MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Civil and Environmental Engineering

1.017 Computing and Data Analysis for Environmental Applications

---

Practice Quiz 3
December 5, 2001

Please answer all questions on a separate piece(s) of paper with your name clearly identified:

Problem 1 (15 points)

Answer each of the following questions in a few sentences:

a) Suppose that y is a random variable (e.g nitrate concentration in a lake). Also, suppose that you have taken 20 nitrate measurements $y_1, y_2, \ldots, y_{20}$ at various times and locations. Explain the difference between $E[y]$, the expectation of y, and m_y, the sample mean of y.
b) Define the concept of a random sample and explain why it is useful
c) Why is the sample mean of $y_1, y_2, \ldots, y_{20}$ a "good' estimate of $E[y]$?

Solution:

a) The random variable y is characterized by its probability density, which describes the likelihood of obtaining an observation in a given range (or interval) of values. The expectation $E[y]$ is the first moment of this density. It is a property of the density rather than any particular set of data. The sample mean m_y is the arithmetic average of a particular set of data and is derived from the data rather than the density.

b) A sample is a set of measurements $y_1, y_2, \ldots, y_n$ of some random variable y with a probability density $f_y(y)$. The n measured values can be viewed as particular outcomes associated with n related random variables, one for each member of the sample. Generally speaking, these n random variables are described by a multivariate probability distribution. If the sample is random, the y_i's are independent and all have the same marginal probability density $f_y(y)$ as the original variable y. These properties greatly simplifies the task of deriving the probability distributions of estimates computed from the sample.

c) The sample mean is a good estimate of $E[y]$ because it is unbiased [$E(m_y)=E(y)$] and consistent [$Var(m_y)$ goes to zero as the sample size approaches infinity]. Together, these properties imply that the sample mean converges to the expected value of y as the sample size increases.

Problem 2 (15 points)

Suppose that the soil porosity y is a random variable that varies over space. Porosity is a fraction defined as the volume of voids (volume not occupied by soil grains) divided by the total volume of the soil sample. Suppose that you wish to estimate the expected value $E[y]$ of the porosity from 5 observations.

a) What is a reasonable probability density to assume for y ? Is it better to estimate this density from the 5 samples or to simply postulate it's form?
b) What would you use as an estimate of $E[y]$?
c) How would you derive the probability distribution (e.g. CDF) of your estimate? What could you do with this distribution?

Solution:

a) Since porosity is constrained between 0.0 and 1.0 a uniform probability over this (or perhaps a more limited range is a reasonable choice. A CDF or PDF estimated from a sample size as small as 5 is unlikely to be very informative. An assumed uniform density is probably a better choice.
b) The sample mean is a good choice, for the reasons given in the solution to Problem 1c.
c) Since the sample is quite small a large sample assumption is not really justified. It is probably better to carry out a stochastic simulation by generating many (thousands) of 5 measurement samples from a uniform random number generator with specified lower and upper bounds. The sample CDF obtained from the simulation could be used to test hypotheses about the properties of y or, if the bounds used in the generator are varied, to construct confidence intervals for these properties.

Problem 3 (20 points)

Consider the following 10 samples of suspended sediment concentration (in mg/L), taken at various distances up an estuary (rounded off to facilitate calculation):

6 10 11 10 14 10 13 4 7 6

a) Derive a **two-sided** 90% confidence interval (10% significance level) for the expected value of sediment concentration, using a **large sample** assumption. Do you think this assumption is justified?

b) Formulate a **large sample two-sided test** to accept or reject the hypothesis that the mean concentration is equal to 12 mg/L. What is the p value obtained with the data set given above?

Solution:

a) The appropriate estimator for the mean concentration is the sample mean m_y, which has the value $\xi = 9.1$ for this particular sample.

