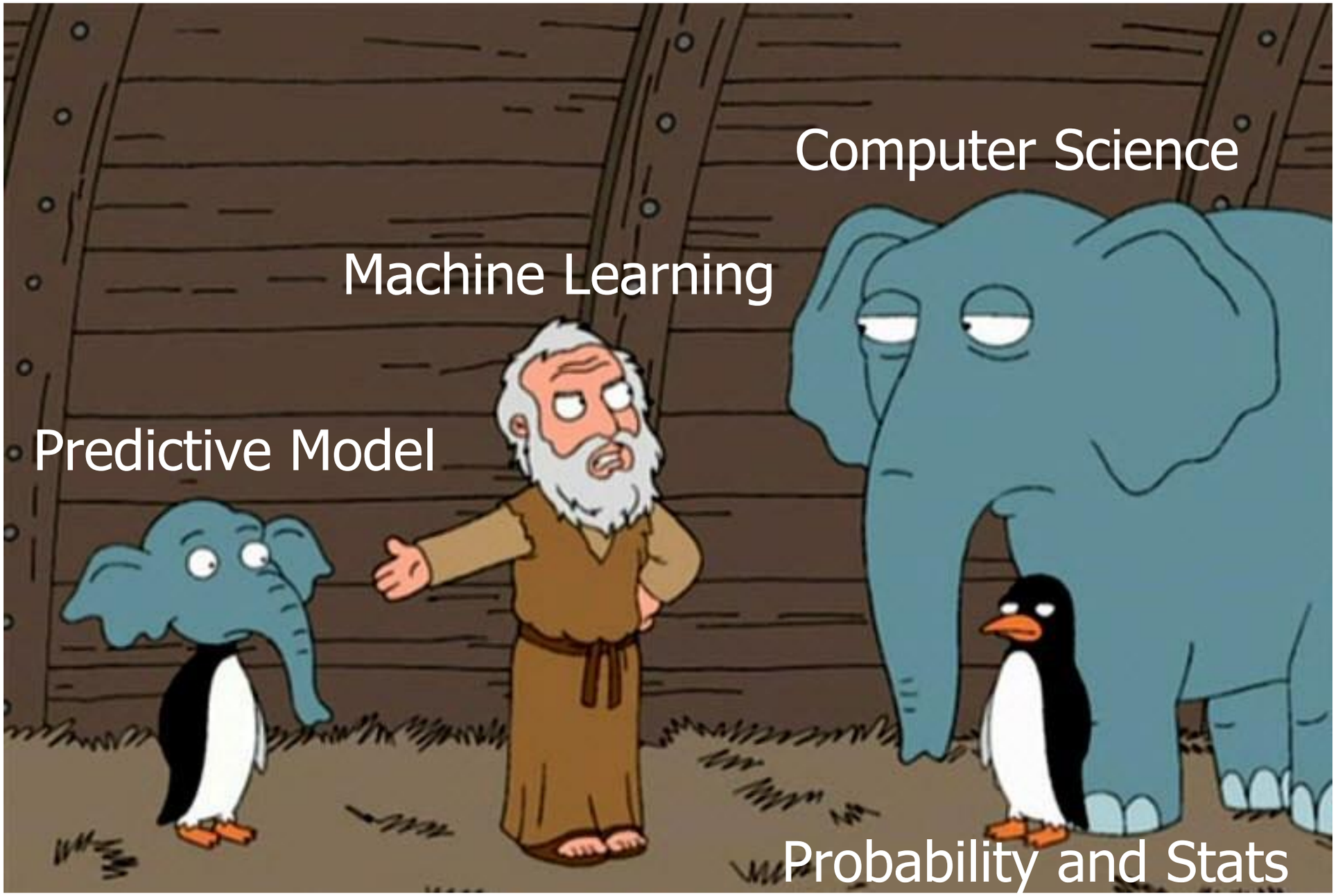

Computer Science
Machine Learning
Predictive Model
Probability and Stats

Georgia Tech

Probability and Statistics

Mahdi Roozbahani
Georgia Tech

These slides are inspired based on slides from Le Song , Sam Roweis, and Chao Zhang.

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Probability

- A sample space S is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (S can be finite or infinite.)
  - E.g., S may be the set of all possible outcomes of a dice roll: S
    $(1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$ $P(X=x) = P(x=1) = \frac{1}{6}$
  - E.g., S may be the set of all possible nucleotides of a DNA site: S
    $(A \quad C \quad G \quad T)$

  - E.g., S may be the set of all possible time-space positions of an aircraft on a radar screen.
- An Event A is any subset of S
  - Seeing "1" or "6" in a dice roll; observing a "G" at a site; UA007 in space-time interval

Three Key Ingredients in Probability Theory

A **sample space** is a collection of all possible **outcomes**

Random variables X represents **outcomes** in sample space

Probability of a random variable to happen $p(x) = p(X = x)$

$p(x) \geq 0$ $0 \leq p(x) \leq 1$

Pdf → probability distribution function

Pdf Pmf

Continuous variable

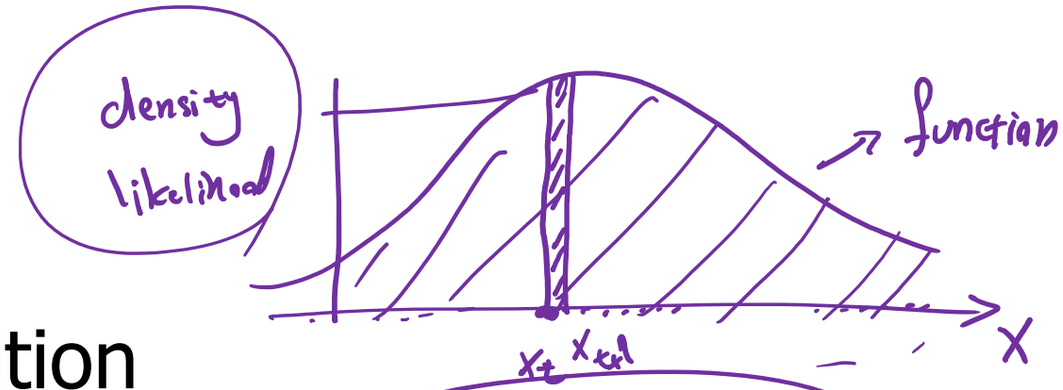
Continuous probability distribution

Probability density function pdf

Density or likelihood value

Temperature (real number)

Gaussian Distribution



$$\int_x p(x)dx = 1$$

$P(x_t) = 0$

Discrete variable

Discrete probability distribution

Probability mass function pmf

Probability value

Coin flip (integer)

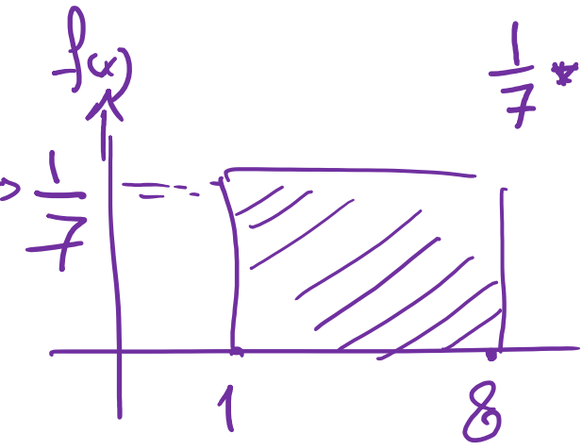
Bernoulli distribution

$$\sum_{x \epsilon A} p(x) = 1$$

Continuous Probability Functions

- Examples:
  - Uniform Density Function:

$$f_x(x) = \begin{cases} \dfrac{1}{b-a} & for\ a \le x \le b \\ 0 & otherwise \end{cases}$$



$\theta \rightarrow$ Parameters

$\{b, a\} \in \theta$

$\frac{1}{7} * 7 = 1$

  - Exponential Density Function:

$$f_x(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \qquad for\ x \ge 0$$

$$F_x(x) = 1 - e^{\frac{-x}{\mu}} \qquad for\ x \ge 0$$

$\{\mu\} \in \theta$

μ as avg { ① Arithmetic avg = ② weighted avg = expected value

  - Gaussian(Normal) Density Function

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\{\mu, \sigma\} \in \theta$

a

b

Discrete Probability Functions

- Examples:
  - Bernoulli Distribution:
    - $\begin{cases} 1-p & for\ x=0 \\ p & for\ x=1 \end{cases}$

In Bernoulli, just a **single** trial is conducted

  - Binomial Distribution:
    - $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$

k is number of successes

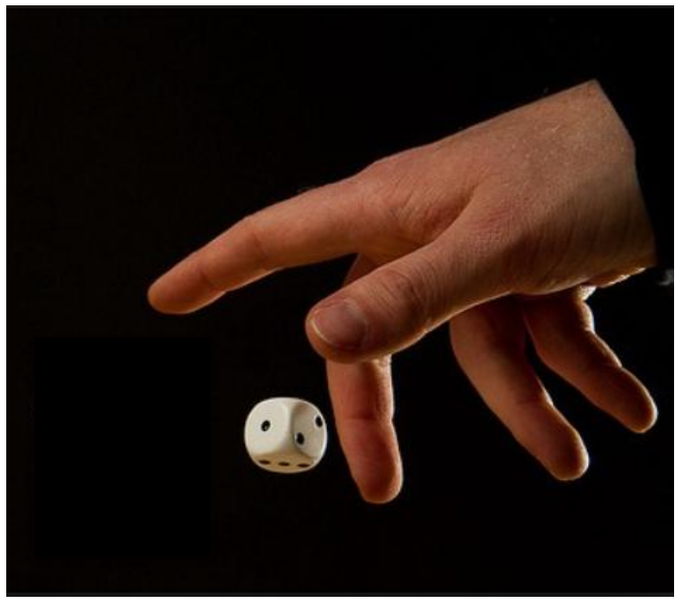
n-k is number of failures

$\binom{n}{k}$ The total number of ways of selection **k** distinct combinations of **n** trials, **irrespective of order**.

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Example



X = Throw a dice

Y = Flip a coin

X and **Y** are random variables

N = total number of trials

n_{ij} = Number of occurrence

X

Y	$x_{i=1}=1$	$x_{i=2}=2$	$x_{i=3}=3$	$x_{i=4}=4$	$x_{i=5}=5$	$x_{i=6}=6$	C_j
$y_{j=2}=tail$	$n_{ij}=3$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=5$	$n_{ij}=1$	$n_{ij}=5$	20
$y_{j=1}=head$	$n_{ij}=2$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=1$	15
C_i	5	6	6	7	5	6	N=35

X

Y	$x_{i=1}=1$	$x_{i=2}=2$	$x_{i=3}=3$	$x_{i=4}=4$	$x_{i=5}=5$	$x_{i=6}=6$	C_j
$y_{j=2}=tail$	$n_{ij}=3$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=5$	$n_{ij}=1$	$n_{ij}=5$	20
$y_{j=1}=head$	$n_{ij}=2$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=4$	$n_{ij}=1$	15
C_i	5	6	6	7	5	6	N=35

