2.3 Sampling plan optimization

We wish to estimate the average weight of patients with hypertension in the age group 60 to 80 years, we know that the weight differs considerably according to gender, denote the average weight asμ_1 for men and μ_2 women. The time and money avaiable for this research allow us to include a sample of size 100. We know that the proportion of men and women with hypertension differs in the population, denote the proportion of men by d. We wish to know how to split our sample size between men and women to ensure the smallest possible standard error. Assume that the standard deviation of the weigth of men is larger than the standard deviation of the weight of women by factor k.

- Find an unbiassed estimator of the population mean

  The population mean μ can be expressed as $\mu = d\mu_1 + (1-d)\mu_2$. Since $E(\bar{X}_1) = \mu_1$ in $E(\bar{X}_2) = \mu_2$, the unbiassed estimator of the mean is given by
  $$M = d\bar{X}_1 + (1-d)\bar{X}_2.$$

- Express the standard error using the subsample sizes (use n_1 to denote the number of men and n_2 to denote the number of women in the sample).

  Taking into account that the average weight of women is independent of the average weight of men, we get

  $$\text{var}(M) = d^2 var(\bar{X}_1) + (1-d)^2 var(\bar{X}_2) = d^2\frac{\sigma_1^2}{n_1} + (1-d)^2\frac{\sigma_2^2}{n_2}.$$

- Let $\sigma_1 = k\sigma_2$. Find the subsample sizes that minimizes the standard error. Calculate n_1 for $k = 1$ and $k = 2$, assume that the proportion of men equals $0,7$.

  The variance of the sampling mean can be written as

  $$var(M) = \sigma^2\left(\frac{d^2k^2}{n_1} + \frac{(1-d)^2}{n-n_1}\right),$$

we thus need to minimize the term in the brackets. Take the derivative with respect to n_1 and equal to 0:

$$\begin{aligned} -\frac{d^2k^2}{n_1^2} + \frac{(1-d)^2}{(n-n_1)^2} &= 0 \\ -(n-n_1)^2d^2k^2 + n_1^2(1-d)^2 &= 0 \\ (d^2k^2 - (1-d)^2)n_1^2 - 2nd^2k^2n_1 + n^2d^2k^2 &= 0 \end{aligned}$$

We now solve the quadratic equation to get the solution (there are actually two solutions, but the other one is larger than n and thus makes no sense):

$$n_1 = \frac{ndk}{dk + 1 - d}$$

For $k = 1$, we get $n_1 = nd$, in our example this implies $n_1 = 70$ and $n_2 = 30$. The standard error is minimised when the proportions of men and women in the sample equal the proportions in the population.
If the standard deviation of the weight in men is twice the standard deviation in women, we get $n_1 = 82, 4$. This implies that if one subgroup is more variable than the other, more individuals from that subgroup have to be sampled.

Understanding the ideas in R:

- Choose sensible values for all the parameters and generate data. Graffically show how values of n_1 affect the quality of your estimate for various values of d and k.

2.4 Simple random sample from a finite population, second attempt

Consider again a simple random sample of size n from population N, denote the values in the population as x_i; $i = 1, \ldots, N$, and the population mean and variance as μ and σ^2, respectively. Define the random variable $I_i = I_{[i \text{ is included in the sample}]}$ and write the estimator of the population mean μ as $C = \frac{1}{n}\sum_{i=1}^{N} I_i x_i$.

- What is the sum $\sum_{i=1}^{N} I_i$? What is the probability $P(I_i = 1)$?

  Since the sample is of size n, the sum equals $\sum_{i=1}^{N} I_i = n$. To calculate the probability that the unit i is chosen, we realize that there are $\binom{N}{n}$ combinations to take a sample of size n from a population of size N. The count the number of samples that include the unit i, we take into account that we already know one of the units and that we choose $n-1$ units out of the remaining $N-1$. The probability thus equals:

$$P(I_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

- Show that the estimator C is unbiased.

