

ECE 3075A  
Random Signals

Lecture 20  
Confidence Interval, Hypothesis Testing and Regression

School of Electrical and Computer Engineering  
Georgia Institute of Technology  
Fall, 2003

## Confidence Interval

- Sample mean is a point estimate of the population mean – a single value that we offer as an estimate of the true parameter.
- An alternative is to offer an interval estimate – e.g., the population mean is estimated to be within a certain interval with a certain probability.
- A  $q$ -percent confidence interval is the interval within which the parameter being estimated would lie with a probability of  $q/100$ . The value  $q$  is called the confidence level and calculated as:

$$\Pr\left\{\bar{X} - \frac{k\sigma}{\sqrt{n}} \leq \hat{X} \leq \bar{X} + \frac{k\sigma}{\sqrt{n}}\right\} = q = 100 \int_{\bar{X} - k\sigma/\sqrt{n}}^{\bar{X} + k\sigma/\sqrt{n}} f_{\hat{X}}(x) dx$$

- $\bar{x} - \frac{k\sigma}{\sqrt{n}}$  and  $\bar{x} + \frac{k\sigma}{\sqrt{n}}$  are called the confidence limits;  $\bar{x}$  is a "realized" sample average.

This is interpreted as:  $\bar{X}$  (which is unknown and being estimated) will lie within the interval  $\left(\bar{x} - \frac{k\sigma}{\sqrt{n}}, \bar{x} + \frac{k\sigma}{\sqrt{n}}\right)$  at  $q\%$  of the times sampling is made.

## More on Confidence Interval

$$\Pr\left\{\bar{X} - \frac{k\sigma}{\sqrt{n}} \leq \hat{X} \leq \bar{X} + \frac{k\sigma}{\sqrt{n}}\right\} = q = 100 \int_{\bar{X} - k\sigma/\sqrt{n}}^{\bar{X} + k\sigma/\sqrt{n}} f_{\hat{X}}(x) dx$$

⇒ This is based on the previous result that for large  $n$

$$\hat{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad E[\hat{X}] = \bar{X}, \quad \text{var}(\hat{X}) = \frac{\sigma_X^2}{n}$$

But after a sample is chosen, the realized sample mean  $\bar{x} = \sum_{i=1}^n x_i$  and the sample variance  $\tilde{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  are already calculated.

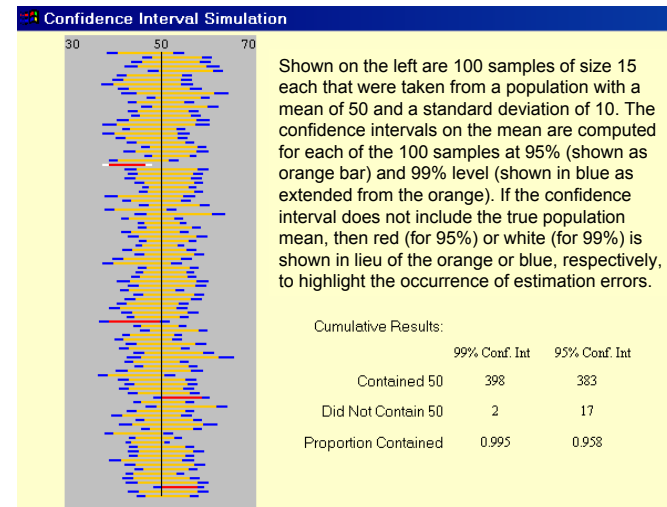
The interval for the purpose of predicting  $\bar{X}$ ,  $\bar{X} - \frac{k\sigma}{\sqrt{n}} \leq \hat{X} \leq \bar{X} + \frac{k\sigma}{\sqrt{n}}$ ,

is equivalent to  $\hat{X} - \frac{k\sigma}{\sqrt{n}} \leq \bar{X} \leq \hat{X} + \frac{k\sigma}{\sqrt{n}}$  or in one realization

$\bar{x} - \frac{ks}{\sqrt{n}} \leq \bar{X} \leq \bar{x} + \frac{ks}{\sqrt{n}}$  where the estimated mean and variance are used.

We thus say that  $\bar{X}$  (the unknown being estimated) will lie within the interval  $\left(\bar{x} - \frac{ks}{\sqrt{n}}, \bar{x} + \frac{ks}{\sqrt{n}}\right)$  at  $q\%$  confidence level.

