## WEB PAGE FOR CHAPTER 9

## MULTIPLE CHOICE QUESTIONS

1 If, from an extremely large population, a very large number of samples were drawn randomly and their mean values calculated, which of the following statements is true?
(a) The sample means would each be equal to the population mean.
(b The sample means would be distributed normally around the population mean
(c) The sample means would form a distribution whose standard deviation is equal to zero.
(d) The sample means would be very different from each other if the sample sizes were very large.

2 Of the following, the one which most often weakens research studies is:
(a) inaccurate statistical analysis
(b) inaccurate computer input
(c) sampling bias
(d) eliberate distortion

3 A common error in selecting a sample is to:
(a) elect whoever or whatever is available
(b) select too large a sample
(c) rely only on stratified sampling
(d) use tables of numbers that are not random

4 All members of a real or hypothetical set of persons or objects or events are called:
(a) a sample
(b) a random sample
(c) a population
(d) a parameter

5 The first step in sampling is to:
(a) determine sample size
(b) define the population
(c) decide what sort of sample is required
(d) define the parameters of the sample

6 If the investigator defines the population in too narrow a fashion, the results will be:
(a) useless
(b) generalizable to a limited population
(c) not generalizable at all
(d) generalizable to the sample

7 A random sample:
(a) is one in which each member or item from the population has an equal chance of being selected
(b) must be large
(c) must be selected by a computer program
(d) must bear at least a one in ten ratio to the population size

8 The main reason for using random sampling is:
(a) to include the correct number of subjects or items
(b) to provide generalizable results
(c) to produce statistically significant research findings
(d) to ensure everyone who should be included is included

9 Systematic sampling may be used instead of simple random sampling if the:
(a) population list is in random order
(b) sample size is large
(c) expected differences are small
(d) population is clearly defined

10 In cluster sampling the unit of sampling is the:
(a) individual
(b) population
(c) subgroups of the population based on characteristics not related to the research
(d) naturally occurring groups

11 The main advantage of cluster sampling compared to random sampling is the:
(a) degree of randomness achieved
(b) accuracy of sampling achieved
(c) improved generalizability
(d) saving of time and money

12 Large samples must be used when:
(a) few uncontrolled variables $\leq$ are present
(b) small differences are expected
(c) subgroup analysis is not going to be conducted
(d) the population is homogeneous

13 A stratified sample is not a probability sample:
(a) true
(b) false
(c) sometimes
(d) depends on N

14 A non-probability sample
(a) allows more freedom in the selection of a representative sample
(b) prevents bias in selection
(c) does not permit generalization from sample to population
(d) provides a balance between randomness and economy of resources

15 A sampling frame is:
(a) the listing of the sample
(b) the listing of the population
(c) the sequence of selection of the sample
(d) a cluster

16 A population is made up of groups that have a wide variation within each group but little variation between groups. Which sampling technique is the most appropriate?
(a) stratified
(b) systematic
(c) cluster
(d) multistage

17 Which words logically go together?
(a) sample parameter
(b population statistic
(c) sample element
(d) population parameter
(e) representative estimate

18 If a university's researchers select 250 students at random from the official student list to take part in a study, this is an example of:
(a) quota sampling
(b convenience sampling
(c) probability sampling
(d) non-probability sampling
(e) snowball sampling

19 Which is the simplest and cheapest type of sampling?
(a) convenience sampling
(b) quota sampling
(c) snowball sampling
(d) simple random sampling
(e) systematic sampling

## FURTHER QUESTIONS

1 What is the difference between a sample and a census?
2 Name two objectives which must be balanced when deciding on the size of a sample.
3 What is the chief disadvantage of non-probability (non-random) sampling?
4 What are the two major advantages of stratified sampling?
5 What sample design is the most appropriate for selecting individuals from the population of a country with the objective of determining the opinions of citizens on the priorities for the allocation of national government expenditure?

6 What is the main criticism of quota sampling?
7 An accountant believes that errors are likely to be more common in high value than low value accounts and classifies accounts whether they are high, medium or low value. How should the accountant proceed to sample accounts from the whole population of accounts in order to obtain a reliable estimate of the frequency of errors?

8 In what situation might cluster sampling be appropriate?
9 For each of the following examples, define the population of interest and specify whether a census or a sample would be more appropriate. Explain your reason.
(a) In order to estimate the durability of a new line of luggage, the company conducts a series of tests, including dropping the luggage from varying heights onto a concrete pad and hitting it with various size sledge hammers on the top and bottom surfaces and on each of the four sides.
(b) A medical researcher wishes to investigate the relationship between shoplifting and the presence of mental illness in the shoplifter.
(c) In two months, voters will be asked to decide on a referendum that specifically sets up separate areas for smoking and non-smoking in all public places. Suppose you want to predict the outcome of this vote.
(d) The automobile industry wishes to conduct a consumer survey of all registered automobile owners in order to determine when they intend to purchase a new automobile.

10 A school board wishes to obtain voter sentiment on plans to construct a new elementary school in the district. Each child is given a questionnaire to take home to his or her parents. What comment would you make on this method?

11 Which of the following use judgemental, convenience, or random sampling? If random sampling, identify the type.
(a) The 30 stocks constituting the Hang Seng Industrials were selected by experts as representative of all industrial stocks.
(b) In order to check conformity to affirmative action, personnel files were subdivided into ethnic groups and random samples were selected from each group.
(c) To obtain overnight ratings of a highly advertised television special, each of a number of interviewers was instructed to make 25 phone calls from numbers they found in a phone book.
(d) To check on the quality of paper clips, 20 boxes were randomly selected from a shipment and all clips in each box were inspected.
(e) To assess employee reaction to an anticipated change in car-pooling incentives, the computer randomly selects 50 names from the 6,500 that are in the personnel files.

12 If the number 2 was the starting point on a selection ratio of 1 to 5 , what are the next four successive selections?

13 There is a fundamental difference between random and systematic sampling. One of the conditions of random sampling is that the selection of one individual should be independent of the selection of another. Systematic sampling does not satisfy this condition. Can you explain why?

14 The local radio station conducts a person-in-the-street interview with five typical citizens each day on topical issues. What sort of sample is this?. Criticize the sampling method.

15 Below are examples of the selection of samples. Decide for each which sampling technique was used.
(a) Restricted to a $5 \%$ sample of the total population, the researcher chose every 20 th person on the electoral register.
(b) A government social worker investigating juvenile theft from shops and broken homes obtained his sample from adolescents appearing at the juvenile court.
(c) A research organization used random digit dialling to obtain their sample.

16 Explain why a random sample from a population in which certain subjects were inaccessible would be a contradiction in terms.

17 In what context would a multi-stage cluster sample be particularly useful?
18 If a sample of business executives were randomly selected from the Yellow Pages in a particular city would you necessarily have a representative sample?

19 How could we get random samples of:
(i) students in your university/college;
(ii) a pile of chemical feedstock which needs analysis before it is fed into the plant;
(iii) a crop growing in a field?