  We wish to estimate $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$E(C) = \frac{1}{n}\sum_{i=1}^{N} E(I_i)x_i$$

  The variable I_i can take values 0 or 1, hence $E(I_i) = P(I_i = 1) = \frac{n}{N}$ (the sample is random, which implies that the probabilities for all i are equal), and we get

$$E(C) = \frac{1}{n}\sum_{i=1}^{N} \frac{n}{N}x_i = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu.$$

- Calculate $\mathrm{var}(I_i)$ and $cov(I_i, I_j)$.

  The variable I_i is Bernoulli, with probability $P(I_i = 1) = \frac{n}{N}$. Its variance thus equals

$$\mathrm{var}(I_i) = \frac{n}{N}(1 - \frac{n}{N}) = \frac{n}{N}\frac{N-n}{N}$$

  To calculate the covariance, we use the usual idea: since $cov(I_1, I_1 + \ldots I_N) = cov(I_1, n) = 0$ and $cov(I_i, I_j)$ is equal for all $i \neq j$, we get

$$\mathrm{cov}(I_i, I_j) = -\frac{\frac{n}{N}\frac{N-n}{N}}{N-1} = -\frac{n(N-n)}{N^2(N-1)}$$

- Show that the variance of the estimator equals $\text{var}(C) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

$$\begin{aligned}
\text{var}(C) &= \frac{1}{n^2} cov(\sum_{i=1}^{N} I_i x_i, \sum_{j=1}^{N} I_j x_j) \\
&= \frac{1}{n^2} \sum_{i=1}^{N} cov(I_i x_i, \sum_{j=1}^{N} I_j x_j) \\
&= \frac{1}{n^2} \sum_{i=1}^{N} \left[cov(I_i x_i, I_i x_i) + \sum_{j=1, j\neq i}^{N} cov(x_i I_i, I_j x_j) \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{N} \left[x_i^2 cov(I_i, I_i) + \sum_{j=1, j\neq i}^{N} x_i x_j cov(I_i, I_j) \right]
\end{aligned}$$

The population variance is defined as:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu^2 \\
\sum_{i=1}^{N} x_i^2 &= N(\sigma^2 + \mu^2)
\end{aligned}$$

We derive the variance as

$$\begin{aligned}
\mathrm{var}(C) &= \frac{1}{n^2}\sum_{i=1}^{N}\left[x_i^2\mathrm{var}(I_i) - \sum_{j=1,j\neq i}^{N} x_i x_j \frac{\mathrm{var}(I_i)}{N-1}\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\sum_{i=1}^{N}\left[(N-1)x_i^2 - x_i\sum_{j=1,j\neq i}^{N} x_j\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\sum_{i=1}^{N}\left[(N-1)x_i^2 - x_i(\sum_{j=1}^{N} x_j - x_i)\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\sum_{i=1}^{N}\left[(N-1)x_i^2 - x_i(N\mu - x_i)\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\sum_{i=1}^{N}\left[Nx_i^2 - x_i N\mu\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\left[N\sum_{i=1}^{N}x_i^2 - N\mu\sum_{i=1}^{N}x_i\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\left[N^2(\sigma^2+\mu^2) - N^2\mu^2\right] \\
&= \frac{\mathrm{var}(I_i)}{n^2(N-1)}\left[N^2\sigma^2\right] \\
&= \frac{N^2\sigma^2}{n^2(N-1)}\frac{n}{N}\frac{N-n}{N} \\
&= \frac{\sigma^2}{n}\frac{N-n}{N-1}
\end{aligned}$$

2.5 A more complex sampling scheme

We wish to estimate the achievement of Ljubljana pupils on a test written in several countries. We split the population of $N = 2800$ 7th grade pupils by schools ($K = 46$). On the first step, we randomly (independently of the number N_i of pupils in school i) sample $k = 10$ schools, on the second step, we choose a sample of $n = 15$ on each of the 10 schools. Let μ denote the population mean test score and let μ_i denote the mean of each school. The two sampling steps are independent.

- Find an unbiased estimator of μ.

