MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Civil and Environmental Engineering

1.017/1.010 Computing and Data Analysis for Environmental Applications/ Uncertainty in Engineering

---

Quiz 2 (Solutions provided at the end of each problem)
Tuesday, November 4, 2003

Problem 1 (15 points)

Sometimes we may want to generate two correlated random variables for stochastic simulation applications. For example, the duration and intensity of rain storm may be highly uncertain but positively correlated. One option for generating correlated variables x and y is to obtain y from:

$$y = ax + b\varepsilon$$

where x and ε are independent random variables with means equal to 0.0 and variances equal to 1.0 and a and b are specified constants.

a: (5 points) What are the mean and variance of y? (expressed as functions of a and b):

b: (10 points) Select a and b so that y has a variance of 2 and the correlation between x and y is 0.5. Show all relevant calculations.

Problem 1 Solution:

a:

$$E[y] = E[ax+b\varepsilon] = aE[x]+bE[\varepsilon] = 0$$
$$Var[y] = Var[ax+b\varepsilon] = a^2Var[x] + b^2Var[\varepsilon] = a^2 + b^2$$

b:

$$a^2 + b^2 = 2$$
$$Correl(x,y) = 0.5 = Cov(x,y)/(Std[x]Std[y])$$
$$Cov(x,y) = E[(x-\bar{x})(y-\bar{y})]$$
$$y-\bar{y} = a(x-\bar{x})+b(\varepsilon-\bar{\varepsilon})$$
so: $Cov(x,y) = E[a(x-\bar{x})^2+b(x-\bar{x})(\varepsilon-\bar{\varepsilon})] = aVar[x] = a$
$\rightarrow$ **$a = .5*\sqrt{2} = 0.707$; $b = 1.22$**

Problem 2 (25 points)

Suppose that the time between failures of a structural component is modeled as an exponentially distributed random variable. You want to use the 10% quantile x_{10} [defined by $F_x(x_{10}) = 0.10$] as an

indication of how often the component should be tested. You have the following 10 recorded times between failures (in hrs):

512 1464 4995 7216 1150 2717 7842 39,898 1967 8103

a: (10 points) Propose a technique for estimating x_{10} from the observed times between failures.

b: (5 points) Compute an x_{10} estimate from the above data.

c: (10 points) Use the above data to derive a **99%** large sample double-sided confidence interval for the true x_{10} value. You may wish to use the unit normal CDF plot provided at the end of this quiz.

Problem 2 Solution:

a: One technique is to use the exponential CDF to obtain x_{10} estimates using m_x estimates from the data:

$F(x) = 1\text{-}exp[\text{-}x/a]$, where m_x approximates a, and $F(x) = 0.1$
$\rightarrow \quad \hat{x}_{10} = 0.1054 m_x$

This estimator is both unbiased and consistent. Estimators that were not unbiased and consistent were also given credit (since the problem does not specify that they should be), but only estimators that actually estimate x_{10} legitimately were given full credit.

b:

$\hat{x}_{10} = .1054*7586 = 799.6$

c: Because the estimator in this case is x_{10} and not m_x, the approximation $SD(\hat{a}) \sim SD(x)/\sqrt{N}$ CANNOT be used!

For the above estimator, $SD(\hat{x}_{10}) \sim 0.1054\ SD(x)/\sqrt{N} = 390.4$
The 99% confidence interval is:

$-2.575 < (799.6 - x_{10})/390.4 < 2.575$
$205.7 < x_{10} < 1804.9$

Problem 3 (40 points)

Suppose that you have reason to believe that fluctuations x around the long-term average tidal velocity normal to a shoreline are **uniformly distributed**, between $-a$ and $+a$, with a mean of 0. The distributional parameter a, the upper limit on the velocity, is unknown. You have N velocity measurements $[x_1, x_2, \ldots, x_N]$, which you assume is a random sample drawn from the postulated uniform distribution. In this problem you will use the **method of moments** to estimate a from the random sample.

a. (10 points) Derive an expression that relates the variance of x to the parameter a.

b. (5 points) Use this expression to suggest an estimator $\hat{\boldsymbol{a}}(x_1, x_2, x_N)$ for a. Do you think your estimator is unbiased and consistent (you do not need to prove that these properties apply --- just state your opinion, with justification)?

c. (10 points) Describe how you would derive a two-sided large sample confidence interval for a from *i*) a specified confidence level 1-α, *ii*) an actual estimate $\hat{a}$ of a computed from the N velocity measurements with your suggested estimator, and *iii*) the standard deviation of $\hat{\boldsymbol{a}}$. Describe your procedure step-by-step so it could be carried out with a real data set.

d. (15 points) Identify how you would obtain approximate values for any unknown quantities appearing in the confidence interval expression. In particular, provide a MATLAB (or pseudocode) program for any virtual experiment/Monte Carlo calculations that you would perform. If you cannot remember the exact name or syntax for a particular internal MATLAB function (such as `exprnd` or `normcdf`), just specify your own syntax and identify what the function does in words. Then include it in the appropriate place in your program. Alternatively, ask us.