20 Is a stratified sample a probability sample?
21 What do we mean by saying the sampling fraction is disproportionate.

## CLASS ACTIVITIES/DISCUSSION

1 Chang's Chicken and Rice is a chain of restaurants. Chang wants to know what consumers' reactions are to his new menu. During the week of 7th April each customer is given a coupon for a free plate of chicken and rice. However, the coupon is not valid until the customer calls in on a free call number and answers some survey questions and obtains a validation code.

In class: Comment on the sample plan. Can you devise a better one?
2 Lesbos Magazine want to conduct a survey on the attitudes of the modern Asian young woman: 'What the Asian Woman thinks and feels'. A questionnaire is written and placed inside one issue of the magazine. It is to be completed anonymously and then returned post-paid. There is a large response.

Discuss in class: What criticism would you make of this plan? How acceptable or unacceptable is the method of data collection given the aims of the magazine's editors.

3 Discuss in class what the target population and sample frame for each of the following should be:
(a) You want to evaluate the effectiveness of a Drug Free Asia anti-drug advertising campaign
(b) You have just introduced the new APEC luxury sports car. There has been national TV advertising. You now want to determine perceptions of the car.

4 Devise examples of situations where a sample would be preferable to a census because of the following reasons:
(a) Population too large.
(b) Costs too large.
(c) Timeliness.
(d) Process involves destruction.
(e) Inaccessibility of members of the population

5 Give examples of situations in which the following types of error could occur in the selection of the sample or the collection of the data:
(a) Bias leading to the non-representative sample.
(b) Sampling error due to uncontrolled factors.
(c) Non-sampling error due to the way the observations are made.

6 Suppose you were interested in comparing the wages of government employees in a certain classification (e.g., data processing analysts) with their counterparts in private industry. Assume all the data from these populations are available to you.

Divide into four class groups and each group plan how they might go about obtaining the following kinds of samples for this study and report back:
(a) Simple random.
(b) Systematic random.
(c) Stratified.
(d) Cluster.

7 Given a population of 100 shoppers in a supermarket, would the following constitute methods for selecting a random sample?
(a) Enter the shop and grab the nearest 20 shoppers?
(b) Stand by the door and take the first 20 shoppers to emerge?

Do these procedures qualify as random sampling? Why or why not?
Discuss your answers in class.

8 Evaluate each of the following hypothetical situations, in terms of whether the method of selecting the sample is appropriate for getting information about the population of interest. How would you improve the sample design?
(a) A principal in a large high school is interested in student attitudes towards a career in business to determine which students should be allowed to take the business studies option. She lists all of the first-period classes, assigning a number to each. Then, using a random number table, she chooses a class at random and interviews every student in that class about their proposed career interests.
(b) An anthropology professor wanted to compare accountancy majors with psychology majors with respect to their attitudes toward premarital sex. She administered a questionnaire to her large class of Anthropology 437, Comparative Human Sexuality. She found no significant difference between the two groups in their attitudes, so she concluded that professionals in the two careers held similar attitudes towards premarital sex.

9 Complete the table below to provide a useful summary of the sampling designs. The type of sample is indicated in column 1. You should complete column 2, describing how each sample is drawn, and column 3, the advantages and disadvantages of each method. In completing column 3, keep in mind such things as representativeness, error, appropriateness, cost, and/or feasibility.

| Type of sample | How sample is drawn | Advantages and disadvantages |
| :--- | :--- | :--- |
| 1 | Simple Random |  |
| 2 | Systematic |  |
| 3 | Stratified |  |
| 4 | Cluster |  |
| 5 | Opportunity |  |
| 6 | Judgement |  |

10 An investigator plans to select a random sample of 1,000 subjects from the population of all employees in a large government organization to assess their levels of job satisfaction. The organization wishes to know whether there are differences between employees who work in different districts and different sections and whether there are differences based on sex and length of work experience. Work out the steps the investigator must take to obtain the sample.

11 Where might you find a sampling frame for:
Handicapped persons; golfers; banks; KFC employees; hotels in a particular tourist area?
12 Explain the differences between a probability and non-probability sample.
13 Comment on the following sampling designs:
(a) A group of skate board enthusiast wanted to generate public interest in pressurizing the local government authority to provide financial support for the building of a skate park. A small printed sheet is placed in all copies of the mid-week edition of the daily local newspaper. Readers are requested to write or phone the local authority in support.
(b) A national department store is concerned over the decline in custom. It surveys all its store credit card holders selecting every 20th name.

14 Explain how you would select:
(a) a stratified sample of 100 clinical psychologists, 100 educational psychologists and 100 counselling psychologists
(b) a quota sample of 50 heavy coffee drinkers and 50 light coffee drinkers in a shopping mall intercept study
(c) A sample of male and female hotel workers in a major holiday area to compare hourly wages

15 A manufacturer's representative organization represents a company that manufactures jewellery. So far sales have mainly been confined to a small group of specialty merchandisers. However, the company decides it wants to expand its market to include retail outlets such as department stores, boutiques, drugstores, and ladies' dress shops. It develops a new line especially designed for this expanded market. It asks its representatives to sample these markets in order to determine whether to go into fullscale production on the new line.

Here are some of the ways the manufacturer's representatives believe they could sample this new population in order to comply with the company's request. In each case, identify the target population, the type of sampling technique used and the adequacy of the technique for this particular problem.
(a) Danny, familiar with department stores and the way they buy, chooses several that he knows from prior experience are the 'trendsetters' in the retail field. Based on their reactions, he feels he can estimate how successful the new line will be in most department stores.
(b) Another representative, Kumar, in charge of drugstores, wrestles with the problem that his accounts are scattered over a fairly broad geographical area. He decides that he will drive over to one particular area that happens to contain a few drugstores that are not too far from each other. He will sample these accounts and then make a few selected phone-calls to a number of other drugstores. He hopes that this sample will enable him to estimate accurately how the new line will 'go' in drugstores.
(c) Jeri, a relatively new addition to the organization, devises a way to obtain sample reaction from boutiques. She divides her territory into shopping areas, and then randomly samples a few boutiques in each shopping centre.
(d) Alice, another new addition to the organization, decides to compile a list of all the ladies' dress shops in her territory. Then, after randomly selecting the first, she checks off every 10th one and schedules visits to each of these shops.
(e) Rosmah, in charge of the organization, decides to test samples across the board. She divides the total territory into a number of different areas. She then randomly selects a few of these areas and visits every department store, boutique, drugstore, and ladies' dress shop in that area. What kind of sampling technique has been left out? Can you see any reasons why this technique was not used?

16 Using the formula for the SEm, fill in the value of SEm at each intersection of the following table. What is the relationship between SEm and N?

|  | Standard deviation |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| N | 1 | 24 | 4 | 8 |
| 1 |  |  |  |  |
| 16 |  |  |  |  |
| 64 |  |  |  |  |

17 The Association of Emeritus University Women (EUW) is a semi-formal liaison group at the University of the West. It meets regularly and makes submissions about issues related to female staff and students. At the end of a recent meeting the secretary thought it would be a good idea to obtain the opinions of members on how effective they perceived EUW to be in enhancing the role of women on campus. She decided she would ask a few quick questions to members as they left through the door. Discuss this case in class and comment on the secretary's procedure. Can you suggest and plan another method?

