2 Sampling

2.1 Sampling - infinite population

We are trying to estimate the decrease of systolic blood pressure in hypertension patient after three months of taking a certain drug. We've collected a sample of 25 patients, let X_i denote the difference in i-th patient of our sample. Assume that the random variables X_i are independent and equally distributed.

- Show that the sample average is an unbiased estimate of the mean decrease in the population of patients (denote it by μ).
- What can we say about $cov(X_i, X_j)$ for $i \neq j$?
- Let the population variance equal σ^2. What is the variance (standard error) of our estimate?
- Based on our sample, we would like to estimate σ^2. Write our estimate as $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$. What should be the value of the constant c to ensure an unbiased estimate?
- We get the following results in our sample: $\bar{x} = 4$, $\hat{\sigma} = 20$. Estimate the sample standard error (i.e. standard error of the sample mean). Do the data support the claim that the pressure is decreasing in the population?

Understanding the ideas in R:

- Generate samples of size 30 from the uniform distribution. Calculate the mean of each sample and observe the distribution of sample means. Estimate the expected value of the sample mean and its standard deviation and compare to the theoretical values.
- Repeat the above procedure with other (perhaps more asymmetric) distributions.

2.2 Sampling - finite population

We wish to estimate the average number of employees at the beginning of this year in companies of a certain branch. The branch is divided into subgroups,

there are only 11 companies in one of the subgroups. We managed to get the data for a random sample of 6 out of these 11 companies. Let X_i denote the number of employees in the i-th company of our sample, let μ denote the population average and σ the population standard deviation.

- Let X_1 and X_2 denote the values of the first two randomly chosen companies. What can we say about the covariance $\text{cov}(X_1, X_2)$? What can we say in general for any $i \neq j$?
- Calculate the correlation $\text{cor}(X_i, X_j)$ for any $i \neq j$. What does it depend on?
- Calculate the standard error of our sample (assume that you know σ^2).
- A second subgroup includes 100 companies. What sample size is needed to ensure approximately the same standard error (assuming that the variance in this subgroup also equals σ^2)? What if we have an extremely large group of companies to sample from?
- What should be the value of the constant c, to let $\widehat{\sigma}^2 = c\sum_{i=1}^{n}(X_i - \bar{X})^2$ be an unbiased estimator of σ^2?
- Find the unbiased estimator for the variance of sample mean

Understanding the ideas in R:

- Define 11 values that represent the population. Generate random samples of size 6 and observe the distribution of sample means. Show that the above derived formula represents an unbiased estimator of the population variance.

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.