Probability and Statistics

Nakul Gopalan
Georgia Tech

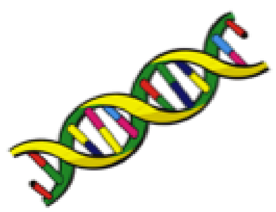
These slides are based on slides from Le Song , Sam Roweis, Mahdi Roozbahani, and Chao Zhang.

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 $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix}$
  - E.g., S may be the set of all possible nucleotides of a DNA site: S $\begin{pmatrix} A & C & G & T \end{pmatrix}$



  - 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 $p(x) = p(X = x)$

$$p(x) \geq 0$$

Continuous variable

Continuous probability distribution

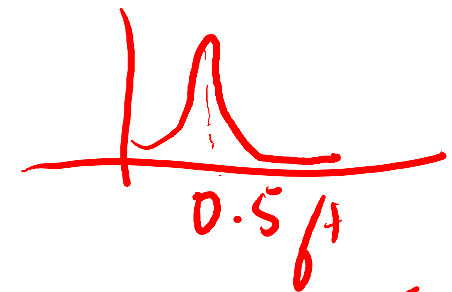
Probability density function

Density or likelihood value

Temperature (real number)

Gaussian Distribution

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



0.5   0.500001

0.500.... 13

Discrete variable

Discrete probability distribution

Probability mass function

Probability value

Coin flip (integer)

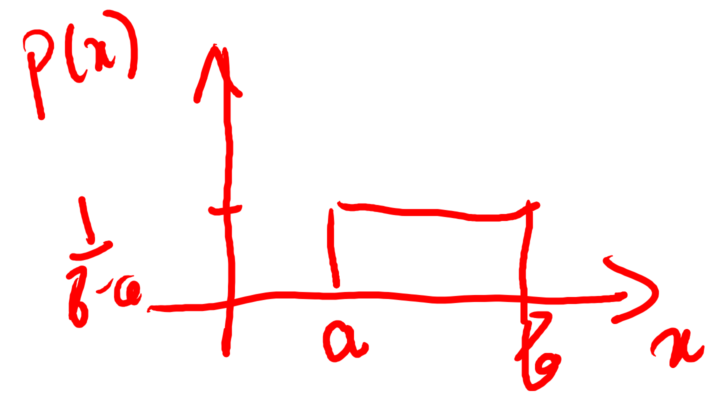
Bernoulli distribution

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

Continuous Probability Functions

- Examples:
  - Uniform Density Function:

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



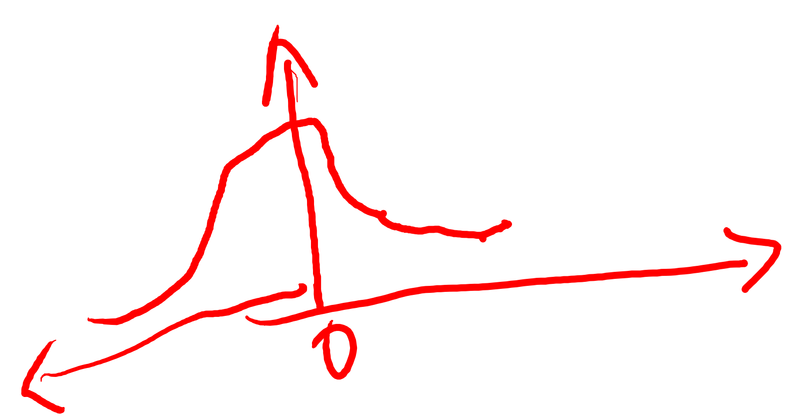
  - Exponential Density Function:

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

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

  - Gaussian(Normal) Density Function

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



Discrete Probability Functions

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

      → 0.5

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

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

P(X = k) =

n → total coin tosses

k is number of successes

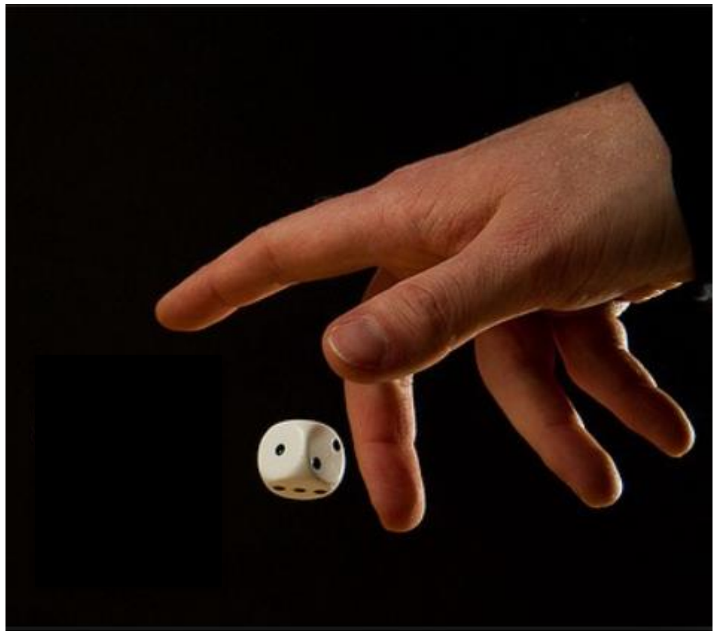
n-k is number of failures

$\binom{n}{k}$ 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

coin toss

die roll

N = total number of trials

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

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						
		$x_{i=1} = 1$	$x_{i=2} = 2$	$x_{i=3} = 3$	$x_{i=4} = 4$	$x_{i=5} = 5$	$x_{i=6} = 6$	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

Pr(x=4, y=h) $= \frac{2}{35}$

Pr(x=5) $= \frac{\sum_Y (x=5)\,Y}{N} = \frac{(x=5, y=T) + (x=5, y=F)}{35} = \frac{5}{35}$

Marginalization

Pr(y=t | x=3) $= \frac{2}{2+4} = \frac{2}{6} = \frac{1}{3} = \frac{n_{ij}}{C_i} = \frac{P(y=t, x=3)}{P(x=3)} = \frac{2/35}{6/35}$

Joint from conditional

Pr(x=xᵢ , y = yᵢ) $= P(y=y_i \mid x=x_i)\, P(x=x_i)$

$1/3 \cdot 6/35 = 2/35$

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

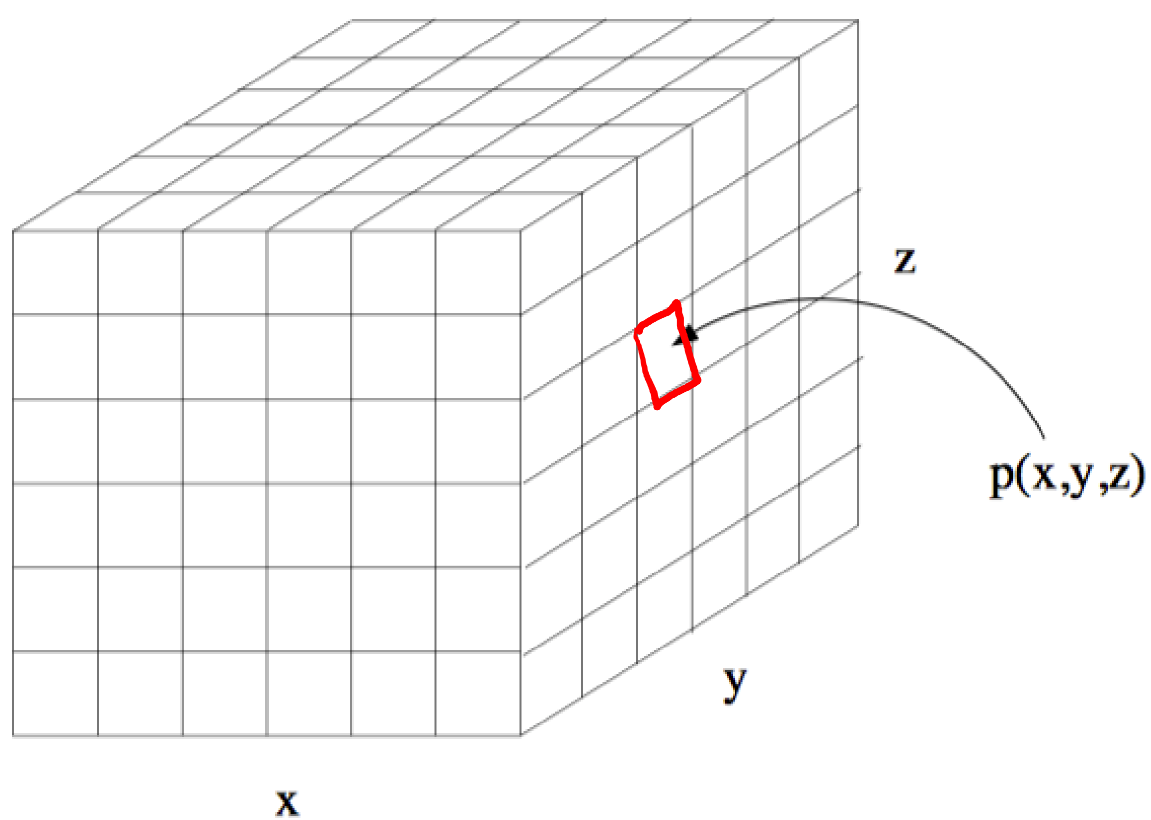
$$p\left(X = x_i, Y = y_j\right) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \frac{c_i}{N} = p\left(Y = y_j \middle| X = x_i\right) p(X = x_i)$$

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

Joint Distribution

- Key concept: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write

$$p(x, y) = \text{prob}(X = x \text{ and } Y = y)$$

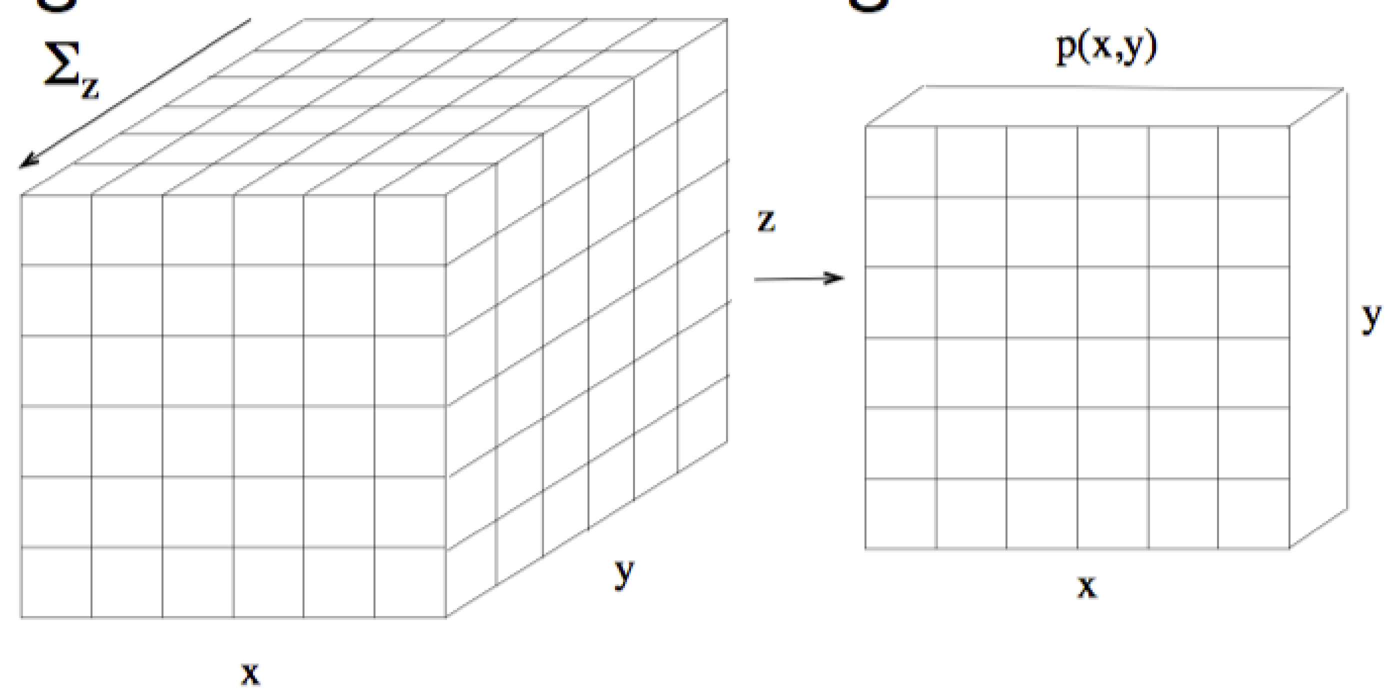


Marginal Distribution

- We can "sum out" part of a joint distribution to get the *marginal distribution* of a subset of variables:

$$p(x) = \sum_y p(x, y)$$

- This is like adding slices of the table together.

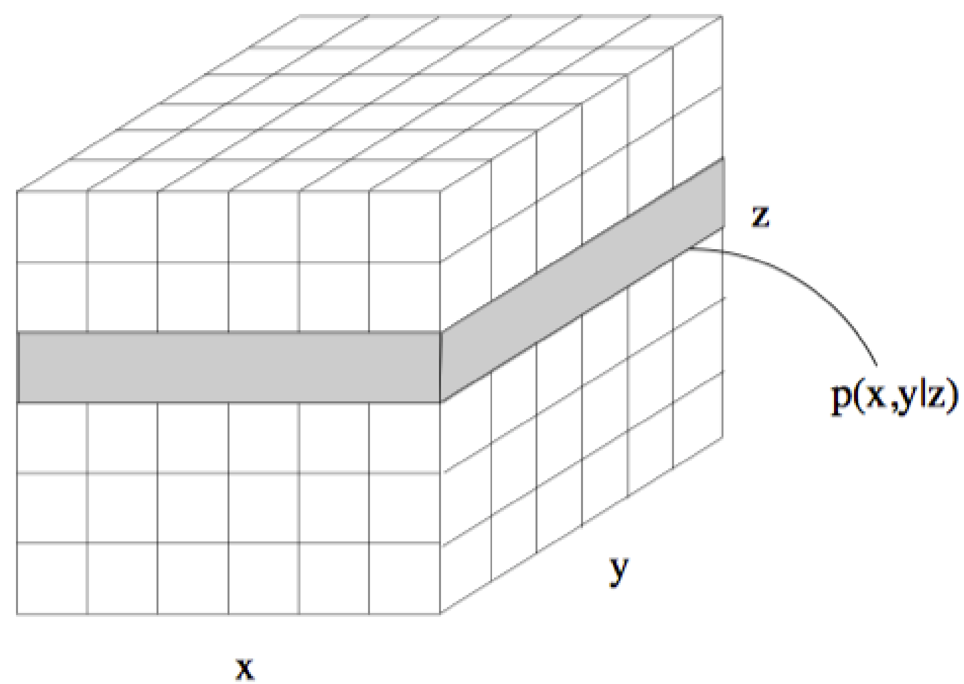


- Another equivalent definition: $p(x) = \sum_y p(x|y)p(y)$.

Conditional Distribution

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a "slice" through the joint table.

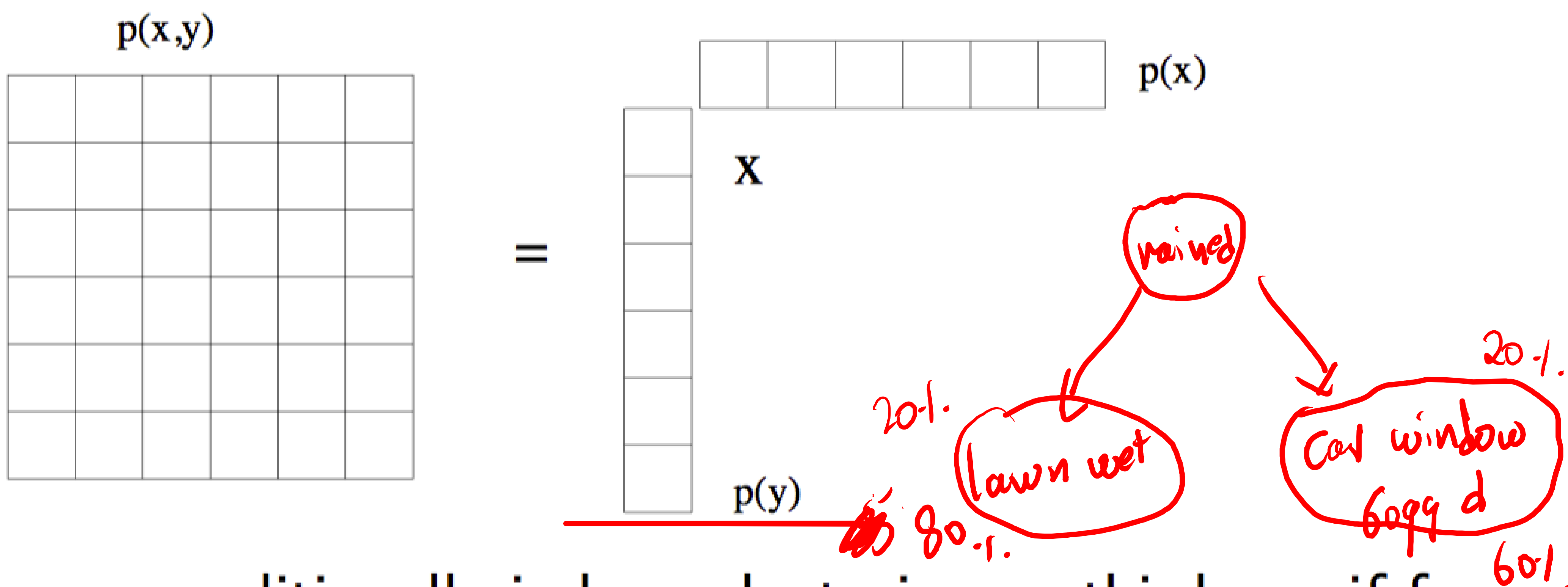
$$\boxed{p(x|y) = p(x,y)/p(y)}$$



Independence & Conditional Independence

- Two variables are independent iff their joint factors:

$$p(x,y) = p(x)p(y)$$



- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

$$p(x,y|z) = p(x|z)p(y|z) \qquad \forall z$$

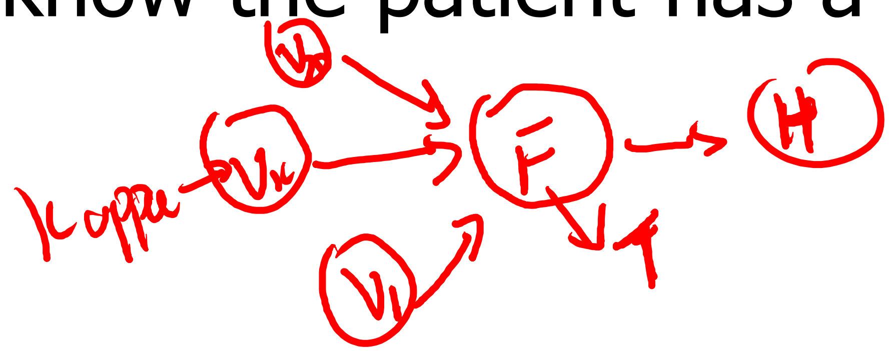
Poll

- A virus Kappa is known to cause a flu. And a flu is know to cause a headache sometimes. A flu can be caused by multiple reasons. A patient shows up with a diagnosis of a flu. We know the patient has flu. Does the probability of the patient having a headache depend on the virus Kappa now or not?

Options:

1) Yes, probability of a headache depends on the virus Kappa. 33%

2) No, probability of a headache is independent of the virus Kappa as we know the patient has a flu. 67%



Conditional Independence

- Examples:

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)~~ 4D
=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)
3 D 2D 1D 1D

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.

- 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?

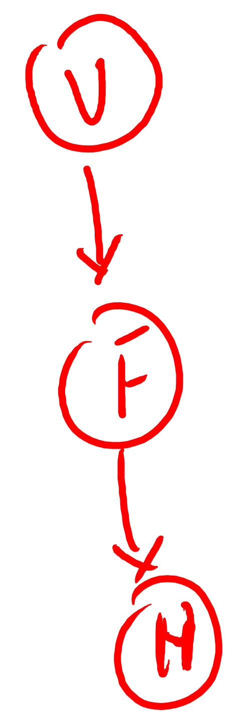
- Definition:

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

Corollary:

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

This is called Bayes Rule



Bayes' Rule

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

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

Other cases:

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

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

Administrative business

- Office hours are live
- Live Q&A parallel to the lectures to TAs can help with answering questions. Use voting as well so I know what needs to be handled here.
- Chris is holding a python tutorial on Thursday at 6 pm
  numpy
- Project questions answered on Thursday's lecture by me, come one, come all.

Outline

- Probability Distributions $P(x)$
- Joint and Conditional Probability Distributions $p(x)$ $P(y)$
- Bayes' Rule $p(x|y) = \dfrac{p(x,y)}{p(y)} = \dfrac{p(y|x)\,p(x)}{p(y)}$
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

Mean and Variance

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

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

$\int_{-\infty}^{\infty} x\, p(x)\, dx$

h_1
h_2
$\vdots$
h_n

$\frac{\sum_i^n h_i}{n}$

- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n$
- Mean: $E_X[X] = \int_{-\infty}^{\infty} x p_X(x)dx$
  - $E[\alpha X] = \alpha E[X]$
  - $E[\alpha + X] = \alpha + E[X]$
- Variance(Second central moment): $Var(x) = E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2$
  - $Var(\alpha X) = \alpha^2 Var(X)$
  - $Var(\alpha + X) = Var(X)$