  We first express the overall mean μ with μ_i. Use x_{ij} to denote the value of the j-th pupil on i-th school:

$$\mu = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{K} N_i \cdot \mu_i \tag{1}$$

  Denote the estimated mean of each school as $\bar{X}_i$ and let I_i be an indicator variable that equals 1, if a school is included in the sample. Let our estimator equal

$$\bar{X} = \sum_{i=1}^{K} c_i I_i \bar{X}_i$$

  We wish to determine the value of c_i to ensure an unbiased estimator. A simple random sample was taken within each school, hence $E(\bar{X}_i) = \mu_i$. Since sampling on the two levels is independent, we have $E(I_i\bar{X}_i) = E(I_i)E(\bar{X}_i)$. Since schools were sampled with equal probability, we have $E(I_i) = \frac{k}{K}$ for each i. Using all these results, we get

$$\begin{aligned} E(\bar{X}) &= \sum_{i=1}^{K} c_i E(I_i\bar{X}_i) = \sum_{i=1}^{K} c_i E(I_i)E(\bar{X}_i) \\ &= \sum_{i=1}^{K} c_i \frac{k}{K}\mu_i \end{aligned}$$

  To get (1), we must have $c_i\frac{k}{K} = \frac{N_i}{N}$, our estimator thus equals

$$\bar{X} = \frac{K}{N}\frac{1}{k}\sum_{i=1}^{K} N_i I_i \bar{X}_i.$$

- How would you estimate the population mean if the all schools were of equal size L?

  Since $N = \sum_{i=1}^{K} N_i$, equal $N_i = L$ give $N = KL$ and thus

$$\bar{X} = \frac{1}{L}\frac{1}{k}\sum_{i=1}^{K} L I_i \bar{X}_i = \frac{1}{k}\sum_{i=1}^{K} I_i \bar{X}_i$$

- Is the sample size at each school important for the bias of the estimator?

  No, $\bar{X}_i$ is an unbiased estimator of μ_i regardless of the sample size. The sample size is important for the standard error of our estimator.

- Express the variance of the estimator using the variance and covariance

$$\begin{aligned}\text{var}(\bar{X}) &= \text{var}(\frac{K}{N}\frac{1}{k}\sum_{i=1}^{K} N_i I_i \bar{X}_i) \\ &= \left(\frac{K}{Nk}\right)^2 \sum_{i=1}^{K}\left[N_i^2 \text{var}(I_i\bar{X}_i) + \sum_{j=1, i\neq j}^{K} N_i N_j \text{cov}(I_i\bar{X}_i, I_j\bar{X}_j)\right]\end{aligned}$$

- Denote the variance within each school as $\sigma_{wi}^2 = \frac{1}{N_i}\sum_{j=1}^{N_i}(x_{ij}-\mu_i)^2$. Find the expressions for $\text{var}(I_i\bar{X}_i)$ and $\text{cov}(I_i\bar{X}_i, I_j\bar{X}_j)$.

  We use the fact that sampling on the two levels is independent and that $I_i^2 = I_i$ ($1^2 = 1,\ 0^2 = 0$):

$$\begin{aligned}\text{var}(I_i\bar{X}_i) &= E(I_i^2\bar{X}_i^2) - E(I_i\bar{X}_i)^2 = E(I_i)E(\bar{X}_i^2) - E(I_i)^2E(\bar{X}_i)^2 \\ &= \frac{k}{K}E(\bar{X}_i^2) - \left(\frac{k}{K}\right)^2\mu_i^2\end{aligned}$$

  Since $E(\bar{X}_i^2) = \text{var}(\bar{X}_i) + E(\bar{X}_i)^2 = \frac{\sigma_{wi}^2}{n}\frac{N_i-n}{N_i-1} + \mu_i^2$, we get

$$\text{var}(I_i\bar{X}_i) = \frac{k}{K}\left(\frac{\sigma_{wi}^2}{n}\frac{N_i-n}{N_i-1} + \mu_i^2\right) - \left(\frac{k}{K}\right)^2\mu_i^2 = \mu_i^2\frac{k(K-k)}{K^2} + \frac{k}{K}\frac{\sigma_{wi}^2}{n}\frac{N_i-n}{N_i-1}$$

