Practice exam problems

1. What will be the approximate value of `M` following execution of each of the following programs?

   (a)

```
X <- array(rnorm(1000), c(10,1000))
A <- apply(X, 2, mean)
M <- mean(A)
```

   Solution: The answer is zero.

   Reasoning: Each element of `A` is the average of 10 iid standard normal values. The expected value of each element of `A` is the same as the expected value of each element of `X`, which is 0. Since `M` is the average of 1000 random values with expected value 0, `M` also has expected value 0. Moreover, `M` has very small variance (1/10000), so the observed value of `M` is likely to be very close to zero.

   (b)

```
X <- array(rnorm(1000), c(10,1000))
A <- apply(X, 2, var)
M <- mean(A)
```

   Solution: The answer is 1.

   Reasoning: Since the individual X_{ij} values have variance 1, the sample variance of any subset of them will be approximately 1. Thus the values in `A` will fluctuate around 1, and their mean will be very close to 1.

   (c)

```
X <- array(rnorm(1000), c(10,1000))
A <- apply(X, 2, mean)
M <- var(A)
```

   Solution: The answer is 1/10.

   Reasoning: The variance of each element of `A` is $\sigma^2/n = 1/10$ since we are averaging 10 independent values from a standard normal population. Since `M` is the sample variance of everything in `A`, and the sample variance estimates the population variance very accurately for large sample sizes (1000 in this case), `M` will be very close to 1/10.

(d)

```
X <- array(rnorm(1000), c(10,1000))
Y <- array(rnorm(2000), c(20,1000))
A <- apply(X, 2, mean)
B <- apply(Y, 2, mean)
M <- sd(A)/sd(B)
```

Solution: The answer is $\sqrt{2}$.

Reasoning: Each element of `A` has mean 0 and variance 1/10. Therefore `SD(A)` is approximately $1/\sqrt{10}$. Each element of `B` has mean 0 and variance 1/20. Therefore `SD(B)` is approximately $1/\sqrt{20}$. The ratio is $(1/\sqrt{10})/(1/\sqrt{20}) = \sqrt{20}/\sqrt{10} = \sqrt{2}$.

2. Do you expect `M` to be positive or negative after running the following program?

```
X <- rexp(10000)
A <- 1/mean(X)
B <- mean(1/X)
M <- A-B
```

Solution: The reciprocal function $f(x) = 1/x$ is convex (for $x > 0$). Therefore $f(EX) \leq Ef(X)$. Since $B \approx Ef(X)$ and $A \approx f(EX)$, we expect M to be negative.

3. If `X` is a vector of `n` observations, what does the following program do?

```
ii <- ceiling(n*runif(1000*n))
B <- X[ii]
B <- array(B, c(n, 1000))
S <- apply(B, 2, sd)
S <- sort(S)
C <- c(S[25], S[975])
```

Solution: This program produces a non-parametric bootstrap 95% CI for the standard deviation of X.

4. If `X` is a vector of `n` observations, what does the following program do?

```
Y <- sort(X)
Y <- Y[6:(n-5)]
M <- mean(Y)
```

Solution: This program calculates the trimmed mean of X, trimming 5 observations from each end.

5. Suppose we wish to estimate the bias, variance, and MSE of the sample variance for standard exponential data of sample size 15 using simulation. What is the appropriate code to put where the `***1***`, `***2***`, and `***3***` are placed below?

```
X <- rexp(15*1000)
X <- array(X, c(15,1000))
U <- apply(X, 2, var)
B <- ***1***
V <- ***2***
MSE <- ***3***
```

Solution:

```
B <- mean(U) - 1
V <- var(U)
MSE <- mean( (U-1)^2 )
```

Now we can check the identity MSE = $\text{bias}^2 + \text{var}$:

```
> B^2 + V
[1] 0.4997875
> MSE
[1] 0.4992879
```

6. Approximately what value will `A` have after executing the following program?

```
X <- rnorm(20*1000)
X <- array(X, c(20,1000))
M <- sqrt(20)*apply(X, 2, mean)
A <- var(M)
```

Solution: Each element of `apply(X, 2, mean)` has mean zero and variance 1/20. Multiplying by `sqrt(20)` yields values with mean zero and variance 1. Thus A will be approximately 1.

7. Suppose we wish to estimate the variance of the sample variance for a set of 10 standard exponential draws. What should `***` be replaced with in the following program so that the value of `V` estimates the quantity of interest?

```
X <- array(rexp(10*1000), c(10, 1000))
M <- apply(X, 2, var)
V <- ***
```

Solution: `V <- var(M)`

8. Approximately what value will `f` have after running the following program?

```
f <- 0
for (r in (1:1000))
{
  X <- rnorm(10)
  m <- mean(X)
  g <- 1.96/sqrt(10)
  if ( (m+g > 0) && (m-g < 0) )
  {
    f <- f+1
  }
}
```

Solution `f` is the number of times that the CI $\bar{X} \pm 1.96/\sqrt{10}$ covered zero. Since zero is the true mean and one is the true variance, this is an exact confidence interval (i.e. all the assumptions underlying the interval are met). Therefore f should be close to 950.

9. Suppose X has a normal distribution with mean 0 and variance 2. Which percentile of the standard normal distribution is equal to the 95^{th} percentile of X?

Solution: Start with the fact that

$$P(X \leq T) = 0.95$$

where T is the 95^{th} percentile of X. Next we standardize, yielding

$$P(X/\sqrt{2} \leq T/\sqrt{2}) = 0.95.$$

Since $X/\sqrt{2}$ has mean zero and variance 1, and is scaled from a normal distribution, it is a standard normal random variable. Therefore $T/\sqrt{2} = 1.64$, so $T = \sqrt{2} \cdot 1.64 \approx 2.32$. Next we find (using R or a normal probability table) that $P(Z \leq 2.32) = 0.9898$. Therefore the 95^{th} percentile of X is the 98.98 percentile of the standard normal distribution.

10. What is the approximate numerical value of `V` after running the following program?

```
X <- array(rnorm(10*1000), c(10,1000))
for (k in (1:5))
{
  X[k,] <- X[k,] * sqrt(k)
}
M <- apply(X, 2, mean)
V <- var(M)
```

Solution: The rows of X all have mean zero, with variances $1, 2, 3, 4, 5, 1, 1, 1, 1, 1$. The average variance is 2. From the notes we have that

$$\mathrm{var}(\bar{X}) = \bar{\sigma}^2/n$$

for independent observations with different variances, where σ_j^2 is the variance of the j^{th} term in the average, and $\bar{\sigma}^2$ is the average of the σ_j^2 values. Therefore in this case the variance of $\bar{X}$ is $2/10 = 0.2$.

11. Suppose A and B are standard normal, C is normal with mean 0 and variance 2, and A, B , and C are independent. What is the value of $\mathrm{cov}(A + B, A + C)$?

Solution:

$$\begin{aligned}
\mathrm{cov}(A + B, A + C) &= \mathrm{cov}(A, A) + \mathrm{cov}(A, C) + \mathrm{cov}(B, A) + \mathrm{cov}(B, C) \\
&= \mathrm{var}(A) \\
&= 1.
\end{aligned}$$

12. Is the following statement true or false: if the MSE of an estimator is zero, it must be unbiased.

Solution: Since MSE = $\mathrm{bias}^2 + \mathrm{var}$, and all the numbers are positive, then when MSE is zero, the bias must also be zero.

13. Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are both estimators of θ, and $\hat{\theta}_1$ is unbiased and has lower variance than $\hat{\theta}_2$. True or false: $\hat{\theta}_1$ has lower MSE than $\hat{\theta}_2$?

Solution: Since MSE = $\mathrm{bias}^2 + \mathrm{var}$, if $\hat{\theta}_1$ has no bias and lower variance than $\hat{\theta}_2$, it must also have lower MSE.

14. Suppose we observe n iid draws $X_1, \ldots, X_n$ and use $\bar{X}$ to estimate EX. What is the MSE?

    Solution: Since $\bar{X}$ is unbiased for EX, the variance and the MSE are the same. Therefore the MSE is σ^2/n.

