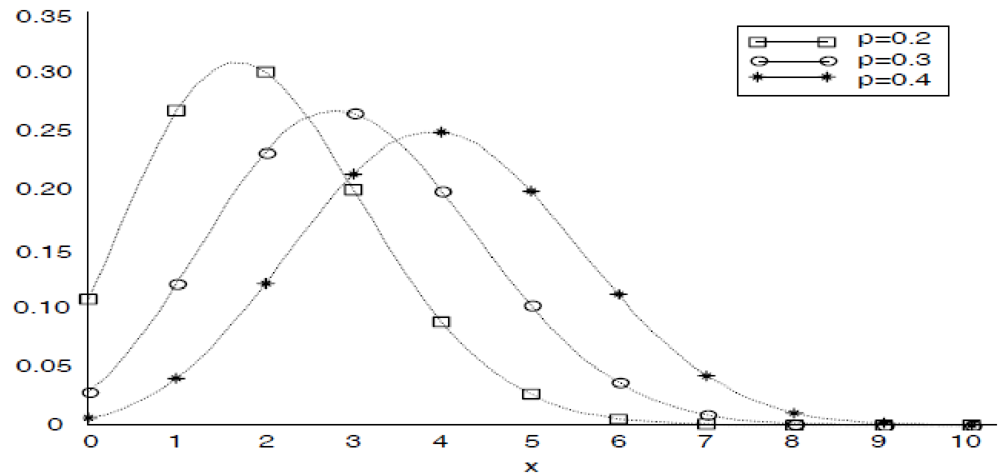
# توابع توزیع . . .

## توزیع دوجمله‌ای (Binomial Distribution) ...

- برای توصیف رویدادهایی با تصمیم‌گیری دودویی
  - آزمایش برنولی: شکست (۰) یا پیروزی (۱)
- مثال: انداختن یک سکه (دو حالت شیر یا خط)
  - احتمال آمدن شیر p و احتمال خط 1-p
- سکه را n بار بیندازیم (n بار تکرار مستقل آزمایش برنولی) و تعداد شیرهای مشاهده شده (مجموع پیروزی‌ها) را با Xنشان دهیم، متغیر تصادفی X دارای تابع احتمال دوجمله‌ای است

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

تعداد x بار از n بار شیر بیاید



# توابع توزیع . . .

## توزیع دوجمله‌ای (Binomial Distribution)

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

- میانگین و واریانس توزیع دوجمله‌ای

$$E(X)=np$$

$$Var(X)=np(1-p)$$

- مثال: اگر یک تیرانداز با احتمال ۰/۷ تیری را به هدف بزند و x تعداد تیرهای به هدف خورده در ۵ شلیک باشد
  - تابع توزیع

$$p(k)=\binom{n}{k}(0.7)^k(0.3)^{5-k}$$

  - احتمال اینکه دقیقاً ۳ تیر به هدف بزند

$$P(X=3)=\binom{5}{3}(0.7)^3(0.3)^2=0.3087$$

  - احتمال اینکه حداکثر ۲ تیر به هدف بزند

$$P(X\le 2)=p(0)+p(1)+p(2)=0.16308$$

- مثال: امتحان چهارگزینه‌ای
  - پیروزی =گزینه درست (۰.۲۵)، شکست =سایر گزینه‌ها(۰.۷۵)



Hadi Veisi (h.veisi@ut.ac.ir)

# توابع توزیع . . .

- توزیع هندسی (Geometric Distributions) . . .
  - بیانگر تعداد آزمایش‌های (زمان) برنولی تا رسیدن به اولین پیروزی در توزیع دوجمله‌ای
  - مثال: انداختن سکه – تعداد آزمایش‌هایی که تکرار می‌شود تا یک بار خط بیاید
  - مثال: تعداد بارهای انداختن تاس تا آمدن شش در منج!
  - مثال: چند مرتبه آزمایش کنیم تا یک رمز عبور ۸ کاراکتری یک کامپیوتر را حدس بزنیم

  - X=متغیر تصادفی زمان (تعداد انداختن‌ها) قبل از آمدن اولین خط
    - احتمال آمدن شیر = p و احتمال آمدن خط = 1-p

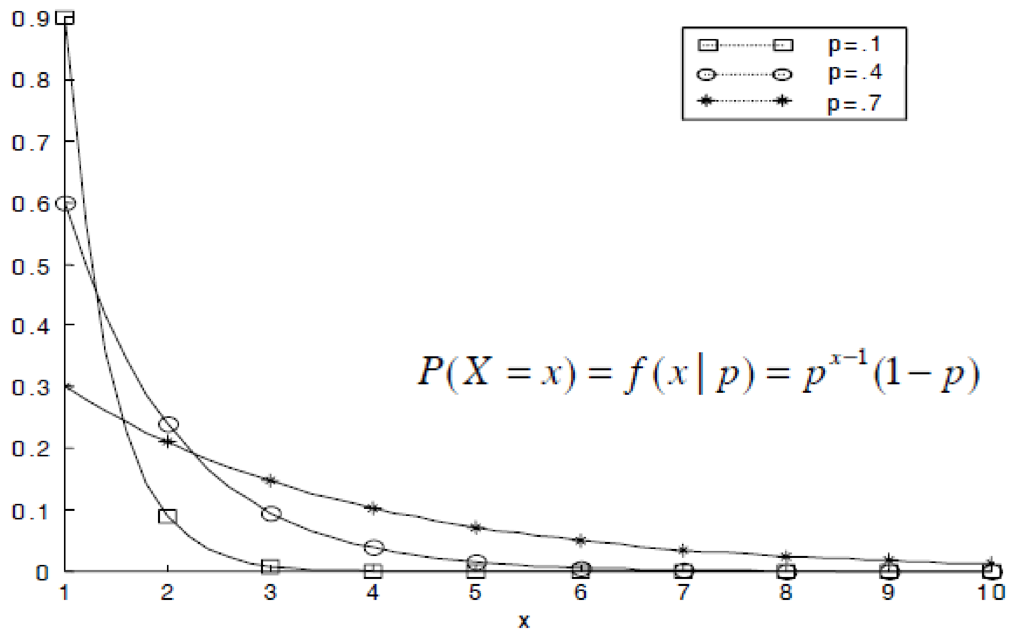
$$P(X=x)=f(x\mid p)=p^{x-1}(1-p) \quad x=1,2,\dots \text{ and } 0<p<1$$

  - میانگین و واریانس

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

# توابع توزیع . . .

- توزیع هندسی (Geometric Distributions)




- مثال: توزیع طول حالت (State Duration) در مدل مخفی مارکوف (HMM)
  - بیانگر مدت زمانی است که در یک حالت خاص باقی می‌مانیم

# توابع توزیع . . .

- توزیع چندجمله‌ای (**Multinomial Distribution**) . . .
  - **حالت کلی‌تر توزیع دوجمله‌ای با k حالت (به جای دو حالت)**
  - **مثال: کیسه‌ای حاوی توپ‌هایی با k رنگ مختلف داریم**
    - نسبت توپ‌های رنگ i برابر با $p_i$ است
    - فرض کنید n توپ به طور تصادفی از کیسه انتخاب شده‌اند
    - فرض کنید $X_i$ بیانگر تعداد توپ‌های انتخاب شده با رنگ i باشد
    - آنگاه بردار تصادفی $\mathbf{X}=(X_1,\ldots X_k)$ دارای توزیع چندجمله‌ای با پارامترهای n و $\mathbf{p}=(p_1,..p_k)$ است

$$P(\mathbf{X}=\mathbf{x}) = f(\mathbf{x}\,|\,n,\mathbf{p}) = \begin{cases} \dfrac{n!}{x_1!,\ldots x_k!} p_1^{x_1},\ldots p_k^{x_k} & \text{where } x_i \ge 0 \ \forall i = 1,\ldots,k \\ & \text{and } x_1 + \cdots + x_k = n \\ 0 & \text{otherwise} \end{cases}$$

