Massachusetts Institute of Technology

Department of Electrical Engineering & Computer Science

6.041/6.431: Probabilistic Systems Analysis

(Fall 2008)

Recitation 23
December 4, 2008

Problem 1 is a continuation of problem 2 in Recitation 22. The first two parts have been solved for you.

1. Let $X_1, \ldots, X_n$ be i.i.d. samples of a Gaussian random variable with an unknown common mean θ, and an unknown variance σ^2. Suppose we have sample values $X_1 = x_1, \ldots, X_n = x_n$. The mean estimator is given by

$$\hat{\Theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

(a) Find the mean and variance of $\hat{\Theta}_n$. Is $\hat{\Theta}_n$ Gaussian?

Solution

$$E[\hat{\Theta}_n] = \frac{1}{n}\sum_{i}^{n} E[X_i] = \theta,$$

$$\mathrm{var}(\hat{\Theta}_\mathrm{n}) = \frac{\mathrm{var(X_i)}}{\mathrm{n}} = \frac{\sigma^2}{\mathrm{n}}.$$

$\hat{\Theta}_n$ is gaussian because it is the sum of independent Gaussian (normal) random variables.

(b) A common approximation (which is not exactly correct, but is close for large values of n) is that the unbiased variance estimator $\hat{S}_n^2$ is exactly equal to σ^2. Using this approximation, find the probability distribution for the random variable

$$T_n = \frac{\sqrt{n}(\hat{\Theta}_n - \theta)}{\hat{S}_n},$$

where $\hat{S}_n = +\sqrt{\hat{S}_n^2}$.

Write the event that θ lies in the confidence interval

$$[\hat{\Theta}_n - z\frac{\hat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + z\frac{\hat{S}_n}{\sqrt{n}}]$$

in terms of a range of possible values for T_n. Using the approximation above, find the 95 % confidence interval for Θ , i.e., find the value of z for which

$$\mathbf{P}_\theta\left(\hat{\Theta}_n - z\frac{\hat{S}_n}{\sqrt{n}} < \theta < \hat{\Theta}_n + z\frac{\hat{S}_n}{\sqrt{n}}\right) = 0.95.$$

Find the confidence interval for $n = 4$ and $n = 16$ in terms of $\hat{S}_n$ and $\hat{\Theta}_n$.

Solution The probability distribution of the random variable T_n under the assumption $\hat{S}_n^2 = \sigma^2$ is that of the standard normal random variable.

The event that θ lies in the confidence interval

$$[\hat{\Theta}_n - z\frac{\hat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + z\frac{\hat{S}_n}{\sqrt{n}}]$$

can be written as the event

$$[-z \leq T_n \leq z].$$

Since we are interested in the 95 % confidence interval we want to find z such that $P([-z \leq T_n \leq z]) \geq 0.95$. Using the CDF of the standard normal, we have $P([-z \leq T_n \leq z]) = \Phi(z) - \Phi(-z) = \Phi(z) - 1 + \Phi(z) = 0.95$ from which we obtain $\Phi(z) = 0.975$. The value of z that attains this value is 1.96.

The confidence interval when $n = 4$ is given by,

$$[\hat{\Theta}_n - 0.98\hat{S}_n, \hat{\Theta}_n + 0.98\hat{S}_n],$$

and when $n = 16$ it is given by,

$$[\hat{\Theta}_n - 0.49\hat{S}_n, \hat{\Theta}_n + 0.49\hat{S}_n].$$

c) When the X_i's are iid normal, the random variable T_n is called the "t-distribution with n-1 degrees of freedom," and it has a known probability distribution. The distribution is symmetric about the origin and broadly resembles the standard normal density, $N(\mu = 0, \sigma = 1)$, but with "fatter tails." See figure on next page. Using the table on the next page, find values of z that give a more accurate 95 % confidence interval for θ for $n = 4$ and $n = 16$. Give the confidence intervals for both values of n in terms of $\hat{S}_n$ and $\hat{\Theta}_n$.

d) Compare your answers to parts (b) and (c). Which method gives a wider confidence interval? How does this behavior depend on n?

2. (Problem 18, see page 514.)We want to estimate Y given x, assuming a linear relation $Y_i = \theta_0 + \theta_1 x_i + W_i$, where $W_1, \ldots, W_n$ are i.i.d. normal random variables with mean zero and variance σ^2. Then the maximum likelihood estimates of θ_0 and θ_1 are given by the linear regression formulas, as discussed in Section 9.2. Find the bias of the estimators $\hat{\Theta}_0$ and $\hat{\Theta}_1$.