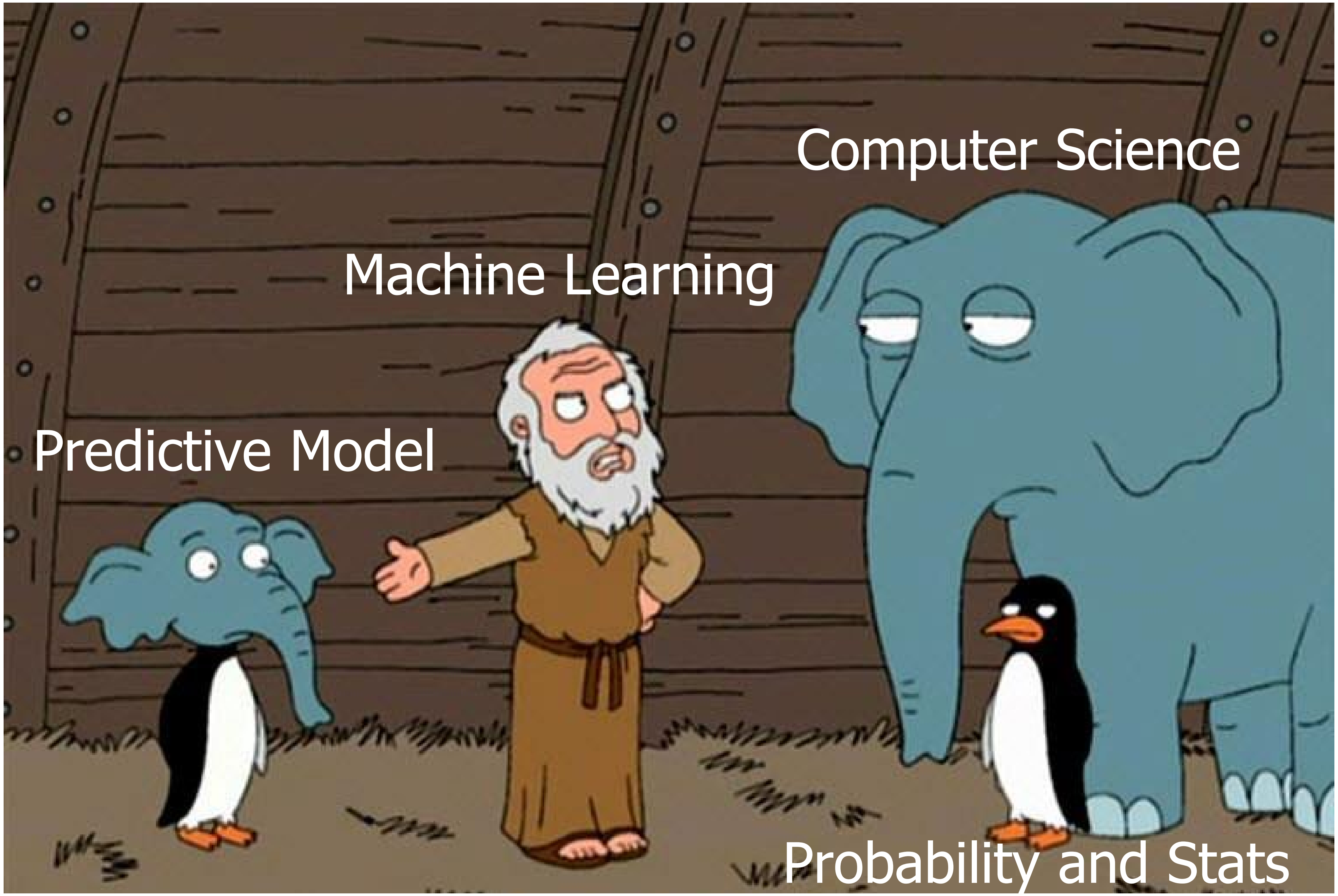

Computer Science
Machine Learning
Predictive Model
Probability and Stats

Machine Learning CS 4641-7641

Probability and Statistics

Mahdi Roozbahani
Georgia Tech

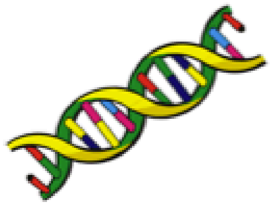
These slides are inspired based on slides from Le Song , Sam Roweis, and Chao Zhang.

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 $(1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$
  - 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**

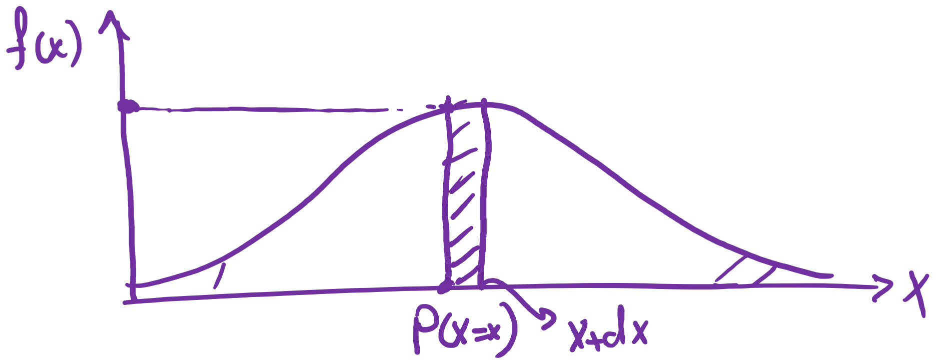
Random variables X represents **outcomes** in sample space

$P(X=2) = \frac{1}{6}$

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

$$p(x) \geq 0$$

density or likelihood



Continuous variable

Continuous probability distribution

Probability density function

Density or likelihood value

Temperature (real number)

Gaussian Distribution

$$\boxed{\int_x p(x)dx = 1}$$

Discrete variable

Discrete probability distribution

Probability mass function

Probability value

Coin flip (integer)

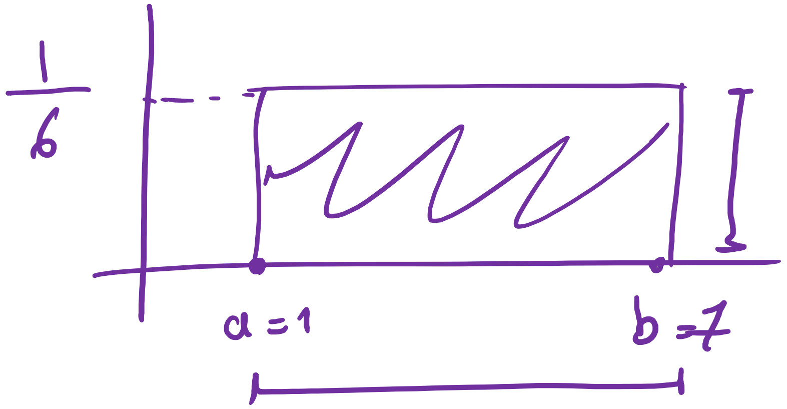
Bernoulli distribution

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

Continuous Probability Functions

- Examples:
  - Uniform Density Function:

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



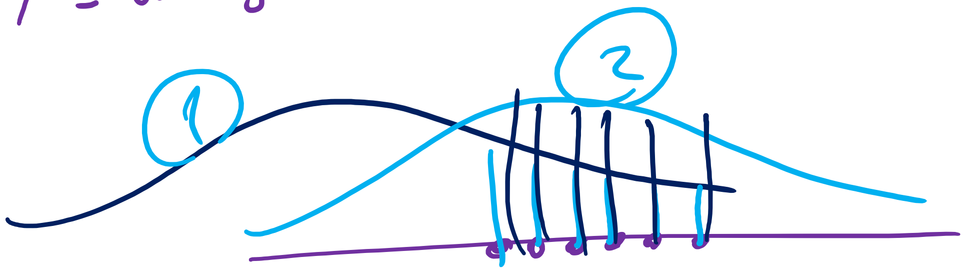
  - Exponential Density Function:

Parameter

μ = average = mean

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

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



  - Gaussian(Normal) Density Function

$$f_x(x) = \frac{1}{\sqrt{2\pi}\underset{b}{\cancel{\sigma}}} e^{-\frac{(x-\overset{a}{\cancel{\mu}})^2}{2\underset{b^2}{\cancel{\sigma^2}}}}$$

