Lecture 03
Probability and Statistics

Chao Zhang
Georgia Tech

These slides are based on slides from Le Song and Sam Roweis.

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

- $\boxed{\text{Random variables } X}$ represents outcomes or states of world.
  Instantiations of variables usually in lower case: x
  We will write $p(x)$ to mean probability$(X = x)$.
- $\boxed{\text{Sample Space}}$: the space of all possible outcomes/states.
  (May be discrete or continuous or mixed.)
- $\boxed{\text{Probability mass (density) function}}$ $p(x) \geq 0$
  Assigns a non-negative number to each point in sample space.
  Sums (integrates) to unity: $\sum_x p(x) = 1$ or $\int_x p(x) dx = 1$.
  Intuitively: how often does x occur, how much do we believe in x.
- Ensemble: random variable $+$ sample space$+$ probability function

Discrete Probability Functions

- A probability distribution P defined on a discrete sample space S is an assignment of a non-negative real number P(s) to each sample s∈S :
  - Probability Mass Function (PMF): $p_x(x_i) = \mathrm{P}[X = x_i]$
  - Properties: $p_x(x_i) \geq 0$ and $\sum_i p_X(x_i) = 1$

- Examples:
  - Bernoulli Distribution:
    - $\begin{cases} 1-p & for\ x = 0 \\ p & for\ x = 1 \end{cases}$
  - Binomial Distribution:
    - P$(X = k)$=$\binom{n}{k} p^k (1-p)^{n-k}$

Continuous Probability Functions

- A continuous random variable X is defined on a continuous sample space: an interval on the real line, a region in a high dimensional space, etc.
  - It is meaningless to talk about the probability of the random variable assuming a particular value --- P(x) = 0
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval, or arbitrary Boolean combination of basic propositions.
  - Cumulative Distribution Function (CDF): $F_x(x) = \mathrm{P}[X \leq x]$
  - Probability Density Function (PDF): $F_x(x) = \int_{-\infty}^{x} f_x(x)\,dx$ or $f_x(x) = \frac{d\,F_x(x)}{dx}$
  - Properties: $f_x(x) \geq 0$ and $\int_{-\infty}^{\infty} f_x(x)\,dx = 1$
  - Interpretation: $f_x(x) = \mathrm{P}[X \in \frac{x,x+\Delta}{\Delta}]$

Continuous Probability Functions

- Examples:
  - Uniform Density Function:

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

  - Exponential Density Function:

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

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

  - Gaussian(Normal) Density Function

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

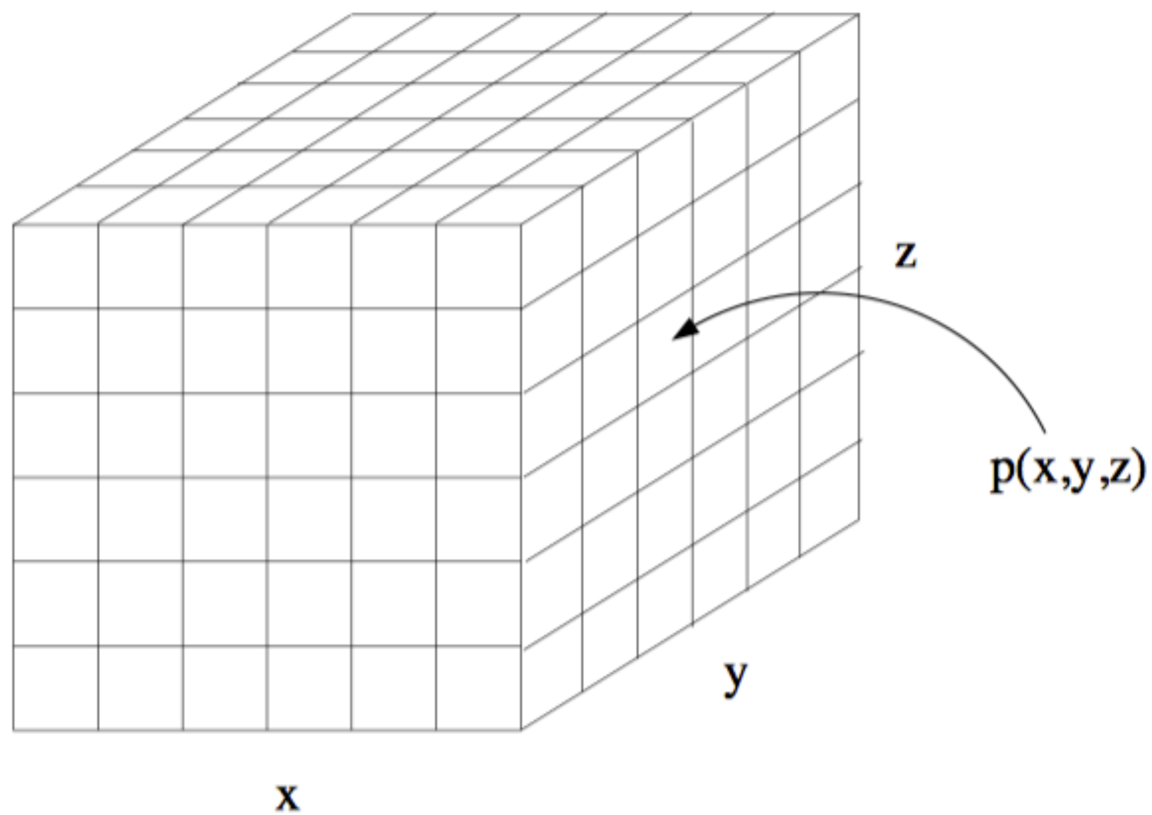
Outline

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

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

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

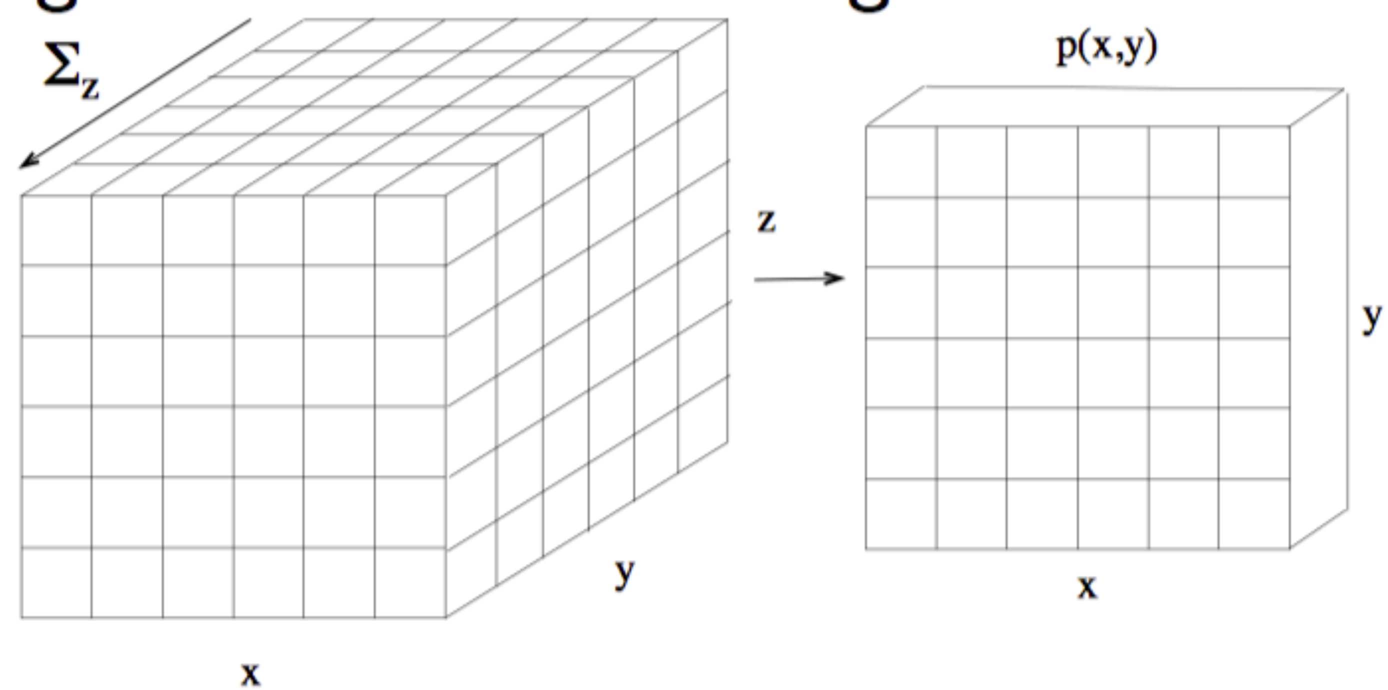


Marginal Distribution

- 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.