18 A national magazine believes that consumers aged 18-30 are becoming the nation's biggest spenders. They decide they must appeal to this group to maintain and increase sales. Discuss and design in class what sort of sampling plan you would use to assess the tastes of this age group.

19 In what context would a multi-stage sample be quite useful?
20 Why might a stratified sample be superior to a simple random sample?
23 What are the problems with the following as ways of obtaining data:
(a) A major multiple supermarket located in a large shopping mall employs interviewers to stay in the parking area for four Saturdays in October and ask the drivers when they have parked their zip or postal code. The supermarket is interested in where their customers come from and whether it would be worth while opening a branch in another location.
(b) To assess the popularity of a new television game show, the Continental Network invite viewers to SMS a particular telephone number if they would not want to watch it again. The call will cost \$1.90.
(c) A supermarket recently undertook a renovation enabling it to expand its display areas and range of items. To gauge customer reaction, each customer was asked several short questions by the counter staff at the till after paying.

## ANSWERS TO MULTIPLE CHOICE QUESTIONS

1 (b), 2 (c), 3 (a), 4 (c), 5 (b), 6 (b), 7 (a), 8 (b), 9 (a), 10 ( d), 11 (d), 12 (b),
13 (b), 14 (c), 15 (b), 16 (c), 17
(d) ), 18
(c), 19
(a).

## ANSWERS TO QUESTIONS IN CHAPTER 9

Qu. 9.3 (a) 98 102; (b) 99-101; (c) 99.33-100.66
Qu. 9.4
(a) The standard error of the proportion is given by:

$$
\mathrm{SEp}=\sqrt{\frac{\mathrm{pq}}{\mathrm{n}}}=\sqrt{\frac{0.144 \times 0.586}{2000}}=0.011 \text { or } 1.10 \%
$$

Using the sample proportion as an estimate of the population proportion we can say, as before, that we are: $68 \%$ certain that the true percentage of motorists whose cars went in for repair and servicing would be contained in the range $41.4 \% \pm 1.10 \%$;
$95 \%$ certain that the true percentage would lie within $41.4 \% \pm(1.10 / 1.96) \%$; and
$99 \%$ certain that it would lie within $41.4 \% \pm(1.10 / 2.58) \%$.
(b)

| Sample size | Standard error |
| :--- | :--- |
| 1000 | 1.56 |
| 4000 | 0.78 |
| 8000 | 0.55 |

This shows that doubling the size of the sample does not halve the error. To reduce the standard error by half it is necessary to increase the sample size fourfold. At the same time, by halving the size of the original sample (i.e., if only 1,000 motorists had been interviewed instead of 2,000 ) the standard error would have been $1.56 \%$ and not $1.10 \%$. It may be that for the purposes of this survey, a sample of 1,000 would have been quite sufficient. Although many of the costs of sampling (e.g., planning, questionnaire preparation, contacting, and tabulation costs), do not go up proportionally with the increase in sample size, the costs of interviewing and questionnaire production and postage or phoning (where applicable) do increase more or less directly. Therefore, the level of accuracy considered acceptable for decision-making purposes has to be set against the cost of getting it.

Qu. 9.6
All six definitions are inadequate because:
1 They fail to describe the geographic boundaries of the test area (the capital city).
2 Definitions A and B are too broad. A 'purchaser' of a magazine is not necessarily the reader. Similarly, a subscriber is not necessarily the reader (for example, if a parent subscribes on behalf of their student son or daughter). Since, it is the reader's opinions that are of interest these definitions are unacceptable.

3 Definitions C and D are vague and ambiguous. In these definitions, a 'reader' is anyone who has ever read the magazine. This includes those who have read the most recent issue and those who read an issue four years ago but not since. Similarly, 'any Body Builder magazine within the past 30 days' does not necessarily mean that the magazine read is one of the more recent issues. These definitions are inadequate because the editor is interested in responses to advertising in recent issues.

4 Definition E solves many problems of the prior definitions. This definition defines a reader in terms of recent reading ('new issues') and the three of four issue criterion helps to assure that the reader has adequate experience with the test issues so that reasonable opinions can be formed. However, the age boundaries make this definition unacceptable. This sample definition would bias the survey because younger and older individuals meeting the readership criterion are excluded from the study.

5 Definition F is the best option among the presented definitions. The lower age boundary is reasonable from a reader population perspective. However, while this definition adequately defines the sample population in terms of recent readership it does not provide any guidance in terms of demographic or geographic information. The term 'individuals,' for example is still quite vague.
A more complete test target population definition might be:
'Men and women over the age of seventeen who have read at least three of the last five issues of The Body Builder magazine within the past three months. These individuals will either (a) obtain their magazine by subscription and reside within the capital city limits or (b) purchase the minimum of three new issues from retail outlets found within the capital city limits. The area of the capital city is that as defined in the capital city Planning Act 1995’.
Qu. 9.7 Yellow pages most likely under-registers the sampling frame as some bookshops will not have advertised there.

Qu. 9.8
(a) The population of library card holders would be the theoretical population but the sampling frame would be better formed from those who had used the various services of the library (e.g. for research or borrowing) in the last three months. Many library card holders may not have used the library services recently.
(b) The company will have a full list of current employees and this will act as both population and sampling frame.

Qu. 9.9 A stratified sample is better than a simple random sample since it increases the chances that relevant strata in the population are accurately represented.

Qu. 9.10 Multi-stage cluster sampling is extremely useful when a population is highly dispersed. It reduces time and cost of interviewing.

Qu. 9.11
(a) Obtain marital and parental status from employees' records. The sampling frame is therefore the employees' records. A stratified random sample is appropriate.
(b) Use a multi-stage sampling. For example, randomly select two big, two medium and two small towns on the east coast then two warehouses from each of the towns. Check pilferage rates by comparing records of supplies, stock sold and balance in warehouse.
(c) Problems here involve identifying drug users and defining antisocial behaviour. Sources of drug abusers may depend on access to medical records and court/police records. Cluster samples in several major centres seem required as the case would be scattered all over the country. With a clear definition of antisocial behaviour, interviews with employers of sample members would then be necessary.

Qu. 9.12 Among the many issues you could raise are:

- Most people are at work and will not be shopping, therefore you are only tapping a very specific population.
- What does 'people' mean? Could a child be interviewed, for example.
- On what basis would a person be stopped for interview?
- There is no consideration of important strata such as age, singles, married, etc.
- Position outside one shop would bias interviewees to those who probably use that shop.
- A selection of times through the day need to be used.
- Other days should also be used. Different groups may shop on different days, depending when paid or pension paid.
- Shoppers who use shopping centres in the suburbs would not be interviewed.


## ANSWERS TO QUESTIONS ON WEB PAGE

1 A census involves the whole population; a sample is a randomly selected set of the population.
2 Precision and cost.

3 It introduces subjective bias which can result in a misleading sample.
4 Increased precision with a smaller sample; additional information on each strata.
5 Multi-stage cluster sampling - e.g. provinces, districts, settlements.
6 At the final stage of quota sampling selection is non-random.
7 Stratified sampling within the three account levels with sample sizes proportional to the number of accounts in each stratum.

8 For example, an accountant might make a random selection of suppliers' accounts. The entries in a selected account form a cluster, all of them being examined.

9 (a) New line of luggage. Sample more appropriate as test will destroy luggage.
(b) Shoplifters who have been caught. Population too dispersed, therefore carefully selected purposeful sample.
(c) All voters; sample as time is a constraint.
(d) Population too large and many will not respond/be inaccessible, therefore carefully selected random sample from previous sales list

10 Children lose, forget, destroy the questionnaire. The survey also leaves out voters who currently have no children or have no children at that school.
$11 \mathrm{a}=$ judgemental; $\mathrm{b}=$ random sampling; $\mathrm{c}=$ convenience; $\mathrm{d}=$ random; $\mathrm{e}=$ random.
$127,12,17,22$.
13 The selection of the first individual controls successive selections.

14 Convenience sample; will not provide a true reflection of what citizens think, i.e. cannot be generalized
$15 \mathrm{a}=$ systematic $; \mathrm{b}=$ convenience; $\mathrm{c}=$ random sample of people with phones.
16 A random sample implies that all members of the population have an equal chance of selection.
17 Time and cost benefits when population is dispersed.
18 No, as not a complete sampling frame of all executives in that city; you would have a random sample of those who advertise in YP.

19 (i) Go through the entire student list and select (say) every tenth name.
(ii) It is important to realize that the pile may not be uniform: for example, the larger lumps of material may have sunk or rolled to the bottom: material on the outside may have weathered and changed its composition. It is therefore necessary to take many small samples from say top, middle and bottom, on the north, south, east and west faces from surface, two feet down and four feet down - perhaps 36 small samples in all - and mix them thoroughly. Chemical firms have evolved quite elaborate procedures for carrying out this type of operation.
(iii) It is similarly important to be aware that the field will not be uniform even over quite a small area; one part may be well-drained and another slightly waterlogged; one part sandy and another clay; one part shaded for part of the day, and so on. To eliminate the effects of these variations, the field is divided into a large number of numbered squares and a set of these is chosen to be the sample using a table of random numbers.
20. Yes, because elements are randomly selected from each stratum.
21. Different selection ratios are used for each strata or subgroup.

## ADDITIONAL MATERIAL - SAMPLE SIZE DETERMINATION

## Some factors relating to sample size

A major question that needs to be satisfied in any sampling is, 'Is the sample large enough?' In a sense, any size sample, properly selected, is big enough to say something relative to the population from which it has been selected. Instead, a more appropriate question is, 'Is the sample adequate for the desired accuracy and confidence required?' Several inter-related factors affect sample size determination.

## The population and characteristics to be measured

The first consideration in determining the size of a sample is the characteristics of the population being sampled and the characteristics or variables to be measured or observed. The more heterogeneous the individuals in a population are with regard to a characteristic to be measured, the larger will be the sample required to measure that characteristic accurately; the more homogeneous they are, the smaller the sample required.

Another general factor affecting the sample size is the number of subgroups of interest in the research. For example, making estimates of average salaries for public servants as a whole group will require a smaller sample than will be required for subgrouping by department (stratification). The more subgroupings required, the greater the total sample size.

## Precision (accuracy) and confidence

The essential goal of any study is to obtain precise or accurate estimates of population parameters within time and resource constraints. What constitutes precise data may be debatable. Often, the researcher can only use their judgement to decide what constitutes the minimum acceptable precision (or allowable error) for data to be gathered. The decision relating to the standards for precision depends upon the answer to the following question:

## How much sampling error can be tolerated in the estimate or what precision is desired?

We don't just want to obtain accurate estimates of the population parameter but also have a strong degree of confidence that this is so. The concept of confidence in sample estimates and generalizations is expressed in terms of the confidence interval and the confidence level.

- A confidence interval is an estimate of the range in which we believe the true population estimate lies. For example, it is common to read or hear that: 'Eighty percent of all adults surveyed agree that there need to be major changes to the income tax rates. The confidence interval is $\pm 2 \%$. This means that the true percentage of voters agreeing with the statement probably lies between $78 \%$ and $82 \%$. Is it permissible for the estimate of the population mean for salaries of bank managers to be within $\$ 200$, $\$ 100$, or $\$ 50$ of the true value? Is it important that the sample percentage favouring a given issue be within two percentage points of the true percentage in the population, or can a sampling error as high as four or five percentage points be accepted? Responses to these questions of permissible sample errors can greatly affect the size of sample required for a survey. In general, the smaller the error (the better the precision), the larger the sample size required. We therefore need to balance accuracy with costs of obtaining the data. The limits established by the acceptable error or desired precision are referred to as 'confidence limits or interval.'
- The confidence level is the mathematical expression of our confidence that the population estimate lies within the confidence interval. It depends upon the answer to the following question:


## How certain does the researcher wish to be that the calculated confidence limits, based upon the sample statistic, do in fact include the parameter being estimated?

Is $99 \%$ confidence necessary? Or will only $95 \%$ certain be acceptable? Or even $90 \%$ ? For example, a confidence level of $95 \%$ means that there is a $95 \%$ probability that the population estimate derived from the sample lies within the identified confidence interval; a confidence level of $99 \%$ infers $99 \%$ probability. We met these intervals in Chapter 8 where they were the significance level boundaries, enclosing the same areas of the normal curve. The confidence level represents the probability, or odds, that the parameter falls within the confidence limits derived from the sample data. So just as with the normal distribution, we are saying that if a statistic like a sample mean falls within the range +2 to -2 standard errors (deviations with the normal curve) of the distribution mean then we know that $95 \%$ of sample means also lie there and the population mean will lie within that area.

There is usually no way to be $100 \%$ confident that the parameter of interest lies within the confidence limits - unless a complete census is taken. Therefore, researchers and the users of sample data must be willing to accept something less than $100 \%$ confidence that estimates of the population parameter made from the sample do not exceed the permissible sampling error. You should not feel bound by convention to use one of the conventional levels of significance just because they are commonly used. Any level may be selected, but it must be reported in the presentation of the results.

Within limits, the greater the confidence level required, the larger the sample size. Therefore, for the sake of efficiency and economy the researcher should select a confidence level no higher than that required to meet the objectives of the study.

Thus sample size is determined in light of confidence intervals and confidence levels. Greater precision in either or both levels requires larger sample sizes. Thus, the most important step in sample size determination occurs when you explicitly state your desired confidence interval and confidence level. Once this is done, there are several ways for determining the appropriate sample size.