μ → mean

σ → Standard deviation

σ^2 → Variance

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

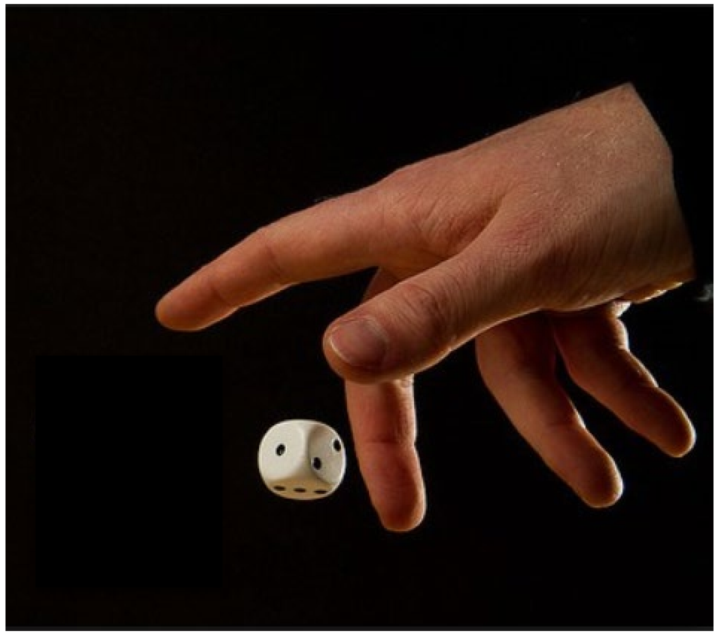
n-k is number of failures

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

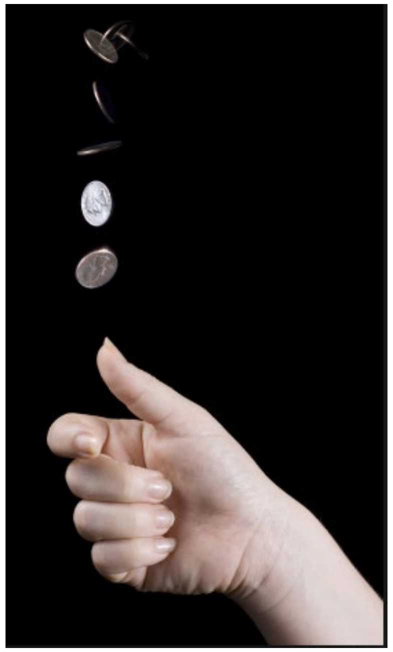
Outline

- 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

$\boldsymbol{n_{ij}}$ = Number of occurrence

X

		$x_{i=1} = 1$	$x_{i=2} = 2$	$x_{i=3} = 3$	$x_{i=4} = 4$	$x_{i=5} = 5$	$x_{i=6} = 6$	C_j
Y	$y_{j=2} = tail$	$n_{ij} = 3$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 5$	$n_{ij} = 1$	$n_{ij} = 5$	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

		$x_{i=1}=1$	$x_{i=2}=2$	$x_{i=3}=3$	$x_{i=4}=4$	$x_{i=5}=5$	$x_{i=6}=6$	C_j
Y	$y_{j=2}=tail$	$n_{ij}=3$	$n_{ij}=4$	$n_{ij}=2$	$n_{ij}=5$	$n_{ij}=1$	$n_{ij}=5$	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

$$P(X=2, Y=\text{tail}) = \frac{4}{35} = \frac{n_{ij}}{N}$$

$$P(Y=\text{head}) = \frac{15}{35} = \frac{C_j}{N}$$

$$P(X=3) = \frac{6}{35} = \frac{C_i}{N} = \sum_{Y} P(X=3, Y=y) \rightarrow \text{Sum Rule}$$

$$P(Y=\text{tail} \mid X=4) = \frac{5}{7} = \frac{n_{ij}}{C_i}$$

$$P(X=4 \mid Y=\text{tail}) = \frac{5}{20} = \frac{n_{ij}}{C_j}$$

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

$$P(x,y) = P(y|x)\,P(x) = P(x|y)\,P(y)$$

↳ product rule

Probability:

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

Joint probability:

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

Conditional probability:

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

Sum rule

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

Product rule

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

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

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

Conditional Independence

- Examples:

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

P(Virus | Drink Beer) = P(Virus)
iff Virus is independent of Drink Beer

P(Flu | Virus,DrinkBeer) = P(Flu|Virus)
iff Flu is independent of Drink Beer, given Virus

P(Headache | Flu,Virus,DrinkBeer) =
P(Headache|Flu,DrinkBeer)
iff Headache is independent of Virus, given Flu and Drink Beer

Assume the above independence, we obtain:
P(Headache;Flu;Virus;DrinkBeer)
=P(Headache | Flu;Virus;DrinkBeer) P(Flu | Virus;DrinkBeer)
P(Virus | Drink Beer) P(DrinkBeer)
=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.

$$P(x,y) = P(x|y)\,P(y) \Rightarrow P(x|y) = \frac{P(x,y)}{\boxed{P(y)}} = \frac{P(y|x)\,P(x)}{P(y)}$$

- 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?

$$P(y) = \sum_{x} P(y, x=x) = \sum P(y|x)\,P(x)$$

- 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

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

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

Other cases:

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

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

$$\rightarrow P(X,Y,Z) = P(X|Y,Z)P(Y,Z)$$

Outline

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

Mean and Variance

$E[g(x)] = \sum g(x)\, p(x)$

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

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

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

$Var(x) = \boxed{E_X[(X - E_X[X])^2]} = \underline{E_X[X^2] - E_X[X]^2}$ $\quad E[x^2] = Var(x) + E[x]^2$

  - $Var(\alpha X) = \alpha^2 Var(X)$
  - $Var(\alpha + X) = Var(X)$

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

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

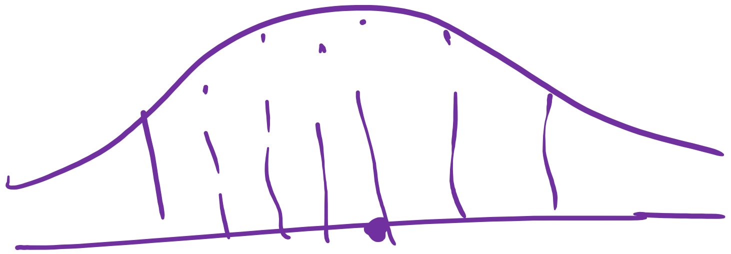
$P(x=1) = \frac{1}{5} \qquad P(x=2) = \frac{2}{5} \qquad P(x=3) = \frac{2}{5}$

$$E[g(x)] = 1 \times \frac{1}{5} + 2 \times \frac{2}{5} + 3 \times \frac{2}{5} = \frac{11}{5}$$

$$\mu = \text{mean} = \frac{1+2+3}{3} = 2$$

$x = [1, 2, 2, 3, 3]$

$$\mu = \frac{1+2+2+3+3}{5} = \frac{11}{5} = E[g(x)]$$



$$X = \overset{\text{height}=h}{\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}} \qquad \mu_h = \frac{1+2+3}{3} = 2 \qquad \bar{X} = \overset{\bar{h}}{\begin{bmatrix} 1-\mu_h \\ 2-\mu_h \\ 3-\mu_h \end{bmatrix}} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

$$\sigma_h^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_h)^2 = \boxed{\frac{1}{N} \sum_{i=1}^{N}} (x_i - E[h])^2$$

$$= E\left[(x_i - E[h])^2 \right]$$

$$\sigma_h^2 = \frac{\bar{X}^T \bar{X}}{N} = \frac{1}{N} \begin{bmatrix} 1-\mu_h & 2-\mu_h & 3-\mu_h \end{bmatrix} \begin{bmatrix} 1-\mu_h \\ 2-\mu_h \\ 3-\mu_h \end{bmatrix}$$

$$= \frac{1}{N} \left[(1-\mu_h)^2 + (2-\mu_h)^2 + (3-\mu_h)^2 \right] = \frac{\sum_{i=1}^{N} (x - \mu_h)^2}{N}$$

$$X = \begin{array}{c} \\ \\ \end{array}\left[\begin{array}{cc} h & weight=w \\ 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{array}\right]_{\substack{3x2 \\ nxd}}$$

$$\mu_h = 2 \qquad \mu_w = 5$$

$$\bar{X} = \left[\begin{array}{cc} \bar{h} & \bar{w} \\ 1-\mu_h & 4-\mu_w \\ 2-\mu_h & 5-\mu_w \\ 3-\mu_h & 6-\mu_w \end{array}\right]$$

$$Cov = \frac{\bar{X}^T \bar{X}}{N} = \frac{1}{N}\begin{bmatrix} 1-\mu_h & 2-\mu_h & 3-\mu_h \\ 4-\mu_w & 5-\mu_w & 6-\mu_w \end{bmatrix}\begin{bmatrix} 1-\mu_h & 4-\mu_w \\ 2-\mu_h & 5-\mu_w \\ 3-\mu_h & 6-\mu_w \end{bmatrix}$$

$$Cov = \begin{bmatrix} \sigma_h^2 = \sigma_{hh} & \sigma_{hw} \\ \sigma_{wh} & \sigma_w^2 = \sigma_{ww} \end{bmatrix}_{\substack{2x2 \\ dxd}}$$

$$X - \mu_h \perp X - \mu_w$$

$$\begin{bmatrix} 1-\mu_h \\ 2-\mu_h \\ 3-\mu_h \end{bmatrix} \qquad \begin{bmatrix} 4-\mu_w \\ 5-\mu_w \\ 6-\mu_w \end{bmatrix}$$

For Joint Distributions

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

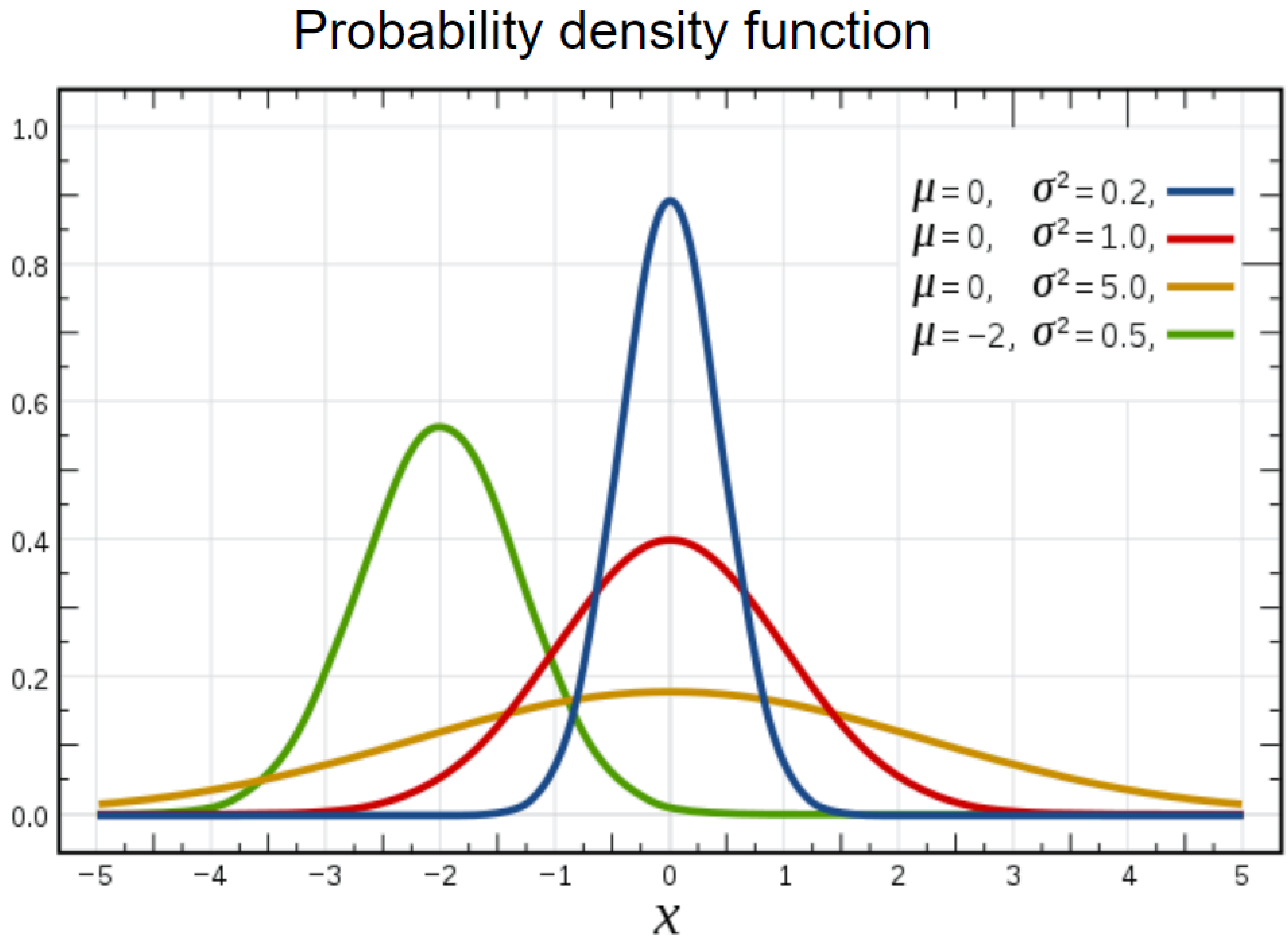
Uncorrelated vs Independent RV

Outline

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

Gaussian Distribution

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



Probability versus likelihood

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:

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

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

$$\mathrm{Y}=\mathrm{X}_1+\mathrm{X}_2,\ \mathrm{X}_1\perp X_2\ \rightarrow \mu_y=\mu_1+\mu_2,\Sigma_y=\Sigma_1+\Sigma_2$$

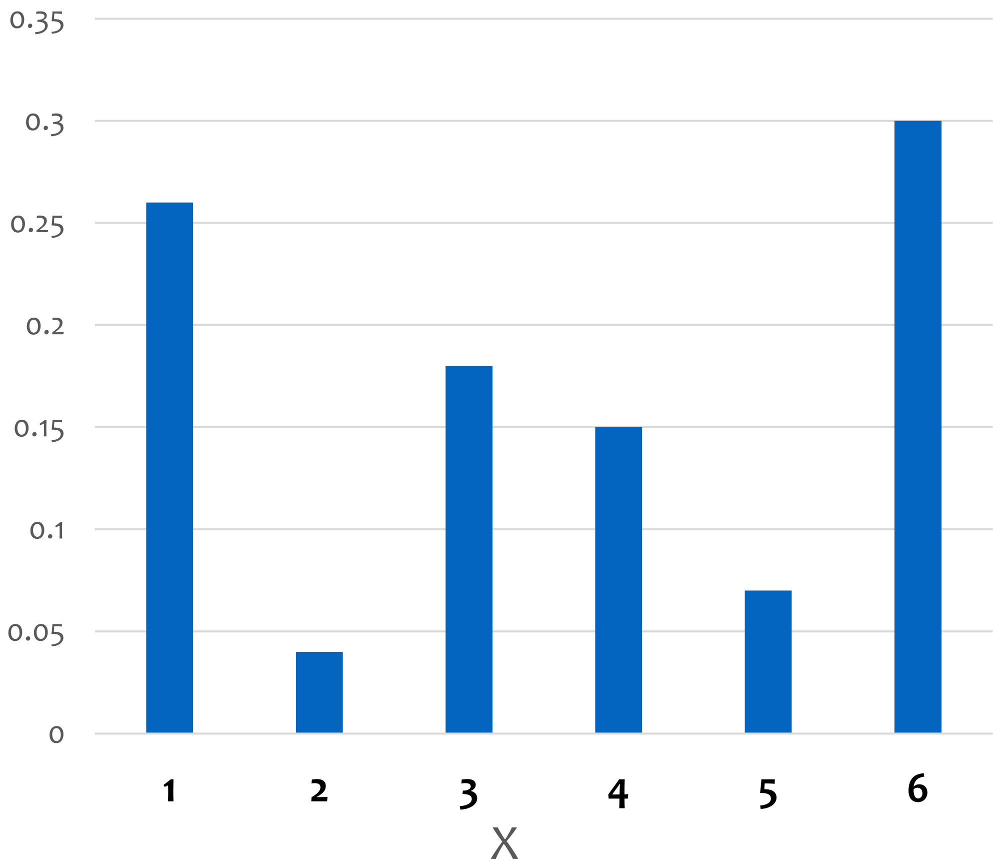
- The multiplication of two Gaussian functions is another Gaussian function (although no longer normalized)

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

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

Central Limit Theorem

Probability mass function of a **biased** dice



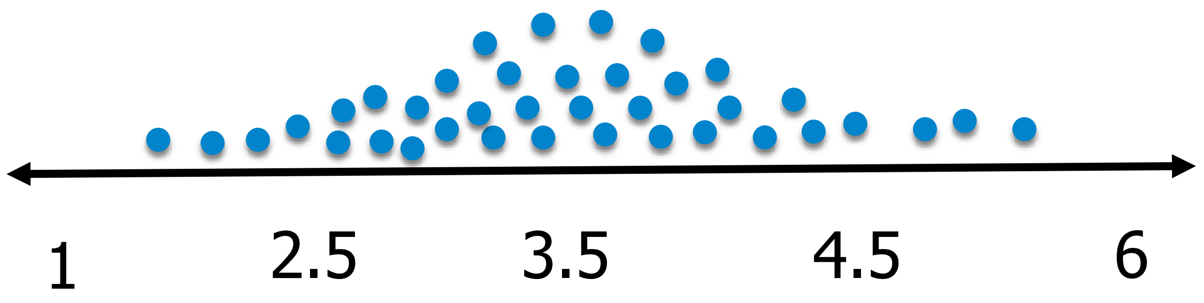
Let’s say, I am going to get a sample from this pmf having a size of $\boldsymbol{n = 4}$

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

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

$$\vdots$$

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

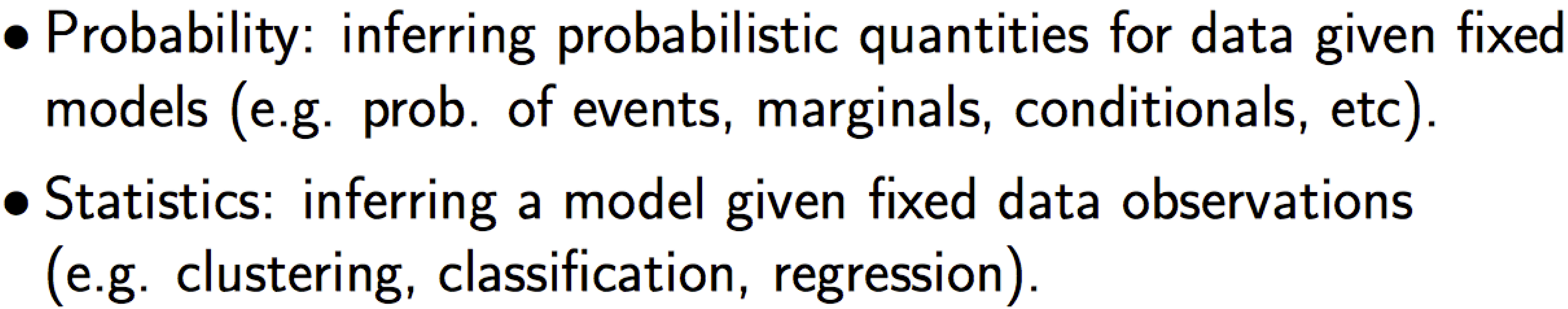


According to CLT, it will follow a bell curve distribution (normal distribution)

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

Maximum Likelihood Estimation

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

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

$$L(\theta|X) = L(\theta|X = x_1, X = x_2, X = x_3, \dots, X = x_n)$$

i.i.d assumption

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

$$L(\theta|X) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$L(\theta|X) = \theta^{x_1}(1-\theta)^{1-x_1} \times \theta^{x_2}(1-\theta)^{1-x_2} \dots \times \theta^{x_n}(1-\theta)^{1-x_n} =$$
$$= \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)}$$

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

What’s the trick? Take the log

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

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

How to optimize θ?

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

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$