

Statistics 406 Midterm Exam

Fall 2007

No calculators, formula cards, computers, or notes may be used.

It is best to try every question. Partial credit will be given.

1. Describe in 2-3 sentences what the following program is doing. Focus on the statistical purpose of the program and the approach being used to achieve the purpose. You do not need to predict its result or describe each line of code individually.

```
N = c(10,20,30)
F = NULL

for (k in 1:3) {
  n = N[k]
  R = 0

  for (j in 1:1e3) {
    X = rexp(n)
    m = mean(X)
    s = sd(X)
    Q = m + s*rnorm(n*1e3)
    Q = array(Q, c(1e3,n))
    Q = apply(Q, 1, mean)
    Q = sort(Q)
    c1 = Q[25]
    c2 = Q[975]
    if ((c1<1) & (1<c2)) { R=R+1 }
  }
  F[k] = R/1e3
}
```

Solution: This is a simulation study to estimate the coverage probability of the parametric bootstrap approach to constructing a confidence interval for the expected value. Coverage probabilities are estimated for sample sizes $n = 10, 20, 30$. The simulation considers the case where the actual data-generating distribution is exponential, but the parametric bootstrap is applied assuming the normal distribution family.

2. Suppose we have an iid sample X_1, \dots, X_n from population A with expected value μ_A and an iid sample Y_1, \dots, Y_n from population B with expected value μ_B . Our interest is in the ratio μ_A/μ_B . We intend to use \bar{X}/\bar{Y} as an estimate of this ratio. In order to better understand how this estimate performs, we will use simulation to estimate its bias and mean squared error when the sample size is $n = 10, 20$, and 30. What R code should **** A **** and **** B **** be replaced with below to achieve this aim?

```

N = c(10,20,30)

nrep = 1e4

mu1 = 1
mu2 = 2

MSE = NULL
Bias = NULL

for (k in 1:3) {

  n = N[k]

  X = rnorm(nrep*n, mean=mu1)
  X = array(X, c(nrep,n))

  Y = rnorm(nrep*n, mean=mu2)
  Y = array(Y, c(nrep,n))

  R = apply(X,1,mean)/apply(Y,1,mean)

  MSE[k] = **** A ****
  Bias[k] = **** B ****
}

```

Solution:

```
**** A **** = mean( (R-1/2)^2 )
**** B **** = mean(R - 1/2)
```

3. Suppose we observe iid data X_1, \dots, X_n following this density function

$$f(X; s) = \sqrt{\frac{s}{2\pi}} \cdot \frac{\exp(-s/(2X))}{X^{3/2}},$$

where s is an unknown parameter.

(a) What is the log-likelihood function?

Solution:

$$n \log(s/2\pi)/2 - s \sum_i 1/2X_i - 1.5 \sum_i \log X_i$$

(b) Suppose we want to numerically calculate the MLE of s using Newton's method. What do we replace **** A ****, **** B ****, and **** C **** with in the following program to do this? You can assume that X is a vector containing the data.

```
s = 0.1 ## Starting value.

while (TRUE) {

  D1 = **** A ****

  if (abs(D1) < 1e-10) { break }

  D2 = **** B ****

  s = **** C ****
}
```

Solution: Here is the code:

```
s = 0.1

while (TRUE) {

  D1 = n/(2*s) - sum(1/(2*X))
```

```

if (abs(D1) < 1e-10) { break }

D2 = -n/(2*s^2)

s = s - D1/D2
}

```

- (c) Let \hat{s} be the maximum likelihood estimate obtained from part (b). How can we construct an approximate 95% confidence interval for s , centered at \hat{s} ?

Solution: The second derivative of the log-likelihood is $-n/2s^2$. Therefore the approximate standard deviation for \hat{s} is $s\sqrt{2/n}$. We can use $\hat{s} \pm 2s\sqrt{2/n}$ as an approximate 95% confidence interval.

4. Suppose U_i, V_i are independent with standard normal distributions. Let $X_1 = U_i$, $X_2 = U_i + V_i$ and $X_3 = U_i - V_i$. Let $\bar{X} = (X_1 + X_2 + X_3)/3$.

- (a) What is $E\bar{X}$?

Solution: $E\bar{X} = (EX_1 + EX_2 + EX_3)/3 = 0$.

- (b) What is $\text{var}\bar{X}$?

Solution:

$$\text{cov}(X) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

$\text{var}\bar{X} = 9/3^2 = 1$. Alternatively, $\bar{X} = U_i$, which has variance 1.

