MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Civil and Environmental Engineering

1.017 Computing and Data Analysis for Environmental Applications

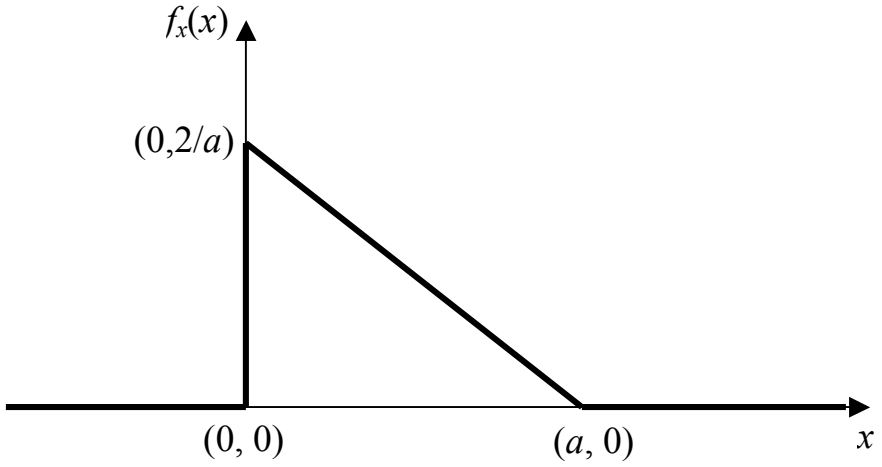
---

Quiz 2
Thursday, November 7, 2002

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

Problem 1 (60 points – you will receive partial credit for each part successfully completed)

Consider a random variable $\boldsymbol{x}$ with the triangular PDF shown below:



Note that this distribution has one unknown parameter a. Its mean and variance are:

$$E[\boldsymbol{x}] = \frac{a}{3} \qquad Var[\boldsymbol{x}] = \frac{a^2}{18}$$

You will receive 15 extra credit points if you prove that these expressions are correct.

Extra Credit:

$E[x]=\int_0^a xf(x)dx=\int[(-2/a^2)x^2+2x/a]dx = a/3$

$Var[x] = \int_0^a (x-mu)^2 f(x)dx = \int (x-a/3)^2(2/a-2x/a^2)dx = a^2/18$
(work in between is required)

Suppose that you obtain a random sample of 8 $\boldsymbol{x}$ values: $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_8]$.

Perform the following steps:

a) Devise an unbiased estimator $\hat{a}$ for the parameter a that depends only on the sample mean $m_x = \frac{1}{8}\sum_{i=1}^{8} x_i$.

b) Derive the mean and variance of this estimator. Is the estimator consistent? Why?

c) Suppose that the 8 values of your random sample are:

[0.15 0.32 0.65 0.08 0.28 0.28 0.05 0.12]

Derive a **large sample two-sided 95% confidence interval** for the parameter a. Note that a is related to but **not equal** to the population mean $E[x]$ (see expression for $E[x]$ above).

Hint: The two-sided 95% values for a unit normal distribution are –1.96 and +1.96.

d) Test the hypothesis H0: $a = 1$ by deriving a **p value** from the sample values given above and the unit normal CDF attached to the quiz (the CDF is plotted on normal probability paper). Assume that the sample is "large".

Hint: Please note that H0: $a = 1$ is **not the same** as H0: $E[x] = 1$ for this problem (since $E[x] \neq a$)

Solution:

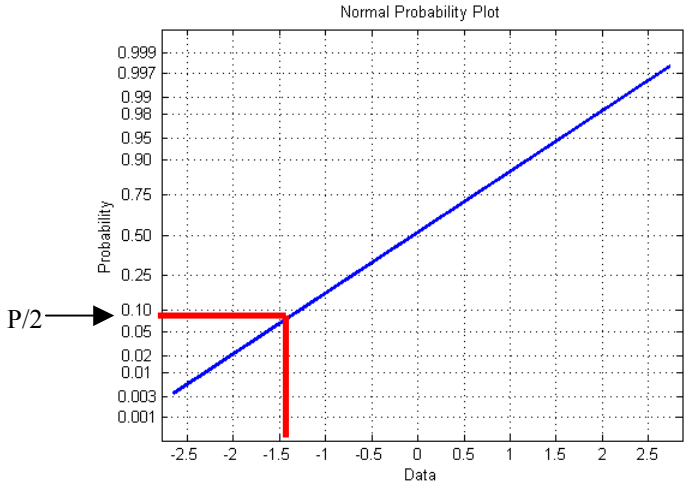
a.) estimator: $\hat{a} = 3*m_x$

b.) $E[\hat{a}]=E[3*m_x] = 3*E[x] = a$
$Var[\hat{a}] = Var[3*m_x]=9*Var[m_x]=9/N*\sigma^2_x \approx 9/N*S^2_x$
The estimator is consistent because its variance goes to zero as the number of values goes to infinity.

c.) sample mean=0.241 $\hat{a} = 3*m_x=0.723$ $S^2_x =0.0327$ $S_x=0.181$
$\sigma^2_{\hat{a}} =9/8* S^2_x = 0.0368$ $\sigma_{\hat{a}} = 0.192$

$-1.96 < (0.723-a)/0.192 < 1.96$ 95% Confidence Interval
$0.35 < a < 1.1$

d.)H_0: $a_0=1$
$Z=(0.723-1)/0.192= -1.44$
Reading off the graph: $p/2\approx0.075$ $p=0.15$



Problem 2 (25 points)

Provide specific one-sentence answers to the following questions:

a) Define a random sample.
b) What are the implications of making a "large sample" assumption when deriving confidence intervals or testing a hypothesis ?
c) How is the variance of a sample mean estimate of the population mean related to the sample size?
d) How does a two-sided confidence interval for the population mean change as the sample size changes?
e) How does the two-sided rejection region for a hypothesis test of the population mean change as the test significance level changes?

Solution:

a.) A random sample is one in which the Xi's are independent, and each Xi has the same probability distribution.
b.) A large-sample assumption means that the sample is large enough to assume that the estimate of interest is normally distributed.
c.) The variance of a sample mean estimate of the population mean is inversely related to the sample size. Therefore, the variance decreases as N increases. **Note**: the variance of the sample mean estimate is NOT the same as the variance of the data.
d.) As the sample size increases, the confidence interval width decreases. This makes sense, since the more data points you have, the more confident you are in your estimate, so you can make your interval smaller, ie you are really sure the true value is not far from your estimated value.
e.) As the significance level increases, the rejection region also increases.

Problem 3 (15 points)

Write a brief MATLAB program to display the histogram and estimate the variance of the sample mean $\boldsymbol{m}_{\boldsymbol{x}} = \frac{1}{4}\sum_{i=1}^{4} \boldsymbol{x}_i$ of a random sample $[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_2, \boldsymbol{x}_4]$ of size $N = 4$. The desired histogram and variance should be computed from a population of 1000 replicates. Assume that $\boldsymbol{x}$ is exponentially distributed with parameter $a = E[\boldsymbol{x}] = 1$. Do you think the histogram will closely resemble a normal distribution? Why?

Please write your code neatly and precisely so we can enter it into MATLAB to see if it works.

Solution:

Matlab code:

```
nrep=1000;
a=1;
n=4;
xsamples=exprnd(a,n,nrep);
sampmeans=mean(xsamples);
hist(sampmeans)
var(sampmeans)
```

The histogram will not be normal since there are a small number of samples being averaged (4), and the original distribution is exponential. It is, however, more normal than the exponential distribution because of the central limit theorem, which states that if we add a large number of random variables together, the sum (or mean) will be normal. But our N is only 4 so we can't expect it to be completely normal.

Note that the number of nreps does not really matter, except to visualize the distribution and get the variance. It is the actual mean of the four numbers that may or may not be normal, so the 4 is what counts.

Here's the histogram from the above program: