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
- 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).
- 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 $0,7$.

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 $P(I_i = 1)$?

- Calculate $\text{var}(I_i)$ and $cov(I_i, I_j)$.
- Show that the variance of the estimator equals $\text{var}(C) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

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.
- How would you estimate the population mean if the all schools were of equal size L?
- Is the sample size at each school important for the bias of the estimator?
- Express the variance of the estimator using the variance and covariance
- Denote the variance within each school as $\sigma^2_{wi} = \frac{1}{N_i}\sum_{j=1}^{N_i}(x_{ij} - \mu_i)^2$. Find the expressions for $\text{var}(I_i\bar{X}_i)$ and $\text{cov}(I_i\bar{X}_i, I_j\bar{X}_j)$.
- 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^2_b.

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 $i \neq j$.
- Denote $\gamma = cov(X_i, Y_i)$. Calculate the covariance $cov(X_i, Y_j)$ for $i \neq j$.
- How would you estimate the correlation? What do we know about the bias of this estimator?

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.