  We now express the covariance:

$$\text{cov}(I_i\bar{X}_i, I_j\bar{X}_j) = E(I_iI_j\bar{X}_i\bar{X}_j) - E(I_i\bar{X}_i)E(I_j\bar{X}_j)$$

  Independence of sampling and independence of sample means gives:

$$\begin{aligned}\text{cov}(I_i\bar{X}_i, I_j\bar{X}_j) &= E(I_iI_j)\mu_i\mu_j - E(I_i)E(I_j)\mu_i\mu_j \\ &= \mu_i\mu_j cov(I_i, I_j) = -\mu_i\mu_j\frac{k(K-k)}{K^2(K-1)}\end{aligned}$$

- Derive the formula for the variance of the estimator in the case when all values N_i equal L and the variance within schools is the same for all the schools. Denote the between schools variance as σ_b^2.

$$\begin{aligned}
\text{var}(\bar{X}) &= \left(\frac{1}{Lk}\right)^2 \sum_{i=1}^{K} \left[L^2 \text{var}(I_i \bar{X}_i) + \sum_{i=1, i\neq j}^{K} L^2 \text{cov}(I_i \bar{X}_i, I_j \bar{X}_j) \right] \\
&= \left(\frac{1}{k}\right)^2 \sum_{i=1}^{K} \left[\mu_i^2 \frac{k(K-k)}{K^2} + \frac{k}{K} \frac{\sigma_w^2}{n} \frac{L-n}{L-1} \right. \\
&\quad \left. - \sum_{j=1, i\neq j}^{K} \mu_i \mu_j \frac{k(K-k)}{K^2(K-1)} \right]
\end{aligned}$$

The between schools variance can be expressed as :

$$\begin{aligned}
\sigma_b^2 &= \frac{1}{K} \sum_{i=1}^{K} [\mu_i - \mu]^2 \\
&= \frac{1}{K} \sum_{i=1}^{K} \left[\mu_i^2 - 2\mu\mu_i - \mu^2 \right]^2 \\
&= \frac{1}{K} \sum_{i=1}^{K} \mu_i^2 - \mu^2
\end{aligned}$$

We get:

$$
\begin{aligned}
&\sum_{i=1}^{K} \mu_i^2 \frac{k(K-k)}{K^2} - \sum_{i=1}^{K} \sum_{j=1, i\neq j}^{K} \mu_i \mu_j \frac{k(K-k)}{K^2(K-1)} \\
&= \frac{k(K-k)}{K^2(K-1)} \sum_{i=1}^{K} \left[(K-1)\mu_i^2 - \mu_i (\sum_{j=1}^{K} \mu_j - \mu_i) \right] \\
&= \frac{k(K-k)}{K^2(K-1)} \sum_{i=1}^{K} \left[(K-1)\mu_i^2 - \mu_i (K\mu - \mu_i) \right] \\
&= \frac{k(K-k)}{K^2(K-1)} \sum_{i=1}^{K} \left[K\mu_i^2 - K\mu\mu_i \right] \\
&= \frac{k(K-k)K}{K^2(K-1)} \left[\sum_{i=1}^{K} \mu_i^2 - \mu \sum_{i=1}^{K} \mu_i \right] \\
&= \frac{k(K-k)K}{K^2(K-1)} \left[K(\sigma_b^2 + \mu^2) - K\mu^2 \right] \\
&= \frac{k(K-k)K}{K^2(K-1)} K\sigma_b^2 \\
&= \frac{k(K-k)}{(K-1)} \sigma_b^2
\end{aligned}
$$

and therefore

$$
\begin{aligned}
\mathrm{var}(\bar{X}) &= \frac{1}{k^2} \left[\frac{k(K-k)}{(K-1)} \sigma_b^2 + \sum_{i=1}^{K} \left(\frac{k}{K} \frac{\sigma_w^2}{n} \frac{L-n}{L-1} \right) \right] \\
&= \frac{1}{k^2} \frac{k(K-k)}{(K-1)} \sigma_b^2 + \frac{K}{k^2} \frac{k}{K} \frac{\sigma_w^2}{n} \frac{L-n}{L-1} \\
&= \frac{1}{k} \frac{K-k}{(K-1)} \sigma_b^2 + \frac{1}{k} \frac{\sigma_w^2}{n} \frac{L-n}{L-1}
\end{aligned}
$$

Understanding the ideas in R:

- Try checking all the results in R.

