Sampling

How data are generated?

Quantitative reasoning involves the use of data for decision making. It is thus very important to know the relevancy and reliability of the data. One critical area that we need to pay particular attention to is the way that data are generated.

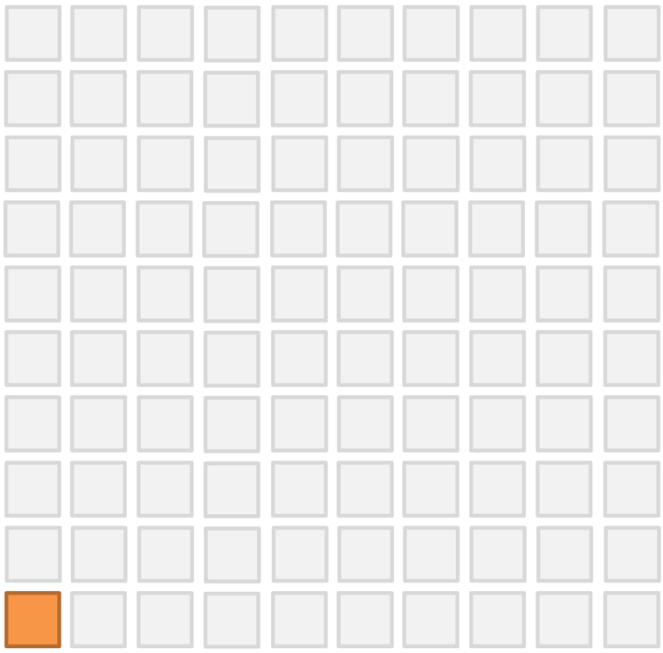
Learning Outcome of this chapter

We would be able to

- Generate our own data set as primary source of information
- Identify good existing data set as secondary source of information for our own use.

When we embark on a quantitative study, we need to decide whether we want to generate our own data set or use the one available in the market. Two critical factors that affect our decision are cost and time. It takes long time and costs more to generate our own data; but we can tailor it to suit our need. Even though it is cheaper to use data set that is already available in the market; however the data set might be collected for other purposes and it might not meet exactly what we are looking for.

Defining a Common Language



Unit

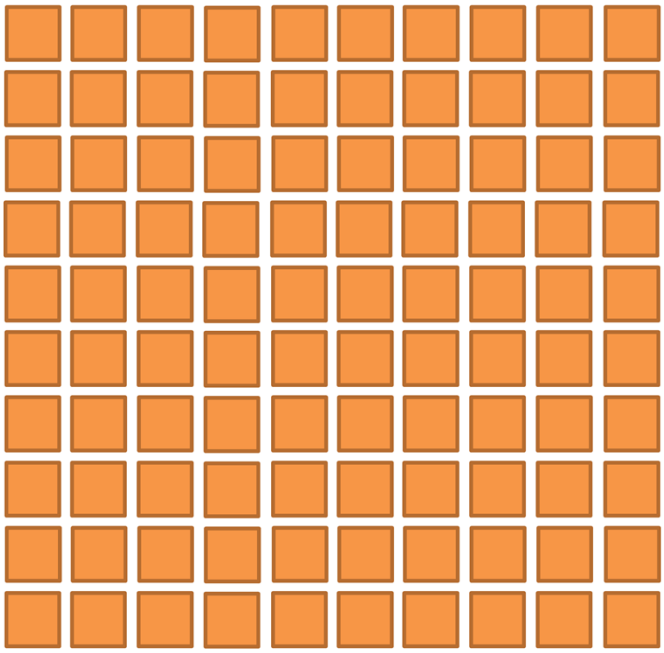
Object/individual

Based on the objective of our study, we first need to identify elements from which measurements are to be taken. These elements are called units.

Defining a Common Language

Population

Collection of units



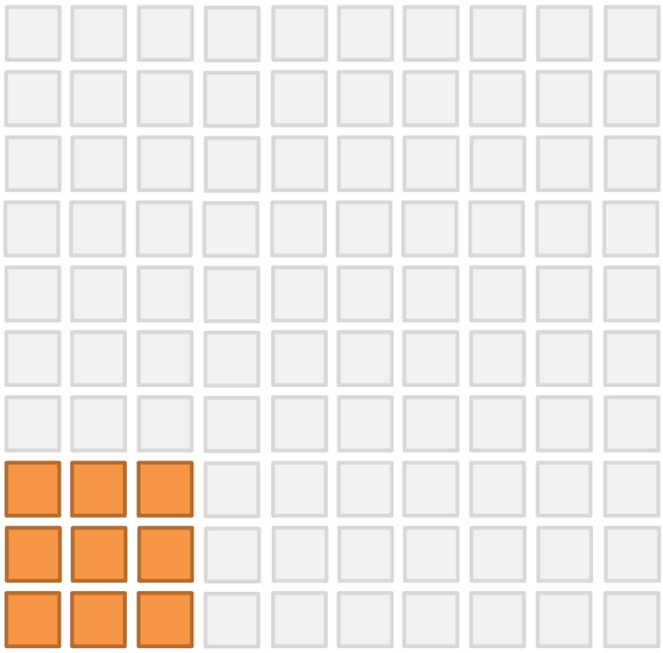
And the collection of all these units is called population. It would be great if we have sufficient resources to take measurements on all the units in the population.

aka census

Defining a Common Language

Sample

Subset



In most of the practical situation only a portion of the population would be used for our study. This portion is called a sample.

