Statistics 406 Problem Set 4
Due in lab, Tuesday October 9

1. Suppose we observe $X_1, \ldots, X_{2n}$ such that

$$X_i = U + Z_i \quad i \leq n$$

and

$$X_i = V + Z_i \quad i > n$$

where $U, V, Z_1, \ldots, Z_{2n}$ are independent standard normal values.

(a) Calculate the covariance matrix of $X_1, \ldots, X_{2n}$ analytically.

**Solution:**

To be extremely thorough about it, you can consider six cases:

$i = j (i, j \leq n)$:

$$\text{cov}(X_i, X_j) = \text{cov}(U + Z_i, U + Z_j)$$

$$= \text{cov}(U, U) + \text{cov}(U, Z_i) + \text{cov}(U, Z_j) + \text{cov}(Z_i, Z_j)$$

$$= 1 + 0 + 0 + 1$$

$$= 2$$

$i = j (i, j > n)$:

$$\text{cov}(X_i, X_j) = \text{cov}(V + Z_i, V + Z_j)$$

$$= \text{cov}(V, V) + \text{cov}(V, Z_i) + \text{cov}(V, Z_j) + \text{cov}(Z_i, Z_j)$$

$$= 1 + 0 + 0 + 1$$

$$= 2$$

$i \neq j (i, j \leq n)$:

$$\text{cov}(X_i, X_j) = \text{cov}(U + Z_i, U + Z_j)$$

$$= \text{cov}(U, U) + \text{cov}(U, Z_i) + \text{cov}(U, Z_j) + \text{cov}(Z_i, Z_j)$$

$$= 1 + 0 + 0 + 0$$

$$= 1$$

1
\[
i \neq j \ (i, j > n):
\]
\[
\text{cov}(X_i, X_j) = \text{cov}(V + Z_i, V + Z_j)
\]
\[
= \text{cov}(V, V) + \text{cov}(V, Z_i) + \text{cov}(V, Z_j) + \text{cov}(Z_i, Z_j)
\]
\[
= 1 + 0 + 0 + 0
\]
\[
= 1
\]

\[
i \neq j \ (i \leq n, j > n):
\]
\[
\text{cov}(X_i, X_j) = \text{cov}(U + Z_i, V + Z_j)
\]
\[
= \text{cov}(U, V) + \text{cov}(U, Z_i) + \text{cov}(V, Z_j) + \text{cov}(Z_i, Z_j)
\]
\[
= 0 + 0 + 0 + 0
\]
\[
= 0
\]

\[
i \neq j \ (i > n, j \leq n):
\]
\[
\text{cov}(X_i, X_j) = \text{cov}(V + Z_i, U + Z_j)
\]
\[
= \text{cov}(U, V) + \text{cov}(U, Z_i) + \text{cov}(V, Z_j) + \text{cov}(Z_i, Z_j)
\]
\[
= 0 + 0 + 0 + 0
\]
\[
= 0
\]

The covariance matrix has this pattern:
\[
\begin{pmatrix}
2 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 \\
1 & 2 & \cdots & 1 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 2 & 1 & 0 & \cdots & 0 \\
1 & 1 & \cdots & 1 & 2 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 2 & \cdots & 1 \\
0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 2 \\
0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 1 \\
\end{pmatrix}
\]
(b) Use simulation to estimate the covariance matrix of $X_1, \ldots, X_{2n}$. You can use the \texttt{cov} function in R, where \texttt{cov(X)} calculates the matrix of covariances between the columns of $X$.

Solution:

Here are two approaches:

\begin{verbatim}
nrep = 1e3
n = 5

## Non-vectorized approach.
X = array(0, c(nrep,2*n))
for (k in 1:nrep) {
    U = rnorm(1)
    V = rnorm(1)
    X[k,] = rnorm(2*n)
    X[k,1:5] = X[k,1:5] + U
    X[k,6:10] = X[k,6:10] + V
}
C1 = cov(X)

## Vectorized approach.
U = rnorm(nrep)
V = rnorm(nrep)
X = rnorm(nrep*2*n)
X = array(X, c(nrep,2*n))
X[,1:n] = X[,1:n] + array(U, c(nrep,n))
X[,,(n+1):(2*n)] = X[,,(n+1):(2*n)] + array(V, c(nrep,n))
C2 = cov(X)
\end{verbatim}

