


Probability and Statistics

Mahdi Roozbahani
Georgia Tech

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 2 3 4 5 6)
 - E.g., S may be the set of all possible nucleotides of a DNA site: S
(A C G 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 $p(x) = p(X = x)$

$$p(x) \geq 0$$

Continuous variable

Continuous probability distribution

Probability density function

Density or likelihood value

Temperature (real number)

Gaussian Distribution

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

Discrete variable

Discrete probability distribution

Probability mass function

Probability value

Coin flip (integer)

Bernoulli distribution

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

Continuous Probability Functions

- Examples:

- Uniform Density Function:

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

- Exponential Density Function:

$$f_x(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \quad \text{for } x \geq 0$$

$$F_x(x) = 1 - e^{-\frac{x}{\mu}} \quad \text{for } x \geq 0$$

- 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:

- $$\begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$$

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

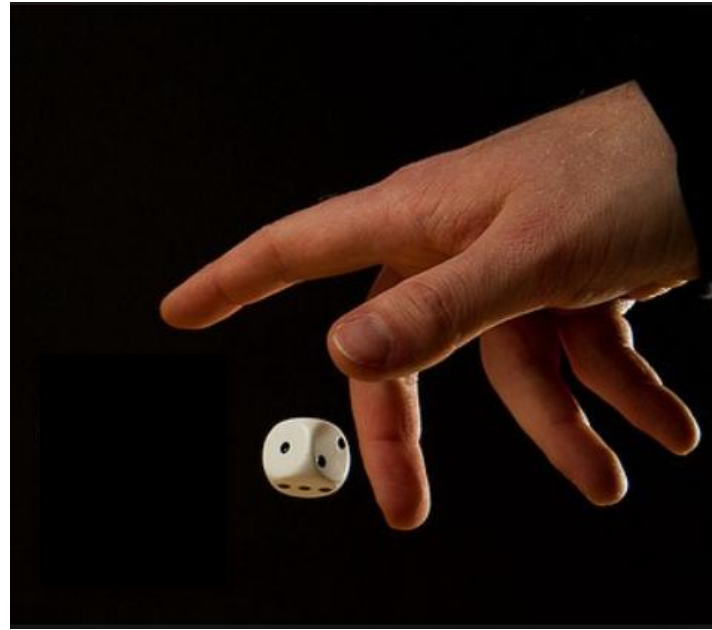
n-k is number of failures

$\binom{n}{k}$ 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

\mathbf{X} and \mathbf{Y} are random variables

\mathbf{N} = total number of trials

n_{ij} = Number of occurrence

		\mathbf{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
\mathbf{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	

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

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

Product rule

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

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

Conditional Independence

- Examples:

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

iff **Virus** is independent of **Drink Beer**

$$P(\text{Flu} | \text{Virus}, \text{DrinkBeer}) = P(\text{Flu} | \text{Virus})$$

iff **Flu** is independent of **Drink Beer**, given **Virus**


$$\begin{aligned} P(\text{Headache} | \text{Flu}, \text{Virus}, \text{DrinkBeer}) \\ = P(\text{Headache} | \text{Flu}, \text{DrinkBeer}) \end{aligned}$$

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

Assume the above independence, we obtain:

$$\begin{aligned} P(\text{Headache}, \text{Flu}, \text{Virus}, \text{DrinkBeer}) \\ = P(\text{Headache} | \text{Flu}, \text{Virus}, \text{DrinkBeer}) P(\text{Flu} | \text{Virus}, \text{DrinkBeer}) \\ P(\text{Virus} | \text{DrinkBeer}) P(\text{DrinkBeer}) \\ = P(\text{Headache} | \text{Flu}, \text{DrinkBeer}) P(\text{Flu} | \text{Virus}) P(\text{Virus}) P(\text{DrinkBeer}) \end{aligned}$$

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"
 - $P(Headache|Flu)$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?
- Definition:

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

Corollary:

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

This is called **Bayes Rule**


Bayes' Rule

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

Other cases:

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

Outline

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

Mean and Variance

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

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

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

Mean and average

Variance and average:


Covariance:

Correlation:

For Joint Distributions

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

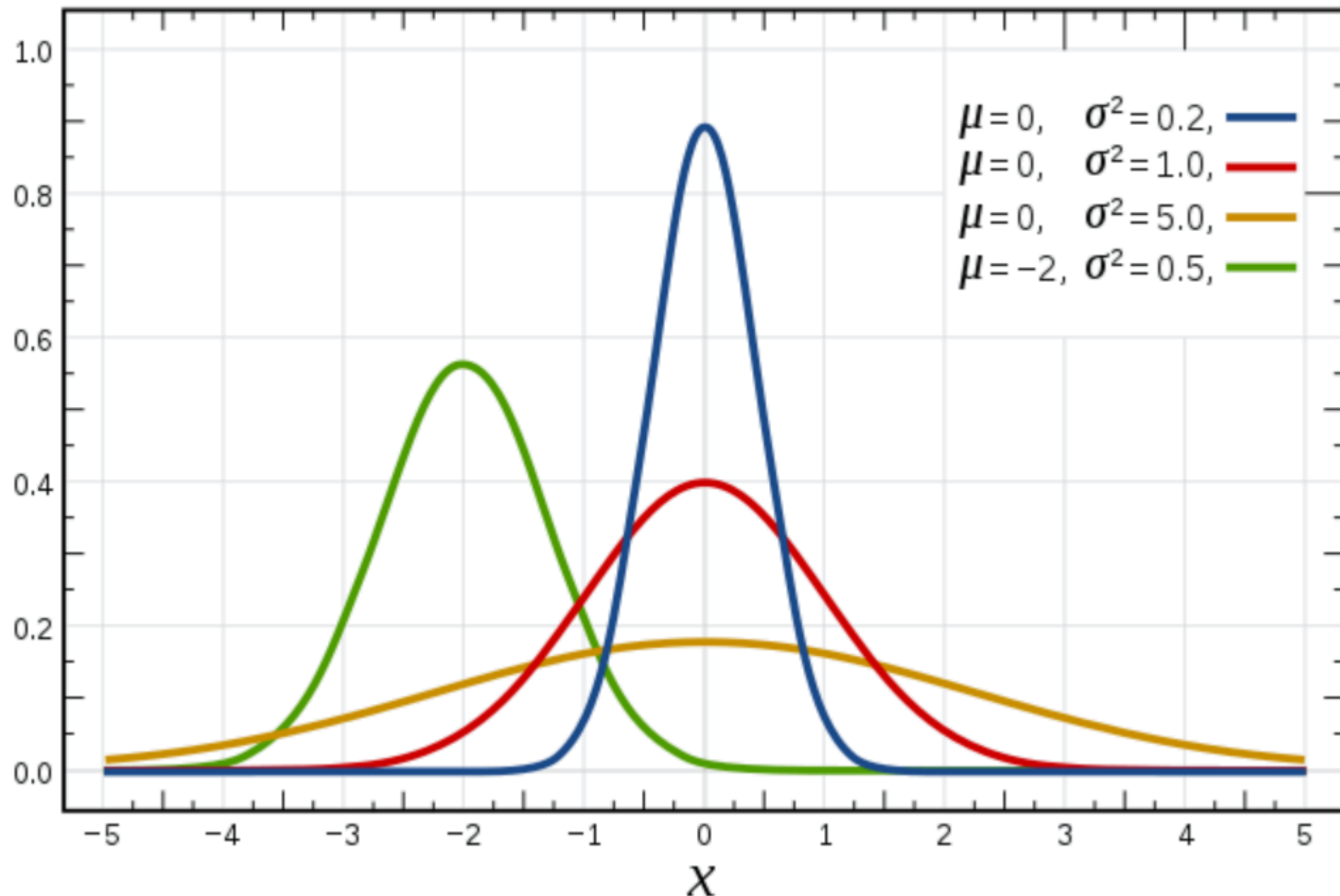
Outline

- 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



Probability versus likelihood

Prob vs Likelihood

Prob vs Likelihood

Multivariate Gaussian Distribution

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

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

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

- 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

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

$$\text{Cov}(AX + b) = A\text{Cov}(X)A^T$$

this means that for Gaussian distributed quantities:

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

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

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

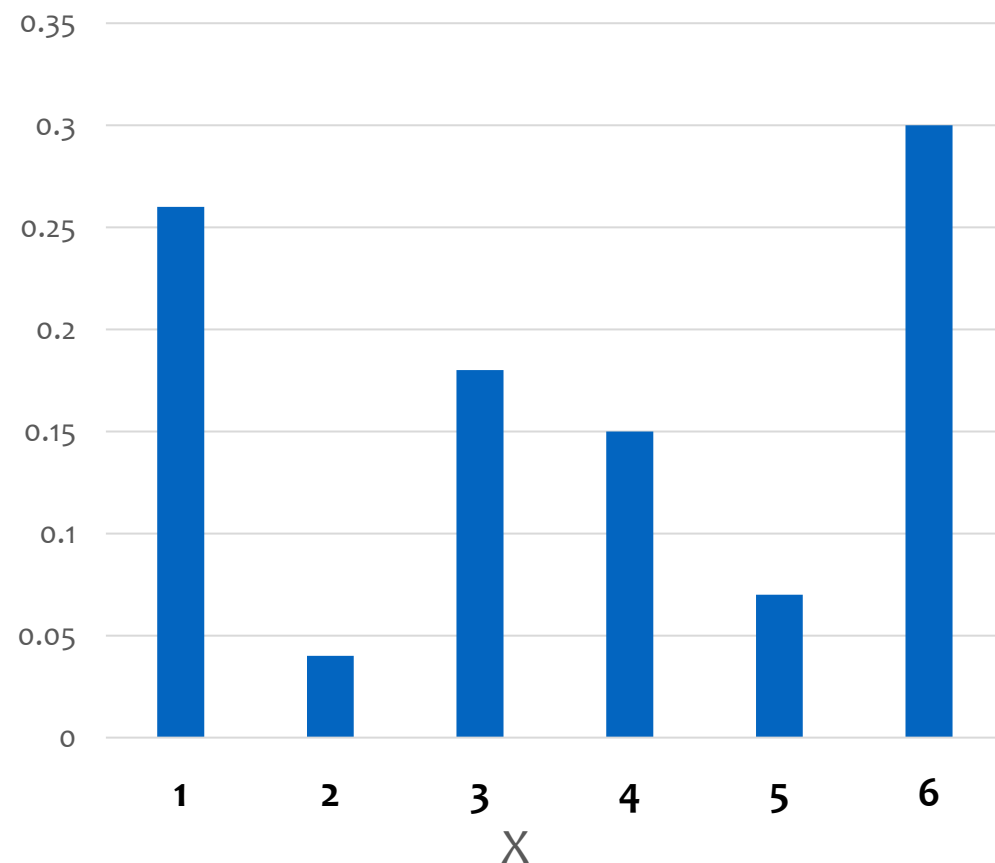
- The **multiplication** of two Gaussian functions is another Gaussian function (although no longer normalized)

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

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

Central Limit Theorem

Probability mass function of a **biased** dice



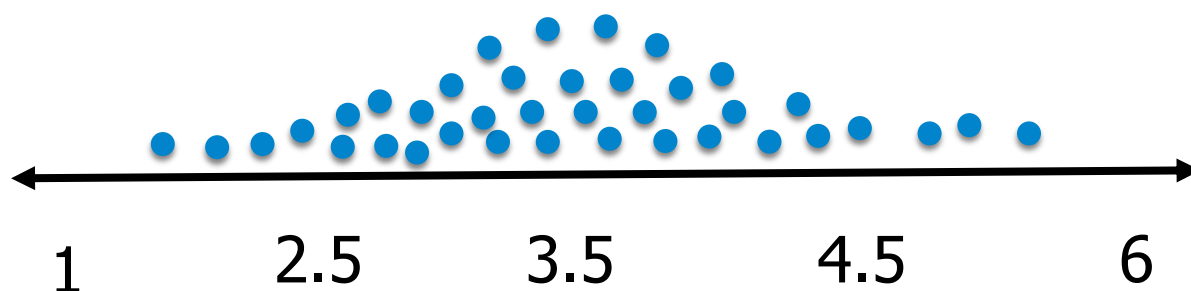
Let's say, I am going to get a sample from this pmf having a size of **$n = 4$**

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

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


\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)

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

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

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

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

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

What's the trick?

Take the log

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

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

How to optimize θ ?

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

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