① $P(Y=t, X=5) = \frac{1}{35} = \frac{n_{ij}}{N}$

② $P(Y=t) = \frac{20}{35} = \frac{C_j}{N}$ $\quad P(X=4) = \frac{7}{35} = \frac{C_i}{N}$

③ $P(Y=t \mid X=4) = \frac{5}{7} = \frac{n_{ij}}{C_i}$ $\quad P(X=5 \mid Y=t) = \frac{1}{20} = \frac{n_{ij}}{C_j}$

$P(Y=y, X=x) = \frac{n_{ij}}{N} = \left(\frac{n_{ij}}{C_j}\right)\frac{C_j}{N} = P(X=x \mid Y=y)\,P(Y=y)$

$= \frac{n_{ij}}{C_i}\frac{C_i}{N} = P(Y=y \mid X=x)\,P(X=x)$

} product rule

$P(A,B) = P(A \mid B)\,P(B) = P(B \mid A)\,P(A)$

Probability:

$$p(X = x_i) = \frac{c_i}{N}$$

Joint probability:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

$P(A,B,C) = P(A|B,C)\,P(B,C) = P(A,B|C)\,P(C)$

Conditional probability:

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Sum rule

$$p(X = x_i) = \sum_{j=1}^{L} p\left(X = x_i, Y = y_j\right) \Rightarrow \quad p(X) = \sum_{Y} P(X, Y)$$

Product rule

$$p\left(X = x_i, Y = y_j\right) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} = p\left(Y = y_j \middle| X = x_i\right)p(X = x_i)$$

$$p(X, Y) = p(Y|X)p(X)$$

Conditional Independence

$P(H,F,V,D) = P(H|F,V,D)P(F,V,D) =$
$P(H|F,D)P(F,V,D)$
$P(H|F,D)P(F|V,D)P(V,D)$
$P(H|F,D)P(F|V)P(V|D)P(D)$
$P(H|F,D)P(F|V)P(V)P(D)$

- Examples:

$\boldsymbol{P(Virus|\ DrinkBeer) = P(Virus)}$

iff Virus is independent of Drink Beer

$\boldsymbol{P(Flu\mid Virus, DrinkBeer) = P(Flu\mid Virus)}$

iff Flu is independent of Drink Beer, given Virus

Statistical independence = independence

$y = x^2$ $y = 2x$

$\boldsymbol{P(Headache\mid Flu, Virus, DrinkBeer)}$
$\boldsymbol{= P(Headache|Flu, DrinkBeer)}$

iff Headache is independent of Virus, given Flu and Drink Beer

Assume the above independence, we obtain:

$\boldsymbol{P(Headache, Flue, Virus, DrinkBeer)}$
$\boldsymbol{= P(Headache|Flu, Virus, DrinkBeer)\ P(Flu|Virus, DrinkBeer)}$
$\mathbf{P(Virus|DrinkBeer)P(DrinkBeer)}$
$\mathbf{= P(Headache|Flu, DrinkBeer)P(Flu|Virus)P(Virus)P(DrinkBeer)}$

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Bayes' Rule

- P(X|Y)= Fraction of the worlds in which X is true given that Y is also true.

$$P(Y,X) = P(Y|X)P(X) \Rightarrow P(Y|X) = \frac{P(Y,X)}{P(X)}$$

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

- For example:
  - H="Having a headache"
  - F="Coming down with flu"
  - $P(Headche|Flu)$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?

- Definition:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{\mathrm{P(Y|X)P(X)}}{\mathrm{P(Y)}}$$

Corollary:

$$P(X,Y) = P(Y|X)P(X)$$

This is called Bayes Rule

Bayes' Rule

$$P(a|b,c) = \frac{P(a,b,c)}{P(b,c)} = \frac{P(a,b|c)\,P(c)}{P(b|c)\,P(c)} = \frac{P(a,b|c)}{P(b|c)}$$

$$P(a|b,c)\,P(b|c)$$

- $P(Headache|Flu) = \frac{P(Headache,Flu)}{P(Flu)}$

$$= \frac{P(Flu|Headache)P(Headache)}{P(Flu)}$$

Other cases:

$$P(a,b|c)$$

- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y)+P(X|\neg Y)P(\neg Y)}$

$$P(a,b) = P(a|b,c)P(b|c)$$

- $P(Y = y_i|X) = \frac{P(X|Y)P(Y)}{\sum_{i\in S} P(X|Y = y_i)P(Y=y_i)}$
- $P(Y|X,Z) = \frac{P(X|Y,Z)P(Y,Z)}{P(X,Z)} =$

$$\frac{P(X|Y,Z)P(Y,Z)}{P(X|Y,Z)P(Y,Z)+P(X|\neg Y,Z)P(\neg Y,Z)}$$

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance 
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

$E[\cdot]$ ⇝ expected value

$E(\cdot)$ ⇝ function

Mean and Variance

- Expectation: The mean value, center of mass, first moment:

$$E_X[g(X)] = \int_{-\infty}^{\infty} g(x)p_X(x)dx = \mu$$

$g(x) = x$

$$E[g(x)] = \sum_{i=1}^{N} p(x^{[i]})\, g(x^{[i]})$$

- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n$
- Mean: $E_X[X] = \int_{-\infty}^{\infty} xp_X(x)dx$
  - $E[\alpha X] = \alpha E[X]$
  - $E[\alpha + X] = \alpha + E[X]$

$E[\alpha] = \alpha$

$E[E[(1,2,3)]] = E[\quad]$

$E[(1,2,3)] = 2$

$$Var(x) = \frac{1}{N}\sum_{i=1}^{N}(x^{[i]} - M)^2$$

- Variance(Second central moment): $Var(x) = E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2$
  - $Var(\alpha X) = \alpha^2 Var(X)$
  - $Var(\alpha + X) = Var(X)$

$E[(X - E[X])^2]$ ✓

Mean and average

$X = [1, 2, 3]$ $\quad g(x) = x$

$P(x) = [\frac{1}{6}, \frac{3}{6}, \frac{2}{6}] \rightsquigarrow$ pmf

$$E[g(x)] = \sum_{i=1}^{N} p(x)\, g(x)$$

$$E[g(x)] = \frac{1}{6} * 1 + \frac{3}{6} * 2 + \frac{2}{6} * 3 = \frac{13}{6}$$

$$Avg(x) = \frac{1}{N} \sum_{i=1}^{N} g(x) = \frac{1+2+3}{3} = 2 \qquad Avg(x) \neq \frac{13}{6}$$

$$X = [1, 2, 2, 2, 3, 3] = \frac{1+2+2+2+3+3}{6} = \frac{13}{6}$$

Variance and average:

h = height = x_1

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}_{n \times d=1} \qquad x_1^{[2]}$$

$$\mu_h = \frac{1+2+3}{3} = 2$$

$$\sigma_h^2 = \frac{1}{N} \sum_{i=1}^{N} (x_1^{[i]} - \mu_h)^2$$

$$\frac{1}{N} \sum (\cdot) = \text{avg}(\cdot) = E[\cdot]$$

$$E[(x - \mu_h)^2]$$

$$E[(x_1 - E[x_1])^2]$$

$X \rightarrow \bar{X}$ (centering)

$$\bar{X} = \begin{bmatrix} \bar{h} = h - \mu_h \\ 1 - \mu_h \\ 2 - \mu_h \\ 3 - \mu_h \end{bmatrix}_{n \times d} = \begin{bmatrix} \bar{h} \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

$$\mu_{\bar{h}} = 0$$

$$\text{Var}(x) = \frac{1}{N} \underset{d \times n}{\bar{X}^T} \underset{n \times d}{\bar{X}} = \frac{1}{N} \begin{bmatrix} 1-\mu_h & 2-\mu_h & 3-\mu_h \end{bmatrix} \begin{bmatrix} 1-\mu_h \\ 2-\mu_h \\ 3-\mu_h \end{bmatrix} = \frac{1}{N} \left((1-\mu_h)^2 + (2-\mu_h)^2 + (3-\mu_h)^2 \right)$$

Covariance:

$$X = \begin{matrix} h & w \\ \end{matrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

$$\mu_h = \frac{1+2+3}{3} = 2 \qquad \mu_w = \frac{4+5+6}{3} = 5$$

$$\sigma_h^2 = \frac{1}{N} \sum_{i=1}^{N} (x_1^{(i)} - \mu_h)^2$$

$$\sigma_w^2 = \frac{1}{N} \sum_{i=1}^{N} (x_2^{(i)} - \mu_w)^2$$

$$X \rightsquigarrow \bar{X} = \begin{matrix} \bar{h} & \bar{w} \\ \end{matrix} \begin{bmatrix} 1-\mu_h & 4-\mu_w \\ 2-\mu_h & 5-\mu_w \\ 3-\mu_h & 6-\mu_w \end{bmatrix} \Rightarrow \quad COV = \left(\frac{1}{N} \bar{X}^T \bar{X} \right)_{d \times d}$$

Symmetric

$$\mu_{\bar{h}} = 0 \qquad \mu_{\bar{w}} = 0$$

$$COV = \frac{1}{N} \begin{bmatrix} 1-\mu_h & 2-\mu_h & 3-\mu_h \\ 4-\mu_w & 5-\mu_w & 6-\mu_w \end{bmatrix} \begin{bmatrix} 1-\mu_h & 4-\mu_w \\ 2-\mu_h & 5-\mu_w \\ 3-\mu_h & 6-\mu_w \end{bmatrix} = \begin{matrix} \\ h \\ w \end{matrix} \begin{matrix} \begin{matrix} h & \quad w \end{matrix} \\ \begin{bmatrix} \sigma_h^2 & \sigma_{hw} \\ \sigma_{wh} & \sigma_w^2 \end{bmatrix} \end{matrix}$$

$\sigma_{hw} = \sigma_{wh}$

Correlation:

Standardization → Standard deviation → σ

$$X \rightarrow \bar{X} \rightarrow \bar{X}^{*} = \begin{bmatrix} \frac{h - \mu_h}{\sigma_h} & \frac{w - \mu_w}{\sigma_w} \\ \frac{1 - \mu_h}{\sigma_h} & \frac{4 - \mu_w}{\sigma_w} \\ \vdots & \vdots \end{bmatrix}$$

$$COR = \frac{1}{N} \bar{X}^{*T} \bar{X}^{*} = \begin{matrix} h \\ w \end{matrix} \begin{bmatrix} h & w \\ \frac{\sigma_h^2}{\sigma_h^2} = 1 & -1 \leq \sigma_{hw} \leq 1 \\ -1 \leq \sigma_{hw} \leq 1 & 1 = \frac{\sigma_w^2}{\sigma_w^2} \end{bmatrix}$$