For Joint Distributions

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

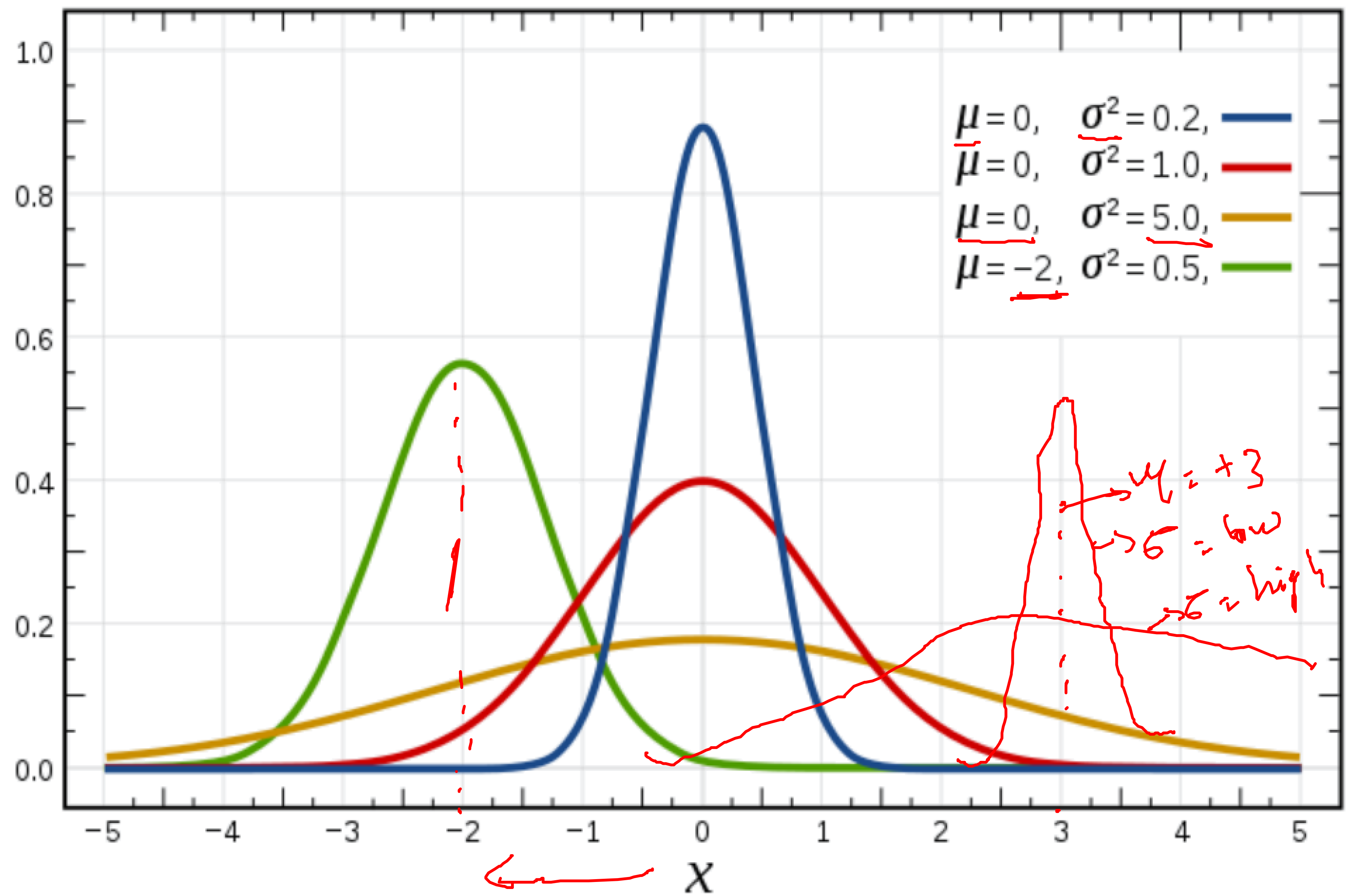
Outline

- 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 density function



Multivariate Gaussian Distribution

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

x_1 μ_1 x_1 x_2

$\vdots$

x_n μ_n

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

$$\Sigma = Cov(X) = E[(X-\mu)(X-\mu)^{\mathsf{T}}]$$

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

- 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

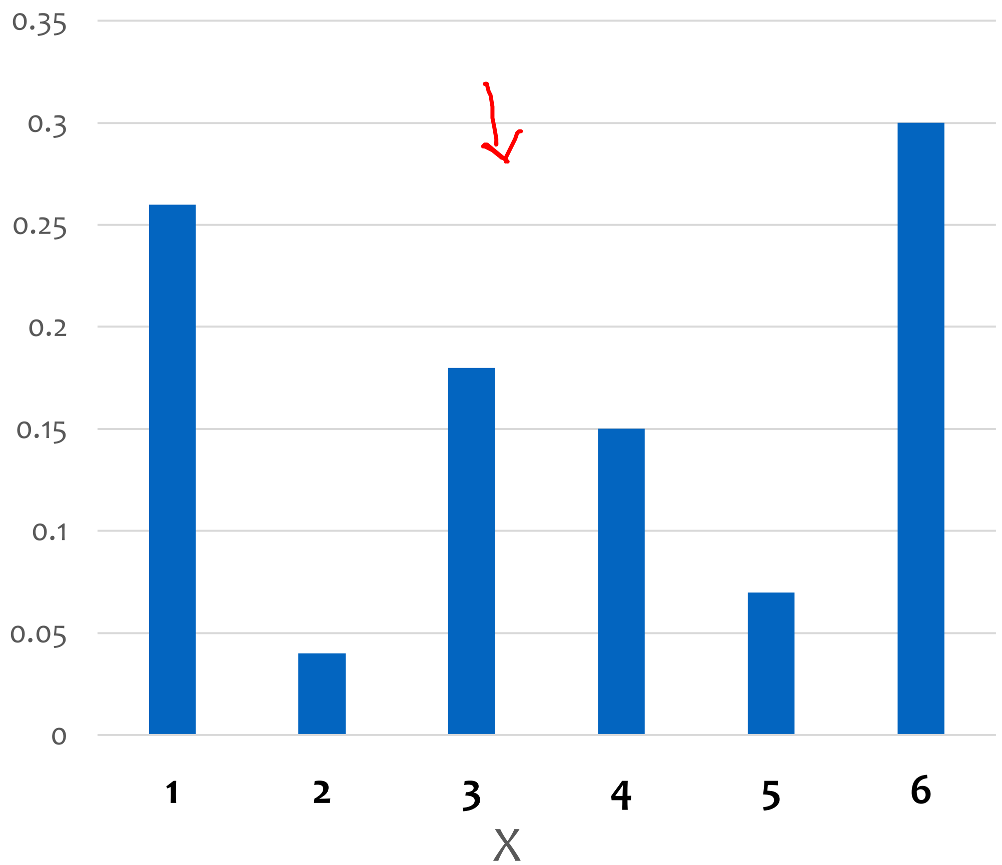
$$\mathrm{Y} = \mathrm{X}_1 + \mathrm{X}_2,\ \mathrm{X}_1 \perp X_2 \;\rightarrow \mu_y = \mu_1+\mu_2, \Sigma_y = \Sigma_1+\Sigma_2$$

