ECE3075 - Random Signals

Chapter 4: Elements of Statistics

Chin-Hui Lee

School of Electrical and Computer Engineering

Georgia Institute of Technology

Atlanta, GA 30332, USA

chl@ece.gatech.edu

Introduction to Statistics

- Statistics: The science of assembling, classifying, tabulating, and analyzing data or facts
  - Descriptive statistics: the science of grouping, and presenting data to be easily understood or assimilated
  - Inductive statistics or statistical inference: uses of data to draw conclusion about, or estimate parameters of, the environment from which the data came
- Branches of Statistics (studied in most universities)
  1. Sampling Theory
  2. Estimation Theory
  3. Hypothesis Testing
  4. Curve Fitting and Regression
  5. Analysis of Variance and Experimental Design

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 (for statistical inference)
  1. Sample Mean (not the mean parameter):

     $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ or $\hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (X_i is any r. v. with a pdf $f(x)$)
  2. Sample Variance (a 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 the 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 the 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: $E[X_i X_j] = E[X^2]\ (i = j)$, and $E[X_i X_j] = (E[X])^2 = \bar{X}^2\ \ (i \neq j)$

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):

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

- 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 [Eq. (4-5) vs. Eq. (4-4)]
- Large Sample Theory ($n > 30$, depending on individual cases)
- Textbook Illustrations: Exercises 4-2.1, 4-2.2 and 4-3.1, 4-3.2

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

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

Confidence Intervals

- Sample Mean : a point estimate related to sample size
  - How about an interval estimate? How to choose *n*?
- *q*-percent confidence interval: *e.g. quartile, median*
  - Example: sample mean for Gaussian samples, known variance
  - For the sample mean: $\boxed{[\bar{X}-k\sigma/\sqrt{n}, \bar{X}+k\sigma/\sqrt{n}]}$

$$\boxed{P(\bar{X} - k\sigma / \sqrt{n} < \hat{\bar{X}} < \bar{X} + k\sigma / \sqrt{n}) = q / 100}$$

- Confidence interval for other statistics can also be established if the distribution of the point estimate of interest can be evaluated (e.g. *t*-distribution).
- Illustrations: Tables 4-1, 4-2, and Exercises 4-4.1, 4-4.2

Hypothesis Testing

- Testing Statistical Hypothesis
  - 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 a statistical test
  - one-sided test: mean = 1000 vs. mean > 1000
  - two-sided test: mean = 1000 vs. mean > 1000 or <1000
- Confidence interval and confidence level in testing
  - larger level of significance corresponds to a more severe test
- Textbook Illustrations: Examples on pp.174-176

One- and Two-Sided Tests: Summary

One-sided (one-tailed) Test

$H_0: \bar{X} = \mu_0$ vs. $H_1: \bar{X} = \mu_1 > \mu_0$

- *Large-sample test statistic:*

  $z \approx (\bar{x} - \mu_0)/(S_2/\sqrt{n})$

- *Small-sample test statistic:*

  $t = (\bar{x} - \mu_0)/(S_2/\sqrt{n})$

- Region of Rejection

$z > z_\alpha$ $(z < -z_\alpha)$ and $t > t_\alpha$ $(t < -t_\alpha)$

$P(z > z_\alpha) = \alpha$ or $P(t > t_\alpha) = \alpha$

Two-sided (two-tailed) Test

$H_0: \bar{X} = \mu_0$ vs. $H_1: \bar{X} = \mu_1 \neq \mu_0$

- *Large-sample test statistic:*

  $z \approx (\bar{x} - \mu_0)/(S_2/\sqrt{n})$

- *Small-sample test statistic:*

  $t = (\bar{x} - \mu_0)/(S_2/\sqrt{n})$

- Region of Rejection

  $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

  and $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

$P(z > z_{\alpha/2}) = \alpha/2$ or $P(t > t_{\alpha/2}) = \alpha/2$

One-Sided Test: An Example

- Testing of known Gaussian mean (known variance)

$$\text{Test statistic } z = [\bar{x} - \bar{X}]/[\sigma/\sqrt{n}] = [290 - 300]/[40/\sqrt{100}] = -2.5$$

$$\text{Accept } \bar{X} = 300 \text{ if } z > z_c \text{ with confidence } C(z_c) = \int_{z_c}^{\infty} f(z)dz = 1 - \Phi(z_c) \text{ or significance } \alpha = 1 - C(z_c)$$

$$\text{If } C(z_c) = 0.99 \Rightarrow z_c = -2.33, \text{ we reject the hypothesis } \bar{X} = 300 \text{ with 99\% confidence}$$
$$\text{and if } C(z_c) = 0.995 \Rightarrow z_c = -2.575, \text{ we accept the hypothesis } \bar{X} = 300 \text{ with 99.5\% confidence}$$

- Higher confidence level implies large acceptance region
  - a higher level of significance α implies a more severe test
- T-test: for smaller sample sizes (known variance)

$$\text{Test statistic } t = [\bar{x} - \bar{X}]/[\tilde{s}_1/\sqrt{n}] = [290 - 300]/[40/\sqrt{9}] = -0.75$$

$$\text{If } C(t_c) = 0.99 \Rightarrow t_c(8) = -2.896, \text{ we accept the hypothesis } \bar{X} = 300 \text{ with 99\% confidence}$$

Two-Sided Test: An Example

- Testing of known Gaussian mean (known variance)

$$\text{Test statistic } z = [\bar{x} - \bar{X}]/[\sigma/\sqrt{n}] = [10.3 - 10]/[1.2/\sqrt{100}] = 2.5$$

$$\text{Accept } \bar{X} = 10 \text{ if } -z_c < z < z_c \text{ with confidence } C(z_c) = \int_{-z_c}^{z_c} f(z)dz = 1 - 2\Phi(z_c) \text{ or significance } S(z_c) = 1 - C(z_c)$$

$$\text{If } C(z_c) = 0.95 \Rightarrow z_c = 1.96 \text{ (Table 4-1), we reject the hypothesis } \bar{X} = 10 \text{ with 95\% confidence}$$

- T-test: for smaller sample sizes (known variance)

$$\text{Test statistic } t = [\bar{x} - \bar{X}]/[\tilde{s}_1/\sqrt{n}] = [10.3 - 10]/[1.2/\sqrt{9}] = 0.75$$

$$\text{If } C(t_c) = 0.95 \Rightarrow t_c(8) = 2.306 \text{ (Table 4-2), we accept the hypothesis } \bar{X} = 10 \text{ with 95\% confidence}$$

  - small sample test is not as severe as a large sample one
- Critical Value: z_c and t_c are critical values of the tests
- Textbook Illustrations: Exercises 4-5.1 and 4-5.2

Statistical Hypothesis Testing

- In essence, a hypothesis test partitions the entire observation space into two disjointed sets, *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,* often defined by critical values as discussed earlier
- *Type I error* (also called *false rejection error*) of a test:

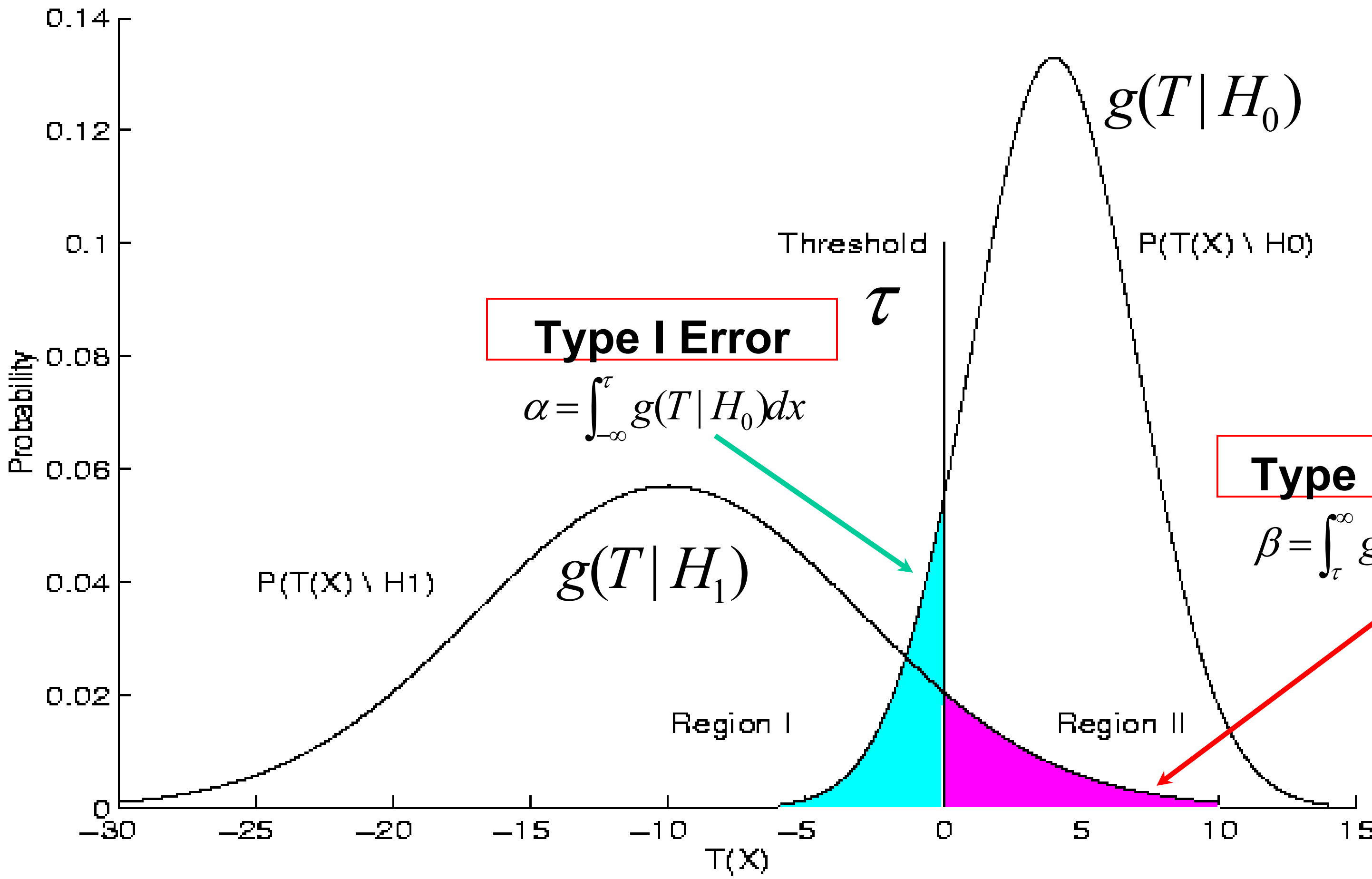
$$\alpha = \mathrm{P}(E_1) = \mathrm{P}(X \in C \mid H_0) \Rightarrow \text{ level of significance}$$

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

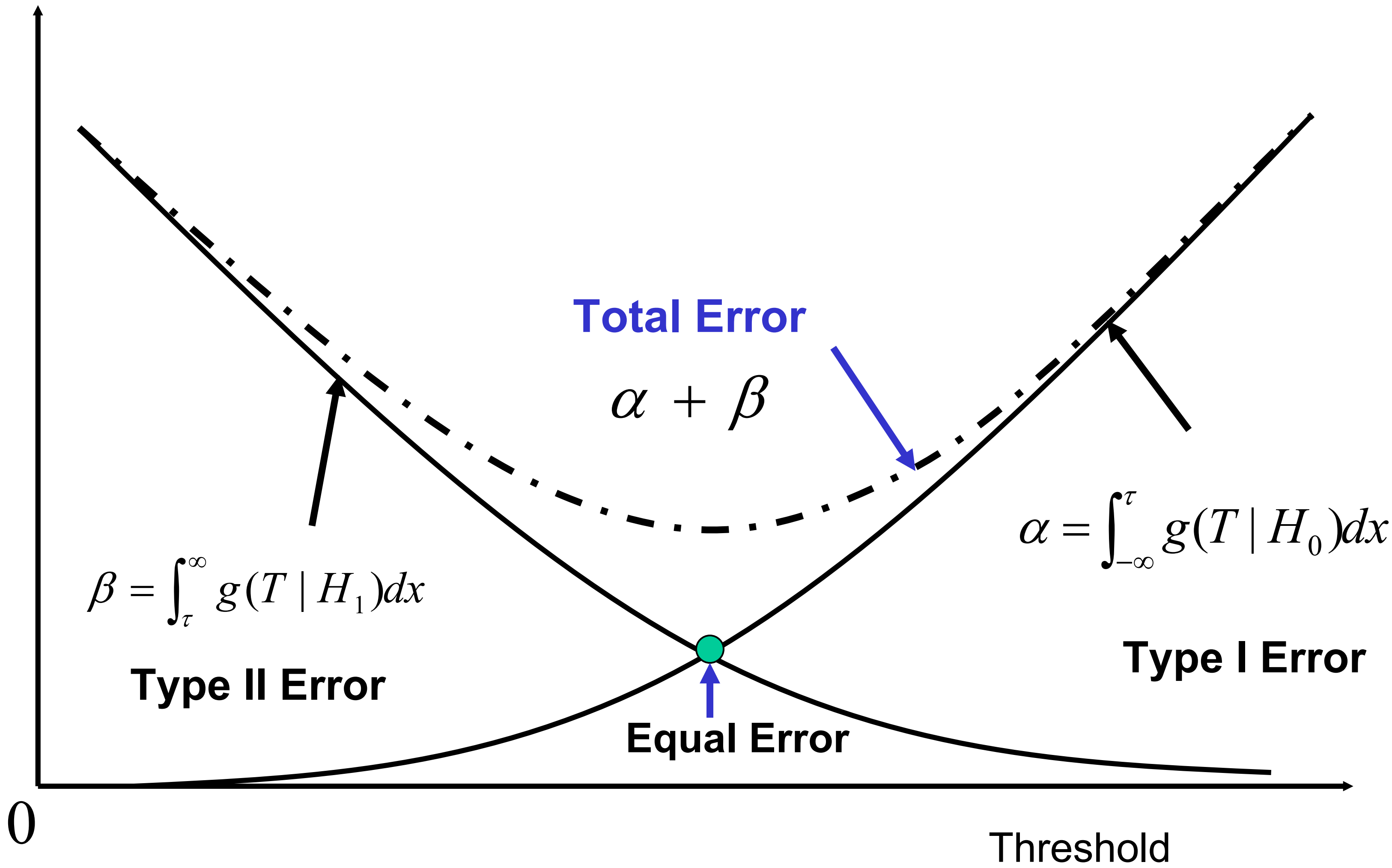
$$\beta = \mathrm{P}(E_2) = \mathrm{P}(X \in D \mid H_1) = 1 - \mathrm{P}(X \in C \mid H_1) = 1 - \gamma$$