## Non-statistical considerations

These factors are focused around resources available, such as available time and finance, the existence of an adequate population listing or sampling frame and the accessibility to the population and sample. These do form very realistic constraints.

Qu. 1
What is the difference between precision and confidence and how are they related to sample size? Check your answer below.

## Sample size in probability samples

It may appear logical that the size of sample needed to provide precise estimates about population characteristics is directly related to the size of the population, but that is not true. An examination of the formula for computing the standard errors of various statistics show that the emphasis should be placed not on the number of individuals in the population, but on the number of individuals in the sample. Or, to put it another way, the standard error of a statistic depends on the size of the sample not on the size of the population.

Generally, larger samples permit greater confidence in population estimates and generalisations. But, increases in confidence do not increase in a linear, one-to-one relationship with increases in sample size. Quite large increases in sample size are required for small increases in confidence. The goal in determining sample size, therefore, is the determination of the minimal sample size that will provide the desired degree of confidence in the population estimates.

## Sample size determination when estimating a population mean

To compute the sample size required we need:

- the population standard deviation of the variable
- a chosen confidence level and
- a confidence interval range.

These last two choices are subjective based on the perceived needs or demands of the study. Because the population standard deviation is rarely known, it is typically estimated in one of three ways:
1 Estimate the standard deviation from a prior, similar study conducted among the same population,
2 Estimate it from a small pilot study,
3 Divide the range of response by six. You recall from Chapter 3 that the usual spread of scores along the baseline of a normal curve is around 6 SD's. This procedure is fairly accurate if the population is large and tends to normal distribution, with the largest number of items clustering in the middle of the distribution.

Once the standard deviation is estimated and the confidence interval and confidence level are chosen, sample size can be estimated through the formula:

$$
\text { Sample size }=\left(\frac{\sigma \mathrm{Z}}{\mathrm{e}}\right)^{2}
$$

where:
$\sigma$ is the estimate of the population standard deviation;
Z is the two-tailed value of the standardized deviation of the normal distribution associated with the desired level of confidence;
e is the desired precision or acceptable error.

## Example

Suppose a finance director of a super-annuation fund wishes to estimate the average monthly retirement income from the large population of retired person's drawing on the fund. An acceptable error in the sample mean, as an estimate of the population mean, is $\pm \$ 75$ per month, with an associated confidence level of $90 \%$.

The research director must specify values for $\mathrm{e}, \mathrm{Z}$, and $\sigma$ in advance. In this example, $\varepsilon$ has been set equal to $\$ 75$, and the appropriate value of $Z$ (at the $90 \%$ confidence level) is 1.645 . Usually, the difficult value to estimate is $\sigma$. If no prior studies are available, the research director could estimate $\sigma$ by assuming a largest and smallest value for monthly retirement income being drawn. If the largest is $\$ 3,600$ and the smallest is $\$ 600$, the estimate of the population standard deviation is:

$$
(\$ 3,600-\$ 600) / 6=\$ 500
$$

Substituting the appropriate values yields the following estimated sample size:

$$
\begin{aligned}
\text { Sample size required } & =\left[\frac{\sigma \mathrm{Z}}{\mathrm{e}}\right]^{2} \\
& =\left[\frac{(\$ 500)(1.645)}{75}\right]^{2} \\
& =153.8(\text { or a sample of } 154)
\end{aligned}
$$

If we reduce the acceptable error to $\pm \$ 25$, with the confidence level and estimate of $\sigma$ being the same, the estimate of the required sample size then becomes:

$$
\begin{aligned}
\text { Sample size } & =\left[\frac{(\$ 500)(1.645)}{25}\right]^{2} \\
& =1383.8(\text { or a sample of } 1384)
\end{aligned}
$$

Decreasing the error or wanting to be more accurate causes a considerable increase in sample size as sample size estimates are highly sensitive to the variability in all three elements of the formula. Similarly increasing the population SD and/or the level of confidence also increases sample size.

For the above example, consider what happens to the estimated sample size when the confidence level is increased to $95 \%(Z=1.96)$ :

$$
\begin{aligned}
\text { Sample size } & =\left[\frac{(\$ 500)(1.96)}{25}\right]^{2} \\
& =2209(\text { instead of a sample of } 1384)
\end{aligned}
$$

Thus the estimated sample size increases as both the variability in the population and the desired confidence level increase; conversely, it decreases as the allowable error and desired confidence decrease.

Qu 2
(a) Imagine that you ask a group of respondents: 'On a scale of one to five, please rate the credibility of Koolroom airconditioning product advertising?' You want to be $95 \%$ confident that the average rating estimated by the research is within plus or minus .2 points of the true population mean. You estimate the sample standard deviation to be 1.5, derived by adding the extremes of the rating scale and dividing by 4 (i.e., $5+1+4=1.5$ ). Calculate the required sample size.
(b) Recalculate the sample size for the above Koolroom airconditioning advertising if we raise the confidence level to $99 \%(z=2.56)$ and lower the confidence range to within plus or minus .1 points of the true population mean.
Check your answers below.
In sum, the relationship between sample size, confidence and accuracy is an important consideration when planning an investigation. When evaluating alternative sample sizes you need to ask yourself: 'Is the increased cost of precision justified by the increased costs of data collection from a much larger sample?'

Where the population size is known, e.g. number of employees in a specific company, the formula below can be used:

$$
\text { Sample size }=\sigma^{2} /\left[(\mathrm{e} / Z)^{2}+(\sigma / \mathrm{N})\right]
$$

where:
$\sigma$ is the estimate of the population standard deviation;
Z is the two-tailed value of the standardized deviation of the normal distribution associated with the desired level of confidence;
e is the acceptable error;
N is the known population size.

## Example

Suppose a professional accountancy association is planning a new type of major medical insurance policy for its 20,000 members and needs to know the mean age of these accountants to within one year of the true mean age (i.e. $\pm 1 \mathrm{yr}$ is the acceptable error), with $95 \%$ confidence. How large a sample will be needed to yield this information? Using data from a study made a few years earlier, the standard deviation of accountants' ages in the population is estimated at 7 years. Substitution of this information yields:

$$
\begin{aligned}
\mathrm{n} & =\sigma^{2} /\left[(\varepsilon / \mathrm{Z})^{2}+(\sigma / \mathrm{N})\right] \\
& =7^{2} /\left[(1 / 1.96)^{2}+(7 / 20000)\right] \\
& =49 /[0.26+.00035] \\
& =188
\end{aligned}
$$

If the confidence level is relaxed to $90 \%$, the resultant sample size becomes:

$$
\begin{aligned}
& \mathrm{n}=7^{2} /\left[(1 / 1.645)^{2}+(7 / 20000)\right] \\
& \mathrm{n}=132
\end{aligned}
$$

## Sample size determination when estimating a population proportion

Suppose the issue of opening up the local bus service monopoly to other private companies arose and the local politician wished to gauge the feeling about this issue among the registered voters. This situation calls for the estimation of a population proportion $(\pi)$. A listing of 50,000 registered voters is available for sampling. The local politician is interested in having the estimated sample proportion (p) that support the initiative to be within 0.05 or $5 \%$ of the true proportion, with an associated confidence of 0.90 .

$$
\text { Sample size }=\frac{\pi(1-\pi) \mathrm{Z}^{2}}{\mathrm{e}^{2}}
$$

where:
$\pi$ is the estimate of the proportion in the population;
Z is the two-tailed value of the standardized normal deviate associated with the desired level of confidence; and
e is the desired precision, or acceptable error.
This formula requires some preliminary estimate of the value of $\pi$. This may seem unusual since the reason for sampling is to estimate $\pi$. Fortunately, the product of $\pi(1-\pi)$ assumes a maximum value when $\pi$ $=.50$. For any other value of $\pi$, the product is less than .25 . Hence, when no prior information is available for the value of $\pi$, a safe assumption is always to make $\pi=.50$.