- (c) Suppose we observe Z_1, Z_2 , and Z_3 , each with expected value zero, and with $\text{var}Z_k = \text{var}X_k$ for $k = 1, 2, 3$. Is $\text{var}\bar{Z}$ greater than or less than $\text{var}\bar{X}$? Briefly explain your answer.

Solution: The covariance matrix is

$$\text{cov}(Z) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

so $\text{var}\bar{Z} = (1 + 2 + 2)/9 = 5/9$, which is less than $\text{var}\bar{X}$.

5. Suppose we run the following program.

```

D = NULL

nrep = 1e4

q = 5

for (k in 1:nrep) {

  P1 = sum(runif(q) < 2/3)
  P2 = sum(runif(q) < 1/2)

  if (P1 >= P2) {
    D[k] = mean(runif(10) < 2/3)
  } else {
    D[k] = mean(runif(10) < 1/2)
  }
}

F = mean(D)

```

- (a) Do you expect F to be less than $1/2$, between $1/2$ and $2/3$, or greater than $2/3$? Briefly explain your reasoning.

Solution: The value of F will be between $1/2$ and $2/3$. Based on the values of $P1$ and $P2$, we are either sampling from a population with probability $2/3$ of yielding a 1, or from a population with probability $1/2$ of yielding a 1. Thus the proportion of 1's is expected to fall between these values.

- (b) Suppose we run the program with $q=10$ instead of $q=5$. Do you expect the value of F in this case to be greater than, less than, or equal to the value of F from part (a)? Briefly explain your reasoning.

Solution: The value of F is expected to be greater when $q=10$ compared to $q=5$. The reason is that the larger "pilot sample" used to determine which population to sample from at the second stage is more likely to identify the population with success probability $2/3$.

6. Suppose we run the following program, then plot column 2 of R on the vertical axis against column 1 of R on the horizontal axis using $+$ points. On the same axes, we plot column 3 of R on the vertical axis against column 1 of R on the horizontal axis using triangular points. Which of A-D (next page) shows this plot? Briefly explain your reasoning.

```
N = c(5, 10, 20, 40, 80)
```

```
R = array(0, c(4,3))

for (k in 1:4) {

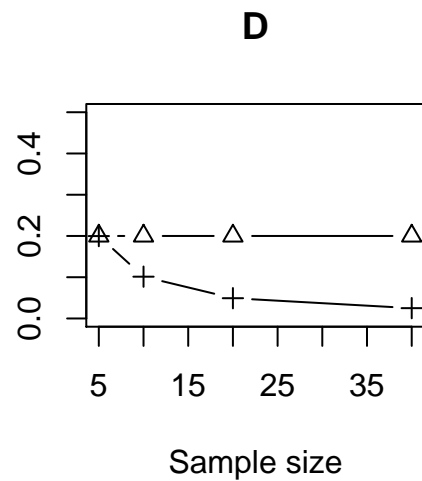
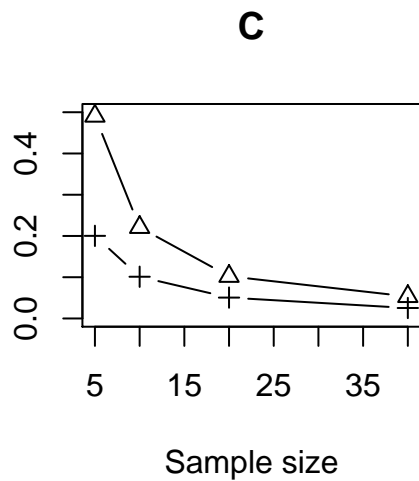
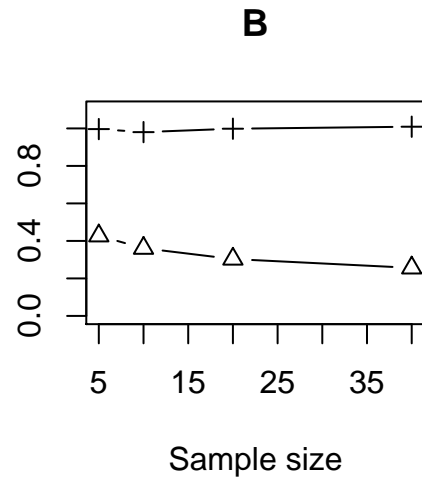
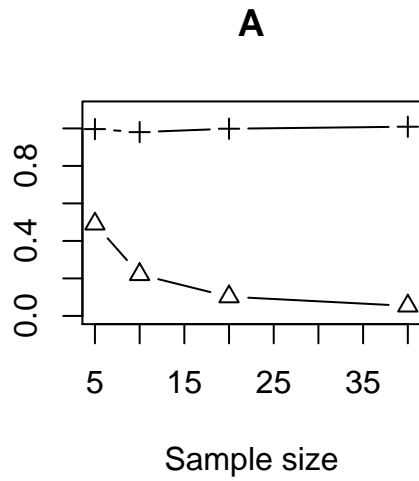
  M = rnorm(N[k]*1000)
  M = array(M, c(1000,N[k]))

  V = apply(M, 1, var)

  R[k,] = c(N[k], mean(V), var(V))
}
```