EDA

$y = x^2$

$\sigma_{hw} = 1 \quad h = 2w \qquad \sigma_{hw} = 0 \quad h \perp w$

$\sigma_{hw} = 0.99 \quad h = 2w + \alpha \underset{=}{E}$ ← noise

For Joint Distributions

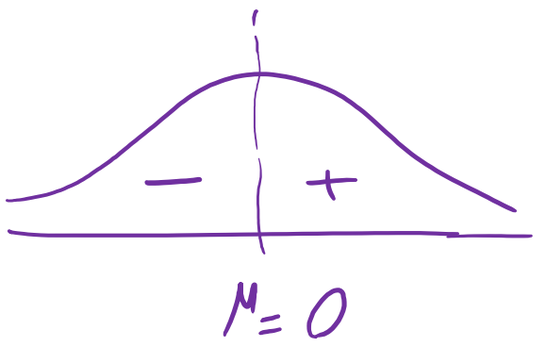
$Cov(X,Y) = E[Z^3] - E[Z]E[Z^2] = 0$

- Expectation and Covariance:
  - $E[X+Y] = E[X] + E[Y]$
  - $cov(X,Y) = E[(X - E_X[X])(Y - E_Y(Y)] = E[XY] - E[X]E[Y]$
  - $Var(X+Y) = Var(X) + 2cov(X,Y) + Var(Y)$

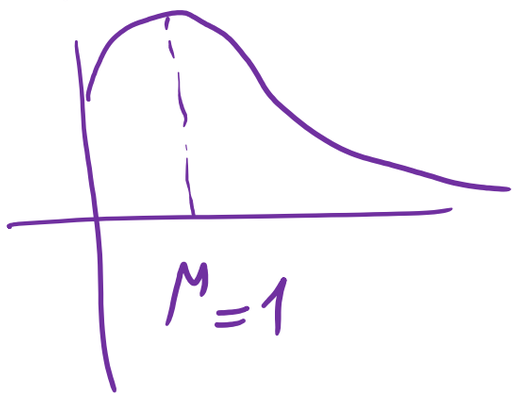
① Z standard gaussian distribution

$\mu = 0 \quad \sigma = 1$

$X = Z$



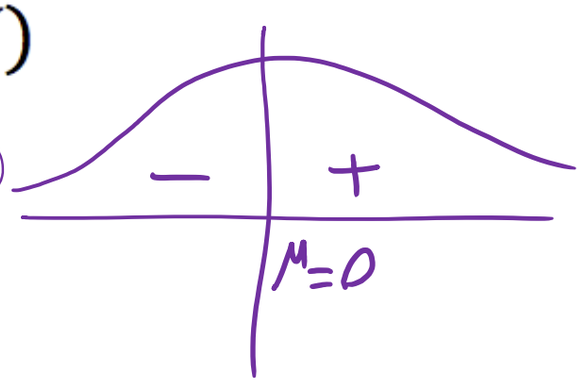
② $Y = Z^2 = X^2$ Chi-squared



$Var(X) = E[X^2] - E[X]^2$

$1 = E[Y] - 0 \Rightarrow E[Y] = 1$

$g = Z^3$



[Uncorrelated vs Independent RV]