If we adopt a large sample approximation we assume that the sample mean is normally distributed with mean $E[m_y] = E[y]$ ((the unknown true mean) and standard deviation $SD[m_y] = s_y/n^{1/2} = 3.25/(10)^{1/2} = 1$. The critical points for a two-sided 90% confidence interval are given by:

$$m_{yl,0.9} = \xi - SD[m_y](1.64) = 9.1 - (1)(1.64) = 7.5$$
$$m_{yu,0.9} = \xi + SD[m_y](1.64) = 9.1 + (1)(1.64) = 10.7$$

So we infer from the data that $E[y]$ lies in the interval between 7.5 and 10.7. In fact, the samples were generated from a log normal distribution with a mean of 8.33.

The large sample assumption is reasonable, even for a sample of 10, for a significance level of 10% but becomes less accurate as this level decreases.

b) The hypothesis is *H0*: $E[y] = 12$. The hypothesis test is based on the same estimate and probability density as the confidence interval calculation. In the two-sided case the decision rule is:

Accept *H0* : if $m_{yl,\alpha} \le m_y \le m_{yu,\alpha}$

Reject *H0* : otherwise

where $m_{yl,\alpha}$ and $m_{yu,\alpha}$ are the critical values for a significance level of α.

For this problem we do not need to carry out the test or evaluate the critical values. Instead, we look for the *p* value. The *p* value is the value of α that we would give an m_y critical value equal to ξ. This can be found by setting the appropriate critical value ($m_{yl,p}$ if $\xi < E[y]$ and $m_{yu,p}$ if $\xi > E[y]$) equal to ξ. In this example, $\xi < E[y]$ so we use $m_{yl,p} = \xi$. This is equivalent to a standardized (unit normal) critical value of:

$$z_{l,p} = \frac{m_{yl,p} - 12}{SD[m_y]} = \frac{\xi - 12}{SD[m_y]} = \frac{9.1 - 12}{1.0} = -2.9$$

The probability that a unit normal variable is less than –2.9 is (from the tables in the text) 0.002. For a two-sided test the *p* value is twice this probability, so $p = 0.004$. This is a very low value, indicating that $m_y = \xi = 9.1$ is quite unlikely if $E[y] = 12$. So we would reject the null hypothesis, unless we wanted to use a significance level less than 0.004 (an unusually low value).

Problem 4 (25 points)

Consider the following 4 values of the independent variable (daily precipitation) and the dependent variable (streamflow the same day) in a small watershed.

Rainfall (mm): 10 15 8 20

Runoff (mm): 6 8.5 4 15

Derive a regression (least-squares) model $\bar{y}(x, a_1, a_2) = a_1 + a_2 x$ which has an intercept of zero (a_1 = 0) and a slope to be determined from the data. Use this model to predict the runoff that would occur when the precipitation is 25.

Sketch the data on an *x*-*y* plot. Include a plot of the regression line over the range $x = 0$ to $x = 25$.

Solution:

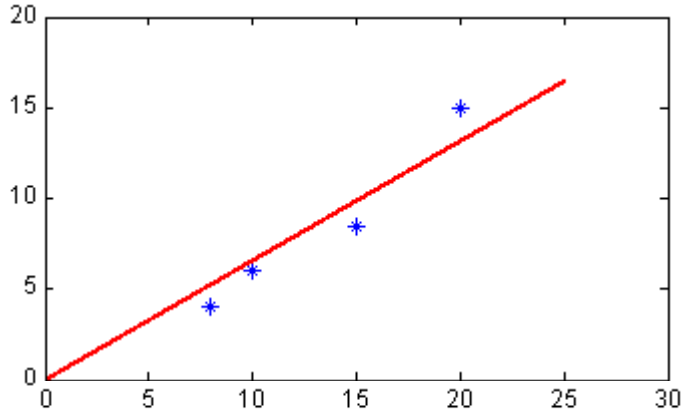
The least-squares estimate minimizes the sum-squared error:

$$SSE = \sum_{i=1}^{n} [y_i - a_2 x_i]^2 = [6 - 10a_2]^2 + [8.5 - 15a_2]^2 + [4 - 8a_2]^2 + [15 - 20a_2]^2$$

with respect to a_2. The minimizing value of a_2 may be found by applying the general least-squares estimation equations or by setting the derivative of *SSE* (taken with respect to a_2) equal to zero and solving for a_2.:

$$\frac{d\ SSE}{d\ a_2} = 20[6 - 10a_2] + 30[8.5 - 15a_2] + 16[4 - 8a_2] + 40[15 - 20a_2] = -1578a_2 + 1039 = 0$$