15. Suppose we observe 10 values from an AR(1) distribution with $\alpha = 0.5$, and the value of τ is set so that $\mathrm{var}(X_i) = 1$ for each i. Is $\mathrm{var}(\bar{X})$ less than, greater than, or equal to 0.1?

    Solution: For an AR(1) process with $\alpha > 0$, every pair of values X_i, X_j has a positive covariance. Since the variances are 1, we have

$$\begin{aligned}
\mathrm{var}(\bar{X}) &= \sum_{ij} C_{ij}/n^2 \\
&= (10 + \sum_{i\neq j} C_{ij})/100 \\
&= 0.1 + \sum_{i\neq j} C_{ij})/100 \\
&\geq 0.1.
\end{aligned}$$

16. Suppose we generate confidence intervals for the expected value when the variance is known and the distribution is known to be normal, using the usual formula $\bar{X} \pm 1.96\sigma/\sqrt{n}$. If we compare the coverage probabilities for sample size 10 and sample size 20, will the coverage probabilities differ, or be approximately the same? If you believe they will differ, state which sample size is expected to have greater coverage probability.

    Solution: Regardless of the sample size, the assumptions of the confidence interval are met. Therefore both will have the nominal coverage probability. However, the interval based on sample size 20 will be narrower.

17. State two reasons why the confidence interval $\bar{X} \pm 1.96\hat{\sigma}/\sqrt{n}$ may not actually cover EX 95% of the time.

    Solution: Here are some reasons: (1) plugging in $\hat{\sigma}$ for σ causes the coverage to be lower than 95%, (2) if the sample size is small, 1.96 is not the right constant to use in the CI, (3) if the X_i have different variances, or are correlated with each other, then the coverage probability is affected.

18. Suppose X and Y are independent and have standard exponential distributions. Which has greater variance, $2X$ or $X + Y$? Do the expected values of $2X$ and $X + Y$ differ?

Solution: $E(2X) = 2EX = 2$, $E(X+Y) = EX+EY = 2$, so the expected values are equal. $\text{var}(2X) = 4\text{var}(X) = 4$, $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) = 2$, so the variances differ. The reason that $X+Y$ has lower variance than $2X$ is that since X and Y are independent, if X is unusually large or small, Y may turn out to be unusually small or large, and therefore the sum falls close to the mean. In the case of $2X$, if X is either unusually small or unusually large, $2X$ will always be unusually far from its mean.

19. What is the numerical result of the following programs? The result may be a scalar or a vector. For *k*-*m* give the expected numerical result.

(a)

```
x <- 1

for (t in c(3,5,6,2))
{
  x <- x*t
}
```

Solution: $1 \cdot 3 \cdot 5 \cdot 6 \cdot 2 = 180$.

(b)

```
x <- 0

for (s in 1:3)
{
   for (t in 1:3)
   {
      x <- x + s*t
   }
}
```

Solution:

$$1 \cdot 1 + 1 \cdot 2 + 1 \cdot 3 + 2 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 3 \cdot 1 + 3 \cdot 2 + 3 \cdot 3 = 36.$$

(c)

```
x <- 1

while (x < 100)
{
  x <- x+x
}
```

Solution: Here is the sequence of values that would be obtained:

```
1
2
4
8
16
32
64
128
```

At this point the program would break since $128 > 100$, so the final result for x would be $x = 128$.

(d)

```
v <- (1:3)
x <- sum(v)
```

Solution: $1 + 2 + 3 = 6$.

(e)

```
A <- array(2, c(3,5))
B <- apply(A, 2, sum)
```

Solution: Since `A` is a 3×5 array consisting of all 2's, it follows that every column sum is $2 + 2 + 2 = 6$. Therefore the value of B is `6 6 6 6 6`.