2.6 Estimation of covariance

A course was given to a random sample of n employees in a company of size N. At the end of the course, the new knowledge was tested. The company wishes to decide whether the course is sensible for all the individuals, so they wish to estimate the correlation between the age of an employee (X_i) and the test score (Y_i).
For each individual from the sample, we have a pair of random variables (X_i, Y_i), $i = 1 \dots n$.

- Explain that the value $cov(X_i, Y_j)$ is equal for any $i \neq j$.

  The sampling procedure can be seen as follows: the units are put in a random order and the first n units represent our sample. Since all the orders have the same probability, all units have the same probability to appear on the ith position. All pairs (X_i, Y_i) thus have the same probability nd the covariance of X_i and Y_j is equal for all i and j.

- Denote $\gamma = cov(X_i, Y_i)$. Calculate the covariance $cov(X_i, Y_j)$ for $i \neq j$.

  We use the same trick as in the previous exercises. The sum of all the population values is constant, hence

  $$\mathrm{cov}(X_i, \sum_{j=1}^{N} Y_j) = \mathrm{cov}(X_i, Y_i) + (N-1)\mathrm{cov}(X_i, Y_j) = 0.$$

  Therefore (for $i \neq j$)

  $$\mathrm{cov}(X_i, Y_j) = -\frac{\gamma}{N-1}.$$

- The covariance of variables X and Y is defined as

  $$\mathrm{cov}(X, Y) = \mathrm{cov}(X_1, Y_1) = \frac{1}{N}\sum_{i=1}^{N}[(x_i - \mu)(y_i - \nu)] = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \mu\nu,$$

  where μ and ν denote the averages $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ in $\nu = \frac{1}{N}\sum_{i=1}^{N} y_i$.

  We would like to estimate the covariance using the estimator $\widehat{\gamma} =$

$c\left[\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right]$. Find the value of c to ensure an unbiased estimator.

The expected value of the estimator equals

$$E(\widehat{\gamma}) = c\left[\sum_{i=1}^{n} E(X_i Y_i) - nE(\bar{X}\bar{Y})\right] \tag{2}$$

Symmetry gives $E(X_i Y_i) = E(X_j Y_j)$ for any i and j. We know that

$$\text{cov}(X_i, Y_i) = E(X_i Y_i) - E(X_i)E(Y_i) = E(X_i Y_i) - \mu\nu$$

Therefore, $E(X_i Y_i) = \mu\nu + \gamma$. The second term on the right side of (2) can be written as:

$$\begin{aligned}
E(\bar{X}\bar{Y}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i \frac{1}{n}\sum_{j=1}^{n} Y_j\right] \\
&= \frac{1}{n^2} E\sum_{i=1}^{n}\left[X_i Y_i + X_i \sum_{j=1, i\neq j}^{n} Y_j\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E(X_i Y_i) + \sum_{j=1, i\neq j}^{n} E(X_i Y_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[E(X_i Y_i) + (n-1)E(X_i Y_j)\right]
\end{aligned}$$

We use the result

$$\begin{aligned}
\text{cov}(X_i, Y_j) &= E(X_i Y_j) - E(X_i)E(Y_i) \\
E(X_i Y_j) &= \mu\nu - \frac{\gamma}{N-1}
\end{aligned}$$

and thus

$$\begin{aligned}
E(\bar{X}\bar{Y}) &= \frac{1}{n^2} n\left[\mu\nu + \gamma + (n-1)(\mu\nu + \frac{-\gamma}{N-1})\right] \\
&= \frac{1}{n}\left[n\mu\nu + \gamma(1 - \frac{(n-1)}{N-1})\right] \\
&= \frac{1}{n}\left[n\mu\nu + \gamma\frac{N-n}{N-1}\right]
\end{aligned}$$