Returning to the example, the following values are substituted:
$\pi=.50$ (the value assumed since nothing is known a priori);
$Z=1.645$ (the two-tailed standardized deviation of the normal distribution associated with $90 \%$ confidence);
$\mathrm{e}=.05$ (the acceptable error)

$$
\begin{aligned}
\text { Sample size } & =\frac{.50(1-.50)(1.645)^{2}}{(.05)^{2}} \\
& =\frac{(.25)(2.706)}{(.0025)}=.6765 / .0025=270.6(\text { or a sample of } 271)
\end{aligned}
$$

Hence, a random sample of 271 persons from the listing of 50,000 registered voters will provide a maximum error of .05 with $90 \%$ confidence. (Since rounding will affect the sample size, at least four significant digits should be used.)

Suppose a similar issue in another district received $35 \%$ of the votes in favour. If the researcher is willing to assume that these results provide an acceptable preliminary estimate of the proportion against the issue, the initial estimate of $\pi$ becomes .35 . (Remember the characteristic of interest is the proportion of registered voters for the issue.) Substitution into the formula yields the following adjusted sample size:

$$
\begin{aligned}
\text { Sample size } & =.35(1-.35)(1.645)^{2} /(.05)^{2} \\
& =(0-35)(.65)(2.706) /(.0025) \\
& =.6156 / .0025=246.2(\text { or a sample of } 247)
\end{aligned}
$$

Hence, if the true value of $\pi$ lies close to .35 , a sample of 247 will provide an estimate for $\pi$ with an error not exceeding . 05 with an associated $90 \%$ confidence.

As a final example, suppose an estimate for $\pi$ is not known and the researcher again wishes the error not to exceed 0.05 but wants an increased confidence of $95 \%$. The only change from the first set of computations is that Z now equals 1.96 .

$$
\begin{aligned}
\text { Sample size } & =.50(1-.50)(1.96)^{2} /(.05)^{2}=(.25)(3.842) /(.0025) \\
& =.9605 / .0025 \\
& =384.2(\text { or a sample of } 385)
\end{aligned}
$$

Increasing the confidence from $90 \%$ to $95 \%$, when the value for $\pi$ is assumed at .50 , raises the sample size from 271 to 385 , or an increase of 114 persons. This is an excellent illustration of the need for researchers to weigh survey constraints against confidence and acceptable error. The question whether the increased confidence is worth the additional cost must be assessed within the framework of the objectives and available resources. The answer to this question is not statistical but practical.

Qu 3
(a) Imagine that you ask a group of respondents 'Are you aware of advertising for Kleen office supplies?' and you anticipate the proportion saying 'yes' to be $35 \%$. Additionally, you want to be $99 \%$ confident that the actual proportion estimated by the research is within $\pm 2 \%$. What is the required sample size?
(b) In the above example, what happens when you change the confidence interval to $\pm 0.4 \%$ and the confidence level is lowered to $95 \%$ ? Check your answers below.

Tables A and B below provides sample size determination for the following conditions:
Allowable errors of $.01, .02, .03, .04$, and .05
Confidence levels of .95 , and .99 .
Table A Random sample sizes for specific permissible errors expressed as absolute proportions when the true proportion in the population is 0.50 and the confidence level is $95 \%$

| Population size | Sample size for permissible error (proportion) |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  | $\mathbf{0 5}$ | $\mathbf{0 4}$ | 86 | $\mathbf{0 3}$ | $\mathbf{. 0 2}$ |
| 100 | 79 | 150 | 91 | 96 | $\mathbf{. 0 1}$ |
| 200 | 132 | 200 | 168 | 185 | 196 |
| 300 | 168 | 240 | 234 | 267 | 291 |
| 400 | 196 | 273 | 291 | 343 | 384 |
| 500 | 217 | 300 | 340 | 414 | 475 |
| 600 | 234 | 323 | 384 | 480 | 565 |
| 700 | 248 | 343 | 423 | 542 | 652 |
| 800 | 260 | 360 | 457 | 600 | 738 |
| 900 | 269 | 375 | 488 | 655 | 823 |
| 1,000 | 278 | 462 | 516 | 706 | 906 |
| 2,000 | 322 | 500 | 696 | 1,091 | 1,655 |
| 3,000 | 341 | 522 | 787 | 1,334 | 2,286 |
| 4,000 | 350 | 536 | 842 | 1,500 | 2,824 |
| 5,000 | 357 | 546 | 879 | 1,622 | 3,288 |
| 6,000 | 361 | 553 | 906 | 1,715 | 3,693 |
| 7,000 | 364 | 558 | 926 | 1,788 | 4,049 |
| 8,000 | 367 | 563 | 942 | 1,847 | 4,364 |
| 9,000 | 368 | 566 | 954 | 1,895 | 4,646 |
| 10,000 | 370 | 577 | 964 | 1,936 | 4,899 |
| 15,000 | 375 | 583 | 996 | 2,070 | 5,855 |
| 20,000 | 377 | 586 | 1,013 | 2,144 | 6,488 |
| 25,000 | 378 | 588 | 1,023 | 2,191 | 6,938 |
| 30,000 | 379 | 591 | 1,030 | 2,223 | 7,275 |
| 40,000 | 381 | 593 | 1,039 | 2,265 | 7,745 |
| 50,000 | 381 | 595 | 1,045 | 2,291 | 8,056 |
| 75,000 | 382 | 1,052 | 2,327 | 8,514 |  |
| 100,000 | 383 |  | 1,056 | 2,345 | 8,762 |
|  |  |  |  |  |  |

