Block 10: Sampling - determining sample size

(Activity solutions can be found at the end of the document.)

Here we consider **sample size determination** in simple random sampling. Properties of the **sampling distribution** are discussed. We describe the required adjustments to statistically determined sample sizes to account for the **incidence rate** and **completion rate**. Non-response issues in sampling are also covered, with ways of improving response rates.

Learning Objectives

- understand the concepts of the sampling distribution, statistical inference and standard error
- discuss the statistical approach to determining sample size based on simple random sampling and the construction of confidence intervals
- derive the formulae to determine statistically the sample size for estimating means and proportions
- discuss the importance of non-response issues in sampling
- appreciate approaches for improving response rates and adjusting for non-response.

Reading List

Malhotra, N.K., D. Nunan and D.F. Birks. Marketing Research: An Applied Approach. (Pearson, 2017) 5th edition [ISBN 9781292103129] Chapter 15.

10.1 Sampling - determining sample size

For each section of *Sampling - determining sample size*, use the LSE ELearning resources to test your knowledge with the Key terms and concepts flip cards.

Definitions and symbols

Parameter:

- A parameter is a summary description of a fixed characteristic or measure of the target population.
- A parameter denotes the true value which would be obtained if a census, rather than a sample, was undertaken.

Statistic:

- A statistic is a summary description of a characteristic or measure of a sample.
- A sample statistic is used as an estimator of the population parameter.

Finite population correction (FPC):

- The FPC is a correction for the overestimation of the variance when estimating a population parameter, such as a mean or a proportion, when the sample size is 10% or more of the population size.

Precision level:

- When estimating a population parameter by using a sample statistic, the precision level is the desired size of the estimating interval.

- This is the maximum permissible difference between the sample statistic and the population parameter.

Confidence interval:

- A confidence interval is the range into which the true population parameter will fall, assuming a given level of confidence.

Confidence level:

- The confidence level is the probability that a confidence interval will cover/span the population parameter.

Symbols used in basic statistical inference are:

Variable	**Population**	**Sample**
Mean	$\mu\mu$	$X^{-}X^{-}$
Proportion	$\pi\pi$	PP
Variance	$\sigma 2\sigma 2$	$S2S2$
Standard deviation	$\sigma\sigma$	SS
Size	NN	nn
Standard error of the mean	$\sigma X^{-}\sigma X^{-}$	$SX^{-}SX^{-}$
Standard error of the proportion	$\sigma P\sigma P$	$SPSP$
Standardised variate (zz)	$(X-\mu)/\sigma(X-\mu)/\sigma$	$(X-X^{-})/S(X-X^{-})/S$
Coefficient of variation	$\sigma/\mu\sigma/\mu$	S/X^{-}

Symbols for population and sample variables

Sampling distributions

A **sampling distribution** is the distribution of sample statistic values computed for each possible sample which could be drawn from the target population under a specific sampling scheme.

A key purpose of market research is to estimate population parameters (such as a mean or a proportion) and perform statistical inference - generalising sample results to the target population. Although *in practice only one sample is drawn* (rather than all possible samples), the concept of a sampling distribution is still relevant.

We can use *probability theory* to make inferences about population values.

For *large samples* (as a rule-of-thumb, say n≥30n≥30), the important properties of the mean and proportion sampling distributions are as follows.

We have:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

If sampling from normal populations, otherwise invoke the **central limit theorem** (CLT). For proportions we have:

$$P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

- The **sample mean** is simply $\bar{\mathrm{X}} = \sum_{i=1}^{n} Xi\,/n$
- The **standard errors** (standard deviations of the sampling distributions) for means and proportions are:

Mean: $\sigma_{\bar{\mathrm{X}}} = \frac{\sigma}{\sqrt{\mathrm{n}}}$ Proportion: $\sigma_P \sqrt{\frac{\boldsymbol{\pi}(1-\boldsymbol{\pi})}{\mathrm{n}}}$

- Often σ is unknown, hence it is estimated using the sample standard deviation, SS, where:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 \frac{(\sum_{i=1}^{n} X_i)}{n}\right)}$$

resulting in the *estimated* standard error, $S/\sqrt{n}$

- Similarly, for proportions we estimate π with P to give the estimated standard error:

$$S_p \sqrt{\frac{P(1-P)}{n}}$$

- The area under the sampling distribution between any two points can be calculated in terms of the zz**-value -** the number of standard errors a point is away from the mean:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

- When n ≥ 0.1N the standard error will be (non-negligibly) overestimated, hence we require a **finite population correction factor** defined by:

$$\sqrt{\frac{N-n}{N-1}}$$

In which case:

$$\boldsymbol{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}}$$

The confidence interval approach

Calculation of a confidence interval involves determining a distance below ($\bar{x}_L$) and above ($\bar{x}_u$) the population mean (μ), which contains a specified area of the normal curve.