- Recall the modem example in Chapter 1 and HW#1

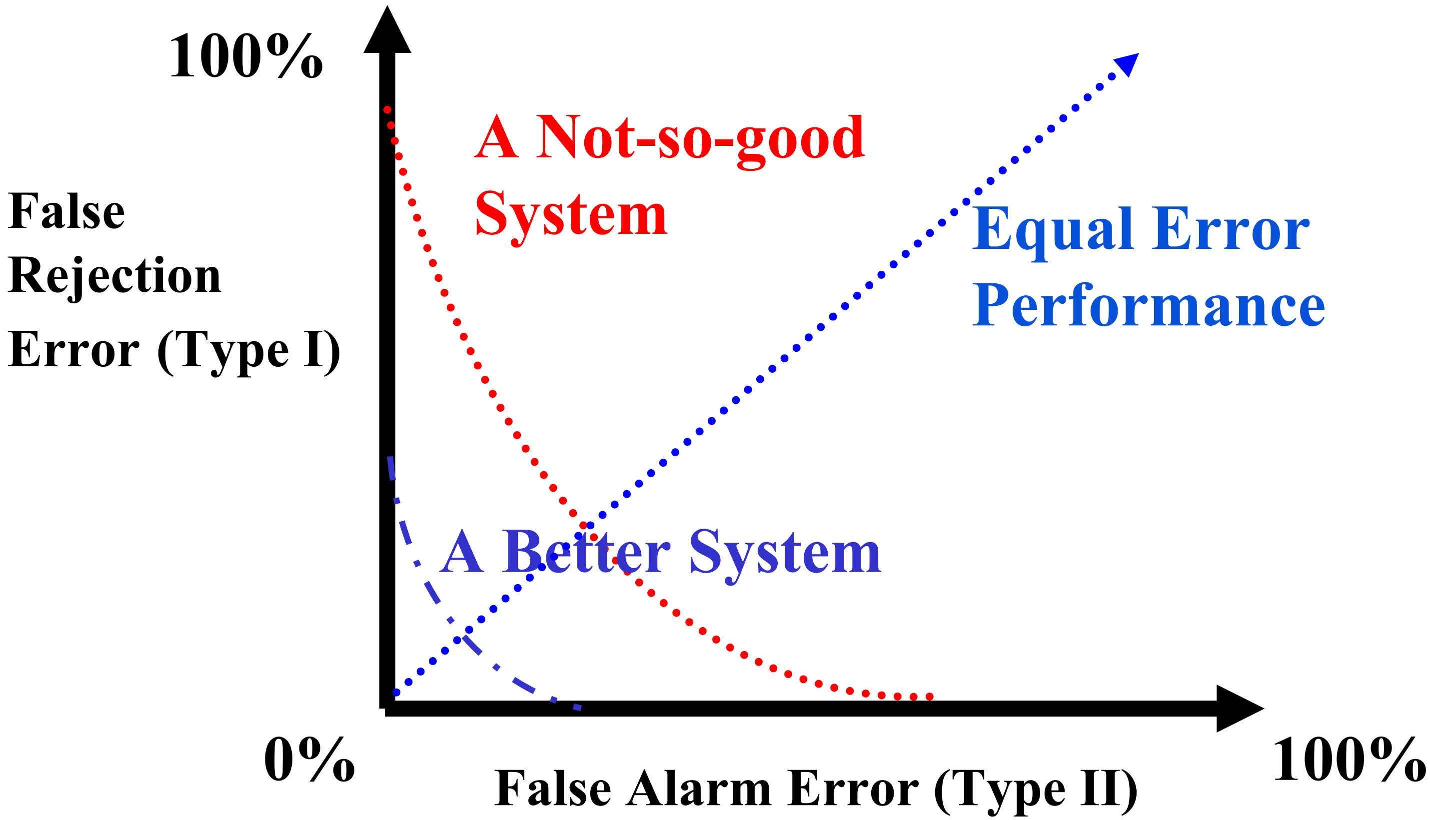
Densities of One-Sided Test Statistic



Evaluating Verification (I)



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



Another important application is biometric authentication.