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

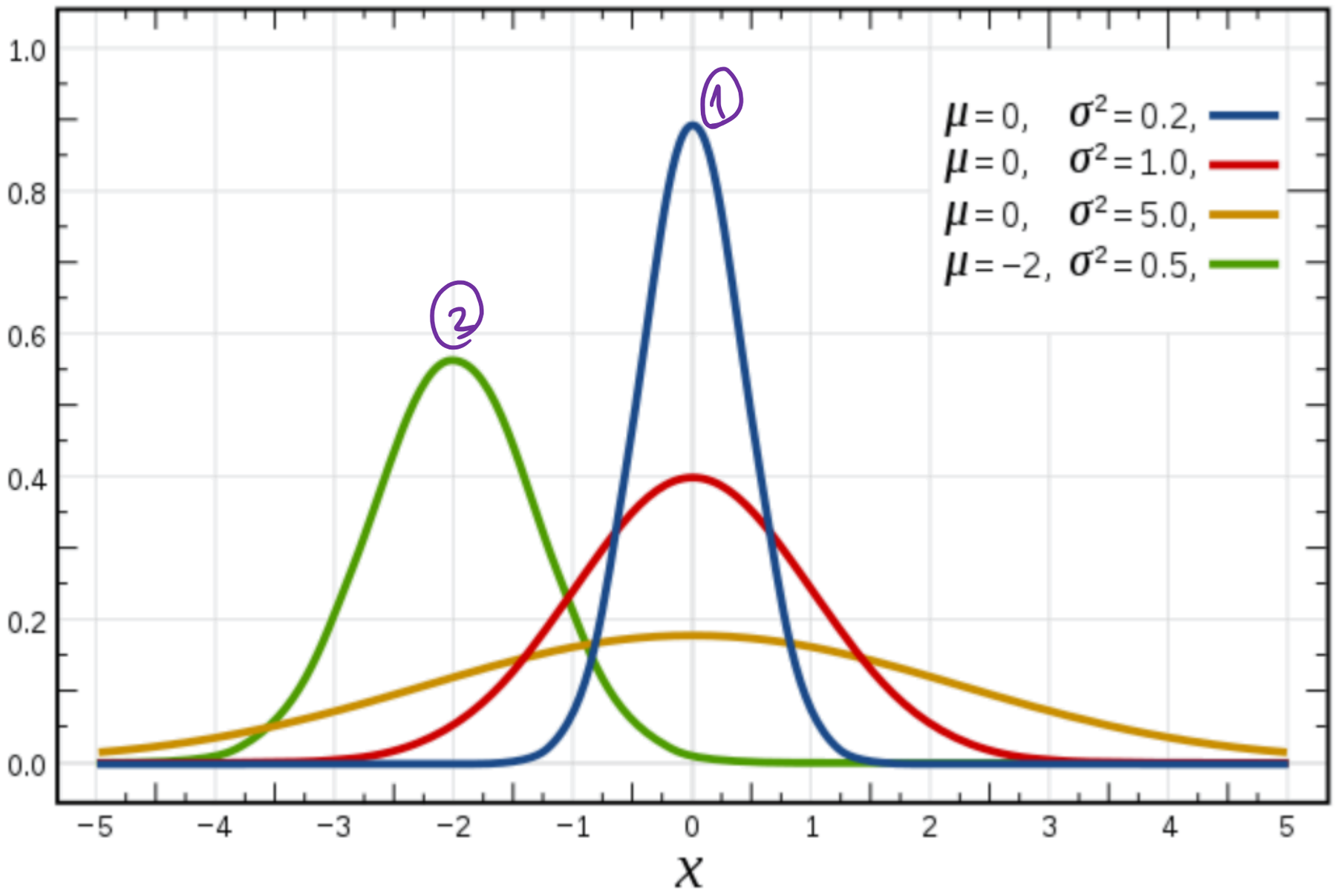
Gaussian Distribution

Univariate

- Gaussian Distribution: $f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

a b

Probability density function



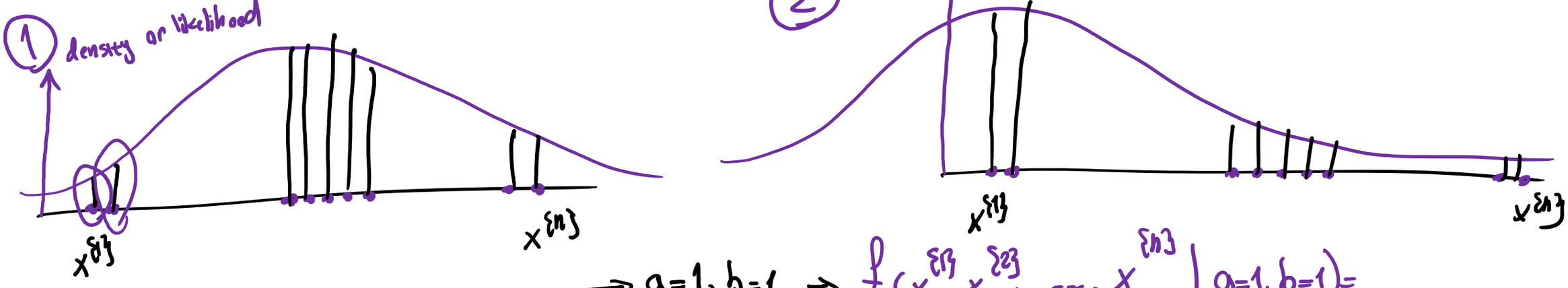
[Probability versus likelihood]

Prob vs Likelihood

$$P(X=h) = \frac{1}{2} \Rightarrow \underbrace{H, H, H, T}_{\text{Probability}} \Rightarrow \underbrace{P(X=H) = \frac{3}{4}}_{\text{Statistic world}}$$

most Lazy form Lazy form

$$\underbrace{f(x)}_{\text{function}} = f(x|\theta) = f(x|a,b) = \frac{1}{\sqrt{2\pi a^2}} \exp\left(-\frac{(x-b)^2}{2a^2}\right)$$



Objective function $\underset{\text{Max}}{L(a,b|X)}$

$\rightarrow a=1, b=1 \Rightarrow f(x^{[1]}, x^{[2]}, \dots, x^{[n]} | a=1, b=1) =$

$\rightarrow a=1, b=2 \Rightarrow f(x^{[1]}, \dots, x^{[n]} | a=1, b=2) = f(x^{[1]}|a,b) f(x^{[2]}|a,b) \dots f(x^{[n]}|a,b)$

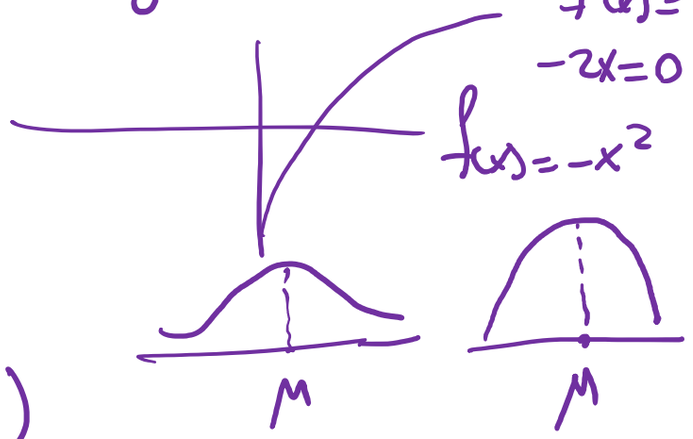
Prob vs Likelihood

→ Likelihood function

$$L(a,b \mid x) = \prod_{i=1}^{N} f(x^{\{i\}} \mid a,b)$$

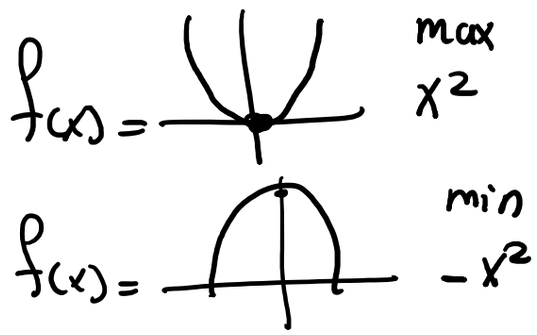
Objective function

$$\log(a*b) = \log a + \log b$$



max

$$l(a,b \mid x) = \log L(a,b \mid x) = \sum_{i=1}^{N} \log f(x^{\{i\}} \mid a,b)$$



$$\frac{\partial l(a,b \mid x)}{\partial a} = 0 \quad \Rightarrow \quad a = \sigma^2$$