The z-values corresponding to $\bar{x}_L$ and $\bar{x}_u$ may be calculated as:

$$z_L \frac{\bar{x}_L - \mu}{\sigma_{\bar{X}}} \quad \text{and} \quad z_U \frac{\bar{x}_{u} - \mu}{\sigma_{\bar{X}}}$$

where $z_L = -z$ and $z_u = +z$. Therefore, the lower value of $\bar{x}$ is:

$$\bar{x}_L = \mu - z \times \sigma_{\bar{X}}$$

Similarly, the upper value of x¯ is:

$$\bar{x}_U = \mu + z \times \sigma_{\bar{X}}$$

Note that μ is estimated by $\bar{X}$, and a 100(1 – α)% confidence interval is given by:

$$\bar{X} \pm z_{a/2} \times \sigma_{\bar{X}}$$

where $z_{a/2}$ is the z-value which cuts off 100(α/2)% probability in the upper tail of the standard normal distribution.

We can now set a 95% confidence interval around a sample mean of, say, $182.

As a first step, we compute the standard error of the mean. For example, if σ=55 and n=300, then:

$$\sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{55}{\sqrt{300}} = 3.18$$

It can be seen that the central 95% of a normal distribution lies within ±1.96 z-values of the mean. Hence a 95% confidence interval for μ is given by:

$$\bar{x} \pm 1.96 \times \sigma_{\bar{X}} = 182 \pm 1.96 \times 3.18 = 182 \pm 6.23.$$

Therefore, a 95% confidence interval ranges from $175.77 to $188.23, or ($175.77, $188.23).

Under repeated sampling, 95% of such confidence intervals are expected to *cover*, or *span*, the true population parameter.

Sample size determination

The question ‘how large a sample do I need?’ is a common one when sampling. The answer to this depends on the *quality of inference* which the researcher requires from the data. In the estimation context, this can be expressed in terms of the accuracy of estimation.

If the researcher requires that there should be a 95% chance that the *estimation error* should be no larger than *e* units (we refer to *e* as the *tolerance on the sampling error*), then this is equivalent to having a 95% confidence interval of width 2*e*. Note here *e* represents the *half-width* of the confidence interval since the point estimate is, by construction, at the centre of the confidence interval (when estimating means or proportions).

To estimate μ to within *e* units with 100(1−α)% confidence, then:

$$z_{a/2} \times \frac{\sigma}{\sqrt{n}} \leq e.$$

Hence we require a sample size of:

$$n \geq \frac{\left(z_{a/2}\right)^2}{e^2}$$

where $z_{a/2}$ is the z-value which cuts off 100(α/2)% probability in the upper tail of the standard normal distribution.

To estimate π to within e units with 100(1−α)% confidence, then:

$$z_{a/2} \times \sqrt{\frac{\pi(1-\pi)}{n}} \leq e.$$

Hence we require a sample size of:

$$n \geq \frac{\left(z_{a/2}\right)^2 \pi(1-\pi)}{e^2}$$

where $z_{a/2}$ is the z-value which cuts off 100(α/2)% probability in the upper tail of the standard normal distribution.

Note that π should be an *approximate* value of π (since π itself is unknown, hence why we are estimating it!), perhaps obtained from a pilot study, or alternatively we make an assumption of this value based on judgement and/or experience. If a pilot study is not feasible and a value cannot be assumed, then set $\pi = 0.5$ as a ‘conservative’ choice, as this value gives the maximum possible standard error.

To see this, note that the standard error of the sample proportion is proportional to:

$$\pi\ (1-\pi)$$

which, with simple calculus, reaches a maximum at 0.5.

If the resulting sample size, *n*, represents 10% or more of the population size, *N*, the finite population correction should be applied. The required sample size should then be:

$$n_c = \frac{nN}{N + n - 1}$$

where n_c denotes the corrected sample size.

Adjusting the statistically determined sample size

The **incidence rate** refers to the rate of occurrence, or the percentage, of persons *eligible to participate* in a study.

In general, if there are c *qualifying factors* with an incidence of Q1, Q2, Q3, …, Qc, each expressed as a proportion, then:

$$\text{Incidence rate} = \text{Q1} \times \text{Q2} \times \text{Q3} \times \cdots \times \text{Qc}.$$

The **completion rate** is the percentage of qualified participants who complete the interview, enabling researchers to account for anticipated refusals by people who qualify.

In light of incidence and completion rates, the initial sample size is:

$$\text{Initial sample size} = \frac{final\ sample\ size}{incidence\ rate \times completion\ rate.}$$

Adjusting for non-response

Subsampling of non-participants - the researcher contacts a subsample of the non-participants, usually by means of telephone or personal interviews.

In **replacement**, the non-participants in the current survey are replaced with non-participants from an earlier, similar survey. The researcher attempts to contact the non-participants from the earlier survey and administer the current survey questionnaire to them, possibly by *offering a suitable incentive*.

In **substitution**, the researcher substitutes for non-participants *other elements from the sampling frame* who are expected to respond.

The sampling frame is divided into subgroups which are internally *homogeneous in terms of participant characteristics*, but *heterogeneous in terms of response rates*. These subgroups are then used to identify substitutes who are similar to particular non-participants, but dissimilar to participants already in the sample.

Subjective estimates - when it is no longer feasible to increase the response rate by subsampling, replacement or substitution, it may be possible to arrive at subjective estimates of the nature and effect of non-response bias. This involves evaluating the likely effects of non-response *based on experience and available information*.

