

## Statistics 406 Problem Set 3

Due in lab, Tuesday October 2

1. Suppose  $X$  is uniform on  $(c, 1 + c)$ , and we use the approximation

$$\text{var} f(X) \approx \text{var}(X) f'(EX)^2$$

to approximate  $\text{var} \log(X)$  for  $c = 0, 0.1, 0.5, 1, 2$ . For each value of  $c$ , calculate the approximate result, and an unbiased estimate of the exact value using simulation. Also calculate the “relative error”, which is  $|(A - E)/E|$ , where  $A$  is the approximate result and  $E$  is the exact result from the simulation. Note that the distribution of  $X$  is the same as the distribution of  $U + c$ , where  $U$  is uniform on  $(0, 1)$ . Briefly describe your findings.

### Solution:

The expected value of  $X$  is  $0.5 + c$ , the variance of  $X$  is  $1/12$ , and  $f'(X) = 1/X$ . Therefore the mathematical approximation is

$$\text{var} \log(X) \approx \frac{1}{12(c + 1/2)^2}.$$

Here is the code:

```
F = array(0, c(5,3))
C = c(0, 0.1, 0.5, 1, 2)

for (k in 1:length(C)) {
  c = C[k]
  F[k,1] = 1/(12*(c+1/2)^2)
  F[k,2] = var(log(runif(1e4, min=c, max=1+c)))
  F[k,3] = abs(F[k,2]-F[k,1]) / F[k,2]
}
```

Here is the result, showing that the relative error goes down as the value of  $c$  increases. The reason for this is that the log function is better approximated by a quadratic function at points further from zero, where log has a singularity.

```
> F
      [,1]      [,2]      [,3]
[1,] 0.33333333 1.00280380 0.66759865
[2,] 0.23148148 0.36252127 0.36146787
[3,] 0.08333333 0.09452180 0.11836920
[4,] 0.03703704 0.03881198 0.04573194
[5,] 0.01333333 0.01361756 0.02087203
```

2. We looked at the crude approximation  $Ef(X) \approx f(EX)$  in class. A more accurate approximation can usually be obtained using a higher order Taylor expansion of  $f$ :

$$f(X) \approx f(X_0) + (X - X_0)f'(X_0) + (X - X_0)^2 f''(X_0)/2.$$

If we set  $X_0 = EX$  and take the expected value of both sides we get

$$Ef(X) \approx f(EX) + \text{var}(X)f''(EX)/2.$$

Suppose our goal is to approximate  $E \exp(c \cdot X)$ , where  $X$  is uniform on  $(0, 1)$ , and  $c = 1, 2, 4, 8, 10$ . Generate a  $5 \times 3$  table containing the exact value of  $Ef(X)$  (estimated using simulation), the crude approximation  $f(EX)$ , and the higher order approximation given above, for each value of  $c$ .

**Solution:**

Here is the code:

```
F = array(0, c(5,3))
C = c(1, 2, 4, 8, 10)

for (k in 1:length(C)) {
  c = C[k]
  F[k,1] = mean(exp(runif(1e4)*c))
  F[k,2] = exp(c/2)
  F[k,3] = exp(c/2) + c^2*exp(c/2)/24
}
```

Here is the result, showing that higher order approximation (column 3) is always closer to the truth (column 1) than the crude approximation (column 2). Both approximations perform poorly for larger values of  $c$ , where  $\exp(cx)$  cannot be accurately approximated by a quadratic function.

```
> F
      [,1]      [,2]      [,3]
[1,]  1.718680  1.648721  1.717418
[2,]  3.213516  2.718282  3.171329
[3,] 13.161841  7.389056 12.315093
[4,] 365.608394 54.598150 200.193217
[5,] 2177.915513 148.413159 766.801322
```

3. Suppose we are using the sample median

$$\hat{m} = \text{med}(X_1, \dots, X_n)$$

to estimate the population median from an iid sample  $X_1, \dots, X_n$ . We are interested in the standard deviation of  $\hat{m}$ . The result  $\text{var}(\bar{X}) = \sigma^2/n$  does not apply if we use  $\hat{m}$  in place of  $\bar{X}$ , but we hypothesize that a formula of the form  $\text{var}(\hat{m}) \approx c/n$ , for some constant  $c$ , is at least a good approximation to the variance. Design and implement a simulation study to assess whether this can be done. Consider at least two different distributions, and at least four different sample sizes. Discuss whether the value of  $c$  appears to be approximately independent of the distribution and/or the sample size, and discuss how it compares to  $\sigma^2$ .

**Solution:**

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

## The sample sizes to consider.
SS = c(10,20,40,80)

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

## Consider three distributions for the data.
for (K in 1:3) {

  ## Loop over the sample sizes.
  for (j in 1:length(SS)) {

    ## The current sample size.
    n = SS[j]

    ## Simulate the data.
    if (K == 1) { X = rnorm(nrep*n) }
    else if (K == 2) { X = rexp(nrep*n) }
    else if (K == 3) { X = runif(nrep*n) }
    X = array(X, c(nrep,n))

    ## Estimate the sampling variance of the sample median.
    M = apply(X, 1, median)
    R[j,K] = n*var(M)
  }
}
```

Here is my result:

```
> R
      [,1]      [,2]      [,3]
[1,] 1.362531 0.9802636 0.1876681
[2,] 1.457417 0.9808092 0.2155396
[3,] 1.484117 1.0049008 0.2346541
[4,] 1.542182 1.0032661 0.2419411
```

Since  $n\text{var}(\hat{m})$  is approximately constant, we can conclude that a formula of the form  $\text{var}(\hat{m}) \approx c/n$  holds. The value of  $c$  appears to be independent of the sample size, but evidently it depends strongly on the distribution of the data.