$$\frac{\partial l(a,b \mid x)}{\partial b} = 0$$

$$b = \mu = \frac{1}{N} \sum_{i=1}^{N} x^{\{i\}}$$

MLE

Multivariate Gaussian Distribution

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\}$$

- Moment Parameterization $\mu = E(X)$

$$\Sigma = Cov(X) = E[(X-\mu)(X-\mu)^\top]$$

- Mahalanobis Distance $\Delta^2 = (x-\mu)^\top \Sigma^{-1}(x-\mu)$

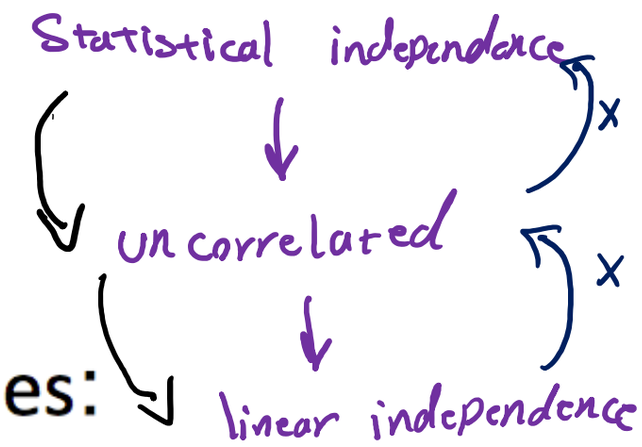
- Tons of applications (MoG, FA, PPCA, Kalman filter,…)

Properties of Gaussian Distribution

- The linear transform of a Gaussian r.v. is a Gaussian. Remember that no matter how x is distributed

$$E(AX + b) = AE(X) + b$$

$$Cov(AX + b) = ACov(X)A^{\top}$$

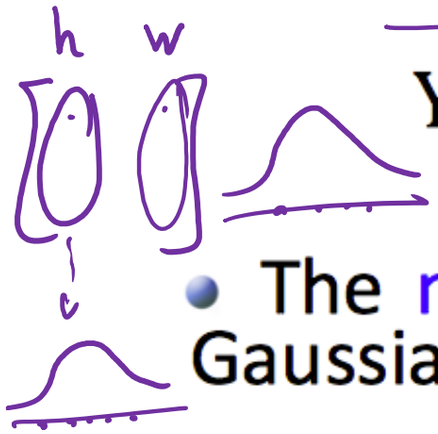


this means that for Gaussian distributed quantities:

$$X \sim N(\mu, \Sigma) \rightarrow Z = AX + b \sim N(A\mu + b, A\Sigma A^{\top})$$

- The sum of two independent Gaussian r.v. is a Gaussian

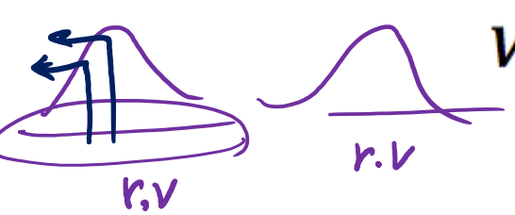
$$Y = X_1 + X_2,\ X_1 \perp X_2 \rightarrow \mu_y = \mu_1 + \mu_2, \Sigma_y = \Sigma_1 + \Sigma_2$$



- The multiplication of two Gaussian functions is another Gaussian function (although no longer normalized)

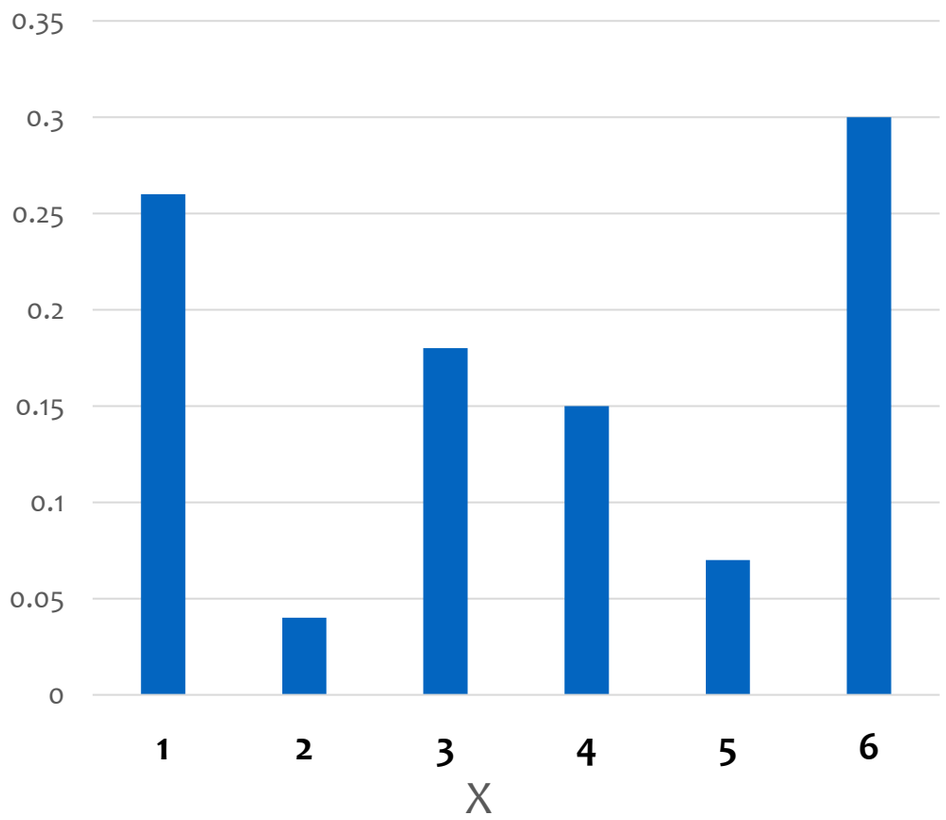
$$N(a, A)N(b, B) \propto N(c, C),$$

$$where\ C = (A^{-1} + B^{-1})^{-1}, c = CA^{-1}a + CB^{-1}b$$



Central Limit Theorem CLT

Probability mass function of a **biased** dice



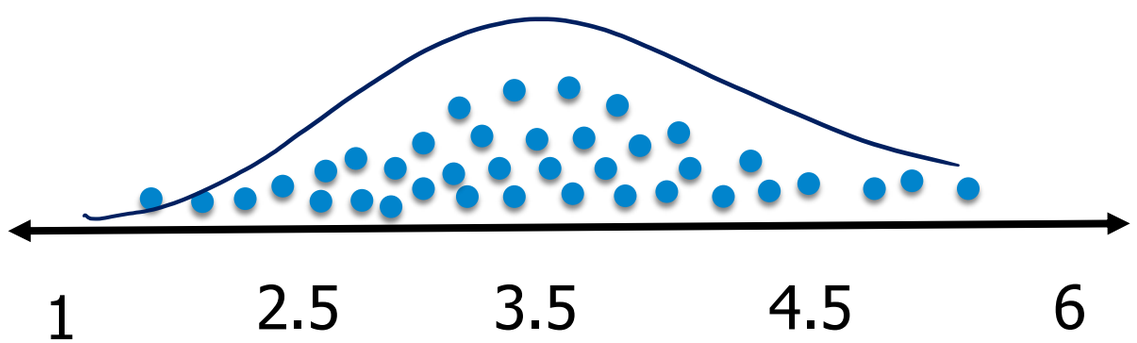
Let's say, I am going to get a sample from this pmf having a size of $\boldsymbol{n = 4}$

$S_1 = \{1,1,1,6\} \Rightarrow E(S_1) = 2.25$

$S_2 = \{1,1,3,6\} \Rightarrow E(S_2) = 2.75$

$\vdots$

$S_m = \{1,4,6,6\} \Rightarrow E(S_m) = 4.25$



According to CLT, it will follow a bell curve distribution (normal distribution)

Outline

- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Maximum Likelihood Estimation

- Probability: inferring probabilistic quantities for data given fixed models (e.g. prob. of events, marginals, conditionals, etc).
- Statistics: inferring a model given fixed data observations (e.g. clustering, classification, regression).

Main assumption:

Independent and identically distributed random variables

i.i.d

$$p(x^{[i]}) = \underline{\underline{\theta}}^{x^{[i]}} (1-\theta)^{(1-x^{[i]})} \rightsquigarrow pmf = P(x^{[i]}|\theta)$$

$$\Rightarrow \text{Objective function } p(x^{[1]}, x^{[2]}, \dots, x^{[n]}|\theta) = p(x^{[1]}|\theta)\,p(x^{[2]}|\theta) \dots p(x^{[N]}|\theta)$$

$$\max L(\theta|x) = \prod_{i=1}^{N} p(x^{[i]}|\theta)$$

$$l(\theta|x) = \log L(\theta|x) = \sum_{i=1}^{N} \log p(x^{[i]}|\theta) = \sum_{i=1}^{N} \log\left(\theta^{x^{[i]}} (1-\theta)^{(1-x^{[i]})}\right)$$

Maximum Likelihood Estimation

For Bernoulli (i.e. flip a coin):

Objective function: $P(x_i|\theta) = \theta^{x^{\{i\}}}(1-\theta)^{1-x^{\{i\}}} \quad x^{\{i\}} \in \{0,1\}\ or\ \{head, tail\}$

$$L(\theta|X) = L(\theta|X = x^{\{1\}}, X = x^{\{2\}}, X = x^{\{3\}}, \dots, X = x^{\{n\}})$$

i.i.d assumption

$$L(\theta|X) = \prod_{i=1}^{n} P\left(x^{\{i\}}|\theta\right)$$

$$L(\theta|X) = \prod_{i=1}^{n} P\left(x^{\{i\}}|\theta\right) = \prod_{i=1}^{n} \theta^{x^{\{i\}}}(1-\theta)^{1-x^{\{i\}}}$$

$$L(\theta|X) = \theta^{x^{\{1\}}}(1-\theta)^{1-x^{\{1\}}} \times \theta^{x^{\{2\}}}(1-\theta)^{1-x^{\{2\}}} \dots \times \theta^{x^{\{n\}}}(1-\theta)^{1-x^{\{n\}}} =$$
$$= \theta^{\sum x^{\{i\}}}(1-\theta)^{\sum(1-x^{\{i\}})}$$

We don't like multiplication, let's convert it into summation

What's the trick? Take the log

$$L(\theta|X) = \theta^{\sum x^{\{i\}}}(1-\theta)^{\sum(1-x^{\{i\}})}$$

$$\log\theta^{\sum x^{\{i\}}} = \sum x^{\{i\}} \log\theta$$

$\log\theta$ ↑ $\log a + \log b$

$$logL(\theta|X) = l(\theta|X) = \log(\theta)\sum_{i=1}^{n} x^{\{i\}} + \log(1-\theta)\sum_{i=1}^{n}(1-x^{\{i\}})$$

How to optimize θ?

$$\frac{\partial l(\theta|X)}{\partial\theta} = 0 \qquad \frac{\sum_{i=1}^{n} x^{\{i\}}}{\theta} - \frac{\sum_{i=1}^{n} 1 - x^{\{i\}})}{1-\theta} = 0$$

$x^{head} = 1$

$x^{tail} = 0$

H, H, H, T, T

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x^{\{i\}} \qquad \theta = \frac{1+1+1+0+0}{5} = \frac{3}{5}$$