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

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

Central Limit Theorem

Probability mass function of a **biased** dice



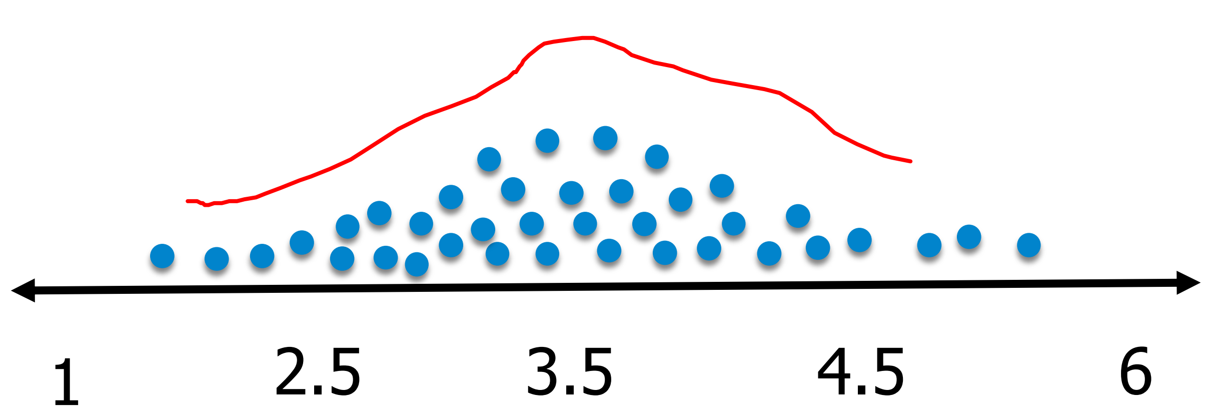
Let's say, I am going to get a sample from this pmf having a size of $n = 4$

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

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

$$\vdots$$

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



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

CLT Definition

- Statement: The central limit theorem (due to Laplace) tells us that, subject to certain mild conditions, the sum of a set of random variables, which is of course itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases (Walker, 1969).

