

Statistics 406 Problem Set 2

Due in lab, Tuesday October 10

1. Let X_i be a sequence of *iid* Bernoulli trials with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$, and let

$$A_k = (X_1 + \cdots + X_k)/k$$

be their partial averages. Using simulation, approximate how large k must be so that $Z_k = \sqrt{k}(A_k - EA_k)/\text{sd}(X_i)$ has its 80th, 90th and 95th percentiles within 0.05 of the same percentiles for the standard normal distribution. Use values 0.05, 0.1, and 0.2 for p , and explain the differences among the results.

Solution:

```
for (p in c(0.05, 0.1, 0.2))
{
  k <- 1

  ## The population standard deviation.
  sdk <- sqrt(p*(1-p))

  while (TRUE)
  {
    ## Generate Bernoulli data.
    X <- array(runif(k*1000), c(k,1000))
    X <- 1*(X < p)

    ## Get the sample means in blocks of k.
    A <- apply(X, 2, mean)

    ## The standardized sample means.
    Z <- sqrt(k)*(A - p)/sdk

    ## The quantiles we are monitoring.
    Q1 <- quantile(Z, 0.8)
    Q2 <- quantile(Z, 0.9)
    Q3 <- quantile(Z, 0.95)

    ## Check if we are sufficiently close to standard Gaussian.
    if ( (abs(Q1 - qnorm(0.8)) < 0.05) & (abs(Q2 - qnorm(0.9)) < 0.05) &
```

```

      (abs(Q3 - qnorm(0.95)) < 0.05) ) { break }

    ## If not close enough, increase the sample size and try again.
    k <- k+1
  }
  print(k)
}

```

My results are:

```

[1] 120
[1] 61
[1] 35

```

Smaller values of p produce more skewed values of the sample mean. Therefore it takes longer to reach approximate normality.

2. Suppose we generate a sequence of random variables as follows. First, let

$$\epsilon_1, \dots, \epsilon_{n+2}$$

be independent normal random variables with mean 0 and variance 1. Then let $X_1 = \epsilon_1 - \epsilon_2 + \epsilon_3$, $X_2 = \epsilon_2 - \epsilon_3 + \epsilon_4$, and so on.

- (a) What is the expected value and variance of each X_i ?

Solution: Expectations always add, so

$$EX_i = E\epsilon_i - E\epsilon_{i+1} + E\epsilon_{i+2} = 0$$

Variances add if the terms are independent. Since the ϵ_i are independent,

$$\text{var}(X_i) = \text{var}(\epsilon_i) + \text{var}(\epsilon_{i+1}) + \text{var}(\epsilon_{i+2}) = 3$$

- (b) Are the X_i identically distributed? Briefly explain your reasoning.

Solution: Since the ϵ_i are *iid* and each X_i is formed from the ϵ_i in the same way, it follows that the X_i are identically distributed.

- (c) Are the X_i independent? Briefly explain your reasoning.

Solution: Intuitively, a consecutive pair X_i, X_{i+1} should be dependent since they both contain two of the same ϵ 's. On the other hand, X_i and X_j should be dependent if $|i - j| > 2$ since in that case four distinct ϵ values are involved.

(d) What is the value of $\text{cov}(X_i, X_j)$ for each pair of indices i, j ?

Solution: There are four cases:

- If $i = j$ then $\text{cov}(X_i, X_j) = \text{var}(X_i) = 3$, as calculated above.
- If $j = i + 1$, or $i = j + 1$,