Examples

1. Quarterly unemployment rate in Singapore
   - Unit: every adult (aged 15 or over) who resides in Singapore.
   - Population: the collection of all the units defined above.
   - Sample: a portion of adults would be selected to provide their employment status.

In Singapore, Ministry of Manpower reports the unemployment rate of the country quarterly. The employment status of Singapore residents, aged 15 or over, is of great interest. Therefore, anyone who meets the criteria is the unit and the collection of all the units will form the population. It is unrealistic to collect the employment information from every unit in the population, and thus every quarter only a portion of the units from the population will provide their employment status.

Examples

2. Drug testing for a disease
   - Unit: individual who has the disease.
   - Population: the collection of all individuals defined above.
   - Sample: a collection of these individuals who will be placed under the experiment.

To determine the effectiveness of a drug for a certain disease, patients who have the disease would be the units. It is almost impossible to ask every patient to go through the drug testing. Therefore only a portion of them will take part in our study.

Census vs sample

- Census

  Measurement would be taken from every unit in the population.

- Sample

  Measurement would be taken from some selected units in the population.

In general we would like to collect measurements from every unit in the population whenever it is possible. But the task is just too huge. The most visible census that we would encounter is the population census of a country. It is conducted every 10 years as recommended by the United Nations. In most studies, only a sample would be used. Actually, there are quite a number of advantages of using sample over census in most of the quantitative studies.

Advantages of Taking a Sample Over Census

- When a census is not possible
  - Blood testing for certain disease
  - Determine the effectiveness of a drug
- Speed
- Cost
- Accuracy

There are situations when a census is not possible. For blood testing, we only provide a sample, a portion, of our blood instead of all of our blood. To determine the effectiveness of a drug we can only assign a portion of the population to the drug and another portion to placebo. Each of the two portions constitutes a sample of the population of patients. As measurements are taken from only a portion of the population, the cost will be smaller, and it is faster to obtain the results. Also because we can devote more resources to train a better task force to take measurements, we would obtain more accurate outcomes.

How to Get a Good Sample

1. Sampling Frame

When we have made a decision to take a sample instead of a census we need to put in place a mechanism to generate the sample. But first we need to understand the purpose of using sample information.

The Purpose of Using a Sample

- The main purpose of using a good sample in a study is that the results can be extended to the population from which the sample was drawn.

Since not every unit in the population is measured, we are hoping that measurements collected from units in the sample will provide adequate information such that results can be extended to the whole population. It should be noted that different samples will give different results; and therefore it is important to find a good selection scheme that provides results that are close to that of the population.

What is a Good Sample?

- Every unit in the population has a possibility of being selected.
- Selection process is not biased.

To generate a good sample, we need to make sure that every unit in the population has a chance to be selected. Units that cannot be selected should not be part of the population. Next we need to make sure that bias is not introduced to the selected process. The results derived from biased sample are skewed and distorted.

Example

- How sweet is your coffee?
  - After we add sugar to a cup of coffee, we stir the coffee before we scoop up a tea spoon of it for tasting.

To determine whether the sweetness of the coffee is just right for us after sugar is added, we first stir the coffee before we scoop up one tea spoon of coffee for tasting. The stirring action is to make sure that every drop of coffee has a chance of being selected and also to make sure that the selection process is not biased. If we do not do that, the sweetness might be too light when the coffee comes from the top part of the cup and the sweetness could be too strong when it comes from the bottom of the cup.

Basic Steps of Taking a Good Sample

1. Sampling Frame
   - Sampling frame is a list of sampling units intended to identify all units in the population.
   - The simplest sampling frame consists of a list of units in the population.

In order to make sure that every unit in the population has a positive chance of being selected, we need to be able to identify these units either directly or through a variable that has direct link to them. For example, we might not be able to identify people who live in HDB housing units directly, but we could construct a list of HDB housing addresses that link to these people. The construction of this list might be costly, however, it is feasible. This list is called sampling frame.
Of course the best sampling frame is the list of population units.

if cannot get to them directly, then must find something in common that can get to the group directly

Example

Annual Fresh Graduate Employment Survey

- Universities in Singapore conduct this study at the second half of each year to find out the employment situations of their fresh graduates.
- The sampling frame is a list of fresh graduates obtained from the Office of Registrar of the universities.

Local universities in Singapore are asked by Ministry of Education to conduct an annual study on employment status of their fresh graduates. Ministry of Manpower also wants to use the outcomes of the studies to compare with its quarterly unemployment study. The Office of Registrar of each university maintains the record of fresh graduate of the year. The sampling frame in this case would be the list of all the fresh graduates. This is the most ideal frame that we can get.

