ECE8813
Statistical Language Processing

Lecture 2: Probability Theory Foundations

Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
chl@ece.gatech.edu

Course Information

- Subject: Statistical Language Processing
- Prerequisite: ECE3075, ECE4270
- Background Expected
  - Basic Mathematics and Physics
  - Digital Signal Processing
  - Basic Discrete Math, Probability Theory and Linear Algebra
- Tools Expected:
  - MATLAB and other Programming Tools
  - Language-specific tools will be discussed in Class
- Teaching Philosophy
  - Textbooks and reading assignments: your main source of learning
  - Class Lectures: exploring beyond the textbooks
  - Homework: hand-on and get-your-hands-dirty exercises
  - Class Project: a good way to go deeper into a particular topic
- **Website:** http://users.ece.gatech.edu/~chl/ECE8813.sp09

Probability Tools: An Overview

- Why probabilistic approach?
  - probabilistic vs. deterministic description of *events*
  - model-based vs. rule-based *inference* (scores)
  - natural way to summarize a large collection of data with a small set of parameters (*corpus-based*)
  - taking advantage of existing theory and methods
  - moving from subjective to *objective evaluation*
  - moving from theory to *computation* and *realization*
- Historic Perspective
  - speech science vs. statistical approach
  - new trend in computational linguistics
  - combining rules and models: a win-win story

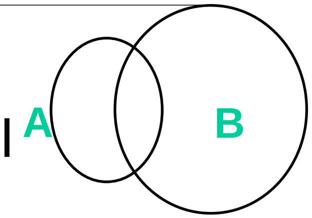
Some Definitions

- Sample Space: Ω
  - collection of all possible observed outcomes
- Sample Event: $A \in \Omega$ including null event ϕ
- σ-field: set of all possible events $A \in F_\Omega$
- Probability Function (Measurable) $P : F_\Omega \to [0,1]$
- Three Axioms:
  - $P(\phi) = 0 \quad P(\Omega) = 1$
  - If $A \subseteq B$ then $P(A) \leq P(B)$
  - If $A \cap B = \phi$ then $P(A \cup B) = P(A) + P(B)$

Some Examples

- Sample Space:
  - Ω_c = {x: x is the height of a person on earth}
  - Ω_d = {(y, z): y is the age and z is resident city}
- Sample Event:
  - A={x: x>200cm}
  - B={x: 120cm<x<130cm}
  - C={(teens; Shenzhen or Hong Kong)}
  - D={(over 70; Japan)}
- σ-field: set of all possible events F_Ω
- Probability Function (Measurable) $P: F_\Omega \to [0,1]$
  - measuring A, B, C and D; computing P(A), P(B), P(C) and P(D); inference about A, B, C, and D

Conditional Events



- *Prior Probability*
  - probability of an event before considering any additional knowledge or observing any samples: P(A)
- *Conditional Probability* $P(A|B) = P(A \cap B) / P(B)$
  - updated probability of an event given some knowledge about another event: P(A|B)
- Prove the *Addition Rule*: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- From *Multiplication Rule*, Show *Chain Rule:*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2 | A_1) \cdots P(A_n | \bigcap_{i=1}^{n-1} A_i)$$

- *Approximating Language Probabilities:*

$$P(W) = P(w_1)P(w_2 | w_1) \cdots P(w_{|W|} | w_1, \ldots, w_{|W|-1})$$

$$\approx P(w_1)P(w_2 | w_1) \prod_{i=3}^{|W|} P(w_i | w_{i-1}, w_{i-2}) \quad n - gram$$

Bayes' Theorem

- Swapping dependency between events
  - calculate P(B|A) in terms of P(A|B) that is available and more relevant in some cases

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

- In many cases, not important to compute P(A)

$$\arg\max_B \frac{P(A|B)P(B)}{P(A)} = \arg\max_B P(A|B)P(B)$$

- Another Form of Bayes' Theorem (try n=2)
  - If a set B partitions A, i.e. $A = \bigcup_{i=1}^{n} B_i \quad B_i \cap B_k = \phi$

$$P(B_j | A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Random Variable (Vector)

- A function that maps sample space to a n-dimensional space of real numbers for easy manipulation (sample space can be irregular)
  - linking events to numerical values $X : \Omega \rightarrow \Re^n$
- Discrete Random Variable
  - mapping events to a subset of integer numbers, e.g. *Bernoulli trial*: 0 for success and 1 for failure (*binomial* distribution)
- Probability Mass Function (pmf)

$$p(x) = p(X = x) = P(A_x) \quad \text{with} \quad A_x = \{\omega \in \Omega : X(\omega) = x\}$$

$$\sum_i p(x_i) = \sum_i P(A_{x_i}) = P(\Omega) = 1$$

- Exercise: define a random variable as the product of the dots on two dices, define the outcome space of the r.v. and derive the pmf

Continuous Random Variable (Cont.)

- Mapping events to real numbers
- Probability Density Function (pdf)

$$P(a \le x \le b) = \int_a^b f(x)dx \quad P(\Omega) = \int_{-\infty}^{\infty} f(x)dx = 1$$

- Probability Distribution Function

$$F(y) = \int_{-\infty}^{y} f(x)dx \quad F(-\infty) = 0 \quad F(\infty) = 1$$

- Expectation of Random Functions

$$E(q(X)) = \int_{-\infty}^{\infty} q(x) f(x)dx \quad \text{or} \quad \sum_i q(x_i) p(x_i)$$

- Mean and Variance

$$\mathrm{Mean}(X) = \mathrm{E}(X) \quad \mathrm{Var}(X) = \mathrm{E}([X - \mathrm{E}(X)]^2)$$

- Covariance between two r.v.'s

$$\mathrm{Cov}(X, Y) = \mathrm{E}([X - \mathrm{E}(X)][Y - \mathrm{E}(Y)])$$

Joint and Conditional Distribution

- Joint Event and Product Space $\Omega_c \times \Omega_d$
  - e.g. E=(A,B)=(200cm<height, live in Pakistan)
- Joint pmf and pdf of two random variables

$$p(x,y)=P(X=x,Y=y) \quad P(a\le x\le b, c\le y\le d)=\int_a^b\int_c^d f(x,y)dydx$$

- Marginal pmf and pdf $p(x)=\sum_y p(x,y) \quad f(x)=\int f(x,y)dy$
- Conditional pmf and pdf

$$p(x|y)=p(x,y)/p(y) \quad f(x|y)=f(x,y)/f(y)$$

- Conditional Expectation

$$E(q(X)|Y=y)=\int_{-\infty}^{\infty} q(x)f(x|y)dx \text{ or } \sum_i q(x_i)p(x_i|y)$$

- Conditional Mean: $\mathrm{Mean}(X|Y=y)=\int xf(x|y)dx$
- Independence: $f(x,y)=f(x)f(y) \quad f(x|y)=f(x)$

Some Useful Distributions (I)

- *Binomial* Distribution: *B(R=r; n, p)*
  - probability of *r* successes in *n* trials with a success rate *p*

$$B(r;n,p)=\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} \quad \text{where} \quad 0 \le r \le n$$

- *Multinomial* Distribution

$$M(r_1,\ldots,r_m;n,p_1,\ldots,p_m)=\frac{n!}{r_1!\cdots r_m!}\prod_{i=1}^m p_i^{r_i} \quad 0 \le r_i \quad \sum_{i=1}^m r_i = n$$

- Show:

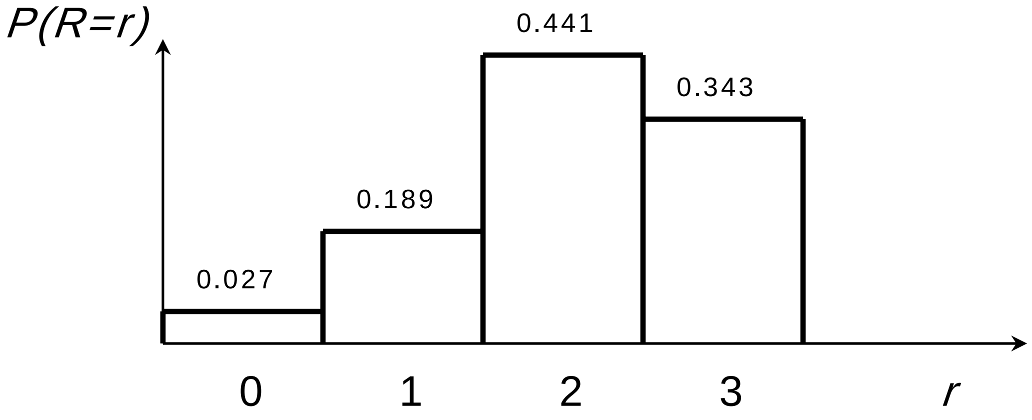
$$\sum_{r=0}^n B(r;n,p)=1 \quad \text{and} \quad \mathrm{E}_B(R)=\sum_{r=0}^n rB(r;n,p)=np$$

- Can you compute Var(*R*) ? Any explanation?

Plot of Probability Mass Function

- Binomial distribution: n=3, p=0.7

$$B(r;n,p)=\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} \quad \text{where} \quad 0\le r\le n$$



Can you plot for other value pairs of (n,p) ?

Some Useful Distributions (II)

- *Uniform* Distribution: *U(X=x; a, b)*

$$U(x;a,b)=\begin{cases} 1/(b-a) & a\le x\le b \\ 0 & \text{otherwise} \end{cases} \quad \text{with } a<b$$

- *Normal* (or *Gaussian*) Distribution: *Bell Curve*

$$N(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2} \quad -\infty<x<\infty \quad \sigma>0$$