Here is what I get, which agrees well with part (a):

\begin{verbatim}
> round(C1,2)
[1,] 2.04 1.01 1.08 1.02 1.04 -0.10 -0.04 -0.06 0.07 -0.01
[2,] 1.01 2.10 1.02 1.06 1.03 -0.07 -0.05 0.02 0.02 -0.10
[3,] 1.08 1.02 1.99 1.02 1.00 -0.07 -0.03 -0.09 0.01 -0.06
[4,] 1.02 1.06 1.02 1.99 1.00 -0.07 -0.07 -0.04 0.01 -0.02
[5,] 1.04 1.03 1.00 1.00 2.01 -0.18 -0.06 -0.10 -0.04 -0.06
[6,] -0.10 -0.07 -0.07 -0.07 -0.18 2.07 1.02 1.03 1.06 1.07
[7,] -0.04 -0.05 -0.03 -0.02 -0.06 1.02 2.00 0.97 1.06 1.08
[8,] -0.06 0.02 -0.09 -0.04 -0.10 1.03 0.97 2.03 1.00 1.06
[9,] 0.07 0.02 0.01 0.01 -0.04 1.06 1.06 1.00 2.07 1.05
[10,] -0.01 -0.10 -0.06 -0.02 -0.06 1.07 1.08 1.06 1.05 2.11
\end{verbatim}
(c) Based on (a), give an exact expression for $\text{var}(\bar{X})$.

Solution:
Looking at the pattern in part (a), there are $2n$ 2's and $2n^2 - 2n$ 1's, so the variance of $\bar{X}$ is

\[
\frac{(2n^2 + 2n)}{(2n)^2} = \frac{1}{2} + \frac{1}{2n}.
\]

(d) Use simulation to assess your answer to part (c). Assess whether the formula continues to hold when the $U, V,$ and $Z_i$ are standard exponential rather than standard normal.

Solution:

```r
## Number of simulation replications.
nrep = 1e4

## Sample sizes.
N = c(5, 10, 15, 20)

## Storage for the results.
R = array(0, c(4, 3))

## Normal case.
for (k in 1:4) {
  n = N[k]
  U = rnorm(nrep)
  V = rnorm(nrep)
  X = rnorm(nrep*2*n)
  X = array(X, c(nrep, 2*n))
  X[,1:n] = X[,1:n] + array(U, c(nrep, n))
  X[, (n+1):(2*n)] = X[, (n+1):(2*n)] + array(V, c(nrep, n))
  M = apply(X, 1, mean)
  R[k,1] = 1/2 + 1/(2*n) ## The analytic formula.
  R[k,2] = var(M) ## The simulation estimate for normal data.
}

## Exponential case.
for (k in 1:4) {
  U = rexp(nrep)
  V = rexp(nrep)
  X = rexp(nrep*2*n)
  X = array(X, c(nrep, 2*n))
  X[,1:n] = X[,1:n] + array(U, c(nrep, n))
  X[, (n+1):(2*n)] = X[, (n+1):(2*n)] + array(V, c(nrep, n))
}
```
\[ X[, (n+1):(2*n)] = X[, (n+1):(2*n)] + \operatorname{array}(V, c(nrep, n)) \]
\[ M = \operatorname{apply}(X, 1, \text{mean}) \]
\[ R[k, 3] = \text{var}(M) \quad \text{## The simulation estimate for exponential data.} \]

(e) If the \( X_i \) were independent, with the same variances as the \( X_i \) defined here, what would the variance of \( \bar{X} \) be?

Solution:

\[ 2/(2n) = 1/n \]

2. The sample variance is

\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2. \]

For iid normal data with population variance \( \sigma^2 \), the sample variance has mean \( \sigma^2 \) and variance

\[ \frac{2\sigma^4}{n-1}. \]

(a) Derive an approximate confidence interval for \( \sigma^2 \), centered at \( \hat{\sigma}^2 \), based on an iid sample of size \( n \). You should begin by standardizing \( \hat{\sigma}^2 \), then treat this standardized value as a standard normal value.

Solution:

\[ 0.95 \approx P(-1.96 \leq \frac{\hat{\sigma}^2 - \sigma^2}{\sqrt{2\hat{\sigma}^2}} \leq 1.96) \]
\[ = P(\hat{\sigma}^2 - \sqrt{2}\hat{\sigma}^2 \cdot 1.96/\sqrt{n-1} \leq \sigma^2 \leq \hat{\sigma}^2 + \sqrt{2}\hat{\sigma}^2 \cdot 1.96/\sqrt{n-1}) \]

(b) Use simulation to assess the coverage probabilities of your interval when the data are standard normal, with variances 1, 2 and 3.