Outline

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

Likelihood, what is it? Biased coin from a stranger

Heads Biased

Tails biased

T T

T T

H H

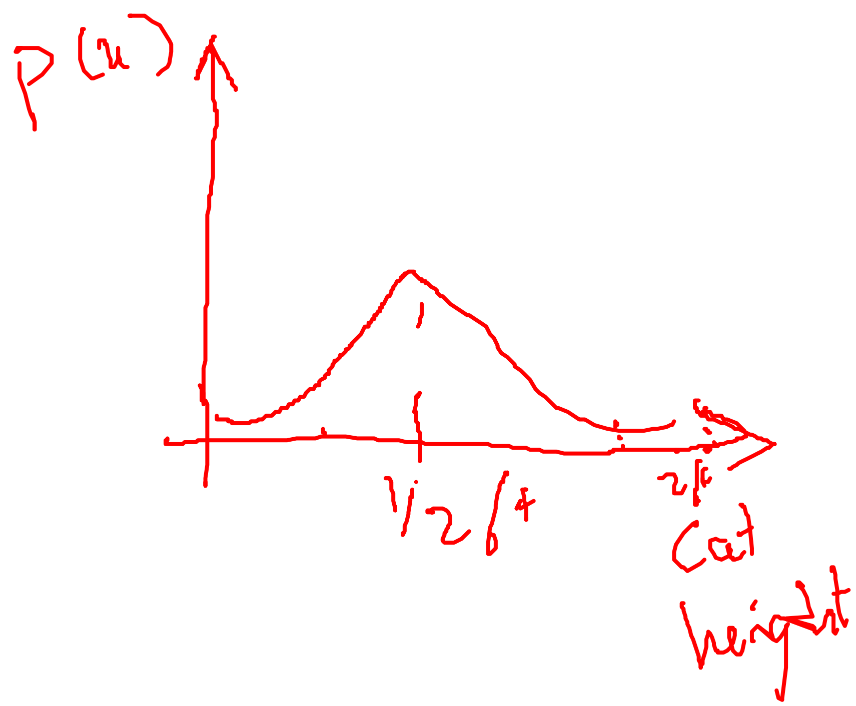
H H

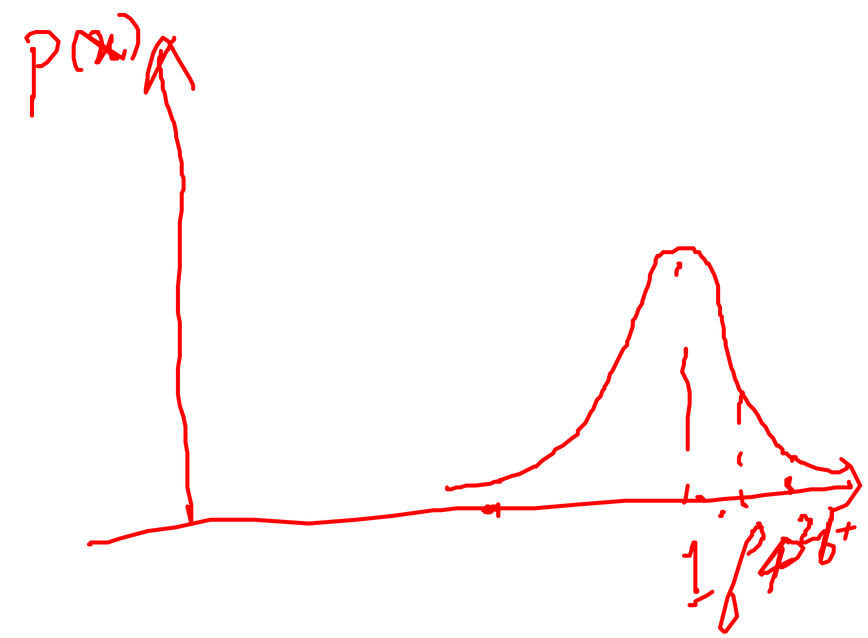
H H

H H

H H

Likelihood, what is it? or Cat or Dog? Do we know? Let's find out!





heard → [2 ft, 1.5 ft, 0.2 ft, 1.2 ft ..]

$$L(x_{1:4} \mid cat) = p(x_1 \mid cat) \cdot p(x_2 \mid cat) \cdots p(x_n \mid cat)$$

$$\underline{0.001} \times 0.003 \times 0.05 \times 0.005$$

$$= C_{cat}$$

$$L(x_{1:4} \mid dog) = p(x_1 \mid dog) \cdot p(x_2 \mid dog) \cdot p(x_n \mid dog)$$

$$= C_{dog} \qquad C_{cat} << C_{dog}$$

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: $f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$ $\qquad x_i \in \{0,1\}\ or\ \{head, tail\}$

$$L(p) = Pr(X = x_1, X = x_2, X = x_3, \dots, X = x_n)$$

We want to know what is the most "likely" value for the probability of success *p* given *n* observations???

Maximum Likelihood Estimation

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

Objective function: $f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$ $\qquad x_i \in \{0,1\}\ or\ \{head, tail\}$

$$L(p) = Pr(X = x_1, X = x_2, X = x_3, \dots, X = x_n)$$

i.i.d assumption

$$= Pr(X = x_1)Pr(X = x_2) \dots Pr(X = x_n) = f(p;x_1)f(p;x_2) \dots f(p;x_n)$$

$$L(p) = \prod_{i=1}^{n} f(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$L(p) = p^{x_1}(1-p)^{1-x_1} \times p^{x_2}(1-p)^{1-x_2} \dots \times p^{x_n}(1-p)^{1-x_n} =$$

$$= p^{\Sigma x_i}(1-p)^{\Sigma(1-x_i)}$$

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

What's the trick? Take the log

$$L(p) = p^{\sum x_i}(1-p)^{\sum(1-x_i)}$$

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

How to optimize p?

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

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