(f)

```
v <- (1:3)
A <- array(v, c(3,5))
B <- apply(A, 2, sum)
```

Solution: The value of A is

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    1    1    1    1
[2,]    2    2    2    2    2
[3,]    3    3    3    3    3
```

The column sums are: `6 6 6 6 6`.

(g)

```
v <- (1:3)
A <- array(v, c(5,3))
B <- apply(A, 2, sum)
```

Solution: The array `A` is

```
> A
     [,1] [,2] [,3]
[1,]    1    3    2
[2,]    2    1    3
[3,]    3    2    1
[4,]    1    3    2
[5,]    2    1    3
```

So the column sums are `9, 10, 11`.

(h)

```
v <- (1:15)
A <- array(v, c(3,5))
B <- exp(apply(log(A), 2, sum))
```

Solution: Since

$$\exp(\log(A) + \log(B) + \log(C))) = A \cdot B \cdot C$$

it follows that `B` contains the column-wise products of the values in `A`. The values in `A` are

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
```

Solution: Therefore the values in `B` are `6 120 504 1320 2730`.

(i)

```
v <- (1:15)
A <- array(v, c(3,5))
v <- (16:30)
B <- array(v, c(3,5))
C <- A + B
```

Solution: The matrices `A`, `B`, and `C` are as follows:

```
> A
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    4    7   10   13
[2,]    2    5    8   11   14
[3,]    3    6    9   12   15
> B
     [,1] [,2] [,3] [,4] [,5]
[1,]   16   19   22   25   28
[2,]   17   20   23   26   29
[3,]   18   21   24   27   30
> C
     [,1] [,2] [,3] [,4] [,5]
[1,]   17   23   29   35   41
[2,]   19   25   31   37   43
[3,]   21   27   33   39   45
```

(j)

```
v <- (1:100)
n <- sum(v/3 == round(v/3))
```

Solution: `n` is a count of the integers between 1 and 100 that are evenly divisible by 3. There are 33 such integers $(1 \cdot 3, 2 \cdot 3, \ldots, 33 \cdot 3)$.

(k)

```
X <- rnorm(1000) + 1
M <- mean(X^2)
```

Solution: The variance is

$$\begin{aligned} E(Z+1)^2 &= E(Z^2 + 2Z + 1) \\ &= EZ^2 + 2EZ + 1 \end{aligned}$$

and for the standard normal distribution, $EZ = 0$ and $EZ^2 = \text{var}(Z) = 1$. Therefore $E(Z+1)^2 = 2$. Since M is the average of many simulated values of $(Z+1)^2$, M will be very close to 2.

(l)

```
X <- rexp(1000)
Y <- rexp(1000)
M <- mean(X*Y)
```

Solution: For independent random variables, $EXY = EX \cdot EY$. In this case, $EX = EY = 1$. Therefore $EXY = 1$. Since M is the average of many simulated values of XY, M will be very close to 1.

(m)

```
X <- rnorm(1000)
Y <- 10*rnorm(1000)
B <- (runif(1000) < 0.05)
Z <- (1-B)*X + B*Y
M <- mean(Z)
```

Solution: `Z` consists of 1000 values, each one of which is generated from either a standard normal distribution (`X`), or from a normal distribution with mean 0 and variance 100 (`Y`). The mean of `Z` is therefore 0, since the fact that $E\bar{X} = EX_i$ does not depend on the variances being equal.

(n)

```
v <- array((1:100), c(2,50))
w <- apply(v, 2, min)
```

Solution: The `v` matrix is:

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1    3    5    7    9   11   13
[2,]    2    4    6    8   10   12   14
```

Therefore `w` is `1 3 5 7 9 ...`

(o)

```
v <- array((1:100), c(2,50))
w <- apply(v, 1, sum)
```

Solution: The `v` matrix is the same as the previous problem. The overall sum of `v` is $10 \cdot 101/2 = 5050$. Since the second row must be 50 greater than the first row, we have an equation $f + (f + 50) = 5050$, where f is the sum of the first row. Solving for f yields $f = 2500$. Therefore the result is `2500`, `2550`.

(p)

```
v <- array((1:10), c(5,50))
w <- apply(v, 1, mean)
```

Solution: The first 10 columns of the `v` matrix are

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    6    1    6    1    6    1    6    1     6
[2,]    2    7    2    7    2    7    2    7    2     7
[3,]    3    8    3    8    3    8    3    8    3     8
[4,]    4    9    4    9    4    9    4    9    4     9
[5,]    5   10    5   10    5   10    5   10    5    10
```

The row means are 3.5, 4.5, 5.5, 6.5, and 7.5.