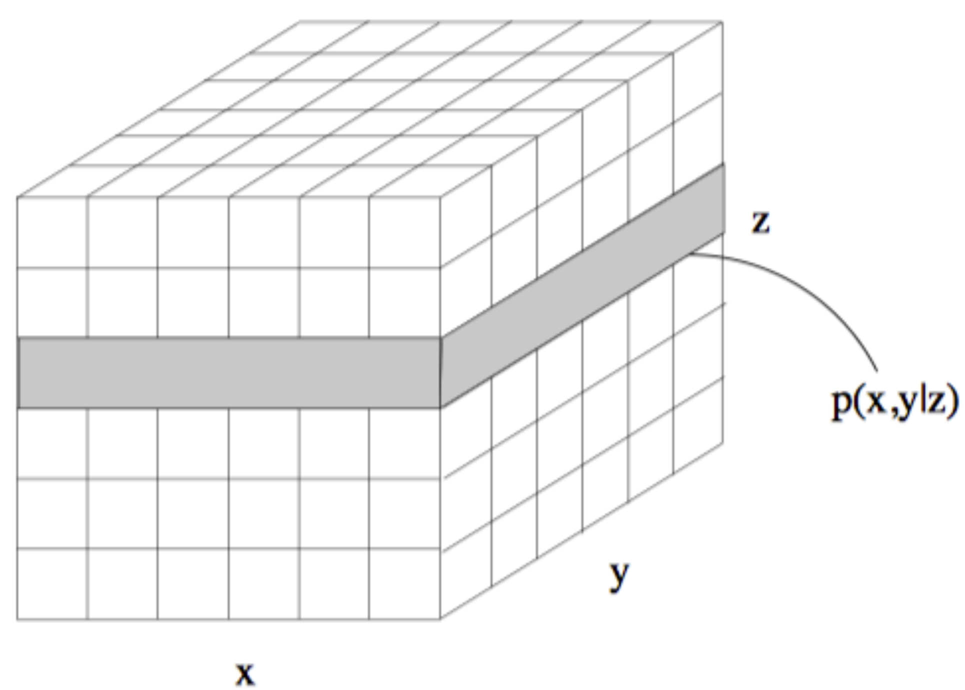


- Another equivalent definition: $p(x) = \sum_y p(x|y)p(y)$.

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.

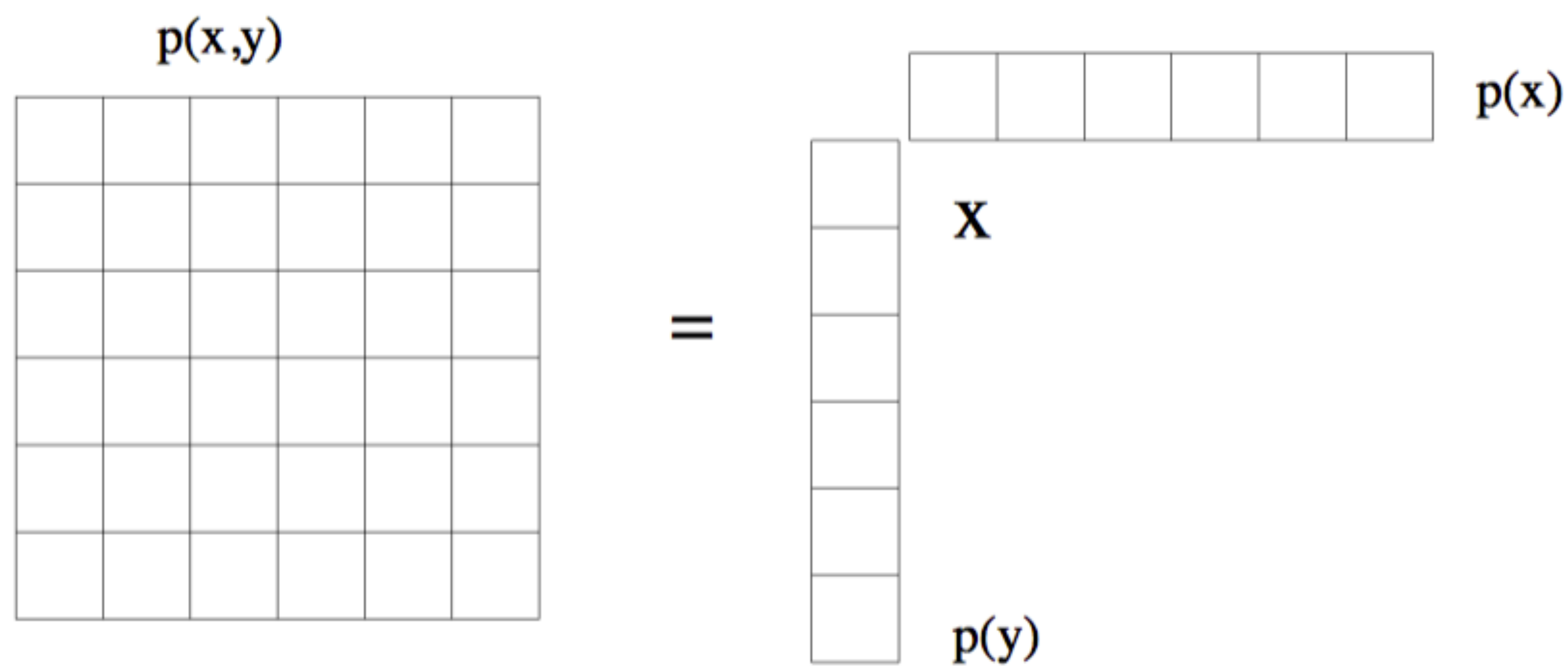
$$\boxed{p(x|y) = p(x,y)/p(y)}$$



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:

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

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)
=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.

- 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:

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

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

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

For Joint Distributions

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

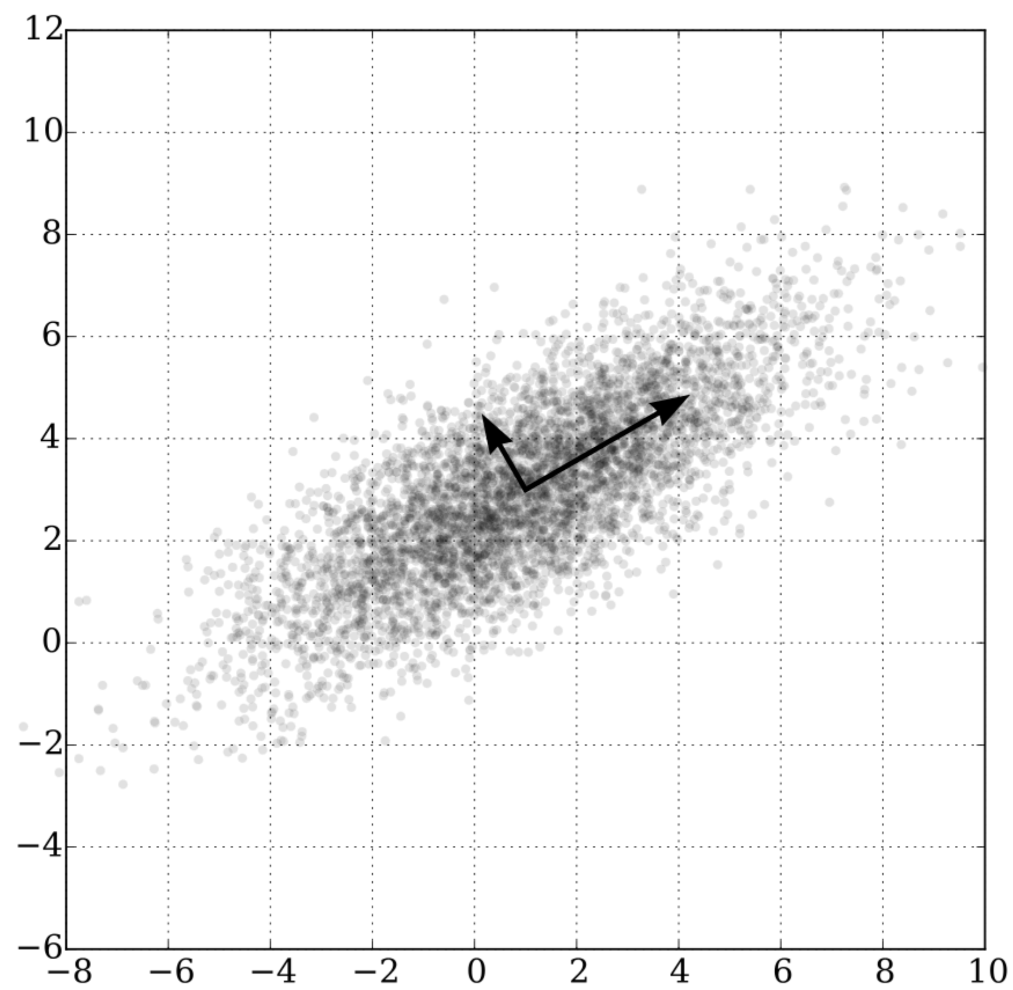


Image source: Wikipedia.