  - **توزیع‌های چندجمله‌ای معمولاً با آزمون مربعات-کای به کار می‌رود**
    - از پرکاربردترین روش‌های آزمون فرضیه‌های نیکویی برازش (Goodness-of-Fit)

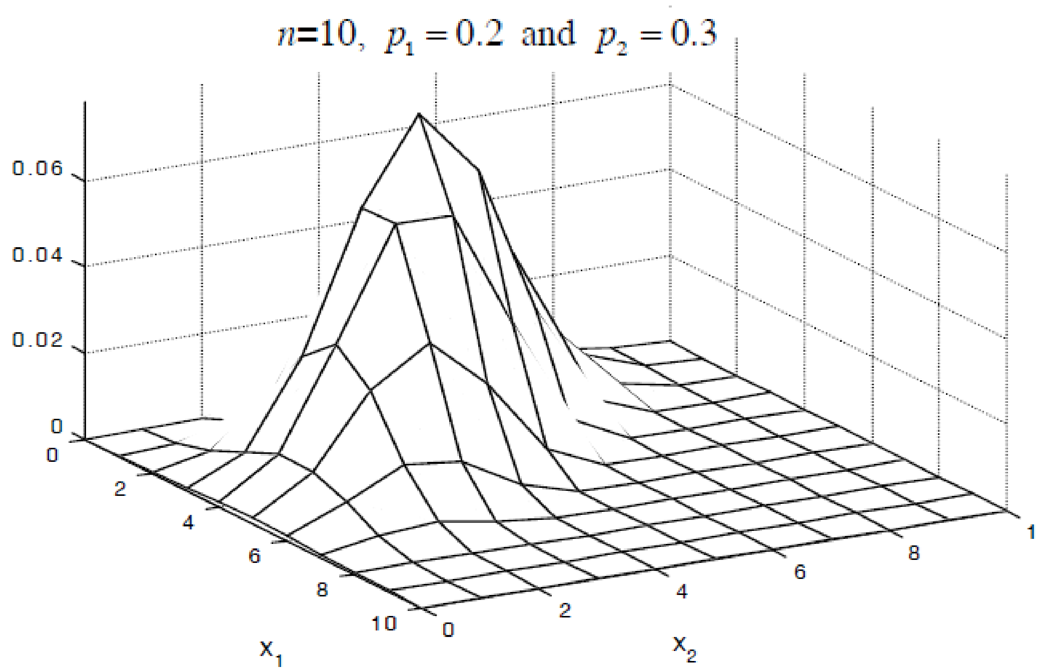
# توابع توزیع . . .

- توزیع چندجمله‌ای (Multinomial Distribution)
  - میانگین، واریانس و کواریانس

$$E(X_i) = np_i$$

$$Var(X_i) = np_i(1-p_i) \quad \forall i = 1,\ldots,k$$

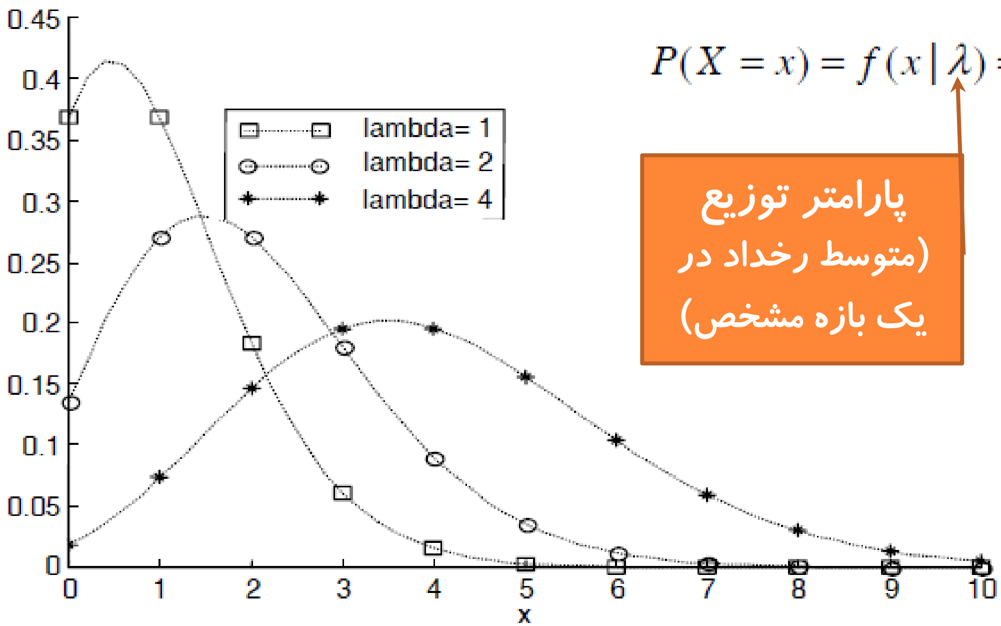
$$Cov(X_i, X_j) = -np_i p_j$$



# توابع توزیع . . .

## توزیع پواسون (Poisson Distribution)

- از توزیع‌های گسسته متداول
- X= تعداد کل رخدادهای یک پدیده در طول یک مدت زمان یا یک منطقه مکانی ثابت
  - تعداد تماس‌های تلفنی دریافت شده توسط یک مرکز مخابراتی در یک مدت زمان ثابت




$$P(X=x)=f(x\mid\lambda)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{x}}{x!} & \text{for } x=0,1,2,\ldots\\ 0 & \text{otherwise}\end{cases}$$

پارامتر توزیع (متوسط رخداد در یک بازه مشخص)

- میانگین و واریانس

$$E(X)=Var(X)=\lambda$$

- کاربرد در پردازش گفتار
  - برای مدل‌سازی دیرش (duration) یک واج

# توابع توزیع . . .

- توزیع گاما (Gamma Distribution) . . .
  - متغیر تصادفی پیوسته X

پارامتر شکل (shape)

پارامتر مقیاس (scale)

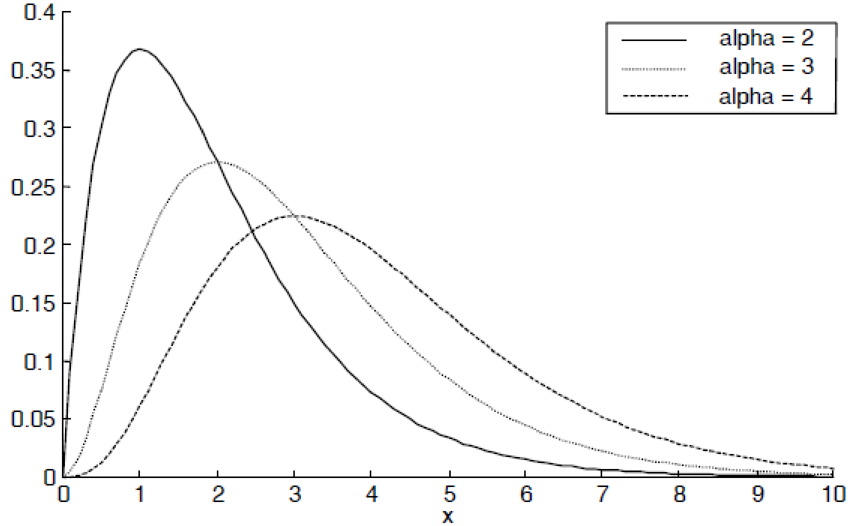
$$f(x|\alpha,\beta)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} & x>0\\ 0 & x\le 0\end{cases}$$

تابع گاما

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

$$\Gamma(n)=\begin{cases}(n-1)! & n=2,3,\dots\\ 1 & n=1\end{cases}$$

تابع فاکتوریل برای n حقیقی




- میانگین و واریانس

$$E(X)=\frac{\alpha}{\beta}\ \text{ and }\ Var(X)=\frac{\alpha}{\beta^2}$$

Hadi Veisi (h.veisi@ut.ac.ir)