This gives $\hat{a}_2 = 0.658$. Substituting a_1=0, $\hat{a}_2 = 0.658$, and a precipitation of $x = 25$ into the original model, we obtain the prediction $\bar{y}(25, 0.0, 0.658) = (0.658)(25) = 16.46$. The requested plot is shown below:



Problem 5 (25 points)

Consider the following random sample of 8 measurements of a chi-squared distributed variable y:

7.0295 6.1062 2.1105 7.8423 10.7457 6.2128 9.4727 6.569

The mean of a chi-squared random variable is $E[y] = n$, the number of degrees of freedom used in the probability distribution.

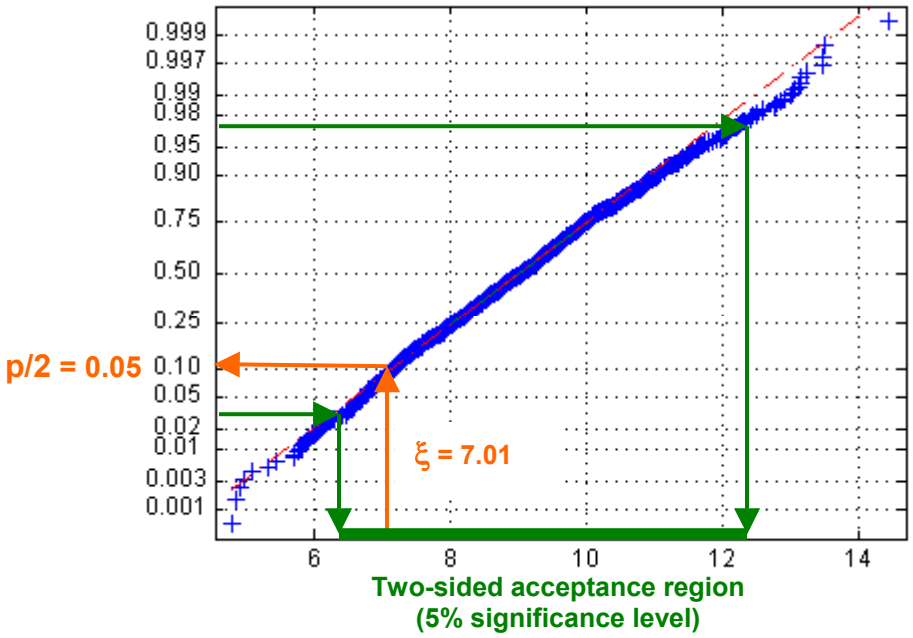
a) Write a stochastic simulation program that evaluates and plots on a normal probability scale the cumulative distribution function of an appropriate estimate of the number of degrees of freedom n of the distribution that generated the above measurements. How would you pick a value for the unknown value of n required in the random number generator of the stochastic simulation?

b) Explain how you would use your stochastic simulation program to test the null hypothesis that $n = 9$ (use a 5% significance level). Also, explain how you would obtain the p value.

Solution

a) The natural (method of moments) estimate of n is the sample mean m_y. The required program is given below with a typical output:

```
function chi2meas
close all
ysamp=[7.0295    6.1062    2.1105    7.8423   10.7457 ...
    6.2128    9.4727    6.5692]
nhatsamp=mean(ysamp)
nrep=1000
ndof=9
for i=1:nrep
    y=chi2rnd(ndof,1,8);
    ndofhat(i)=mean(y);
end
normplot(ndofhat)
return
```



b) Derive and plot the CDF for $\hat{n}$ using a random number generator with $n = 9$ (the null hypothesis value). Then read off critical values for a two-sided acceptance region from the 2.5% and 97.5% CDF values. Also, read off the CDF value corresponding to $\hat{n}$ =7.01 (the estimate derived from the sample). This is the p value.