

# Statistics 406 Midterm Exam

Fall 2006

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

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

1. The four R code fragments below include exactly one correct evaluation of the binomial probability

$$P(X = k) = \binom{10}{k} 2^{-10}, \quad k = 0, 1, \dots, 10.$$

Identify the correct implementation.

1

```
p <- -10*log(2)
if ( (k>0) & (k<10) )
{
  p <- p + sum(log(seq(1,10))) - sum(log(seq(1,k))) - sum(log(seq(1,10-k)))
}
p <- exp(p)
```

2

```
p <- 2^(-10)
if ( (k>0) & (k<10) )
{
  p <- p + sum(log(seq(1,10))) - sum(log(seq(1,k))) - sum(log(seq(1,10-k)))
}
p <- exp(p)
```

3

```
p <- -2*log(10)
if ( (k>0) & (k<10) )
{
  p <- p + sum(log(seq(1,10))) - sum(log(seq(1,10-k)))
}
p <- exp(p)
```

4

```
p <- -10*log(2)
if ( (k>0) & (k<10) )
{
  p <- p + log(sum(seq(1,10))) - log(sum(seq(1,k))) - log(sum(seq(1,10-k)))
}
p <- exp(p)
```

**Solution:** The correct answer is 1.

2. Suppose RD is an R function that generates iid data from a particular distribution. To solve this problem you don't need to know anything more about RD except that it has a finite mean and variance. Approximately what values will  $V1/V2$ ,  $V1/V3$ , and  $V2/V3$  have after running the following code?

```
Z1 <- array(RD(40*1000), c(40,1000))
Z1 <- apply(Z1, 2, mean)
V1 <- var(Z1)
```

```
Z2 <- array(RD(80*1000), c(80,1000))
Z2 <- apply(Z2, 2, mean)
V2 <- var(Z2)
```

```
Z3 <- sqrt(3)*array(RD(40*1000), c(40,1000))
Z3 <- apply(Z3, 2, mean)
V3 <- var(Z3)
```

**Solution:** If  $\sigma^2$  is the population variance of RD, then  $V1 \approx \sigma^2/40$ ,  $V2 \approx \sigma^2/80$ , and  $V3 \approx 3\sigma^2/40$ . Therefore  $V1/V2 \approx 2$ ,  $V1/V3 \approx 1/3$ , and  $V2/V3 \approx 1/6$ .

3. Approximately what value will  $u$  have after running the following program? Explain your reasoning.

```
u <- 0
for (r in (1:1000))
{
  X <- rnorm(10)
  m <- mean(X)
  s <- sd(X)
  q <- qt(0.975, 9)
  g <- q*s/sqrt(10)
  if ( (m+g < 0) | (m-g > 0) )
  {
    u <- u+1
  }
}
```

**Solution**  $u$  is the number of times that the 95% CI does not cover the expected value of  $X$  (zero). This is an exact confidence interval (all the assumptions underlying the interval are met). Therefore  $u$  should be close to  $0.05 \cdot 1000 = 50$ .

4. A simulation was carried out to estimate the actual level of the two-sided, two-sample t-test with nominal level 0.05. The sizes of the two samples will be denoted  $n_1$  and  $n_2$ . The variance of the population from which the first sample was drawn was always 1. The second sample was drawn from a population with variance  $\sigma_Y^2$ . The estimated levels from the simulation were as follows.

$n_1$	$n_2$	$\sigma_Y^2$							
		1/5	1/3	1/2	1	2	3	5	
10	10	0.0579	0.0549	0.0512	0.0505	0.0524	0.0514	0.0522	
20	10	0.0140	0.0195	0.0263	0.0433	0.0708	0.0876	0.1152	
100	10	0.0001	0.0012	0.0066	0.0369	0.1188	0.1788	0.2545	

What do you conclude from this simulation?

**Solution:** If the sample sizes are equal, or if the population variances are equal, the actual coverage and nominal coverage are quite similar. If the sample sizes are quite different and the population variances are quite different, the actual coverage and nominal coverage can disagree substantially. If the smaller group is less variable than the larger group, the actual coverage is less than 0.05. If the smaller group is more variable than the larger group, the actual coverage is greater than 0.05.

5. Suppose we have a population “ $X$ ” and a population “ $Y$ ,” both with the same expected value, which we will denote  $E$ . We do not know that the two populations have the same variance. Based on iid samples from the two populations, possibly of different sizes, point estimates  $\bar{X} = 7$  and  $\bar{Y} = 6$  of  $E$  were obtained. Then 95% confidence intervals  $7 \pm 1$  and  $6 \pm 2$  for  $E$  were constructed, using the “approximate Z” approach; e.g. for the  $X$  data the CI is  $\bar{X} \pm 2\hat{\sigma}_X/\sqrt{n_X}$  where  $\hat{\sigma}_X$  is the sample standard deviation, and  $n_X$  is the sample size.