# توابع توزیع . . .

## توزیع گاما (Gamma Distribution)

- اگر متغیرهای تصادفی $X_1, \ldots, X_k$ مستقل باشند
  و هر متغیر تصادفی $X_i$ یک توزیع گاما با پارامترهای $\alpha_i$ و $\beta$ داشته باشد،
  آنگاه مجموع $X_1 + \ldots + X_k$ نیز یک توزیع گاما با پارامترهای $\alpha_1 + \ldots + \alpha_k$ و $\beta$ دارد

- در پردازش گفتار
  - سیگنال گفتار نویزی دارای توزیع گاما است (دامنه طیف)
  - فرکانس‌های بالا ی سیگنال گفتار در حوزه DCT و بخش‌های حقیقی و موهومی طیف

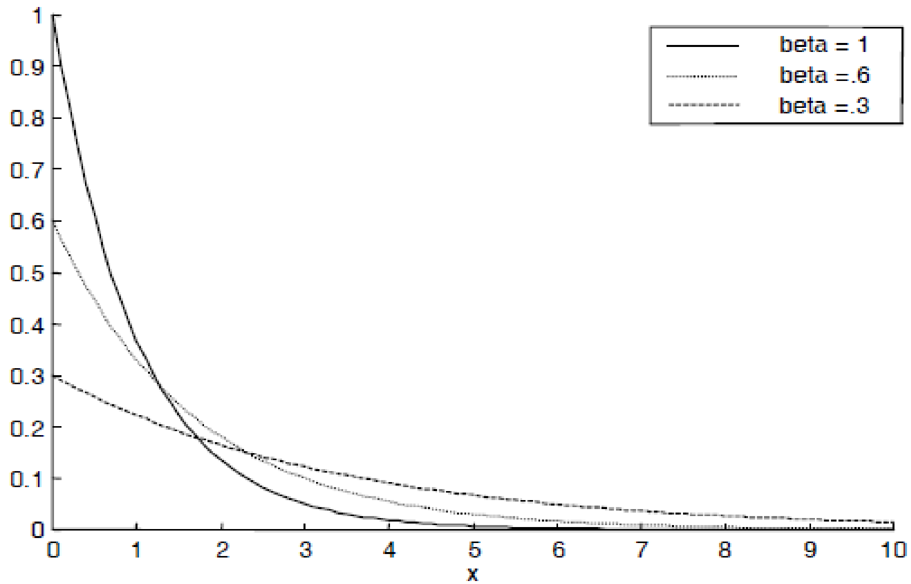
# توابع توزیع . . .

## توزیع نمایی (Exponential Distribution)

- حالت خاصی از توزیع گاما که $\alpha=1$

$$f(x|\beta)=\begin{cases}\beta e^{-\beta x} & x>0\\ 0 & x\le 0\end{cases}$$

- تخمین زدن مدت زمان لازم برای رخداد یک پیشامد خاص
  - زمان ورود مشتری‌ها
  - طول عمر یک وسیله
  - زمان بین دو رویداد در فرایند پواسن




- میانگین و واریانس

$$E(X)=\frac{1}{\beta}\quad \text{and}\quad Var(X)=\frac{1}{\beta^2}$$

# توابع توزیع . . .

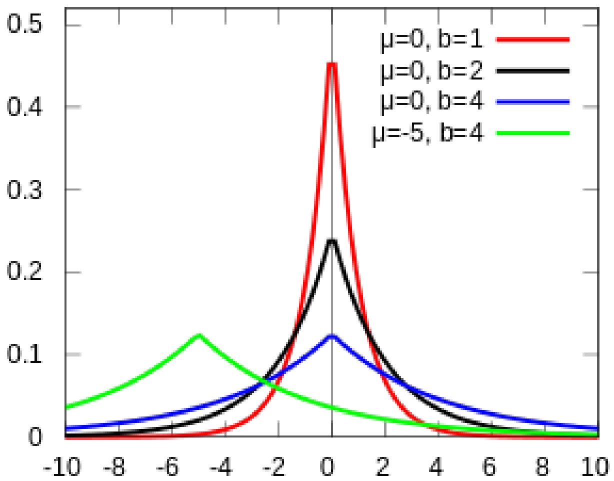
## توزیع لاپلاس (Laplace Distribution)

- برای متغیر پیوسته
- ترکیب دو توزیع گاما

پارامتر مقیاس (scale)

پارامتر مکان (location)

$$f(x|\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) = \frac{1}{2b}\begin{cases}\exp\left(-\frac{\mu-x}{b}\right) & \text{if } x<\mu\\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x\geq\mu\end{cases} = \frac{1}{2\,b}\exp\left(-\frac{|x-\mu|}{b}\right)$$




- میانگین=میانه=نما=μ
- واریانس = $2b^2$

- کاربرد در پردازش گفتار
  - سیگنال گفتار تمیز دارای توزیع لاپلاس است (دامنه طیف)
  - فرکانس‌های پایین سیگنال گفتار در حوزه DCT و بخش‌های حقیقی و موهومی طیف

# توابع توزیع . . .

- توزیع گاوسی (Gaussian Distribution) . . .
  - یا توزیع نرمال (Normal Distribution)
  - مهم‌ترین توزیع احتمال
    - متغیرهای تصادفی مطالعه شده در آزمایش‌های مختلف فیزیکی (از جمله سیگنال‌های گفتاری) دارای توزیع‌هایی هستند که تقریباً گاوسی است
    - محاسبات آن آسان است (به ویژه در تخمین‌ها)
    - قضیه حد مرکزی (Central Limit Theorem)
  - برای یک متغیر تصادفی پیوسته

میانگین   واریانس

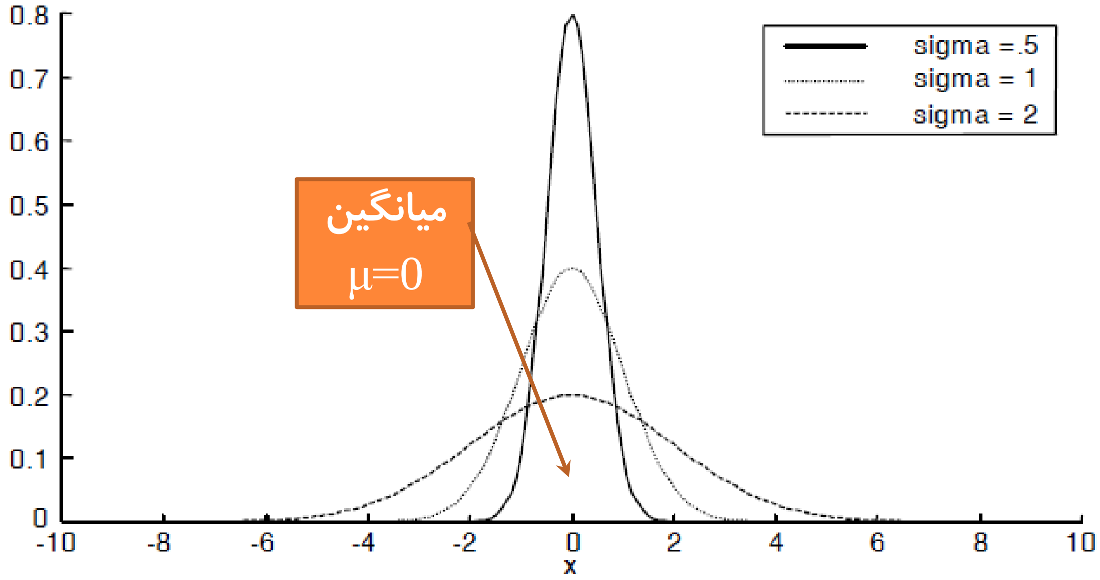
$$f(x|\mu,\sigma^2) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

  - تابع گاوسی حول میانگین تقارنی است
  - میانگین، میانه (نقطه تقارن) و نمای (بیشینه مقدار) توزیع یک نقطه است

# توابع توزیع . . .

- توزیع گاوسی (Gaussian Distribution) . . .