Combining all the results into (2) we get

$$\begin{aligned} E(\widehat{\gamma}) &= c\left[\sum_{i=1}^{n}(\mu\nu+\gamma) - n\frac{1}{n}\left[n\mu\nu+\gamma\frac{N-n}{N-1}\right]\right] \\ &= c\left[n\mu\nu+n\gamma-n\mu\nu-\gamma\frac{N-n}{N-1}\right] \\ &= c\left[n\gamma-\gamma\frac{N-n}{N-1}\right] \\ &= c\gamma\left[\frac{nN-n}{N-1}-\frac{N-n}{N-1}\right] \\ &= c\gamma\frac{N(n-1)}{N-1} \end{aligned}$$

c must therefore equal $\frac{1}{n-1}\frac{N-1}{N}$.

- How would you estimate the correlation? What do we know about the bias of this estimator?

  We use the formulas for estimation of variances and covariance:

$$\begin{aligned} \widehat{\rho} &= \frac{\widehat{\mathrm{cov}}(X,Y)}{\widehat{\sigma}_X\widehat{\sigma}_Y} \\ &= \frac{\frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i-\bar{X})^2\frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \\ &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \end{aligned}$$

  We know nothing about this bias - the expected value of the quotient does not equal the quotient of expected values.

Understanding the ideas in R:

- To check whether the estimate is unbiased, we use a simulation:

  We first generate a population of size $N = 300$, the samples shall be of size $n = 10$. Let the age (X) be distributed uniformly between 25 and

65, and the success negatively associated with age, so that it on average equals $100 - age$ (we assume that the residuals from this average are normally distributed with standard deviation 20)

```
> set.seed(1)
> xi <- runif(300)*40+25                #300 individuals, aged 25-65
> yi <- 100 - xi + rnorm(300)*20        #the result on the test for the population
> cov(xi,yi)                            #population covariance
[1] -136.8110
> cor(xi,yi)                            #population correlation
[1] -0.5207052

> runs <- 10000                         #number of simulation runs
> cova <- cora <- rep(NA,runs)          #prepare the space for the results
> for(it in 1:runs){                    #in each simulation run, do
+ inx <- sample(1:length(xi),size=10,replace=F)     #choose a sample of 10
+ xa <- xi[inx]                         #look at their ages
+ ya <- yi[inx]                         #look at their test results
+ cova[it] <- 1/9*299/300*
+   sum( (xa-mean(xa))*(ya-mean(ya)))    #get the covariance
+ cora[it] <-  sum( (xa-mean(xa))*(ya-mean(ya)))/
+   sqrt(sum( (xa-mean(xa))^2)*sum((ya-mean(ya))^2))   #get the correlation
+ }

> mean(cova)                            #average covariance
[1] -135.4745
> mean(cora)                            #average correlation
[1] -0.5034081
```

- We see that both values are smaller than the population values. We check whether the departures are important with respect to the standard error that can be expected in this number of simulations.

  We first consider the average covariance (`mean(cova)`) and compare it to the true (population) value (`cov(xi,yi)`). The average covariance is a random variable, if we repeated the simulation (all the 10000 runs), we would get a different value. Assume that the distribution of the average covariance is approximately normal, we estimate its variance (variance of the mean of n i.i.d variables is the variance of the variable, divided by n. In our case, n represents the number of simulation steps). The null hypothesis to be checked is: H_0 : the average covariance equals the

population value. The departure from this null hypothesis is checked using the t test.

```
> (mean(cova)-cov(xi,yi))/sqrt(var(cova)/runs)
[1] 1.509540
```

This results was expected - we have theoretically shown that the estimator of covariance is unbiased. We repeat the procedure for the covariance:

```
> (mean(cora)-cor(xi,yi))/sqrt(var(cora)/runs)
[1] 6.66459
```

The departure for the correlation is statistically much more significant, we conclude that it has not happened due to random variation, but rather due to the fact that the estimator is biased.