## Confidence Interval - Example



From Rice University

## Hypothesis Testing

- A hypothesis is an assertion or a theory that one makes about a random observation, e.g., the lifetime of a component, the mean of a population, the presence of a signal together with a random noise, the distribution of a random variable.
- Hypothesis testing is a procedure one follows to accept or to reject said assertion in the context of statistical evidence.

Null Hypothesis – the theory

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than another (say currently used) drug.

$H_0$ : there is no difference between the two drugs on average.

Alternative Hypothesis – opposite of the theory

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than another (say currently used) drug.

$H_1$ : there is difference between the two drugs on average.

## Likelihood Ratio Test

Gather evidence  $\xi$  and evaluate the likelihood of the hypotheses.

Likelihood:  $\Pr\{\xi | H_0\}$  and  $\Pr\{\xi | H_1\}$

Likelihood Ratio Test:

Accept  $H_0$  if  $\log \frac{\Pr\{\xi | H_0\}}{\Pr\{\xi | H_1\}} > \tau$ ; otherwise, reject  $H_0$ .

	Decision	
	Reject $H_0$	Accept $H_0$
$H_0$ is true	Type I error (miss)	Correct decision
$H_1$ is true	Correct decision	Type II error (false alarm)

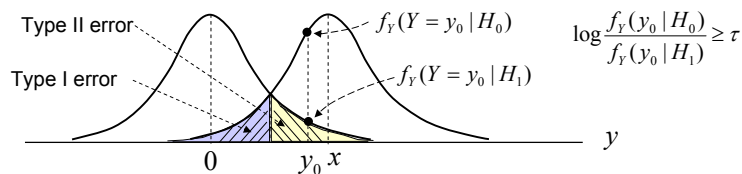
## Simple Likelihood Ratio Test - Illustration

$H_0$ : Signal  $x$  is present in the noisy observation  $y$  in which the additive noise  $v$  is independent gaussian with zero mean and variance  $\sigma_v^2$

$$f_Y(y | H_0) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{(y-x)^2}{2\sigma_v^2}\right\}$$

$H_1$ : Signal  $x$  is not present in the noisy observation  $y$  with the same noise characteristics as in  $H_0$

$$f_Y(y | H_1) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{y^2}{2\sigma_v^2}\right\}$$



## Curve Fitting

- Observations are made on two or more variables that may have a certain relationship – as in study of physics, for example. These data points, when plotted in a multi-dimensional space, form a scatter diagram.
- We use curve fitting to find a mathematical relationship, often simplified and parameterized, among the variables. The mathematical equation that relates the variables is called the regression equation which defines a regression curve, which fits the given observation data in some optimal sense, i.e., so-called criterion of goodness-of-fit.

Given  $n$  data points, in a 2-D example,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  we choose to fit a curve defined by  $y = g(x)$  to the data such that the fitting error defined as  $D = \sum_{i=1}^n [y_i - g(x_i)]^2$  is minimized. This leads to a **least-square regression curve**. We can choose other criterion if we like.

## Least-Square Regression

The function  $y = g(x)$  is usually parameterized. For example, in linear regression

$$y = g(x) = a + bx$$

Then,  $D = \sum_{i=1}^n [y_i - g(x_i)]^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2$

$$\frac{\partial D}{\partial a} = 2 \sum_{i=1}^n -(y_i - a - bx_i) = 0 \Rightarrow \sum_{i=1}^n y_i = \sum_{i=1}^n (a + bx_i) = na + b \sum_{i=1}^n x_i$$

$$\frac{\partial D}{\partial b} = 2 \sum_{i=1}^n -x_i(y_i - a - bx_i) = 0 \Rightarrow \sum_{i=1}^n x_i y_i = \sum_{i=1}^n (ax_i + bx_i^2) = a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2$$

Let  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\overline{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2$ ,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  and  $\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i$

$$\implies b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \quad a = \bar{y} - b\bar{x}$$

We can also generalize to  $m^{\text{th}}$  order polynomial  $y = g(x) = \sum_{m=0}^M c_m x^m$ . But, beware of the problem with over-fitting.

## Correlation between Data Sets

- It is often of interest to determine if two sets of data demonstrate observable statistical dependency or correlation – including if they come from the same source.
- The linear correlation coefficient (used as estimate of the true correlation coefficient) is obtained as:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\tilde{C}_{X,Y}}{\tilde{\sigma}_X \tilde{\sigma}_Y} \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\tilde{C}_{X,Y} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \tilde{\sigma}_X = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \tilde{\sigma}_Y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

When  $n$  is large,  $r$  is approximately Gaussian. In general, the correlation is considered significant if  $|r| > 0.5$ .