Outline

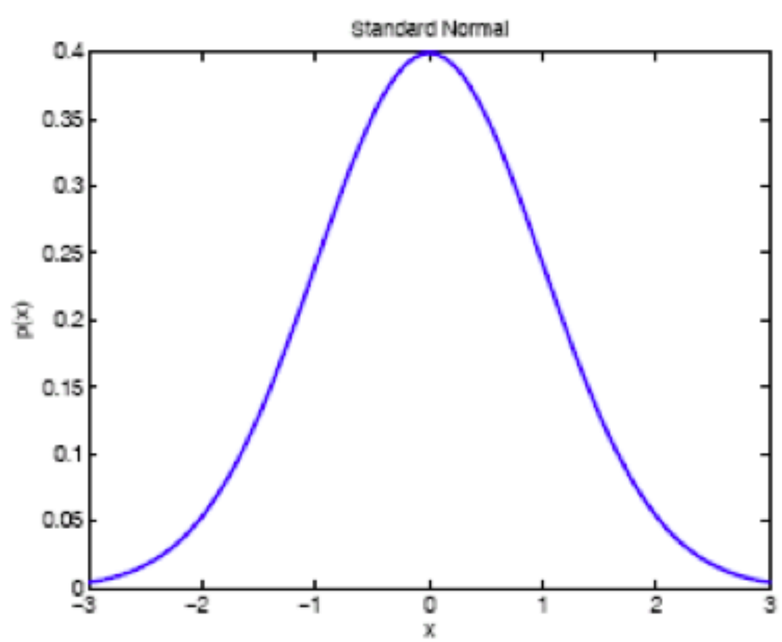
- Probability Distributions
- Joint and Conditional Probability Distributions
- Bayes' Rule
- Mean and Variance
- Properties of Gaussian Distribution
- Maximum Likelihood Estimation

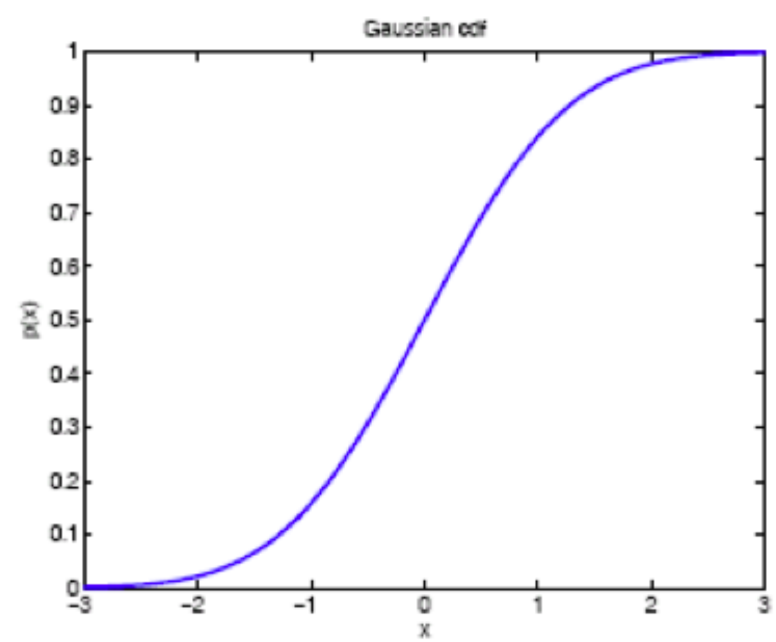
Gaussian Distribution

- Gaussian Distribution:
  - If Z~N(0,1)

$$F_x(x) = \Phi(x) = \int_{-\infty}^{x} f_x(z)dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{\frac{-z^2}{2}}dz$$

