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).

  The sample average equals $\frac{1}{n}\sum_{i=1}^{n} X_i$, the population average is denoted by μ. We assume that the sample is random, i.e. that X_i are i.i.d. with $E(X_i) = \mu$:

$$E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

- What can we say about $cov(X_i, X_j)$ for $i \neq j$?

  Since the values of different patients are independent, the covariance equals 0

- Let the population variance equal σ^2. What is the variance (standard error) of our estimate?

$$\begin{aligned}
\text{var}[\bar{X}] &= \text{cov}[\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{j=1}^{n} X_j] = \frac{1}{n^2}\text{cov}[\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{cov}[X_i, \sum_{j=1}^{n} X_j] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \{\text{cov}[X_i, X_i] + \sum_{j=1, j\neq i}^{n} \text{cov}[X_i, X_j]\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \{\text{cov}[X_i, X_i] + (n-1)\text{cov}[X_i, X_j]\}
\end{aligned}$$

We use the fact that the values are independent, i.e. that $\text{cov}[X_i, X_j] = 0$ for any $i \neq j$.

$$\begin{aligned}
\text{var}[\bar{X}] &= \frac{1}{n^2}\sum_{i=1}^{n}\{\text{cov}[X_i, X_i]\} \\
&= \frac{1}{n^2}\cdot n\text{cov}[X_i, X_i] = \frac{1}{n}\text{var}[X_i] \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Standard error equals $SE = \frac{\sigma}{\sqrt{n}}$.

- 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 know that $\sigma^2 = E(X^2) - E(X)^2$ and hence $E(X^2) = \sigma^2 + \mu^2$. Similarly, $SE^2 = \text{var}(\bar{X}) = \frac{\sigma^2}{n} = E(\bar{X}^2) - E(\bar{X})^2$ and $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}$. We take into account that $\sum_i X_i\bar{X} = \bar{X}\sum_i X_i = n\bar{X}^2 = \sum_i \bar{X}^2$ to get:

$$\begin{aligned}
E[c\sum_{i=1}^{n}(X_i - \bar{X})^2] &= cE[\sum_{i=1}^{n}\{X_i^2 - 2X_i\bar{X} + \bar{X}^2\}] \\
&= cE[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2] \\
&= cE[\sum_{i=1}^{n}\{X_i^2 - \bar{X}^2\}] \\
&= c\sum_{i=1}^{n}\{E[X_i^2] - E[\bar{X}^2]\} \\
&= c\sum_{i=1}^{n}\left[(\mu^2 + \sigma^2) - (\mu^2 + \frac{\sigma^2}{n})\right] \\
&= cn\left[\sigma^2(1 - \frac{1}{n})\right] \\
&= \sigma^2(n-1)c
\end{aligned}$$

  Since we wish that $E(\widehat{\sigma}^2) = \sigma^2$, the constant must be equal to $c = \frac{1}{n-1}$.

- 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?

  The estimated standard error equals $\widehat{SE} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{20}{5} = 4$. The decrease is similar in size to the standard error - a possible true difference cannot be distinguished from random variability. We would need a substantially larger sample to make any conclusions.

Summary: if population is infinite and the sample random, we have:

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n};\ \widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

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 $i \neq j$?

  The population if finite, denote the values in the population as x_k,

$k = 1, \ldots, 11$. We put the all the 11 companies in a random order, and let the first 6 represent our sample. X_i denotes the number of employed in the i-th chosen company. Since each of the X_i presents one of the x_k values and all have the same probability of being chosen, $\text{cov}(X_1, X_2) = \text{cov}(X_i, X_j)$ for any non-equal i and j. But X_i and X_j are no longer independent - if $X_i = x_k$, X_j cannot be equal to x_k. We use the following trick:

$$\text{cov}(X_i, \sum_{j=1}^{N} X_j) = \text{cov}(X_i, \sum_{k=1}^{N} x_k) = 0$$

Since the sum of all values is a constant, the above term equals 0 and hence

$$\text{cov}(X_i, \sum_{j=1}^{N} X_j) = \text{cov}(X_i, X_i) + (N-1)\text{cov}(X_i, X_j) = 0$$

Therefore (for $i \neq j$)

$$\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

The variables are negatively associated.

- Calculate the correlation $\text{cor}(X_i, X_j)$ for any $i \neq j$. What does it depend on?

  The correlation between the covariates can be expressed as

$$\text{cor}(X_i, X_j) = -\frac{\sigma^2}{(N-1)\sigma^2} = -\frac{1}{N-1}$$

  As the population gets larger, the correlation gets very small. It depends solely on the size of the population.

- Calculate the standard error of our sample (assume that you know σ^2).

$$\begin{aligned}
\text{var}[\bar{X}] &= \text{cov}[\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \cdot \text{cov}[X_i, \frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \{\text{cov}[X_i, X_i] + (n-1)\text{cov}[X_i, X_j]\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \{\sigma^2 - (n-1)\frac{\sigma^2}{N-1}\} = \frac{\sigma^2}{n}\frac{N-n}{(N-1)}
\end{aligned}$$

  Standard error equals $SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{(N-1)}}$.

- 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?

  For $N = 11$ in $n = 6$, we get $SE^2 = \frac{\sigma^2}{6}\frac{11-6}{(10)} = \frac{\sigma^2}{12}$. In a population of size 100, sample of size 10 has the standard error equal to $SE^2 = \frac{\sigma^2}{11}$, while the sample of size 11 results in $SE^2 = \frac{\sigma^2}{12,2}$.
  If population is larger, the same standard error requires more units in the sample. If the population is very large, the term $\frac{N-n}{(N-1)}$ approximately equals 1 and we hence need 12 companies. The size of the required sample increases as the population increases, but this increase quickly becomes negligible.

- 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?

  We repeat the calculation of the previous exercise, taking into account

that $E(\bar{X}^2) = \mu^2 + \frac{\sigma^2}{n}\frac{N-n}{N-1}$:

$$\begin{aligned} E[c\sum_{i=1}^{n}(X_i - \bar{X})^2] &= cE[\sum_{i=1}^{n}\{X_i^2 - \bar{X}^2\}] \\ &= c\sum_{i=1}^{n}\left\{(\mu+\sigma^2) - (\mu^2 + \frac{\sigma^2}{n}\frac{N-n}{N-1})\right\} \\ &= cn\left\{\sigma^2(1 - \frac{N-n}{n(N-1)})\right\} \\ &= \sigma^2 c\frac{N(n-1)}{N-1} \end{aligned}$$

Since we wish to have $E(\widehat{\sigma}^2) = \sigma^2$, our constant should equal $c = \frac{1}{n-1}\frac{N-1}{N}$, i.e.

$$\widehat{\sigma}^2 = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- Find the unbiased estimator for the variance of sample mean

  Putting the results together, we get

$$\begin{aligned} \widehat{SE}^2 &= \frac{\widehat{\sigma}^2(N-n)}{N-1} = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2\frac{(N-n)}{N-1} \\ &= \frac{\widehat{\sigma}^2}{n}\frac{N-n}{N}, \end{aligned}$$

  where $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Summary: for a finite population, we have:
The values chosen in the sample are not independent despite random sampling, the covariance between two units equals:

$$\text{cov}(X_i, X_j) = -\frac{\sigma^2}{N-1},$$

the variance of the sample mean and the unbiased estimator of the population variance equal

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{(N-1)};\ \widehat{\sigma}^2 = \frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

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.