Solution:

Here is the code:

```
## Number of simulation replications.
nrep = 1e4
## The sample sizes.
```
n = 30

## The variances to consider.
sigma2 = c(1,2,3)

## Storage for the coverage probabilities.
CP = NULL

for (k in 1:3) {

## Generate the data.
X = rnorm(nrep*n, sd=sqrt(sigma2[k]))
X = array(X, c(nrep,n))

## Get the sample variance for each row of X.
V = apply(X, 1, var)

## Construct the CI.
LB = V - sqrt(2)*1.96*V/sqrt(n-1)
UB = V + sqrt(2)*1.96*V/sqrt(n-1)

## Check whether it covers.
CP[k] = mean( (LB<sigma2[k]) & (sigma2[k]<UB) )
}

I got

> CP
[1] 0.9189 0.9171 0.9106

indicating that the coverage is slightly low. This is expected since we plugged in an
estimate of the population variance in the standard error formula.

(c) Assess the coverage probabilities of your interval for data of the form $X_i = cE_i$, where
the $E_i$ are iid standard exponential values. Choose values of $c$ so that \( \text{var}(X_i) \) is either
1, 2, or 3, as above.

Solution:

## Number of simulation replications.
nrep = 1e4

## Sample sizes.
n = 30
## The population variances to consider.
sigma2 = c(1,2,3)

## Storage for the coverage probabilities.
CP = NULL

for (k in 1:3) {
    ## Generate the data.
    X = sqrt(sigma2[k])*rexp(nrep*n)
    X = array(X, c(nrep,n))

    ## Calculate the sample variance of each data set.
    V = apply(X, 1, var)

    ## Construct the interval.
    LB = V - sqrt(2)*1.96*V/sqrt(n-1)
    UB = V + sqrt(2)*1.96*V/sqrt(n-1)

    ## Check whether it covers.
    CP[k] = mean( (LB<sigma2[k]) & (sigma2[k]<UB) )
}

I got
> CP
[1] 0.6981 0.6867 0.6913

The coverage is poor since the standard error formula given in the problem does not hold for the exponential distribution.

3. Using the NHANES data (see problem set 2), construct 95\% confidence intervals for the data standard deviation $\sigma$ of the log transformed BMI variable, separately for females and males within each 5 year age stratum. To construct a CI for $\sigma$, first construct a CI for $\sigma^2$ of the form $\hat{\sigma}^2 \pm c$, as in problem 2. The CI for $\sigma$ is $\sqrt{\hat{\sigma}^2 - c}, \sqrt{\hat{\sigma}^2 + c}$. Briefly state any conclusion you can draw about the relationship between gender and the variability of BMI.

Solution:

Here is the code:

Z = read.table('NHANES-1', header=TRUE, row.names=1)
BM = Z[,3]  ## Body mass
AG = Z[,1]  ## Age
GD = Z[,4]  ## Gender

age = 20  ## Starting age
k = 1
M = array(0, c(13,8))

## Loop over the age slices.
while (age <= 80) {

    ## Female indices in this age slice.
    fe = which( (AG>=age) & (AG<age+5) & (GD == 2) )

    ## Male indices in this age slice.
    ma = which( (AG>=age) & (AG<age+5) & (GD == 1) )

    ## Log transform the body mass data.
    FD = log(BM[fe])
    MD = log(BM[ma])

    ## Get the sample variances for men and women.
    MV = var(MD)
    FV = var(FD)

    ## Get the standard errors.
    MSE = sqrt(2)*1.96*MV/sqrt(length(ma)-1)
    FSE = sqrt(2)*1.96*FV/sqrt(length(fe)-1)

    ## Save everything we need into the table.
    M[k,] = c(length(fe), length(ma), sqrt(MV), sqrt(FV), sqrt(MV-MSE),
             sqrt(MV+MSE), sqrt(FV-FSE), sqrt(FV+FSE))

    ## Prepare for the next slice.
    k = k+1
    age = age + 5
}

This is what I got:

> round(M,2)
[1,] 8
There is a clear pattern of declining BMI with age, with the female mean being consistently greater than the male mean. The confidence intervals indicate that the age and gender-specific mean BMI's can be estimated within approximately 0.02 log-scale units.