- (a) Based on the information provided above, what is the approximate value of

$$V = \text{var} \left( \frac{\bar{X} + \bar{Y}}{2} \right)?$$

**Solution:** Since  $2\hat{\sigma}_X/\sqrt{n_X} = 1$ ,  $\hat{\sigma}_X^2/n_X = 1/4$ . Similarly, since  $2\hat{\sigma}_Y/\sqrt{n_Y} = 2$ ,  $\hat{\sigma}_Y^2/n_Y = 1$ . Therefore

$$\begin{aligned} \text{var} \left( \frac{\bar{X} + \bar{Y}}{2} \right) &= (\sigma_X^2/n_X + \sigma_Y^2/n_Y)/4 \\ &\approx (\hat{\sigma}_X^2/n_X + \hat{\sigma}_Y^2/n_Y)/4 \\ &= 5/16. \end{aligned}$$

- (b) Based on your answer to part (a), construct an (approximate) 95% CI for  $E$  centered at  $(\bar{X} + \bar{Y})/2$ . You can express your answer in terms of  $V$ , so that you may get full credit for this part regardless of your answer for part a.

**Solution:** The CI is  $6.5 \pm 1.96\sqrt{V}$ .

6. For each of the following short R programs, give the numerical value of  $z$  after the program is run, or in the case of a program involving random values, give the expected numerical value of  $z$  after the program is run.

(a)

```
n <- 0
s <- 0
for (k in 5:9)
{
  s <- s+k
  n <- n+1
}
z <- s/n
```

**Solution:**  $z = 7$

(b)

```
M <- array(rnorm(10*1000), c(10,1000))
U <- apply(M, 2, var)
z <- mean(U)
```

**Solution:**  $Ez = 1$

(c)

```
M <- array(rnorm(10*1000), c(10,1000))
M[1,] <- M[1,] * sqrt(6)
U <- apply(M, 2, mean)
z <- var(U)
```

**Solution:**  $Ez = \bar{\sigma}^2/n = 1.5/10 = 0.15$ .

7. Suppose we observe a set of values, that are then placed into the vector  $X$ . Describe concisely what the following R program is doing. Be sure to discuss both the statistical goal, as well as the method that is being used to achieve the goal.

```
p <- length(X)
mu <- mean(X)
s <- sd(X)
B <- mu + s*rnorm(1000*p)
B <- array(B, c(p,1000))
M <- apply(B, 2, mean)
M <- sort(M)
F <- c(M[50], M[950])
```

**Solution:** This program calculates a parametric bootstrap 90% confidence interval for the expected value of  $X$ , using a normal model for generating the parametric bootstrap data.

8. Suppose we are interested in using the sample standard deviation  $\hat{\sigma}$  to estimate the population standard deviation, based on dependent data. We will focus on the mean squared error (MSE) to evaluate the performance of  $\hat{\sigma}$ .

We will consider two working models for dependent data. To begin with, suppose we have eight iid standard exponential values  $X_1, \dots, X_8$ . Recall that the standard exponential distribution has population variance 1.

For working model 1, we base our estimate of  $\hat{\sigma}$  on the data set

$$X_1, X_2, X_3, X_4, X_5, X_5, X_5.$$

For working model 2, we base our estimate of  $\hat{\sigma}$  on the data set

$$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_7.$$

- (a) In the following R program, what expressions should **\*\*1\*\*** and **\*\*2\*\*** be replaced with to estimate the MSE's for the two models?

```
MSE1 <- 0
MSE2 <- 0
for (rep in 1:1000)
{
  X <- rexp(8)
  S1 <- sd(c(X[1], X[2], X[3], X[4], X[5], X[5], X[5], X[5]))
  S2 <- sd(c(X[1], X[2], X[3], X[4], X[5], X[6], X[7], X[7]))
  MSE1 <- **1**
  MSE2 <- **2**
}
MSE1 <- MSE1/1000
MSE2 <- MSE2/1000
```

**Solution:** **\*\*1\*\*** should be replaced with  $MSE1 + (S1-1)^2$ . **\*\*2\*\*** should be replaced with  $MSE2 + (S2-1)^2$ .

- (b) Which MSE do you expect to be larger? You will not be able to derive or prove this mathematically, but rather you should reason by analogy with other findings discussed in the course. Briefly state your reasoning.

**Solution:** The covariances between duplicated data points are 1, while the covariances between independent data points are zero. We know that the variance of  $\bar{X}$  increases directly with the sum of covariances between pairs of observations. We expect the same to happen with the sample variance. Therefore MSE1 will be larger.