Solution:

The correct answer is plot A. The second column of R contains the expected value of the sample variance, which is equal to the population variance since the sample variance is unbiased. The third column of R contains the variance of the sample variance. We don't have a general formula for this quantity, but like most statistics, the variance decreases by a factor of 2 when the sample size doubles. Plot A is the only one of the four consistent with both these facts.



7. Let

$$f(X) = \frac{X}{1 + X^2},$$

and suppose we are interested in $\text{var}f(X)$, where X has a standard normal distribution.

(a) Using a Taylor's approximation to f , derive an approximation to $\text{var}f(X)$.

Solution: Since

$$f'(X) = \frac{1 - X^2}{(1 + X^2)^2},$$

and $EX = 0$ and $\text{var}(X) = 1$, the approximation is $\text{var}f(X) \approx \text{var}(X)f'(EX)^2 = 1$.

(b) Write a two line R program to unbiasedly estimate $\text{var}f(X)$ using simulation.

Solution:

```
X = rnorm(1e4)
VF = var(X/(1+X^2))
```

8. For each of the following R programs, state the expected result for the final value R. In some cases you can give a specific number, in other cases you will only be able to state whether R is expected to be greater than zero or less than zero. Briefly explain your reasoning.

(a)

```
X = rnorm(50)

m = mean(X)
s = sd(X)

Y = rnorm(50*1e4, mean=m, sd=s)
Y = array(Y, c(1e4,50))
Z = apply(Y, 1, var)

R = var(X) - mean(Z)
```

Solution: The parametric bootstrap distribution of the sample variance is centered at the sample variance of our actual data. Therefore R will be close to zero.

(b)

```
X = rnorm(50, mean=1, sd=2)

m = mean(X)
s = sd(X)

Y = rnorm(50*1e4, mean=m, sd=s)
Y = array(Y, c(1e4,50))
Z = apply(Y, 1, var)

A = rnorm(50*1e4, mean=1, sd=2)
A = array(A, c(1e4,50))
B = apply(A, 1, var)

R = var(Z) - var(B)
```

Solution: The parametric bootstrap distribution of the sample variance has similar variance as the actual sampling distribution of the sample variance. Therefore R will be close to zero.

(c)

```
A = rnorm(10*1e4, sd=2)
A = array(A, c(1e4,10))
A = apply(A, 1, mean)
```

```
B = rnorm(20*1e4, sd=1)
B = array(B, c(1e4,20))
B = apply(B, 1, mean)
```

```
R = var(A)/var(B)
```

Solution: Each value of A has variance $\sigma^2/n = 4/10$, and each value of B has variance $\sigma^2/n = 1/20$. The ratio is 8.

(d)

```
A = rt(1000*20, df=3)
A = array(A, c(1000,20))
A = apply(A, 1, mean)
```

```
B = rt(1000*20, df=10)
B = array(B, c(1000,20))
B = apply(B, 1, mean)
```

```
R = var(A) - var(B)
```

Solution: The t_3 distribution produces more outlying values, so the sample means are more variable. Thus R is expected to be positive.

(e)

```
U = runif(1000, min=0, max=1)
V = runif(1000, min=1, max=2)
```

```
R = var(V) - var(U)
```

Solution: Since the distribution of V is the same as the distribution of $U + 1$, and $\text{var}(U + c) = \text{var}(U)$ when c is a constant, V and U have the same variance. Thus R is expected to be close to zero.

(f)

```
U = runif(1000, min=0, max=1)
```

```
V = runif(1000, min=1, max=2)
```

```
R = var(log(V)) - var(log(U))
```

Solution: Since the $\log(x)$ function gets flatter as x grows, the variance is lower in $[1, 2]$ compared to $[0, 1]$.