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




- کاهش واریانس=تراکم بیشتر حول میانگین

# توابع توزیع . . .

- توزیع گاوسی (Gaussian Distribution) . . .
  - اگر متغیر تصادفی $X$ یک توزیع گاوسی با میانگین $\mu$ و واریانس $\sigma^2$ باشد
    آنگاه هر تابع خطی $Y=aX+b$ نیز یک توزیع گاوسی دارد
    $Y$ یک توزیع گاوسی با میانگین $a\mu+b$ و واریانس $a^2\sigma^2$ دارد
  - حالت کلی: مجموع $X_1+ \dots +X_n$ از متغیرهای تصادفی مستقل $X_1, \dots, X_n$ نیز، که در آن هر متغیر تصادفی $X_i$ یک توزیع گاوسی دارد، یک توزیع گاوسی است
  - توزیع گاوسی استاندارد یا توزیع گاوسی = $N(0,1)$
    - میانگین صفر و واریانس یک
    - رفتار توزیع گاوسی را می‌توان فقط با استفاده از توزیع گاوسی استاندارد توضیح داد
      - تبدیل خطی توزیع گاوسی یک توزیع گاوسی است
    - اگر متغیر تصادفی $X$ یک توزیع گاوسی با میانگین $\mu$ و واریانس $\sigma^2$ باشد، می‌توان نشان داد

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

Hadi Veisi (h.veisi@ut.ac.ir)

# توابع توزیع . . .

- توزیع گاوسی چندمتغیره (Multivariate)
  - برای بردار تصادفی n بعدی

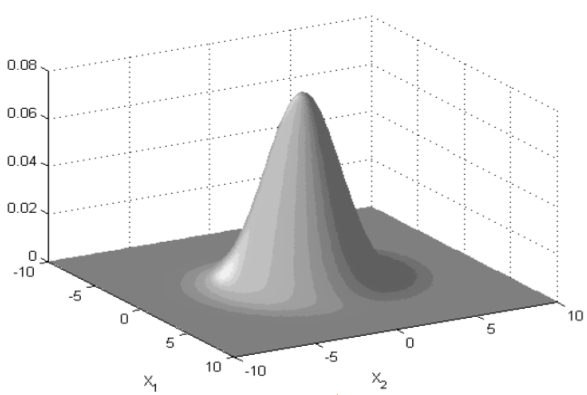
n=تعداد ابعاد بردار x(متغیرها)

$$f(\mathbf{X}=\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}\,|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{t}\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

| . | = دترمینان

t= ترانهاده

-1 = معکوس




برای دو بعد

- میانگین

$$\boldsymbol{\mu} = E(\mathbf{x})$$

- کواریانس

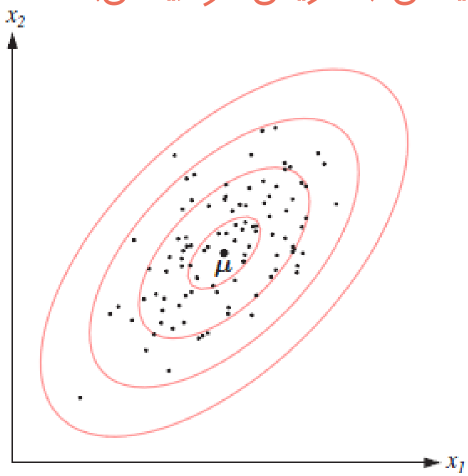
$$\boldsymbol{\Sigma} = E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{t}\right]$$

  - متقارن و مثبت-معین (positive definite) (دترمینان مثبت)
  - عناصر قطر اصلی $\sigma_{ii}$= واریانس متغیر متناسب $x_i$
  - عناصر غیرقطر اصلی $\sigma_{ij}$= کواریانس متغیرهای $x_i$ و $x_j$

$$\sigma_{ij}^{2} = E\left[(x_i-\mu_i)(x_j-\mu_j)\right]$$

# توابع توزیع . . .

- شکل داده‌های دارای توزیع نرمال
  - نمونه‌های داده که از توزیع نرمال پیروی می‌کنند، داخل یک خوشه قرار می‌گیرد
  - مرکز خوشه = میانگین، شکل خوشه (بیضی شکل) = تعیین شده توسط واریانس (ماتریس کوواریانس)
  - بردار ویژه ماتریس کوواریانس = محورهای اصلی بیضی
  - مقادیر ویژه ماتریس کوواریانس = طول محورهای اصلی بیضی




- فاصله ماهالونوبیس (Mahalonobis distance)
  - فاصله بین یک مجموعه داده معین (با پارامترهای میانگین و واریانس) و یک نمونه داده
  - در نظر گرفتن وابستگی بین داده‌ها (متفاوت با فاصله اقلیدسی)
  - تغییرناپذیر با مقیاس (scale invariance)

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

# توابع توزیع . . .

- قضیه حد مرکزی (**Central Limit Theorem**) . . .
  - n متغیر تصادفی $X_1, \ldots, X_n$ که i.i.d هستند (مستقل و با توزیع یکسان) داریم
    - توزیع این متغیرها دارای میانگین μ و واریانس $\sigma^2$ است
  - با افزایش n به سمت بی‌نهایت، داریم
    - متغیر تصادفی Y دارای توزیع گاوسی استاندارد است

$$Y_n = \frac{n(\bar{X}_n - \mu)}{\sqrt{n\sigma^2}} \sim N(0,1)$$

میانگین نمونه‌ای متغیرها

- متغیر تصادفی میانگین نمونه‌ای دارای توزیع گاوسی با میانگین μ و واریانس $\sigma^2/n$ است

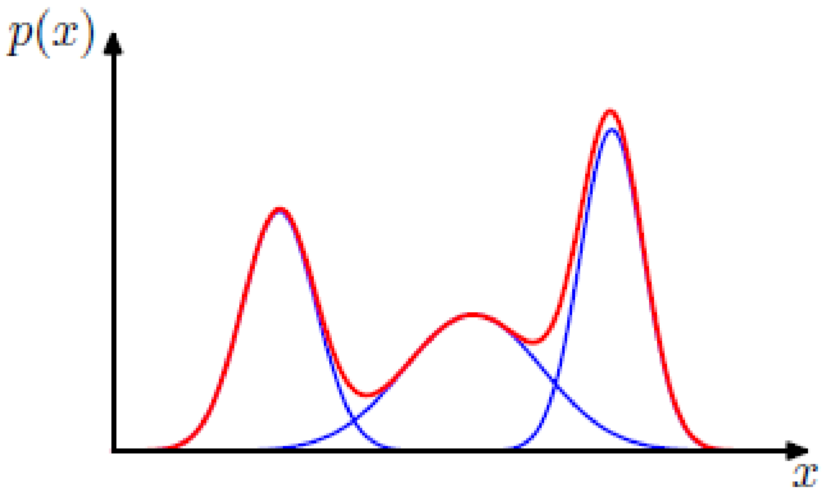
# توابع توزیع . . .

- قضیه حد مرکزی (Central Limit Theorem)
  - توسعه برای حالتی که توزیع‌ها یکسان نیستند (Liapounov 1901)
    - متغیرهای تصادفی $X_1, \ldots, X_n$ مستقل هستند و $E(|X_i - \mu_i|^3) < \infty$
    - آنگاه متغیر زیر دارای توزیع گاوسی است $Y_n = \left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right) / \left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}$
    - مجموع متغیرهای تصادفی $X_1, \ldots, X_n$ دارای توزیع گاوسی با میانگین $\sum_{i=1}^{n} \mu_i$ و واریانس $\left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}$ است

حاصل جمع تعداد زیادی متغیر تصادفی مستقل، صرف‌نظر از توزیع‌های اصلی هر یک از آن‌ها، با بزرگ شدن تعداد متغیرهای تصادفی، دارای توزیع گاوسی است.