Table B Random sample size for specific permissible errors expressed as absolute proportion when the true population is $\mathbf{. 5 0}$ and the confidence level is $\mathbf{9 9 \%}$

| Population <br> (proportion) size | Sample | size | for | permissible | error |
| :--- | :---: | :---: | :---: | :---: | :---: |
|  | $\mathbf{. 0 5}$ | $\mathbf{. 0 4}$ | $\mathbf{. 0 3}$ | $\mathbf{. 0 2}$ | $\mathbf{. 0 1}$ |
| 100 | 87 | 91 | 95 | 98 | 99 |
| 200 | 154 | 168 | 180 | 191 | 198 |
| 300 | 207 | 233 | 258 | 280 | 295 |
| 400 | 250 | 289 | 329 | 365 | 391 |
| 500 | 285 | 337 | 393 | 446 | 485 |
| 600 | 315 | 380 | 453 | 524 | 579 |
| 700 | 341 | 418 | 507 | 599 | 672 |
| 800 | 363 | 452 | 558 | 671 | 763 |
| 900 | 382 | 482 | 605 | 740 | 854 |
| 1,000 | 399 | 509 | 648 | 806 | 943 |
| 2,000 | 498 | 683 | 959 | 1,349 | 1,785 |
| 3,000 | 543 | 771 | 1,142 | 1,741 | 2,541 |
| 4,000 | 569 | 823 | 1,262 | 2,036 | 3,223 |
| 5,000 | 586 | 859 | 1,347 | 2,267 | 3,842 |
| 6,000 | 597 | 884 | 1,410 | 2,452 | 4,406 |
| 7,000 | 606 | 903 | 1,459 | 2,604 | 4,923 |
| 8,000 | 613 | 918 | 1,498 | 2,731 | 5,397 |
| 9,000 | 618 | 930 | 1,530 | 2,839 | 5,835 |
| 10,000 | 622 | 939 | 1,556 | 2,932 | 6,239 |
| 15,000 | 635 | 970 | 1,642 | 3,249 | 7,877 |
| 20,000 | 642 | 986 | 1,688 | 3,435 | 9,068 |
| 25,000 | 646 | 996 | 1,717 | 3,557 | 9,972 |
| 30,000 | 649 | 1,002 | 1,737 | 3,644 | 10,682 |
| 40,000 | 653 | 1,011 | 1,762 | 3,758 | 11,372 |
| 50,000 | 655 | 1,016 | 1,778 | 3,830 | 12,456 |
| 75,000 | 1,023 | 1,799 | 3,930 | 13,585 |  |
| 100,000 | 658 | 1,026 | 1,810 | 3,982 | 14,229 |
| 500,000 | 1,035 | 1,837 | 4,113 | 16,057 |  |
| $1,000,000$ | 1,036 | 1,840 | 4,130 | 16,319 |  |
|  |  |  |  |  |  |

These tables are based on the assumption that the true value of $\pi$ is .50 , hence providing the maximum sample sizes for the given allowable errors and confidence levels. To illustrate the use of Table 10.1 consider the following situation. A conference has just ended that exposed attendees to a number of new computerized techniques. Two months after the conference the sponsor wants, among other items, an estimate of the proportion of the 500 attendees who have implemented at least one of the new techniques. A confidence level of $95 \%$ and an acceptable error of .05 are agreed upon. Also with no prior knowledge about $\pi$, the value of .50 is assumed). Table A suggests that a sample size of 217 be used.

As another example, Table A illustrates, for example, that if 1,000 attendees were at the conference, the sample size, with the same allowable error and confidence level, would rise to 278 . The sample size relative to the population size has decreased from $44.25 \%(217 / 500)$ to $27.8 \%(278 / 1,000)$.

## Sample size in non-probability samples

Sample sizes in this form of sampling are based on different sorts of judgement than for probability samples. Confidence intervals and confidence levels are obviously precluded since probability is not involved. The main methods include:

Unaided judgement. This is the most arbitrary approach to non-probability sampling size. Determination, i.e. 'I feel that 100 (or 500) will do'. It is whatever you feel comfortable with, but there is no assurance that the sample is large enough or representative. This approach should be avoided.

What will the budget allow? Here cost dictates the size obtained. Obviously a realistic constraint, but could reduce the reliability and validity of the results.
Frame of reference. This is a more reasonable approach. First you determine the sample size that others have used for similar studies. Historical precedence is the merit of the approach, but be cautious as there may be no obvious validity in the rationale underlying decisions in previous studies.
Analytic requirements. This is the best method. It is recommended that the minimum number of individuals or observations in subgroups should be 100, while there are at least 25 individuals in minor subgroups.

Because non-probability samples are non-random samples commercial market researchers have found that to obtain the same accuracy as that produced by a random sample of 100 would require a quota sample of 600 . This six-fold increase is because of the greater risk of both statistical and non-statistical errors occurring in quota sampling. Luckily, for many commercial purposes the levels of accuracy required are not so demanding as to require perfect random sampling. As you would expect with commercial researchers, practical considerations strongly influence the choice of research method, which is why quota sampling is popular

It is also fair to say that with the great advances made in quota sampling methods, with greater controls over quota-setting, better selection and training of interviewers, the accumulated experience of professional research organizations in working with quota samples over many years, more experts in the field are coming round to the view that quota sampling is quite reliable, provided it is controlled by experts and provided the most careful attention is paid to establishing proper quota controls.

## ADDITIONAL MATERIAL - NON-RESPONSE AND SAMPLE SIZE

No survey ever obtains a $100 \%$ response. Non-response raises two major problems: (a) the effect of nonresponse on sample size, and (b) the possible effect of non-response on the nature of the findings. Even the best designed projects lose about $10 \%$ of their sample. There is always the likelihood that people have left the area between visits, have been admitted to hospital, or are away on vacation. Losses of this order could seriously affect the representativeness of the sample and cause increased error. The people who are not contactable on the second occasion may have for, all we know, (and we don't know) quite different characteristics from those who are contacted. Non-response can be deliberate too (i.e., refusal to return a postal questionnaire or to answer a question during an interview, which may be indicative of a certain attitude which the interviewer ought to know about).

## Adjustment for non-response in surveys

Suppose the sample size needed to estimate a mean with a stated precision and confidence is 150 . The estimated mean and its precision can be computed only from the data actually received from the sample. If only 130 replies are received, the sample size in the formula used to compute the standard error will actually equal 130 , not 150 . The standard error will be larger and, in turn, the confidence interval will be larger than expected; therefore, the estimate will not meet the precision requirements.

You therefore need to anticipate what the percentage of non-response is likely to be and compensate for it numerically by increasing the sample size. The best clues are usually from similar studies on similar populations, but often you can do little more than guess what the response rate will be. The formula below is useful in adjusting sample size to compensate numerically for expected non-response.

$$
\mathrm{n}_{\mathrm{a}}=\mathrm{n} / \operatorname{Pr}
$$

where:
$\mathrm{n}_{\mathrm{a}}$ is the sample size adjusted for the expected rate of response;
n is the preliminary estimate of sample size;
$P_{r}$ is the expected rate of response expressed as a proportion.
Suppose the preliminary sample size is 150 and only $90 \%$ of the survey instruments are expected to be returned. What should the adjusted sample size be so that 150 instruments will be returned? Substituting in, the adjusted sample size is:

$$
\begin{aligned}
\mathrm{n}_{\mathrm{a}} & =\mathrm{n} / \operatorname{Pr} \\
& =\frac{150}{0.9}=167
\end{aligned}
$$

Qu. 4 Given a sample size of 200 but only $80 \%$ are expected to be returned, what is the adjusted size required to ensure 200 surveys are returned?
Check your answer below.
However, you must not assume that increasing sample size to allow for non-response compensates for non-response bias. Increasing the sample size only helps assure a sample of sufficient size to compute arithmetically confidence limits of estimates with the desired precision for those who respond. Such estimates may or may not contain a non-response bias. The only sure way to eliminate the possibility of non-response bias in sample surveys is to eliminate non-response as far as possible. Devoting sufficient resources and time to effective follow-up procedures are the best ways to increase the survey response rate. An up-to-date sampling frame also helps reduce non-response.