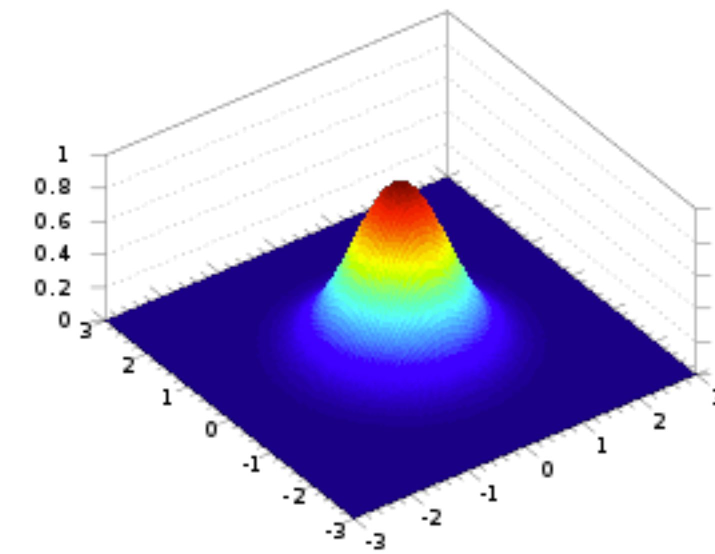
  - This has no closed form expression, but is built in to most software packages (eg. normcdf in matlab stats toolbox).





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,…)

Image source: Wikipedia.

Multivariate Gaussian Distribution

- Joint Gaussian $P(X_1, X_2)$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

- Marginal Gaussian

$$\mu_2^m = \mu_2 \qquad \Sigma_2^m = \Sigma_2$$

- Conditional Gaussian $P(X_1 | X_2 = x_2)$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

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)

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

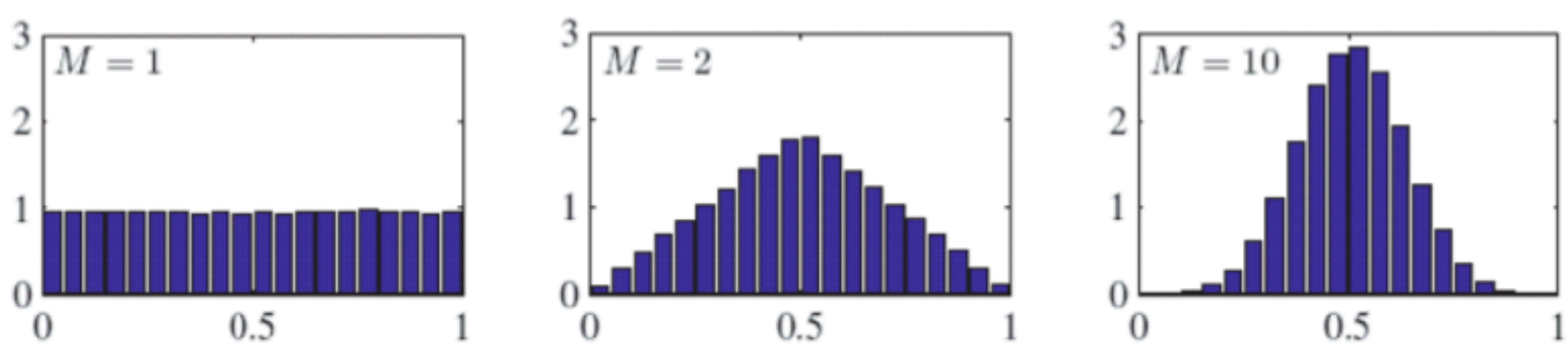
Central Limit Theorem

- If $(X_1, X_2, \ldots X_n)$ are i.i.d. continuous random variables, then the joint distribution is $f(\bar{X})$

- CLT proves that $f(\bar{X})$ is Gaussian with mean $E[X_i]$ and $Var[X_i]$

$$\bar{X} = f(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad as\ n \to \infty$$

- Somewhat of a justification for assuming Gaussian noise



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).

Maximum Likelihood Estimation

- Example: toss a coin
- Objective function:

$$l(\theta; Head) = \log P(Head|\theta) = \log \theta^n (1-\theta)^{N-n} = n \log \theta + (N-n) \log(1-\theta)$$

- We need to maximize this w.r.t. θ
- Take derivatives w.r.t. θ

$$\frac{dl}{d\theta} = \frac{n}{\theta} - \frac{N-n}{1-\theta} = 0$$

$$\longrightarrow \qquad \hat{\theta}_{MLE} = \frac{n}{N}$$