

ECE 3075A
Random Signals

Lecture 12

**Sample Variance, Confidence Interval, Hypothesis
Testing and Regression**

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

Sample Variance

- Do not confuse variance of sample mean with sample variance.
- Sample mean is an estimate of the mean of the population.
- A good estimate of the population mean does not readily represent good knowledge of the statistical properties of the population. Need to know some higher order statistics.
- What would be a reasonable estimate of the **variance of the population?**

With $X_i, i = 1, 2, \dots, n$ being the r.v.s in the sample, the sample variance is defined as

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{X})^2 = \frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right)^2$$

which is the average of the square of the difference between each r.v. and the sample mean.

Expectation of Sample Variance

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{X})^2 = \frac{1}{n} \sum_{i=1}^n \left(X_i - \frac{1}{n} \sum_{j=1}^n X_j \right)^2$$
$$= \frac{1}{n} \sum_{i=1}^n \left(X_i^2 - \frac{2X_i}{n} \sum_{j=1}^n X_j + \frac{1}{n^2} \sum_{j=1}^n X_j \sum_{k=1}^n X_k \right)$$

$$E[S^2] = \left(\frac{1}{n} \sum_{i=1}^n E[X_i^2] \right) - \left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n E[X_i X_j] \right)$$
$$= \overline{X^2} - \frac{1}{n^2} \left[(n^2 - n) \overline{X^2} + n \overline{X^2} \right] = \left(\frac{n-1}{n} \right) \overline{X^2} - \left(\frac{n-1}{n} \right) \overline{X}^2 = \frac{n-1}{n} \sigma^2$$

The expected value of the sample variance is not the true variance, and thus a biased estimate. We make the following adjustment

$$\tilde{S}^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{X})^2, \quad \therefore E[\tilde{S}^2] = \frac{n}{n-1} E[S^2] = \sigma^2$$

Distributions of Parameter Estimates

- Without knowing the true distribution, we may still want to have some idea about the distribution of the parameter estimates.
- If $X_i, i = 1, 2, \dots, n$ are Gaussian and independent with mean \bar{X} and variance σ^2 , then the sample mean is also Gaussian and the normalized random variable

$$Z = \frac{\hat{X} - \bar{X}}{\sigma / \sqrt{n}}$$

is Gaussian with zero mean and unit variance. This true regardless of the size of the sample.

- If $X_i, i = 1, 2, \dots, n$ are not Gaussian, then Z is asymptotically Gaussian as $n \rightarrow \infty$ according to the central limit theorem. As a rule of thumb, the asymptotic result is reached when $n \geq 30$.
- When n is not large enough to ensure normality, then the following normalized random variable Z has a Student's t distribution:

$$Z = \frac{\hat{X} - \bar{X}}{\tilde{S} / \sqrt{n}} = \frac{\hat{X} - \bar{X}}{S / \sqrt{n-1}}$$

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:

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

- $\bar{x} - \frac{k\sigma}{\sqrt{n}}$ 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

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

⇒ 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.

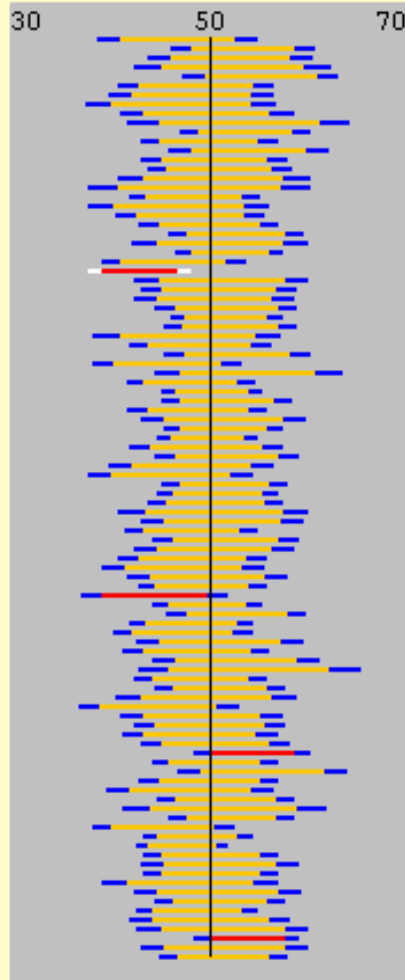
The interval for the purpose of predicting \bar{X} , $\bar{X} - \frac{k\sigma}{\sqrt{n}} \leq \hat{X} \leq \bar{X} + \frac{k\sigma}{\sqrt{n}}$, is equivalent to $\hat{X} - \frac{k\sigma}{\sqrt{n}} \leq \bar{X} \leq \hat{X} + \frac{k\sigma}{\sqrt{n}}$ or in one realization

$\bar{x} - \frac{ks}{\sqrt{n}} \leq \bar{X} \leq \bar{x} + \frac{ks}{\sqrt{n}}$ 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

Confidence Interval Simulation



Shown on the left are 100 samples of size 15 each that were taken from a population with a mean of 50 and a standard deviation of 10. The confidence intervals on the mean are computed for each of the 100 samples at 95% (shown as orange bar) and 99% level (shown in blue as extended from the orange). If the confidence interval does not include the true population mean, then red (for 95%) or white (for 99%) is shown in lieu of the orange or blue, respectively, to highlight the occurrence of estimation errors.

Cumulative Results:

	99% Conf. Int	95% Conf. Int
Contained 50	398	383
Did Not Contain 50	2	17
Proportion Contained	0.995	0.958

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 ξ 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)

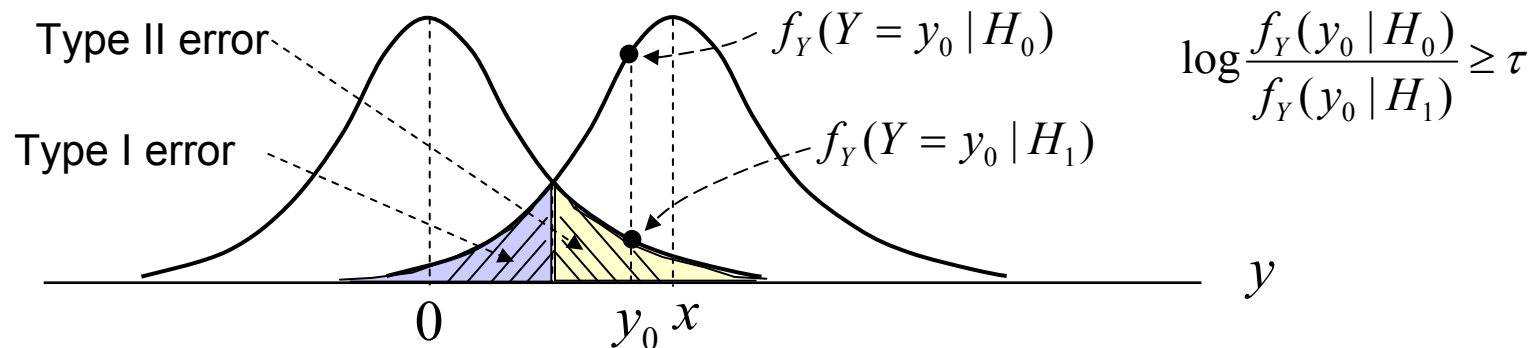
Simple Likelihood Ratio Test - Illustration

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 σ_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

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

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

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

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

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

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

We can also generalize to m^{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 $|r| > 0.5$.