Trend analysis is an attempt to discern a trend between early and late participants. This trend is *projected to non-participants* to estimate where they stand on the characteristic of interest. An example of the use of trend analysis is:

Use of trend analysis in adjusting for non-response

Wave	**Percentage response**	**Average euro expenditure**	**Percentage of previous wave's response**
First mailing	12	412	-
Second mailing	18	325	79
Third mailing	13	277	85
Non-response	(57)	(230)	91
Total	100	275	-

Weighting attempts to account for non-response by assigning differential weights to the data depending on the response rates. For example, suppose that in a survey the response rates were 85%, 70% and 40%, for the high-, medium- and low-income groups, respectively. In analysing the data, these subgroups are assigned weights *inversely proportional* to their response rates. That is, the weights assigned would be 100/85, 100/70 and 100/40, for the high-, medium- and low-income groups, respectively.

Imputation involves imputing, or assigning, the characteristic of interest to the non-participants based on the similarity of the variables available for both non-participants and participants. For example, a participant who does not report brand usage may be imputed the usage of a participant with *similar demographic characteristics*.

Finally, how do we improve response rates? Possible options are summarised in Figure 15.2 of the textbook.

Case study: Electric utility company

A major electric utility company would like to determine the average amount spent per household on air conditioning during summer months. From their own records, they know how much electricity is consumed per household, but not how much is spent on particular appliances and the attitudes toward the use of those appliances. Therefore, the management believes that a survey should be conducted.

Which procedure would you recommend for determining the sample size?

Case study: Restaurants

You work as the market research manager for a chain of themed restaurants. A new menu has been developed based on organic and fair-trade produce. Before the new menu is introduced, the management is concerned about how existing and potential customers will react.

How would you approach the sample size calculations for this task?

Case study: Subaru

Evaluate the reasons for the high response rates to Subaru’s surveys. What lessons of Subaru’s success can be generalised to other survey designs?

Discussion points

Statistical considerations are more important than administrative considerations in determining sample size.

The real determinant of sample size is what managers feel confident with; it has little to do with statistical confidence.

Learning outcomes checklist

Use this to assess your own understanding of the chapter. You can always go back and amend the checklist when it comes to revision!

- ○ Understand the concepts of the sampling distribution, statistical inference and standard error
- ○ Discuss the statistical approach to determining sample size based on simple random sampling and the construction of confidence intervals
- ○ Derive the formulae to determine statistically the sample size for estimating means and proportions
- ○ Discuss the importance of non-response issues in sampling

Block 10: Sampling - determining sample size

Discussion forum and case studies

To access the solutions to these questions and case study, click here to access the printable Word document or click here to go to LSE's Elearning resources.

Commentary on Case study: Electric utility company

The recommended procedure is to use the confidence interval method for means because the population standard deviation can be determined from the records or from a pilot study, and we are concerned about a sample mean.

Commentary on Case study: Restaurants

The preferred method for determining the sample size would be the confidence interval approach for proportions because we want to know if consumers prefer the new menu to the old menu. A random sample of customers at the restaurant could be given samples of the new menu and the old menu (after being told the nature of the two samples) and identify the menu which they prefer.

Commentary on Case study: Subaru

As low response rates increase the probability of non-response bias, an attempt should be made to improve response rates. This is not an issue which should be considered after a survey approach has been decided and a questionnaire designed. Factors which improve response rates are integral to survey and questionnaire design. The market researcher should build up an awareness of what motivates their target participants to participate in a research study. They should ask themselves what their target participants get in return for spending time and effort, answering set questions in a full and honest manner.

Subaru have done all of these things. They have an awareness of the nature of their customers. They are passionate about their products and share this passion with their customers. As such, they understand what issues are important to their customers, the language and logic of their questions are clearly meaningful to Subaru drivers. The questionnaire is a means to share a combined passion, with the participants seeing Subaru as a company with integrity, which values their response and will change things if they are not doing them well, i.e. their voice will be heard.

The key lessons of this for other survey designs lie in the awareness of participants. The motivation to take part, the interest in the topic, the interest in the process of questioning, the language and logic, the respect of the participant, the integrity and belief in the surveying company - all flow from this intimate knowledge of participants. This does not happen overnight or with one survey. It is an attitude toward consumers which is nurtured through viewing market research as an investment rather than as a cost.

Commentary on Discussion points

- Important considerations are the function of quantitative and qualitative factors in determining sample size. Both go hand-in-hand. The quantitative considerations are based on qualitative assumptions or opinions, for example estimating an unknown population mean. While one must certainly be rigid in his/her application of quantitative analysis, if poor qualitative assessment is done, an inappropriate sample size, either too large or too small, may be drawn. In addition, qualitative factors, such as the timeframe for the project or cost considerations, may affect the quantitative decisions made.

- Sample size can be influenced by the average size of samples in similar studies. Sample sizes can be determined based on experience and can serve as rough guidelines, particularly when non-probability sampling techniques are used. Managers can grow used to ‘average’ sample sizes, and be comfortable with them in supporting their decisions. They may ignore and/or may not appreciate the statistical case for sample sizes.