$$\begin{aligned} \text{cov}(X_i, X_{i+1}) &= \text{cov}(\epsilon_i - \epsilon_{i+1} + \epsilon_{i+2}, \epsilon_{i+1} - \epsilon_{i+2} + \epsilon_{i+3}) \\ &= \text{cov}(\epsilon_i, \epsilon_{i+1}) - \text{cov}(\epsilon_i, \epsilon_{i+2}) + \text{cov}(\epsilon_i, \epsilon_{i+3}) \\ &\quad - \text{cov}(\epsilon_{i+1}, \epsilon_{i+1}) + \text{cov}(\epsilon_{i+1}, \epsilon_{i+2}) - \text{cov}(\epsilon_{i+1}, \epsilon_{i+3}) \\ &\quad + \text{cov}(\epsilon_{i+2}, \epsilon_{i+1}) - \text{cov}(\epsilon_{i+2}, \epsilon_{i+2}) + \text{cov}(\epsilon_{i+2}, \epsilon_{i+3}) \\ &= -\text{var}(\epsilon_{i+1}) - \text{var}(\epsilon_{i+2}) \\ &= -2. \end{aligned}$$

- If $j = i + 2$ or $i = j + 2$,

$$\begin{aligned} \text{cov}(X_i, X_{i+2}) &= \text{cov}(\epsilon_i - \epsilon_{i+1} + \epsilon_{i+2}, \epsilon_{i+2} - \epsilon_{i+3} + \epsilon_{i+4}) \\ &= \text{cov}(\epsilon_i, \epsilon_{i+2}) - \text{cov}(\epsilon_i, \epsilon_{i+3}) + \text{cov}(\epsilon_i, \epsilon_{i+4}) \\ &\quad - \text{cov}(\epsilon_{i+1}, \epsilon_{i+2}) + \text{cov}(\epsilon_{i+1}, \epsilon_{i+3}) - \text{cov}(\epsilon_{i+1}, \epsilon_{i+4}) \\ &\quad + \text{cov}(\epsilon_{i+2}, \epsilon_{i+2}) - \text{cov}(\epsilon_{i+2}, \epsilon_{i+3}) + \text{cov}(\epsilon_{i+2}, \epsilon_{i+4}) \\ &= \text{var}(\epsilon_{i+2}) \\ &= 1. \end{aligned}$$

- If $|i - j| > 2$, there are no overlapping ϵ_i terms, so the covariance is zero.

(e) Derive a formula for $\text{var}\bar{X}$. How does your result compare to the case where the X_i are iid X_i with the same mean and variance as you found in part (a)?

Solution: We need to sum all the covariances calculated earlier, so as to apply the formula

$$\text{var}(\bar{X}) = \sum C_{ij}/n^2.$$

Referring to the four cases above, there are n 3's, $2(n - 1)$ -2's, and $2(n - 2)$ 1's. Thus we get

$$\sum C_{ij} = 3n - 4(n - 1) + 2(n - 2) = n,$$

so $\text{var}(\bar{X}) = 1/n$.

If we had iid data with variance 3, then $\text{var}(\bar{X})$ would be $3/n$. In this case, the variance of \bar{X} is lower than the comparable situation with independent data. This is due to the predominance of negative correlations among the terms in the sum.

(f) Write a simulation to assess the accuracy of your formula.

Solution:

```
VE <- NULL
VT <- NULL

## Consider a range of four different sample sizes.
for (n in c(5,10,20,40))
{
  ## Generate the epsilon values.
  E = array(rnorm(1000*(n+2)), c(n+2, 1000))

  ## Generate the X values. Each column of X is a replicated data set
  ## generated using the rule  $X(i) = E(i) - E(i+1) + E(i+2)$ .
  X = array(0, c(n,1000))
  for (k in 1:n) { X[k,] <- E[k,] - E[k+1,] + E[k+2,] }

  ## Get the sample means of each simulated set.
  M <- apply(X, 2, mean)

  ## Save the estimated variance from the simulation in VE, and the
  ## theoretical variance from the formula in VT.
  VE <- c(VE, var(M))
  VT <- c(VT, 1/n)
}
```

I get the following, showing a strong consistency between the simulation and theoretical results.

```
> VE
[1] 0.19136737 0.09906436 0.04949667 0.02418124
> VT
[1] 0.200 0.100 0.050 0.025
```