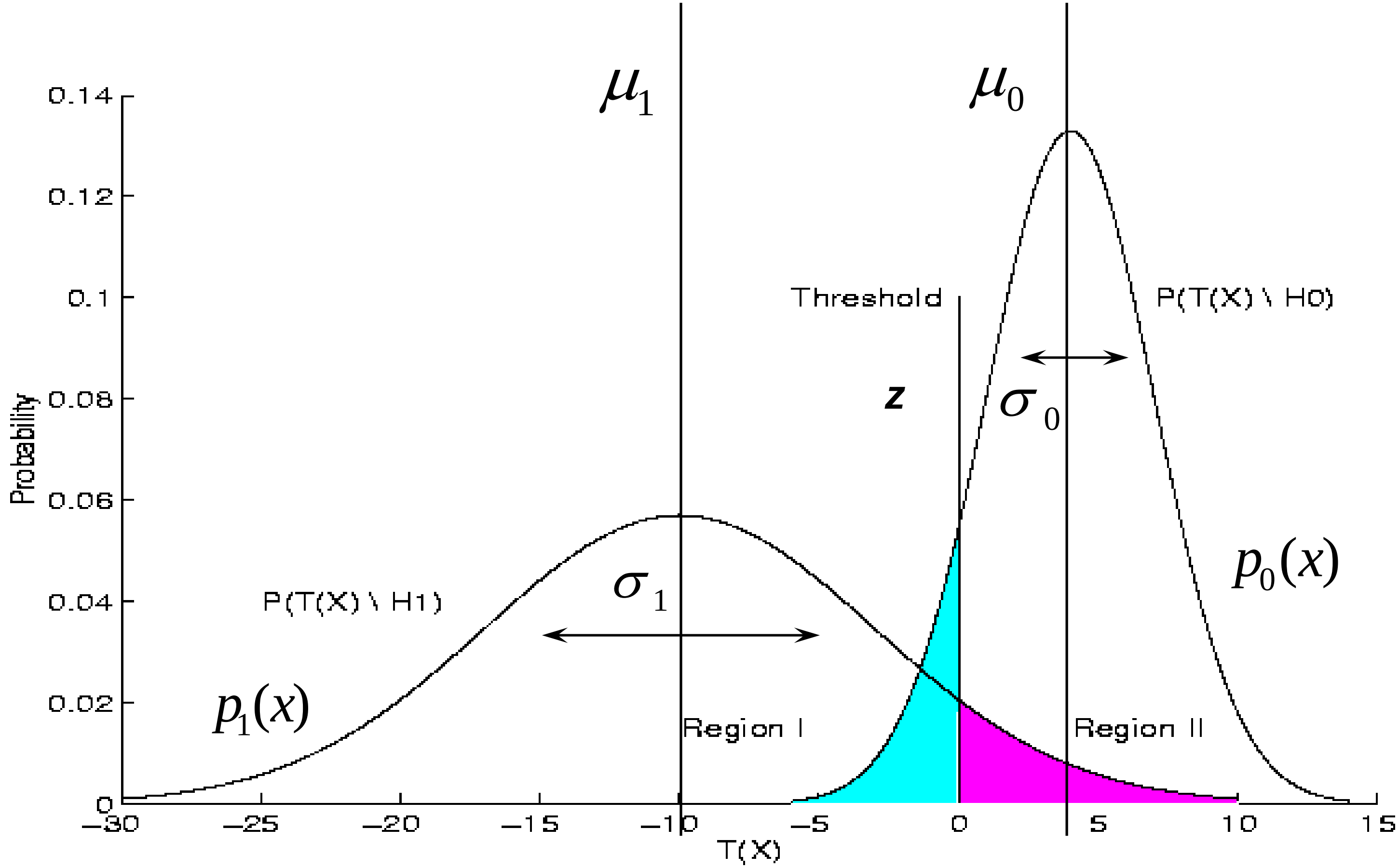
- Show $E_U(X)=\frac{1}{2(b+a)}$ and $E_N(X)=\mu$

- Can you compute their variances? $\mathrm{VAR}_U(X)$ and $\mathrm{VAR}_N(X)$

Center of Signal and Image Processing
Georgia Institute of Technology

CSIP

Typical Normal Distributions



Standard deviation (s.d. or spread): $\sigma_0 < \sigma_1$

Some Useful Distributions (III)

- *2-D Uniform* Distribution:

$$U(x,y;a,b,c,d)=\begin{cases} 1/(b-a)(d-c) & a\le x\le b, c\le y\le d \\ 0 & \text{otherwise} \end{cases} \quad \text{with } a<b, c<d$$

- *Multivariate Normal* Distribution

$$N(\mathbf{x};\boldsymbol{\mu},\mathbf{C})=\frac{1}{\sqrt{(2\pi)^n\,|\mathbf{C}|}}\mathrm{e}^{-(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2} \quad -\infty<\mathbf{x}<\infty$$

- Show $\mathrm{E}_N(\mathbf{X})=\boldsymbol{\mu}$ and $\mathrm{VAR}_N(\mathbf{X})=\mathbf{C}$
- Can you write down the 2-D distribution form, compute Cov(*X*,*Y*), and derive the marginal and conditional densities, *f(y)* and *f(x|y)* ? $\mathbf{x}=\begin{bmatrix} x \\ y \end{bmatrix}$ $\quad \boldsymbol{\mu}=\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ $\quad \mathbf{C}=\begin{bmatrix} \sigma_x^2 & r\sigma_x\sigma_y \\ r\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$

Some Distribution Examples

- Uniform distribution over all directions (radar)

$$U(\theta;-\pi,\pi)=\begin{cases} 1/2\pi & -\pi\le\theta\le\pi \\ 0 & \text{otherwise} \end{cases}$$

- Uniform distribution on a circle (sea surface)

$$U(r,\theta;0,R,-\pi,\pi)=\begin{cases} 1/\pi R^2 & -\pi\le\theta\le\pi, 0\le r\le R \\ 0 & \text{otherwise} \end{cases}$$