# توابع توزیع

- مدل‌های مخلوط (Mixture Model)
  - ترکیب (خطی) چند مدل با همدیگر
  - مدل کردن توزیع‌های پیچیده با بیشینه‌های محلی چندگانه
  - برای توزیع نرمال (گاوسی): مدل مخلوط گاوسی (GMM: Gaussian Mixture Model)
    - از پرکاربردترین روش‌های مدل‌سازی




ضریب مخلوط

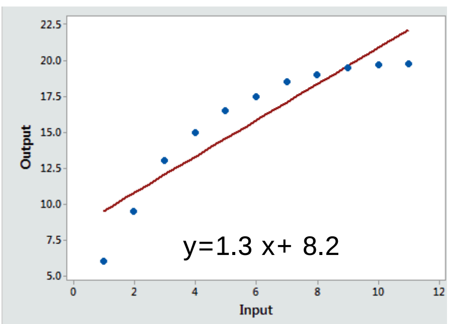
$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k N_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

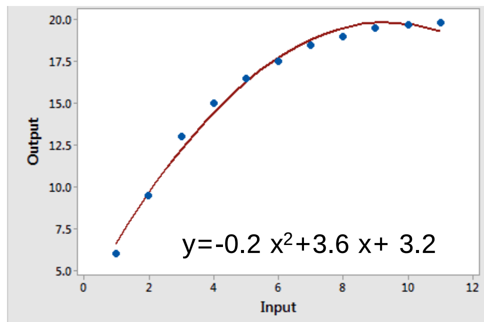
$$c_k \geq 0 \text{ and } \sum_{k=1}^{K} c_k = 1$$

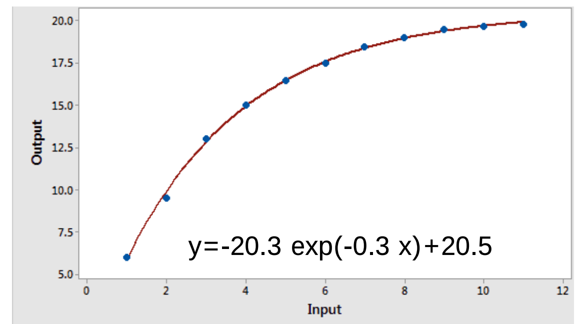
# نظریه تخمین . . .

## مساله

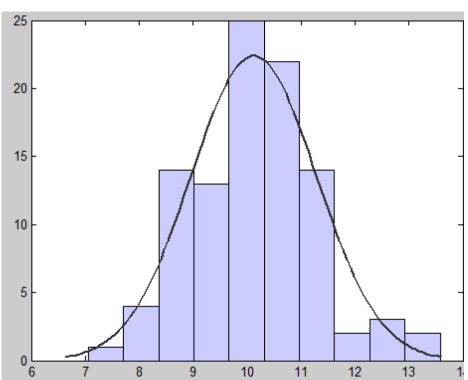
- **تعدادی نمونه داده داریم، می‌خواهیم آنها را با یک تابع (خطی/غیرخطی) مدل کنیم**
  - مثال: وزن افراد بر حسب قد آنها، سیگنال تمیز بر حسب سیگنال نویزی




- **تعدادی نمونه داده داریم، می‌خواهیم تابع توزیع احتمال آنها را بدست آوریم**
  - مثال: توزیع گاوسی به طول کلمات در یک زبان



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

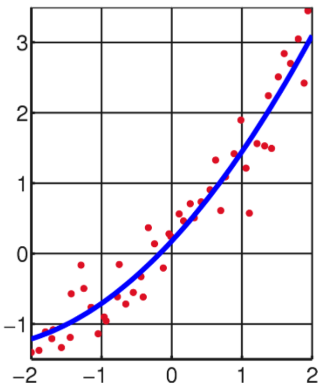
# نظریه تخمین ...

- نظریه تخمین (Estimation theory)
  - در مدل‌سازی آماری یک تابع توزیع (مانند گاوسی) برای داده‌ها فرض می‌کنیم
  - و باید از روی داده‌های آموزش، پارامترهای آن توزیع (مانند میانگین و واریانس) را تخمین بزنیم

  - متغیرهای تصادفی $X_1, \dots, X_n$ را که i.i.d هستند (مستقل و با توزیع یکسان) داریم
  - هدف تخمین پارامترهای $\Phi$
  - تابع تخمین‌گر $\theta(X_1, \dots, X_n)$

- روش‌های تخمین
  - کمینه میانگین مربعات خطا (MMSE: Minimum Mean Square Error)
  - تخمین بیشینه شباهت (MLE: Maximum-Likelihood Estimation)
  - تخمین بیز (Bayesian Estimation)


Hadi Veisi (h.veisi@ut.ac.ir)




# نظریه تخمین . . .

- کمینه میانگین مربعات خطا (MMSE) . . .
  - کمینه کردن امید ریاضی مربعات خطای بین مقدار واقعی و مقدار تخمین زده شده

$$E(Y-\hat{Y})^2 = E(Y-g(X))^2$$

  - فرض کنید هدف ما تخمین مقدار Y با داشتن X باشد، یعنی $\hat{Y} = g(X)$
  - که g(X) تابعی بر حسب پارامترهای Φ است، یعنی g(X,Φ)
    - و با داشتن پارامترهای Φ، تابع g() به صورت کامل مشخص می‌شود
    - پس: هدف تخمین پارامترهای Φ است

$$\hat{\Phi}_{MMSE} = \arg\min_{\Phi}\left[E\left[(Y-g(X,\Phi))^2\right]\right]$$

- تخمین LSE: Least Square Error
  - در عمل به جای تابع توزیع توام X و Y، نمونه‌هایی از $x_i$ و $y_i$ معادل داریم

$$\Phi_{LSE} = \arg\min_{\Phi}\sum_{i=1}^{n}\left[y_i - g(x_i,\Phi)\right]^2$$

  - قانون اعداد بزرگ: وقتی تعداد نمونه‌ها به بی‌نهایت میل می‌کند، LSE و MMSEبرابر می‌شوند

# نظریه تخمین . . .

- کمینه میانگین مربعات خطا (MMSE): برای تابع ثابت . . .
  - تابع ثابت $\hat{Y} = g(x) = c$
    - پارامتر = c
  - هدف کمینه کردن خطاست $E(Y-\hat{Y})^2 = E(Y-c)^2$
    - مشتق گرفتن و برابر صفر قرار دادن $c_{MMSE} = E(Y)$
  - خطای مجذور میانگین کمینه دقیقاً برابر با واریانس Y است

---

- تخمین LSE
  - میانگین نمونه‌ای

$$\min \sum_{i=1}^{n} [y_i - c]^2$$

$$c_{LSE} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

# نظریه تخمین ...

- کمینه میانگین مربعات خطا (MMSE): برای تابع خطی ...
  - تابع خطی $\hat{Y} = g(x) = ax + b$
    - پارامترها: a و b

$$e(a,b) = E(Y-\hat{Y})^2 = E(Y-ax-b)^2$$

$$\frac{\partial e}{\partial a} = 0, \quad \text{and} \quad \frac{\partial e}{\partial b} = 0$$

$$\Rightarrow \quad a = \frac{\operatorname{cov}(X,Y)}{Var(X)} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$$

$$b = E(Y) - \rho_{XY}\frac{\sigma_Y}{\sigma_X}E(X)$$

---

- برای تخمین LSE
  - فرض: بردار x دارای d بعد است و n نمونه داریم

$$\hat{\mathbf{Y}} = \mathbf{XA} \text{ or } \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^d \\ 1 & x_2^1 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^d \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix}$$

$$e(\mathbf{A}) = \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2 = \sum_{i=1}^{n} \left( \mathbf{A}^t \mathbf{x}_i - y_i \right)^2 \Rightarrow \nabla e(\mathbf{A}) = \sum_{i=1}^{n} 2(\mathbf{A}^t \mathbf{x}_i - y_i)\mathbf{x}_i = 2\mathbf{X}^t(\mathbf{XA} - \mathbf{Y})$$