Problem 3 Solution:

a: For this uniform CDF, $Var[x] = (2a)^2/12$ (this can be derived by integration). So **$\mathbf{Var[x] = a^2/3}$**

b: Using this expression (as the problem states), a good estimator is:

$$\hat{a} = \sqrt{3}s_x$$

This is unbiased and consistent since s_x^2 is an unbiased estimator of $Var[x]$ and the variance of $\hat{a}$ approaches zero as N approaches infinity.

c: Confidence interval for a is:

$$\sqrt{3}s_x - SD(\hat{a})F^{-1}(1-\alpha/2) < a < \sqrt{3}s_x - SD(\hat{a})F^{-1}(\alpha/2)$$

where F^{-1} is the inverse of the unit normal CDF (large sample assumption)

d:

```
function test(actual_data)
% This program will simulate replicates to determine the
% SD[ahat], which is the unknown quantity in Part c.
%
% Actual data vector is input as a function argument
N=length(actual_data)  ;       % number of data points per sample
nrep=1000;
ahat=sqrt(3)*std(actual_data)   ;     % estimate of ahat from the
sx of the data.
%  Assume unknown true value of uniform distribution limit
%  a is equal to ahat
% generate nrep replicates from unifrnd
sim_data=unifrnd(-ahat,ahat,N,nrep);
% compute estimate for each replicate
simahats=sqrt(3)*std(sim_data);
% find standard deviation over estimate replicates
```

```
sdahat=std(simahats);

% That's all we need, so just plug in:

lowerbound=ahat-sdahat*norminv(1-alpha/2,0,1)
upperbound=ahat-sdahat*norminv(alpha/2,0,1)
return
```

Problem 4 (20 points)

Consider two groups of engineers 7 years out of MIT: one group of 10 with professional engineer (PE) registration and one group of 8 without. The salaries of each group (in tens of thousand dollars) are as follows:

With PEs: 66 41 77 80 52 98 99 74 81 78

Without PEs: 65 88 55 124 66 72 96 71

Using a large sample assumption and this data set, perform a two-sided test of the hypothesis that the mean salaries of engineers with and without PEs are the same. Summarize your results by reporting the *p* value for the test. When picking the two groups of engineers how could you minimize the impact of factors other than PE registration on your conclusions? You may wish to use the unit normal CDF plot provided at the end of this quiz.

Problem 4 Solution:

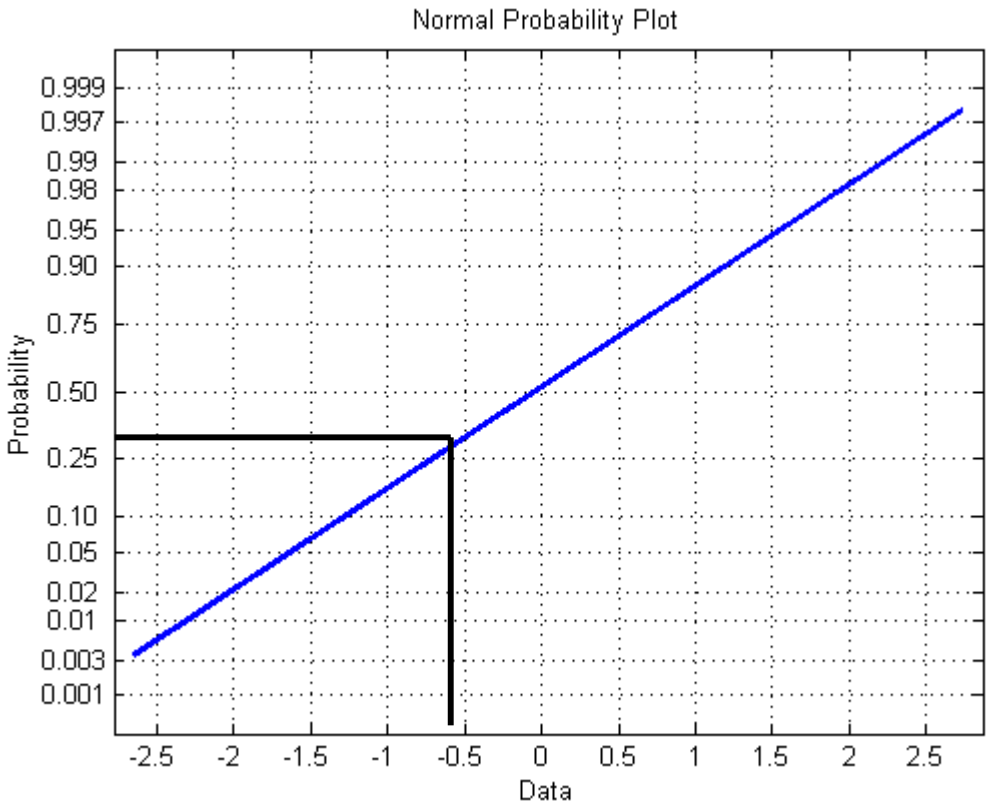
With PE (x): $m_x = 74.6$ $s_x = 18.0874$

Without PE (y): $m_y = 79.625$ $s_y = 22.1871$

$z = (74.6-79.6)\,(18.1^2/10+22.2^2/8)^{-0.5} = -.518$

From the chart below, $F_z(-.518) = 0.3 = p/2$

P=0.6


Normal Probability Plot
Probability
0.999
0.997
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
0.003
0.001
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Data