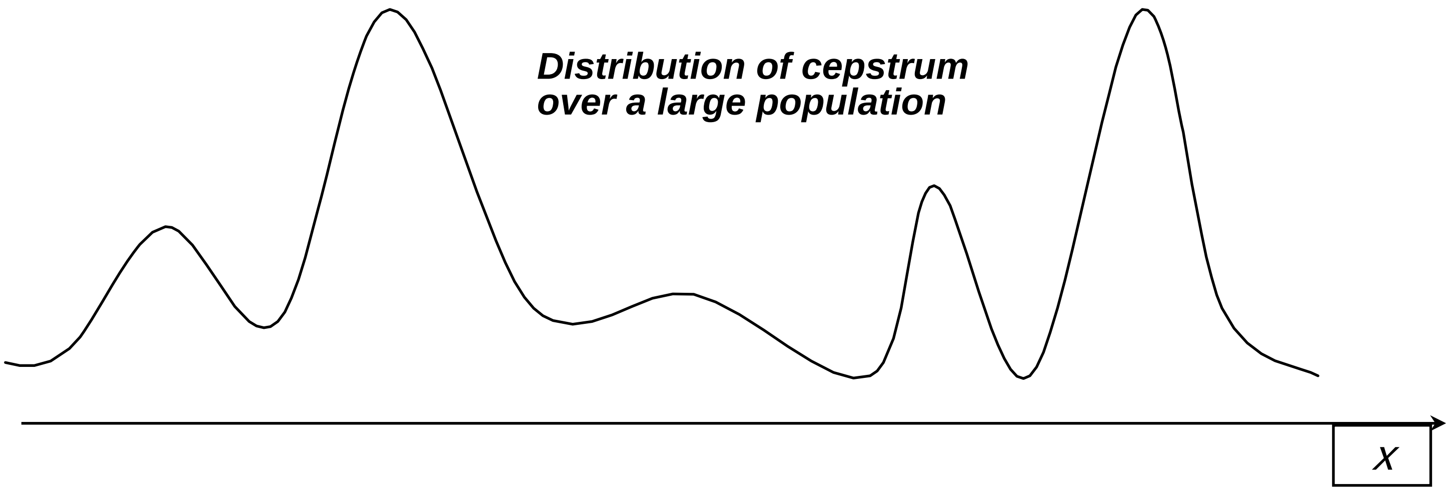
- Mixture Gaussian Distribution

$$MG(x)=\sum_{m=1}^{M}\omega_m N(x;\mu_m,\sigma_m^2) \quad \sum_{m=1}^{M}\omega_m=1 \quad 0<\omega_m<1 \quad \sigma_i>0$$

- Compute: $\mathrm{E}_{\mathrm{MG}}(X)$ and $\mathrm{VAR}_{\mathrm{MG}}(X)$

Properties of Gaussian Mixture

- Mixture Gaussian distribution:



- In theory, *MG(x)* matches any density up to second order statistics (mean and variance)
- Approximating multi-modal densities which is more likely to describe real-world data

Function of Random Variables

- Function of r.v.'s is also a r.v.
  - e.g. $X=U+V+W$, if we know $f(u,v,w)$ how about $f(x)$?
  - e.g. sum of dots on two dices
- Problem easier for known and popular r.v.'s
  - e.g. if U and V are independent Gaussian, so is $X=U+V$
  - e.g. if W and Z are independent uniform, is $Y=W+Z$ uniform?
- Show sample mean of n independent samples of Gaussian r.v.'s is also Gaussian, show that:

$$\mathrm{E}(\bar{X}) = \mu \qquad \mathrm{Var}(\bar{X}) = \sigma^2 / n$$

Distributions of Random Variables

- Parametric distributions
  - r.v. described by a small number of parameters in pdf/pmf
  - e.g. Gaussian (2), Binomial (3), 2-d uniform (3 or 4)
  - many useful and known parametric distributions
  - distributions of independently and identically distributed (i.i.d.) samples from such distributions are easier to derive
- Non-Parametric distributions
  - usually described by the data samples themselves
- Sample distribution & histogram (pmf / bar chart)
  - counting samples in equally-sized bins and plot them

Sum of Many Random Variables

- Show average of two independent samples of uniform r.v.'s form a triangular shape pdf. How about sample mean of *n* samples? Can you imagine what it will be like for very large *n* ?

- *Law of large numbers* – Asymptotic Normal pdf !!

Parametric Distributions

- Parametric Distribution
  - r.v. described by a small number of parameters in pdf/pmf
  - e.g. Gaussian (2), Binomial (3), 2-d uniform (3 or 4)
  - many useful and known parametric distributions
  - distributions of independently and identically distributed (i.i.d.) samples from such distributions are easier to derive
- Non-Parametric Distribution
  - usually described by the data samples themselves
- Sample distribution & histogram (pmf / bar chart)
  - counting samples in equally-sized bins and plot them

Statistics and Probability

- *Statistic*: a function of random samples
  - E.g. sample mean and variance
- *Sufficient Statistics*
  - A minimum set of summary statistics to describe the samples without losing any information, e.g. sample mean, variance, and size for Gaussian samples
  - For some r.v.'s, the sufficient statistics can only be described by the entire set of data samples
  - Distributions of sufficient statistics are often reproducible which are keys to Bayesian estimation with conjugate prior and posterior pairs

Some Useful Descriptive Statistics

- Sampling theory and descriptive statistics
  - From partial observations to overall assessment
- Sample size, mean, variance (margin of error)
- Range, maximum, minimum
- Median, percentile, upper and lower quartiles
- Descriptive statistics are often seen in many articles and reports in our daily lives. Do you know how to evaluate them and judge their validity when certain conclusions are drawn?

The Art and Science of Sampling

- A few examples
    1. Randomly selecting *n* out of *M* vendors in Atlanta for evaluation to award a construction job
    2. Randomly polling *Q* households for TV rating
    3. Randomly selecting parts for error measurement
    4. Opinion polls: done a lot in election seasons
    5. Sending pilot signals to probe a wireless connection
- Questions
    – How many to sample? What's the population like?
    – What can be said about the sampling results?
    – How to use probability theory to help?
    – How to use computer simulation in sampling?

(Empirical) Sample Mean & Variance

- Population: collection of data being studied
  - *N*: Size of the population (typically a large size)
  - (Random) Sample: *n* is the size of the sample set:

$\{x_1, x_2, \ldots, x_n\}$ with x_i's independent samples from the set

- Statistic: function of samples (statistical inference)
  1. Sample Mean (not the mean parameter):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{or} \quad \hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \ \ (X_i \text{ is any r. v. with a pdf } f(x))$$

  2. Sample Variance (r. v., not the variance parameter):

$$\boxed{S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,\ S_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\bar{X}})^2,\ \text{or } \tilde{S}_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\bar{X}})^2}$$

Important Statistics & Expectations (I)

1. Expectation of sample mean:

$$E[\hat{\bar{X}}]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]=\frac{1}{n}\sum_{i=1}^{n}\bar{X}=\bar{X}\ \text{(unbiased statistic of }\bar{X}\text{)}$$

2. Expectation of sample variance (known mean/variance):

$$\mathrm{E}\{S_1^2\}=E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2]=\frac{1}{n}\{\sum_{i=1}^{n}E(X_i)^2-2\sum_{i=1}^{n}E(X_i*\bar{X})+n\bar{X}^2\}$$

$$=\frac{1}{n}\{nE(\bar{X}^2)-n\bar{X}^2\}=\frac{n}{n}[\bar{X}^2-(\bar{X})^2]=\sigma^2$$

