How Students Can Understand Sampling within Finite Populations in Statistics

Tanja Van Hecke
University college Ghent Associated faculty of Applied Engineering Sciences (Belgium)
Tanja.VanHecke@hogent.be

Abstract

Statistics is often perceived by students as a difficult specialization. To overcome their fear for the magic of the probabilities and to make them see the use of statistics, an advantageous approach is to start with a problem setting close to real life. Once the teacher has acquired the attention of the students, a guided tour by means of appealing questions, will give the students a structure which makes the statistical problem easier to manage. We illustrate this approach by a study of the required sample size in relation to the imposed accuracy, the population size and the variance in case of finite populations. The theory of confidence intervals is often presented in case of the theoretical situation of an infinite population. In practice the population can be finite. We will show how a statistics teacher can convince students about the correction of the formulas to calculate the required minimum sample size. An example from the music industry is investigated as a trigger. A positive practical experience with engineering students is presented.

1. Introductory examples

The music industry must adjust to the growing practice of consumers downloading illegally songs instead of paying for downloading them. It therefore is important to estimate the proportion of songs that is currently illegally downloaded. As music is part of everyday life of young people, it can be useful as a motivating example to introduce the influence of finite populations on the required sample size.

The time required and the level of detail to work out the following examples depends on the level of statistical knowledge of the students. It assumes that students are familiar with the concept of confidence intervals.

Example 1

Music company X wants to know the percentage of people downloading music illegally among the total population of people downloading music from the internet. When they start this investigation, the first question that arises is how to determine the size of the sample. Can you give the minimal sample size if we want to be 95% confident that the sample percentage is within ten percentage point of the true population percentage of illegally downloaded songs.

A preliminary investigation gives an estimation \( p = 0.65 \) of the proportion of music illegally downloaded. It is required to estimate the variance. As the sampling error is limited to 0.1, we ask that half the length of the 95% confidence interval does not exceed 0.1, or

\[
t_{0.95} \sqrt{\frac{p(1-p)}{n}} = 1.96 \sqrt{\frac{0.65 \times 0.35}{n}} \leq 0.1.
\]
Solving for \( n \) gives that the music company X will obtain the premised reliability if the sample size is at least 88.

The described work out of this example can also be considered as an exercise on

- solving inequalities
- looking up limit values \( t_{1-\alpha} \) in a table

**Example 2**

Music company Y wants to know the number of downloaded songs a month among people downloading music illegally. At the start of the investigation the sample size is required. The company wants the difference between the sample mean of the number of illegally downloaded songs and its true value, not to exceed 1.

As the sample size depends on the variance, we need an estimation of it. This can be done by a preliminary investigation which yields for example \( s^2 = 16 \). As the sampling error is limited to 1, we ask that half the length of the 95% confidence interval for \( \mu \) does not exceed 1, or

\[
t_{1-\alpha} \frac{s}{\sqrt{n}} = 1.96 \frac{4}{\sqrt{n}} \leq 1.
\]

Solving for \( n \) gives that the music company X will obtain the premised reliability if the sample size is at least 62.

**2. Finite population correction factor**

When we use sampling with replacement, it is like sampling from an infinitely large population. Practically however, populations are not infinite and students should be aware of the consequences of this reality of populations that are limited in number. With sampling without replacement the standard error needs to be corrected, especially when the population size \( N \) is relatively small and when the sample size \( n \) is not small in comparison with \( N \) (i.e. more than 5% of the population is sampled or \( n/N > 0.05 \)). Then we should multiply the standard error of the mean and the standard error of the proportion with a finite population correction factor (fpc). [1] [2]

\[
fpc = \sqrt{\frac{N-n}{N-1}}. \tag{1}
\]

The teacher can use questions as

- Which one is the biggest, the numerator or the denominator of fpc?
- Is this correction factor greater or smaller than 1?
- What is the effect on the standard error, when it is multiplied by the fpc?

to guide the students to the conclusion that multiplying with the fpc, will reduce the standard error.

The \((1-\alpha)100\%\) confidence interval

\[
\left[ \bar{x} - t_{1-\alpha} \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha} \frac{s}{\sqrt{n}} \right]
\]

should be adjusted to

\[
\left[ \bar{x} - t_{1-\alpha} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{x} + t_{1-\alpha} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]
\]
in the case of a finite population.
The sampling error \( e \) is equal to \( t_{1-\alpha} \frac{s}{\sqrt{n}} \) in case of infinite populations. If the population of size \( N \) is finite and \( n_0 \) is the sample size without considering the finite correction factor, than we can solve

\[
\sqrt{n_0} \cdot fpc = \sqrt{n}
\]

for \( n \) (the corrected sample size):

\[
n = \frac{n_0 N}{n_0 + (N-1)}.
\]

**Example 3**

The board of a school of 500 pupils wants to make their pupils aware of the criminal character of illegally downloading music. The board wants to have an idea about the extent of the problem at their school. Advise them under the same conditions as in example 1 about the size of the sample they have to take in order to obtain a 95% confident result on the proportion of pupils downloading music illegally. Give the board a similar advice about the sample size as in example 2, but in case of the pupils of that school as population.

We use formula (2) to adjust the obtained value \( n_0 \) from example 1:

\[
n = \frac{88 \times 500}{88 + (500-1)} = 74.7
\]

So when we restrict the investigation to the school with \( N = 500 \), it is sufficient to have a sample size of at least 75 pupils to be 95% confident that the sample percentage is within ten percentage point of the true population percentage of illegally downloaded songs.

Analogous we find \( n \geq 56 \) is sufficient to be 95% confident that the sampling error for the mean of illegally downloaded songs does not exceed 1.

### 3. Influencing parameters for the sample size

Besides making calculations about the required sample size, it is of major importance that students understand the relations between the influencing parameters. The relation (2) with

\[
n_0 = \left( \frac{t_{1-\alpha} \cdot s}{e} \right)^2
\]

describes the dependency of the sample size \( n \) in case of a finite population size as a function of \( s, e, N \). To make the students understand how \( n \) is related to the population size \( N \), the admitted sampling error \( e \) and the standard deviation \( s \), figures can be used. Figure 1 illustrates that a greater finite population size leads to a greater sample size with given sampling error and variance. Figure 2 makes clear that a small admitted sampling error can only be reached with large sample sizes. Figure 3 shows that a greater variance requires a larger sample size to reach the same confidence level. All figures show that a lower confidence level leads to a lower minimum value for the sample size.
Fig. 1. Required sample size $n$ as a function of the total population size $N$ for different confidence levels, constant variance and constant sampling error.

Fig. 2. Required sample size $n$ as a function of the admitted sampling error $e$, for different confidence levels, constant variance and constant population size.
Fig. 3. Required sample size $n$ as a function of the standard deviation $\sigma$, for different confidence levels, constant population size and constant sampling error.

To get a deeper understanding of these figures, the teacher can try to introduce the student in the meaning of the symbols by means of multiple choice questions as:

**Question 1:**
Instead of estimating the number of downloaded songs among the population of students in my university, I want to estimate this number among the students in my country. To draw conclusions (with a confidence of 90%) for the larger population of students all over the country, with similar variances and admitted sampling error,

A. the required sample size will be less.

B. the required sample size will not be influenced by the population size.

C. the required sample size will be greater.

**Question 2:**
I want to estimate the number of downloaded songs among the population of students in my university. As I decide to require a more precise result than my colleague, I accept only half the error on the estimation of the mean value of downloaded songs as my colleague does. In that case,

A. my required sample size will be less than my colleague’s.

B. my required sample size will be equal to my colleague’s.

C. my required sample size will be greater than my colleague’s.

**Question 3:**
A colleague from a foreign university and I want to estimate the number of downloaded songs among the population of students in our own universities. A preliminary research reveals that the variance for
his numbers of downloaded songs is much bigger. When you can assume that our universities have the same number of students, that we use the same admitted sampling error, the consequence is that

A. my required sample size will be less than my colleague’s.
B. my required sample size will be equal to my colleague’s.
C. my required sample size will be greater than my colleague’s.

As this is just an illustrative example of questions, other questions can be prepared by the teacher.

4. Conclusions and experiences

The author used this example with success in a statistics class for engineering students. The interest in music provided an opportunity to catch students’ attention. This was useful to enhance the teaching of analysing the requested sample size. Besides the classical approach based on infinite populations, an alternative approach for finite populations was given. As the derivation of formulas, does not always motivate the majority of the students, the approach was enriched by means of appealing multiple choice questions and figures made by the students with the mathematics software Maple.

References