شبه معکوس

$$\mathbf{X}^t\mathbf{XA} = \mathbf{X}^t\mathbf{Y} \Rightarrow \mathbf{A}_{LSE} = \boxed{(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t}\mathbf{Y}$$

# نظریه تخمین . . .

- کمینه میانگین مربعات خطا (MMSE): برای تابع غیرخطی
  - تابع غیرخطی

$$\min_{g(\bullet)\in G_{nl}} E\left[Y-g(X)\right]^2$$

$$E_x\left[E_{Y|X}(Y|X)\right]=E_{X,Y}(Y)$$

$$E_{X,Y}\left[Y-g(X)\right]^2=E_X\left\{E_{Y|X}\left[\left[Y-g(X)\right]^2|X=x\right]\right\}$$

$$=\int_{-\infty}^{\infty}E_{Y|X}\left[\left[Y-g(X)\right]^2|X=x\right]f_X(x)dx$$

$$=\int_{-\infty}^{\infty}E_{Y|X}\left[\left[Y-g(x)\right]^2|X=x\right]f_X(x)dx$$

مقدار مثبت، کمینه کردن انتگرال معادل کمینه کردن این مقدار است

$$E(Y-\hat{Y})^2=E(Y-c)^2$$

$$c_{MMSE}=E(Y)$$

$$\min_{g(x)\in R} E_{Y|X}\left[\left[Y-g(x)\right]^2|X=x\right] \Rightarrow \hat{Y}=g_{MMSE}(X)=E_{Y|X}(Y|X)=\int_{-\infty}^{\infty}yf_Y(y|X=x)dy$$

- در بهسازی گفتار

تخمین طیف گفتار تمیز

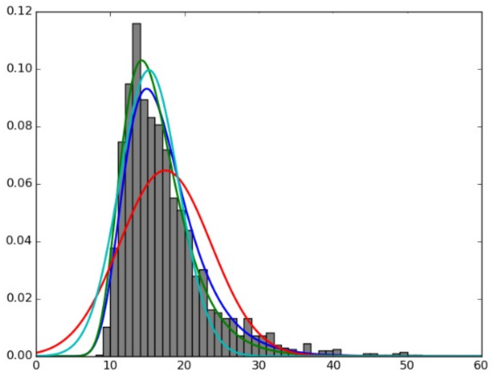
$$\hat{S}(\omega_k)=E[S(\omega_k)|\mathbf{Y}]=\frac{P_{ys}(\omega_k)}{P_{yy}(\omega_k)}=\frac{P_{ss}(\omega_k)}{P_{ss}(\omega_k)+P_{dd}(\omega_k)}Y(\omega_k)$$

سیگنال گفتار نویزی

توان طیف گفتار تمیز

توان طیف نویز

# نظریه تخمین ...




- تخمین بیشینه شباهت (MLE) ...
  - پرکاربردترین روش تخمین پارامتری
  - تخمین توزیع n نمونه داده i.i.d به صورت $X_1, \ldots, X_n$ = $p(x \mid \Phi)$
  - فرض: پارامترهای $\Phi$ دارای مقادیر ثابت، اما نامشخص هستند
    - اگر تابع توزیع گاوسی باشد، $\Phi=\{\mu, \Sigma\}$
  - تخمین پارامترهای توزیع به نحوی که احتمال بدست آوردن نمونه داده‌ها از روی این توزیع بیشینه باشد
  - چون متغیرهای تصادفی مستقل هستند
  - هدف: بیشینه کردن تابع درست‌نمایی

تابع درست‌نمایی

$$p_n(\mathbf{x} \mid \mathbf{\Phi}) = \prod_{k=1}^{n} p(x_k \mid \mathbf{\Phi})$$

$$\mathbf{\Phi}_{MLE} = \underset{\Phi}{\operatorname{argmax}}\, p_n(\mathbf{x} \mid \mathbf{\Phi})$$

- لگاریتم شباهت (Log-Likelihood)
  - عدم تغییر مساله (تابع یکنوای صعودی)
  - ساده کردن محاسبات و فرمول‌ها (تبدیل ضرب به جمع)

$$l(\mathbf{\Phi}) = \log p_n(\mathbf{x} \mid \mathbf{\Phi}) = \sum_{k=1}^{n} \log p(x_k \mid \mathbf{\Phi})$$

# نظریه تخمین . . .

- تخمین بیشینه شباهت (MLE) . . .
  - بیشینه کردن تابع درست‌نمایی (یا لگاریتم آن) با گرادیان
    - مشتق گرفتن برحسب پارامترها و برابر صفر قرار دادن

$$\nabla_{\Phi} = \begin{bmatrix} \frac{\partial}{\partial \Phi_1} \\ \vdots \\ \frac{\partial}{\partial \Phi_k} \end{bmatrix}$$

$$\nabla_{\Phi}\, l(\mathbf{\Phi}) = \sum_{k=1}^{n} \nabla_{\Phi} \log p(x_k \,|\, \mathbf{\Phi}) = 0$$

- مثال ۱: توزیع گاوسی تک متغیره
  - لگاریتم تابع درست‌نمایی

$$p(x\,|\,\mathbf{\Phi}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$\log p_n(\mathbf{x}\,|\,\mathbf{\Phi}) = \sum_{k=1}^{n} \log p(x_k\,|\,\mathbf{\Phi}) = \sum_{k=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]\right) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2$$

$$\frac{\partial}{\partial \mu} \log p_n(x\,|\,\mathbf{\Phi}) = \sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k-\mu) = 0$$

$$\frac{\partial}{\partial \sigma^2} \log p_n(x\,|\,\mathbf{\Phi}) = -\frac{n}{2\sigma^2} + \sum_{k=1}^{n} \frac{(x_k-\mu)^2}{2\sigma^4} = 0$$

$$\Rightarrow$$

$$\mu_{MLE} = \frac{1}{n}\sum_{k=1}^{n} x_k = E(x)$$

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu_{MLE})^2 = E\left[(x-\mu_{MLE})^2\right]$$

همان میانگین و واریانس نمونه‌ای

# نظریه تخمین . . .

## تخمین بیشینه شباهت (MLE)

- مثال ۲: توزیع گاوسی چندمتغیره

$$p(\mathbf{x}|\boldsymbol{\Phi}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\hat{\boldsymbol{\mu}}_{MLE} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\hat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})^t = E\left[(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{MLE})^t\right]$$

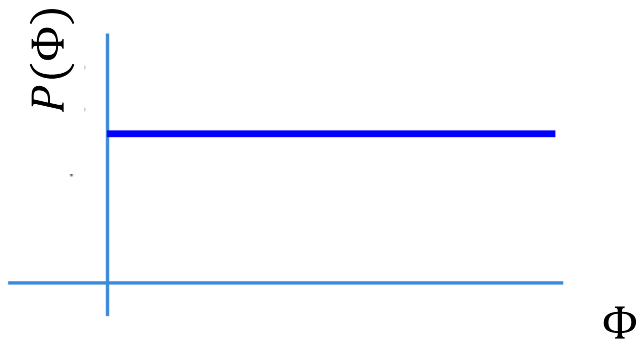
- تخمین ML برای واریانس بایاس شده است
  - امید ریاضی واریانس تخمینی با واریانس واقعی برابر نیست
  - با میل کردن n به سمت بی نهایت اثر بایاس کم می‌شود