Note: $E[X_iX_j]=E[X^2]\ (i=j)$, and $E[X_iX_j]=(E[X])^2=\bar{X}^2\ (i\neq j)$

Important Statistics & Expectations (II)

3. Expectation of sample variance (unknown parameters):

$$\text{Biased statistic: } E\{S_2^2\} = E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\bar{X}})^2] = E\{\frac{1}{n}\sum_{i=1}^{n}[X_i - \frac{1}{n}\sum_{j=1}^{n}X_j]^2\}$$

$$= \frac{1}{n}\{\sum_{i=1}^{n}E[(X_i)^2] - 2\sum_{i=1}^{n}E(X_i * \frac{1}{n}\sum_{j=1}^{n}X_j) + \frac{1}{n^2}\sum_{i=1}^{n}E[(\sum_{j=1}^{n}X_j)(\sum_{k=1}^{n}X_k)]\}$$

$$\frac{1}{n}\{\sum_{i=1}^{n}E[(X_i)^2] - 2\frac{1}{n}\sum_{i=1}^{n}E[(X_i)^2] - \frac{2}{n}E[\sum_{i\neq j}^{n}\sum_{j=1}^{n}X_iX_j] + \frac{1}{n}E[(\sum_{i=1}^{n}X_i)(\sum_{j=1}^{n}X_j)]\}$$

$$= \frac{1}{n}\{nE(X^2) - E(X^2) - (n-1)[E(X)]^2\} = \frac{n-1}{n}\{E[(X-\bar{X})^2]\} = \frac{n-1}{n}\sigma^2$$

4. Unbiased sample variance:

$$E(\tilde{S}_2^2) = \frac{n}{n-1}E(S_2^2) = \sigma^2$$

Other Properties on Statistics

5. Variance of sample variance (unknown parameters):

$$\mathrm{Var}[S_2^2] = E\{[S_2^2 - E(S_2^2)]^2\} = \frac{E[(X-\bar{X})^4] - \sigma^4}{n} = \frac{\mu_4 - \sigma^4}{n} \text{ (your exercise)}$$

- Sample mean and sample variance are correlated random variables useful for statistical inference
  - their joint density can be established (not in ECE3075)
- The same discussion can be extended to multivariate cases (studies have been completed for Gaussian cases)
- Discussion on population size *N* (for your reading)
  - Sampling with or without replacement
- Large sample theory (*n* > 30, depending on individual cases)

Sampling Distributions (I)

- For many applications, it is important to obtain the distribution of a sample statistic. We need to watch for assumptions about the random samples before we work out sample distributions.
  - realize what's known and unknown
- Example 1: Normalized Sample Mean
  - independent Gaussian samples with known variance

$$\boxed{\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ is Gaussian with mean } \bar{X} \text{ and variance } \frac{\sigma^2}{n}}$$

$$\boxed{Z = \frac{\hat{\bar{X}} - \bar{X}}{\sigma/\sqrt{n}} \text{ is Gaussian with mean 0 and variance 1 (standardized r. v.)}}$$

  - note: *Z* can not be defined if we don't know the parameters

Sampling Distributions (II)

- Example 2: Normalized Sample Mean
  - independent Gaussian samples with unknown variance

$$T = \frac{\hat{\bar{X}} - \bar{X}}{\tilde{S}_2 / \sqrt{n}} = \frac{\hat{\bar{X}} - \bar{X}}{S_2 / \sqrt{n-1}} \text{ has a } \mathit{Student's} \text{ t-distribution with n-1 degrees of freedom}$$

- The pdf of *T* (assuming *v=n-1*) is of the form

$$f_T(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(\frac{v}{2})}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}} \quad \text{(Figure 4-2, } v=1,\ \Gamma(v) \text{ is the Gamma function)}$$

  - for large value of *v*, we have an approximate Gaussian

$$\Gamma(v+1) = v\Gamma(v),\ \Gamma(k+1) = k!\ (\text{integer } k),\ \Gamma(2) = \Gamma(1) = 1,\ \Gamma(1/2) = \sqrt{\pi}$$

Center of Signal and Image Processing
Georgia Institute of Technology
CSIP

Correlation between Two Sets of Data

- Linear correlation coefficient (Pearson's *r*)

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad \text{with } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Pearson's *r* approaches Gaussian for large *n*
  - significance of the value of *r*: small *r* is often meaningless unless the sample size *n* is large, and *f(x, y)* is known
  - large *r* implies a tighter coupling between *X* and *Y*

Statistical Inference

- Probability Theory Tools
  - Fuzzy description of phenomena
  - Statistical modeling of data for inference
- Statistical Inference Problems
  - *Classification*: choose one of the stochastic sources
  - *Decision* and *Hypothesis Testing*: comparing two stochastic assumptions and decide on how to accept one of them
  - *Estimation*: given random samples from an assumed distribution, find "good" guess for the parameters
  - *Prediction*: from past samples, predict next set of samples
  - *Regression (Modeling)*: fit a model to a given set of samples
- From theory to many real-world applications

Maximum Likelihood Estimation for Gaussians

- Given iid samples from a normal distribution, what's their joint density (likelihood)?

$$f(x_1,\ldots,x_n)=\prod_{i=1}^{n} f(x_i)=\frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2]$$

- It can been shown that the sample mean has also a normal distribution, can you derive the density?

$$f(\sum_{i=1}^{n} x_i/n)=f(y)=\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp[-\frac{n}{2\sigma^2}(y-\mu)^2]$$

- Suppose the mean needs to be estimated form the iid samples, show the sample mean is the *maximum likelihood* ("best") estimate of μ ?
- ML is the most frequently used estimation method

$$\operatorname{argmax}_\mu f(x_1,\ldots,x_n\,|\,\mu)=\operatorname{argmax}_\mu \log[f(x_1,\ldots,x_n\,|\,\mu)]$$

Maximum Likelihood Estimation of *N*-grams

- Properties of *n*-grams

$$P(w_n \mid w_1,\ldots,w_{n-1}) = \frac{P(w_1,\ldots,w_{n-1},w_n)}{P(w_1,\ldots,w_{n-1})},$$

$$\sum_{w_n \in V} P(w_n \mid w_1,\ldots,w_{n-1}) = 1,$$

$$\sum_i C(e_i) = N_n \quad e_i : i\text{-th event}$$

- *MLE of* Multinomial Distribution Parameters

$$P_{MLE}(w_1,\ldots,w_{n-1},w_n) = \frac{C(w_1,\ldots,w_{n-1},w_n)}{N_n},$$

$$P_{MLE}(w_n \mid w_1,\ldots,w_{n-1}) = \frac{C(w_1,\ldots,w_{n-1},w_n)}{C(w_1,\ldots,w_{n-1})},$$

$$\sum_{W \in V} C(w_1,\ldots,w_{n-1},W) = C(w_1,\ldots,w_{n-1})$$

Hypothesis Testing

- Testing statistical hypotheses
  - Decisions in accepting an assumed distribution from test data
  - What is the level of confidence in accepting right decisions?
  - What is the penalty, if any, for making wrong decisions?
- Formulating statistical tests
  - one-sided test: mean = 1000 vs. mean > 1000
  - two-sided test: mean = 1000 vs. mean > 1000 or <1000
  - Many others (textbooks and handbooks)
- Confidence interval and confidence level in testing
  - Larger level of significance corresponds to a more stringent test
- Confidence measures for assumed theories

Statistical Hypothesis Testing (I)

- In decision, we usuallly need to test a hypothesis based on some observation data. The problem is formulated as a test between two complementary hypotheses:
  - *H0: null hypothesis*
  - *H1: alternative hypothesis*
- Example: Given $X_1, X_2, \ldots, X_n$ as a random sample from a Gaussian distribution $N(\mu, \sigma^2)$, where variance σ^2 is known. We need to verify whether its mean is a given value. Thus we do hypothesis testing:
  - $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$

Statistical Hypothesis Testing (II)

- In essence, a hypothesis test will partition the entire observation space into two disjointed parts, *C* and *D*
- If an observation *X* lies in the region *C*, we reject *H0*; if *X* is in *D*, we accept *H0*. *C* is called the *critical region*
- *Type I error* (also called *false rejection*) of a test:

$$\alpha = \Pr(E_1) = \Pr(X \in C \mid H_0)$$

- *Type II* error (also called *false alarm*) of a test:

$$\beta = \Pr(E_2) = \Pr(X \in D \mid H_1) = 1 - \Pr(X \in C \mid H_1) = 1 - \gamma$$

Statistical Hypothesis Testing (III)

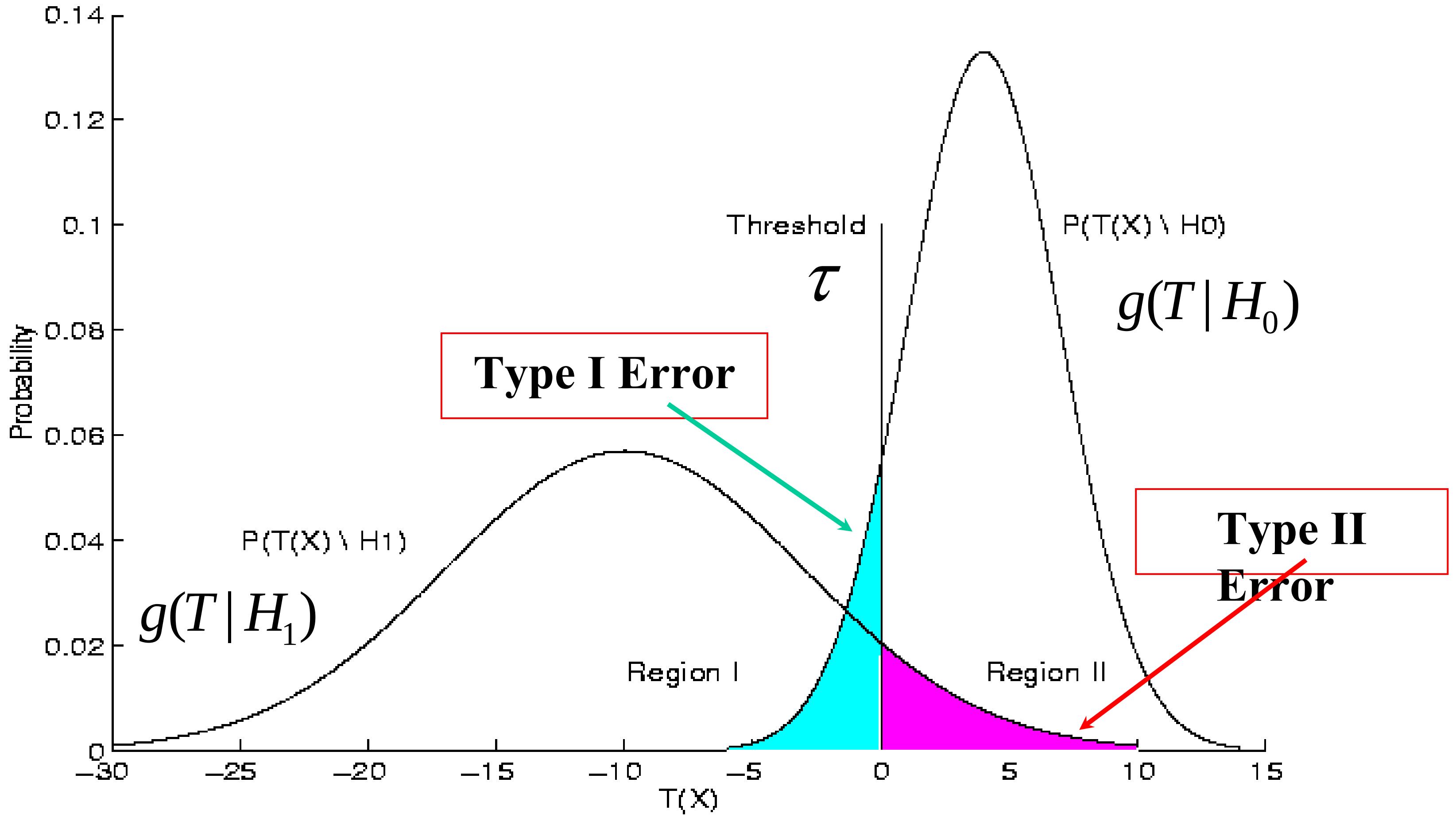
Neyman Pearson Lemma:

For a simple H_0 and simple H_1, if the distributions under both H_0 and H_1 are known, i.e., $f_0(X|\theta_0)$ and $f_1(X|\theta_1)$. Given any iid observation data $X=\{X_1,\ldots,X_T\}$, at any significance level α, the most powerful test is formulated as:

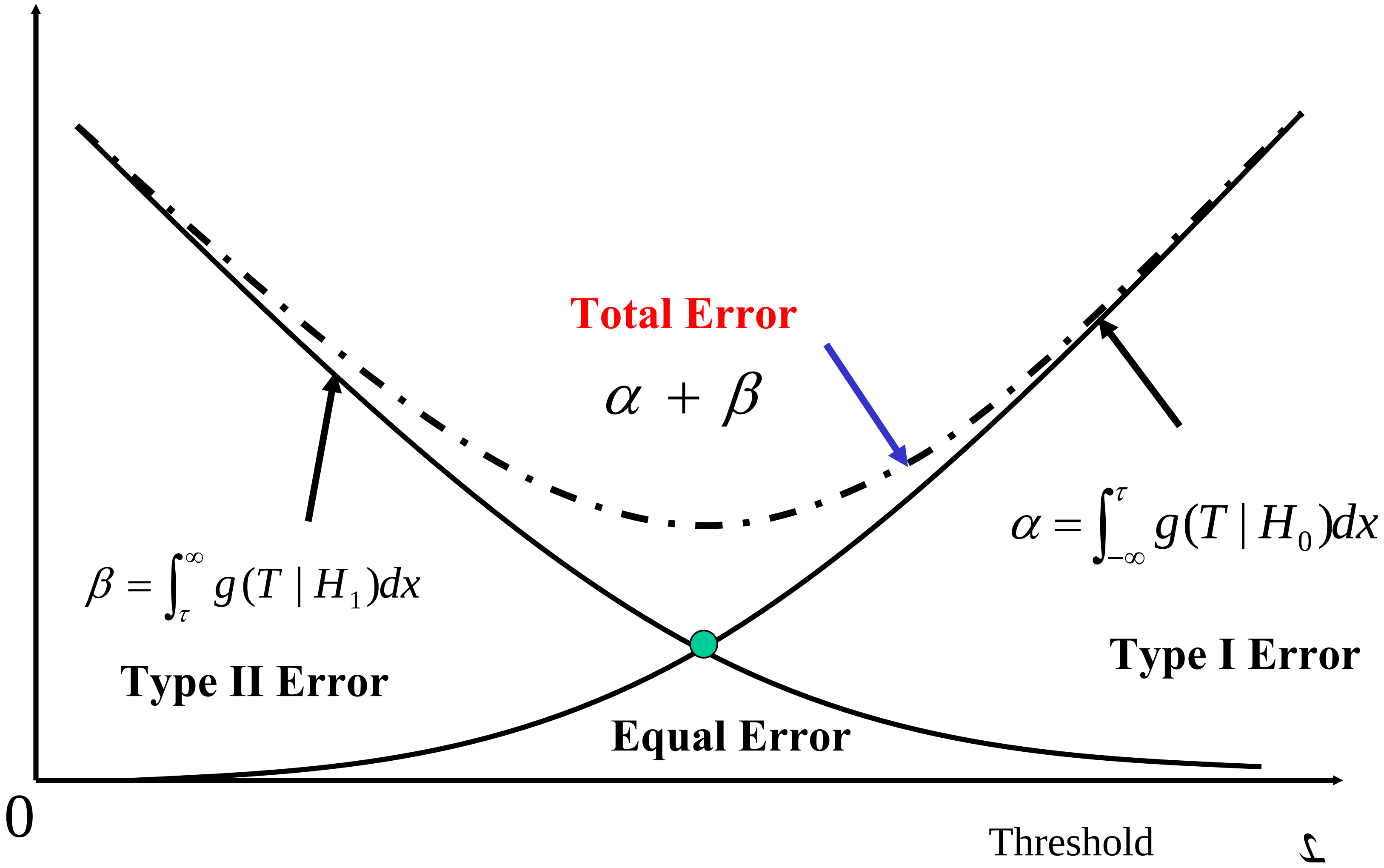
If $LR(X_1^T) = \dfrac{\prod_{t=1}^{T} f_0(X_t \mid \theta_0)}{\prod_{t=1}^{T} f_1(X_t \mid \theta_1)} > \tau$, accept H_0; otherwise reject H_0.

The threshold τ is adjusted to make the significance of the test to be α.
If the both pdf's have the same form, the only difference is parameters,
The ratio is also called likelihood ratio (LR).