Curve Fitting

- Consider fitting *y=r(x)* to a set of pairs of random samples: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - we will have curve fitting errors: $y_t = r(x_t) + d_t$ (cf. Figure 4-4)
  - *r(.)* is a regression function
  - goodness of fit: minimizing least squared errors $D = \sum_{t=1}^{n} d_i^2$
- Polynomial fitting (MATLAB example): $r(x) = \sum_{k=0}^{p} a_k x^k$
- Linear fitting: *y=a+bx*
- Spline fitting
  - local and global optimization
  - various optimization criteria
- Illustrations: Table 4-3, Exercises 4-6.1, 4-6.2

Linear Regression

- Least Squares: Minimizing Sum of Squared Error

$$D=\sum_{t=1}^{n} d_i^2=\sum_{t=1}^{n}[y_i-(a+bx_i)]^2=\text{minimum}$$

- We obtain the following matrix normal equation

$$\boxed{\frac{\partial D}{\partial a}=0 \Rightarrow \sum_{t=1}^{n} y_i=an+b\sum_{t=1}^{n} x_i,\ \ \frac{\partial D}{\partial b}=0 \Rightarrow \sum_{t=1}^{n} x_i y_i=a\sum_{t=1}^{n} x_i+b\sum_{t=1}^{n} x_i^2}$$

- Solving for intercept *a* and slope *b* : y=polyfit(y,x,n)

$$\boxed{\hat{b}=\frac{n\sum_{t=1}^{n} x_i y_i-(\sum_{t=1}^{n} x_i)(\sum_{t=1}^{n} y_i)}{n\sum_{t=1}^{n} x_i^2-(\sum_{t=1}^{n} x_i)^2},\hat{a}=\frac{(\sum_{t=1}^{n} y_i)(\sum_{t=1}^{n} x_i^2)-(\sum_{t=1}^{n} x_i)(\sum_{t=1}^{n} x_i y_i)}{n\sum_{t=1}^{n} x_i^2-(\sum_{t=1}^{n} x_i)^2}=\frac{\sum_{t=1}^{n} y_i-\hat{b}\sum_{t=1}^{n} x_i}{n}=\hat{\bar{Y}}-\hat{b}\hat{\bar{X}}}$$

- Extend to more than one regressor (econometrics)

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*
- Textbook Illustrations: bit error rate (BER) example
  - scatter plot Figure 4-7 (wind velocity versus BER)

An Intuitive Summary

- Simplifying Notations

$$SS_{XY} = \sum_{t=1}^{n} x_i y_i - [(\sum_{t=1}^{n} x_i)(\sum_{t=1}^{n} y_i)] / n,$$

$$SS_{XX} = \sum_{t=1}^{n} x_i^2 - (\sum_{t=1}^{n} x_i)^2 / n \quad \text{and} \quad SS_{YY} = \sum_{t=1}^{n} y_i^2 - (\sum_{t=1}^{n} y_i)^2 / n$$

- We obtain the following solutions

$$\hat{b} = \frac{SS_{XY}}{SS_{XX}}, \hat{a} = \hat{\bar{Y}} - \hat{b}\hat{\bar{X}} \text{ and } r = \frac{SS_{XY}}{\sqrt{SS_{XX} * SS_{YY}}}$$

- Can you extend the above to multiple regression?

Other Topics of Interest

- We did not have time to cover the following:
  1. Comparing two samples means (mean difference): for sampling distributions, confidence interval and hypothesis testing
  2. Multiple Regression (macroeconomics)
  3. Autoregression: Time Series (econometrics)
  4. Parameter Estimation
  5. Decision Theory
- Basic skills learned here can be applied to
  – The above and many other problems

Summary

- **Today's Class**
  – Elements of Statistics
- **Reading Assignments**
  – Cooper & McGillem, Chapter 4
- **Class Next Week**
  – Quiz #1 on 6/8/20 (Chapters 1-3)
  – Finishing Chapter 4