$$\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

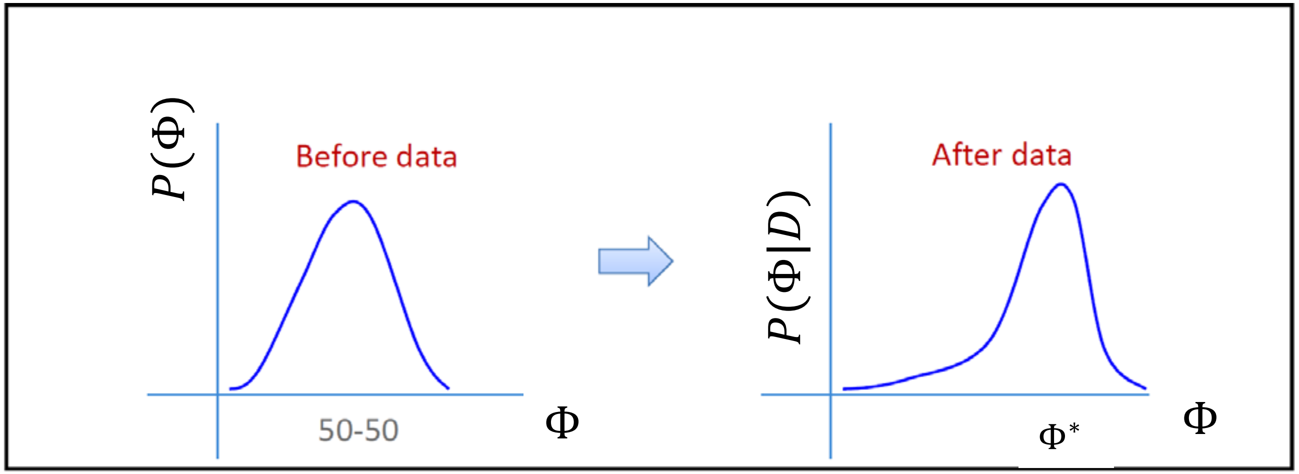
- تخمین غیربایاس

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\hat{\mu})^2$$

# نظریه تخمین . . .




- در تخمین ML
  - فرض = ثابت بودن پارامترهای توزیع
    - توزیع $p(\Phi)$ یکنواخت است
- تخمین بیز (Bayesian)
  - پارامتر $\Phi$ یک متغیر تصادفی است



# نظریه تخمین . . .

## تخمین بیز (Bayesian Estimation) . . .

- پارامتر $\Phi$ یک متغیر تصادفی و نامشخص است
  - در تخمین ML این پارامتر تصادفی نیست و ثابت است
  - تصادفی بودن پارامتر $\Phi$ به معنی وجود احتمال پیشین توزیع $p(\Phi)$ برای آن است
- شکل تابع توزیع $p(x\,|\,\Phi)$ مشخص است (مثلاً توزیع نرمال)
- مجموعه n نمونه داده i.i.d $D=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ دارای توزیع $p(x\,|\,\Phi)$
  - داده‌ها حاوی اطلاعاتی از پارامتر $\Phi$
- با توجه به قانون بیز

احتمال پسین: احتمال پارامترپس از مشاهده داده‌ها

$$p(\Phi\,|\,\mathbf{x}) = \frac{p(\mathbf{x}\,|\,\Phi)\,p(\Phi)}{p(\mathbf{x})} \propto p(\mathbf{x}\,|\,\Phi)\,p(\Phi)$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

- هدف

$$p(x\,|\,D) = \int p(x\,|\,\Phi)\,p(\Phi\,|\,D)\,d\Phi$$

# نظریه تخمین . . .

## تخمین بیز (Bayesian Estimation) . . .

$$p(x \mid D) = \int p(x \mid \Phi) p(\Phi \mid D) d\Phi$$

- نخستین گام در تخمین بیز: محاسبه $p(\Phi \mid D)$

$$p(\Phi \mid D) = \frac{1}{\alpha} p(D \mid \Phi) p(\Phi)$$

  - که $\alpha$ ثابت نرمال کننده است (مستقل از $\Phi$)
    - حذف نمی‌شود
    - در عمل می‌توان مقدار ثابتی را به عنوان تخمین آن قرار داد

- با فرض مستقل بودن نمونه داده‌ها

$$p(D \mid \Phi) = \prod_{k=1}^{n} p(x_k \mid \Phi)$$

# نظریه تخمین . . .

- تخمین بیز (**Bayesian Estimation**) . . .
  - مثال: داده‌ها با توزیع گاوسی، تک متغیره، واریانس معلوم $\sigma^2$ و میانگین نامعلوم $\Phi$

$$p(\mathbf{x}|\Phi)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\Phi}{\sigma}\right)^2\right]\propto\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\Phi}{\sigma}\right)^2\right]$$

$$\sum_{i=1}^{n}(x_i-\Phi)^2=n(\Phi-\bar{x}_n)^2+\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

$$\bar{x}_n=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$p(\mathbf{x}|\Phi)\propto\exp\left[-\frac{n}{2\sigma^2}(\Phi-\bar{x}_n)^2\right]\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right]$$

  - فرض: $p(\Phi)$ نیز توزیع گاوسی با میانگین $\mu$ و واریانس $v^2$ دارد
    - فرض گاوسی بودن توزیع بودن اولیه ضروری نیست و می‌تواند هر توزیع دیگری برای آن فرض شود

$$p(\Phi)=\frac{1}{(2\pi)^{1/2}v}\exp\left[-\frac{1}{2}\left(\frac{\Phi-\mu}{v}\right)^2\right]\propto\exp\left[-\frac{1}{2}\left(\frac{\Phi-\mu}{v}\right)^2\right]$$

  - بنابراین

$$p(\Phi|\mathbf{x})\propto\exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\Phi-\bar{x}_n)^2+\frac{1}{v^2}(\Phi-\mu)^2\right]\right\}$$

# نظریه تخمین ...

- تخمین بیز (Bayesian Estimation)
  - مثال: داده‌ها با توزیع گاوسی، تک متغیره، واریانس معلوم $\sigma^2$ و میانگین نامعلوم $\Phi$

$$p(\Phi \mid \mathbf{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\Phi-\bar{x}_n)^2+\frac{1}{\nu^2}(\Phi-\mu)^2\right]\right\}$$

$$\rho=\frac{\sigma^2\mu+n\nu^2\bar{x}_n}{\sigma^2+n\nu^2} \qquad \tau^2=\frac{\sigma^2\nu^2}{\sigma^2+n\nu^2}$$

$$p(\Phi \mid \mathbf{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\tau^2}(\Phi-\rho)^2+\frac{n}{\sigma^2+n\nu^2}(\bar{x}_n-\mu)^2\right]\right\}$$

مقدار ثابت

$$p(\Phi \mid \mathbf{x})=\frac{1}{\sqrt{2\pi}\tau}\exp\left[\frac{-1}{2\tau^2}(\Phi-\rho)^2\right]$$

- توزیع گاوسی با میانگین $\rho$ و واریانس $\tau^2$
  - ترکیبی از اطلاعات اولیه $(\mu , \nu^2)$ و اطلاعات نمونه‌ها (جمع وزن‌دار میانگین)
  - با افزایش تعداد نمونه‌ها $(n\rightarrow\infty)$، داریم:
    - واریانس به صفر میل می‌کند، عدم قطعیت تخمین میانگین کم می‌شود (واریانس=میزان عدم قطعیت میانگین تخمینی است)
    - میانگین، به میانگین نمونه‌ای نمونه‌ها میل می‌کند
- با داشتن $p(\Phi \mid D)$، محاسبه $p(x \mid D)$ صورت می‌گیرد