Distributions of Test Statistic *T*

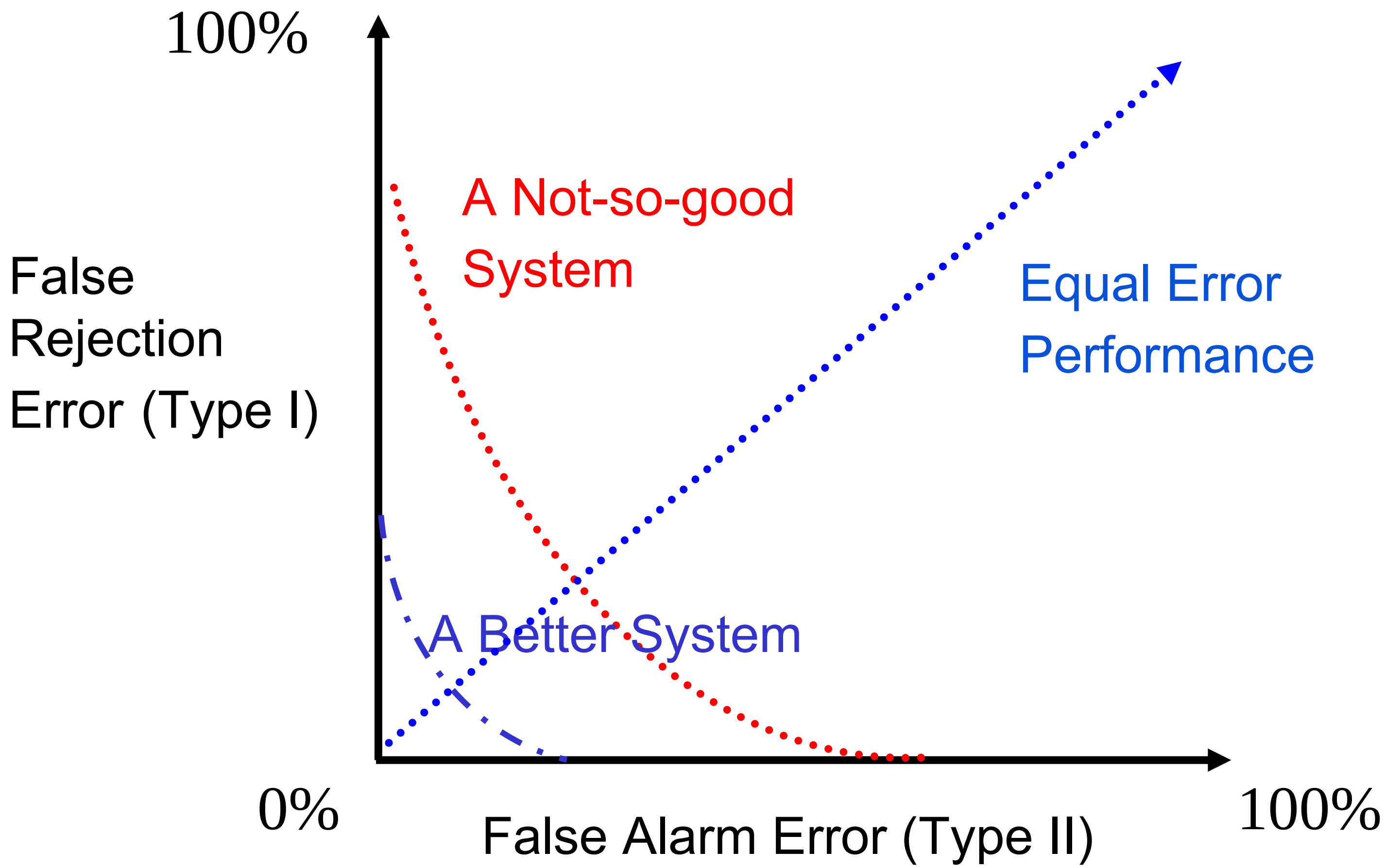


Evaluating Verification (I)



Center of Signal and Image Processing
Georgia Institute of Technology

Evaluating Verification (II): ROC (Receiver Operating characteristic) Curve



Maximum Likelihood Estimation

- Given iid samples from a normal distribution, what's their joint density (likelihood)?

$$f(x_1,\ldots,x_n)=\prod_{i=1}^{n} f(x_i)=\frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2]$$

- It can been shown that the sample mean has also a normal distribution, can you derive the density?

$$f(\sum_{i=1}^{n} x_i/n)=f(y)=\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp[-\frac{n}{2\sigma^2}(y-\mu)^2]$$

- Suppose the mean needs to be estimated form the iid samples, show the sample mean is the *maximum likelihood* ("best") estimate of μ ?

- ML is the most frequently used estimation method

$$\text{argmax}_\mu\, f(x_1,\ldots,x_n\,|\,\mu)=\text{argmax}_\mu \log[f(x_1,\ldots,x_n\,|\,\mu)]$$

Probability Theory Recap

- Probability Theory Tools
  - fuzzy description of phenomena
  - statistical modeling of data for inference
- Statistical Inference Problems
  - *Classification*: choose one of the stochastic sources
  - *Decision* and *Hypothesis Testing*: comparing two stochastic assumptions and decide on how to accept one of them
  - *Estimation*: given random samples from an assumed distribution, find "good" guess for the parameters
  - *Prediction*: from past samples, predict next set of samples
  - *Regression* (*Modeling*): fit a model to a given set of samples

Probability Theory Recap (Cont.)

- Parametric vs. Non-parametric Distributions
  - parsimonious or extensive description (model vs. data)
  - Sampling, data storage and sufficient statistics
- Real-World Data vs. Ideal Distributions
  - "there is no perfect goodness-of-fit"
  - ideal distributions are used for approximation
  - sum of random variables and Law of Large Numbers

Summary

- Today's Class
  - Probability Theory
  - Web: http://www.ece.gatech.edu/~chl/ECE8813.sp09
  - Class web page and data will be ready soon
- Next Class
  - Information Theory on Jan. 13
  - Reading Assignments
    - Manning and Schutze, Chapters 1 & 2