## Effect of subgroups on sample size

The number of subgroups or subcategories into which the findings of the survey are to be grouped for analysis is an extremely important factor in determining sample size. The greater the number of subcategories, the larger the total sample size needed to provide population parameter estimates for the subcategories that will meet the study's minimum precision requirements. Consider a previous example in which a professional accountants association was seeking a new type of major medical insurance policy for its members and needed to know the mean age of accountants in the district with a precision of $\pm 1$ year and a confidence of 95 . A sample of 186 from the 2,000 accountants was determined as adequate for the precision and confidence level requirements.

But suppose the mean ages for each of two subgroups - male and female accountants - were required, and approximately 1,500 male and 500 female accountants are located in the district. The problem now is: How many females are needed in the sample to meet the precision requirements of the study because the smallest subgroup for which data will be analysed must be taken as the basic unit in constructing an estimate of the total sample size required.

The smallest subgroup is treated as though it were a separate population; and the size of the sample needed to estimate the mean of this subpopulation, with the stated accuracy and confidence requirements, is determined. Thus, in this example an estimate of the standard deviation of the subpopulation - judged to be 5 years - is obtained and substituted in to calculate the necessary sample size, as follows:

$$
\begin{aligned}
& \mathrm{n}=\sigma^{2} /\left[(\varepsilon / \mathrm{Z})^{2}+(\sigma / \mathrm{N})\right] \\
& \mathrm{n}=5^{2} /\left[(1 / 1.96)^{2}+(5 / 500)\right] \\
& \mathrm{n}=96
\end{aligned}
$$

A subsample of 96 females is needed to estimate with the desired accuracy and confidence the mean age for the 500 female accountants who comprise the subpopulation. The next concern is: How many accountants from the total 2,000 must be selected to be reasonably sure that about 96 females will be in the sample? To find this number, use this formula:

$$
\mathrm{n}=\mathrm{n}_{\mathrm{s}} \mathrm{~N} / \mathrm{N}_{\mathrm{s}}
$$

where:
$n$ is the estimated total sample size necessary to yield estimates in the smallest subpopulation with the desired precision and confidence level;
$\mathrm{n}_{\mathrm{S}}$ is the estimated sample size for the smallest subpopulation to yield estimates with the desired precision and confidence level;
N is the total population size;
$N_{S}$ is the size of the smallest subpopulation of interest.
Substituting in:

$$
\begin{aligned}
\mathrm{n} & =(96)(2,000) / 500 \\
& =384
\end{aligned}
$$

A total sample of 384 accountants is needed to ensure that at least 96 will be in the smallest subgroup, i.e., the number of women needed to meet the minimum precision requirements.

In most market research studies, rarely is a simple random sample used without wishing to make some type of subgroup analysis of the data. Therefore, the effect of subgroups, even in simple random samples, is an important determinant of sample size. Information about the size and distribution of subpopulations to be sampled and analysed is essential if the total sample size is to be large enough to yield estimates with the desired precision for these subgroups.

## Combining subgroups

Perhaps the single most profitable possibility in reducing the preliminary estimate of sample size is reexamination of the smallest subgroup used in making the preliminary estimate. Is a subgroup that small vital to the study? Could it be enlarged by combining it with one or more contiguous groups? Suppose in the initial planning for a study a researcher decides to analyse staff opinions in a large government department on telecommuting (working from home using computer links rather coming to the office every day) according to the position level of the staff. Five subgroups based on salary scale are identified with the following estimated numbers of staff in each of the subpopulations:


The key subgroup in establishing the sample size for the population is the ' 6 and above' group, which is estimated to have only 50 in the population. Is separate analysis of data for this subgroup vital to the study or could it be combined with position level 5 to form a larger subgroup consisting of 850 designated as 'level 5 and above'? The combining of tentative subgroupings usually results in substantial reductions in sample size requirements.

Qu. 5 Given an estimate of the population $\sigma=20$, a desired level of confidence of $Z=1.96$ (95\% confidence level) and an acceptable error (desired precision) of 5 points calculate the necessary sample size if the population $N=2,000$. (Remember to use the formula with the finite correction factor.)

Qu. $6 \quad$ A survey is to be conducted on the need for turning an area into a national park. The district governor wants to know the proportion of the adult population that support such a move to be within .05 of the true proportion with a confidence of .95 . What size of sample is required?

Qu. $7 \quad$ Required sample size is estimated at 300. But if the investigator believes that there will only be a $75 \%$ response rate what size should the sample be increased to?

## In conclusion

Every business person should ask these basic questions about any investigation or survey using sampling commissioned by them or whose findings may be used by them.

- How has the population or universe been defined?
- What sampling method was used and why?
- How big was the sample? Was it adequate?
- What is the trade-off between cost/time and accuracy/confidence?
- What was the standard error?
- What level of accuracy in the results was looked for?
- What degree of confidence can be placed in the results obtained?
- Has any adjustment been made for non-response?


## Summary

Formulae enable the calculation of sample size to achieve desired levels of confidence and precision in the results when estimating population mean and proportions. Subsample size and non-response are factors that must be taken into account when determining sample size.

## Answers to questions in additional material

Qu. 1 Precision indicates how close the sample statistics are to the population parameters. Usually a bigger sample size will give more precision as most of the population are included. Confidence indicates the percentage of times when our statistic will truly reflect the population parameter. This again is a function of sample size with a larger sample size allowing for a higher level of confidence.

Qu. 2
(a) $\quad$ Sample size $=\left[\frac{(1.5)(1.96)}{.2}\right]^{2}=216$
(b) We will need another 1,258 in our sample.

$$
\text { Sample size }=\left[\frac{(1.5)(2.56)}{.1}\right]^{2}=1474
$$

Qu. 3
(a) $=\frac{.35(1-.35) 2.58^{2}}{.2^{2}}$

$$
=3756
$$

The sample size is large because the confidence level and confidence interval demand a high level of precision.

$$
\begin{align*}
& =\frac{.35(1-.35) 1.96^{2}}{.4^{2}}  \tag{b}\\
& =546
\end{align*}
$$

The sample size decreases considerably by relaxing the confidence intervals and levels.
Qu. 4

$$
\begin{aligned}
\mathrm{n}_{\mathrm{a}} & =\mathrm{n} / \mathrm{P}_{\mathrm{r}} \\
& =\frac{200}{.8}=250
\end{aligned}
$$

Qu. 5 A sample of 34
Qu. 6384
Qu. 7400