# نظریه تخمین . . .

- تخمین بیشینه احتمال پسین (MAP: maximum a posteriori) . . .
  - هدف: بیشینه کردن $p(\Phi | x)$
    - در ML هدف بیشینه کردن $p(x | \Phi)$
    - پارامترها متغیرهای تصادفی با توزیع پیشین $p(\Phi)$ هستند
    - متداول‌ترین تخمین‌گر بیزی

$$p(\Phi|\mathbf{x}) = \frac{p(\mathbf{x}|\Phi)p(\Phi)}{p(\mathbf{x})} \propto p(\mathbf{x}|\Phi)p(\Phi)$$

$$\Phi_{MAP} = \theta_{MAP}(\mathbf{x}) = \underset{\Phi}{\operatorname{argmax}}\ p(\Phi|\mathbf{x}) = \underset{\Phi}{\operatorname{argmax}}\ p(\mathbf{x}|\Phi)p(\Phi)$$

$$\Phi_{MAP} = \underset{\Phi}{\operatorname{argmax}}\ \log p(\mathbf{x}|\Phi) + \log p(\Phi)$$

$$\frac{\partial \log p(\mathbf{x}|\Phi)}{\partial \Phi} + \frac{\partial \log p(\Phi)}{\partial \Phi} = 0$$

$$\left.\frac{\partial \log p(\mathbf{x}|\Phi)}{\partial \Phi}\right|_{\Phi=\Phi_{MAP}} = \left.\frac{-\partial \log p(\Phi)}{\partial \Phi}\right|_{\Phi=\Phi_{MAP}}$$

- تخمین ML و MAP یکسان هستند وقتی توزیع پیشین $p(\Phi)$ یکنواخت باشد

# نظریه تخمین

- تخمین بیشینه احتمال پسین (MAP: maximum a posteriori)
  - مثال (قبلی): داده‌ها با توزیع گاوسی، تک متغیره، واریانس معلوم $\sigma^2$ و میانگین نامعلوم $\Phi$
    - توزیع پیشین $\Phi$: گاوسی با با میانگین $\mu$ و واریانس $v^2$

$$p(\Phi \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi}\tau} \exp\left[\frac{-1}{2\tau^2}(\Phi - \rho)^2\right]$$

    - مشتق‌گیری نسبت به $\Phi$
    - متوسط وزن‌دار میانگین نمونه‌ها و میانگین قبلی

$$\Phi_{MAP} = \rho = \frac{\sigma^2\mu + nv^2\overline{x}_n}{\sigma^2 + nv^2}$$

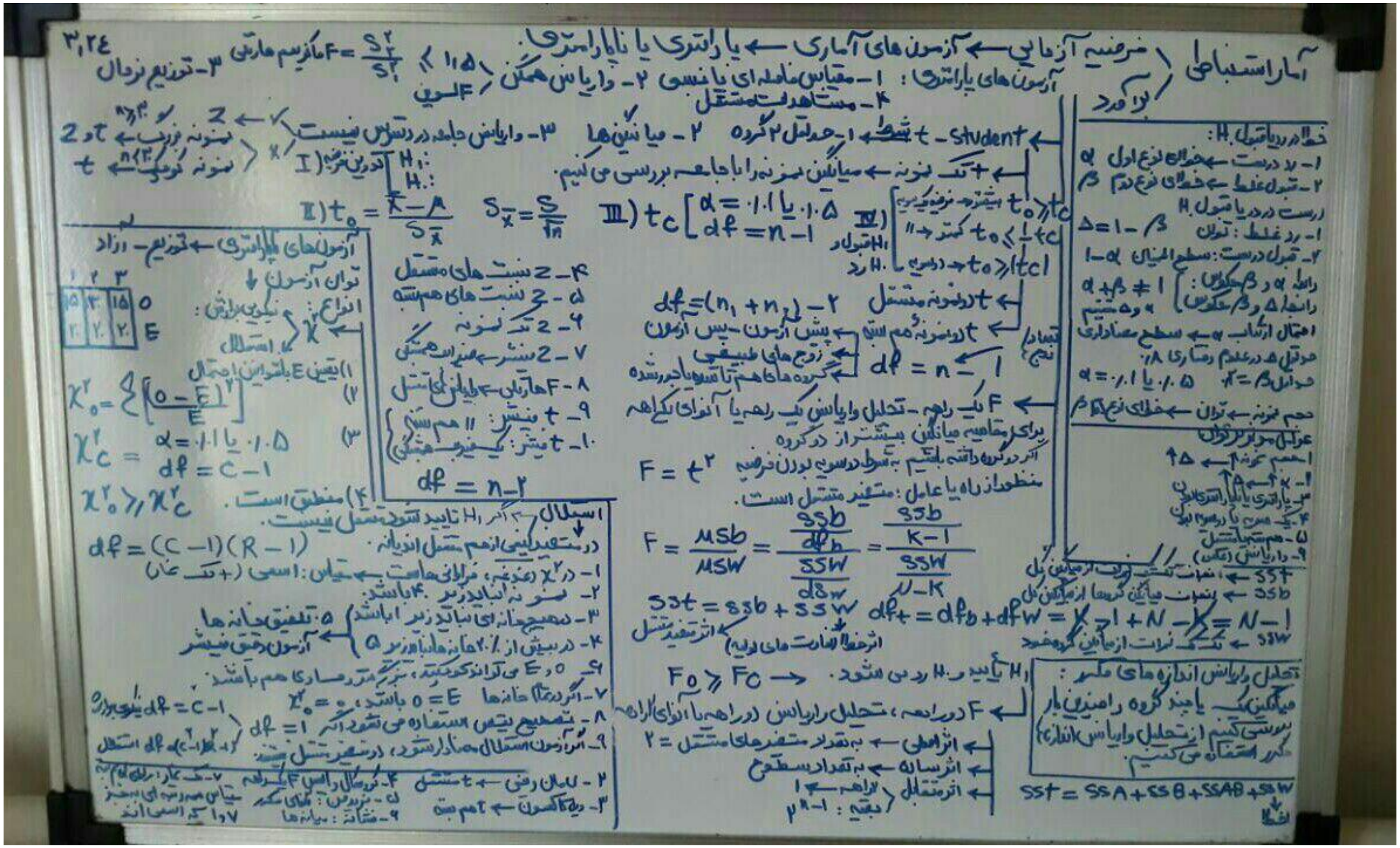
میانگین نمونه‌ها

تعداد نمونه‌ها

  - کاربرد در تطبیق (adaptation)
    - آموزش صدای یک کاربر جدید به سیستم بازشناسی گفتار
    - آموزش پارامترهای مدل با پایگاه داده‌های مستقل از گوینده (با چندین گوینده) = توزیع پیشین
    - تطبیق با محاسبه میانگین نمونه‌های یک گوینده خاص و استفاده از رابطه بالا

# مقایسه تخمین‌گرها

| تخمین‌گر بیز | تخمین‌گر بیشینه شباهت (ML) | تخمین‌گر بیشینه احتمال پسین (MAP) |
|---|---|---|
| **تخمین‌گر توزیع**<br>$p(\mathbf{x}\|\mathcal{D}) = \int p(\mathbf{x}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\mathcal{D})d\boldsymbol{\theta}$ | **تخمین‌گر نقطه**<br>$p(\mathbf{x}\|\mathcal{D}) = p(\mathbf{x}\|\hat{\boldsymbol{\theta}})$ | **تخمین‌گر نقطه**<br>$p(\mathbf{x}\|\mathcal{D}) = p(\mathbf{x}\|\hat{\boldsymbol{\theta}})$ |
| **تخمین توزیع**<br>$p(\boldsymbol{\theta}\|\mathcal{D}) = \frac{1}{\alpha}p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta})$ | **تخمین نقطه**<br>$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathcal{D}\|\boldsymbol{\theta})$ | **تخمین نقطه**<br>$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta})$ |
| 1. استفاده بیشتر از اطلاعات<br>2. کارایی بهتر در صورت عدم سازگاری بین توزیع فرض شده و توزیع واقعی<br>3. در نظر گرفتن بایاس واریانس | 1. تفسیر ساده‌تر (نقطه‌ای)<br>2. محاسبات کم‌تر | 1. استفاده از اطلاعات پیشین پارامتر توزیع |


آمار استنباطی
t - student
F = t²
df = n - 1
Fo ≥ Fc