Example

Drug testing for a particular disease in Singapore

- It is easier or more feasible to identify physicians who specialize in this disease.
- Patients who have the disease would be referred by these physicians.
- The sampling frame is a list of these physicians.

In this example, we see that it is difficult to identify patients with the disease directly. We could use the list of registered physicians as a mean of reaching the units that we want. As this is a dynamic population, with new patients move into and current patients move out of the population, it is important that we set a clear cut-off time to set a fixed target population.

if not will store old data

Characteristics of a Good Frame

- "Good coverage".
- Up-to-date and complete

So what are the attributes of a good sampling frame? It has to cover exactly or bigger than the target population so that every unit in the population has a chance of being selected. If the sampling frame is too big and covers more than what we want, then the cost of getting the right units would be high. Also, the frames need to be updated regularly. This is especially important for those populations that change over time. For example, in business studies, new companies are being set up and old companies have been struck off.

Example

Quarterly unemployment rate of Singapore

- Unit: every adult (aged 15 or over) who resides in Singapore.
- Population: the collection of all the units defined above.
- Sample: a portion of adults would be selected to provide their employment status.
- Sampling Frame: a list of home addresses.

In the study of unemployment rate the unit of interest will be people who are 15 years old or older. To reach out to this group of people, the collection of home addresses in Singapore is the sampling frame.

Example (cont.)

- It is a frame that has the following issues:
  - It includes: (1) people who are under 15 years old; (2) housing units that are not occupied.
  - It excludes housing units that are just completed after the sample is generated.

This frame is not perfect because it includes people who are below 15 years old and it also excludes units that are just completed with new occupants.
We see that finding a perfect sampling frame is not easy. We need to compromise as long as the inclusion of undesired units would not increase the cost and the exclusion of units would not have major impact to the outcomes of the study.

means to say, sending the data collection to households with 15 y/o would not increase the costs than to send to a household w/o 15 y/o

and if leaving out empty households would significantly impact data

How to Get a Good Sample

2. Probability Sampling

A sampling frame that has good coverage guarantees us that every unit in the target population has a chance of being selected. The next step is to find a selection process that will not lead to a biased sample. Some people think that to achieve this we only need to take sample randomly. In other words, as long as we do not prearrange the population units for taking the measurement, then the objective can be achieved. Unfortunately, it isn't so.

Basic Steps of Taking a Good Sample

2. Probability sampling plan

Every unit in the population must have a known probability of being selected into the sample.

To make sure that there is no external influence in our selection process we use probability sampling plan. The most important key word in this plan is that the probability of selecting each unit from the population is known.

Some Probability Sampling Plans

- Simple Random Sampling

  It is the simplest probability sampling plan. For this plan, every possible sample of the same size has the same chance of being selected.

The simplest and easiest in implementation sampling plan is called simple random sampling. It gives equal chance of selecting each of the conceivable samples. This would also lead to equal chance of selection each unit from the population.

How to Selected a Simple Random Sample

1. Assign a number from *1* to *N* to each sampling unit in the sampling frame, where *N* is the number of sampling units.

The procedure of applying simple sampling plan is quite easy. It is like what is done in lucky draw. Names are drawn one by one from a box that is full of names of participants. Of course when the population size is large it would be easier by assigning a number to each sampling unit.

How to Selected a Simple Random Sample

2. Use the function randbetween(1,N) in Microsoft Excel to generate random numbers.
3. The sampling unit that has the assigned number corresponding to the generated number is selected into the sample.

We can then use random number generator, like the ranbetween function in Excel, to pick number for us. The sample unit that is assigned the selected number is entered into the sample.

- Systematic Sampling

  This is a method of selecting units from a list through the application of a selection interval, *K*, so that every *K*th unit on the list, following a random start, is included in the sample

Systematic sampling is a process that produces the same result as simple random sampling if it meets certain conditions. It is also a process that can apply to situation where the exact size of the population is not known at the planning stage. What we need to know is a rough estimate of the population size. The value of K in this sampling plan is chosen based on the exact size or estimated size of the population and also the size of the sample.

Example

Suppose there are 110 sampling units in the population. A study requires us to select a sample of 10 units. Then a random number is selected from 1 to 110/10=11. If the selected number is 6, then units 6, 17, 28, ……105 are selected to form the sample.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110

In this example, we see that the value of K is obtained by dividing the population size 110 by the sample size 10. Since a random number is chosen from 1 to 11, we see that there are actually 11 possible samples.

Some Properties of Systematic Sampling

- A systematic sample is sometimes treated as if it is selected from simple random sampling plan when the sampling units are arranged randomly.

Systematic sample can be taken as a simple random sample if the numbering of the sampling units is done randomly, because it is unlikely that the order of the sampling units is associated with characteristic of interest.

Some Properties of Systematic Sampling

- We might obtain an undesirable sample if the variable of interest in the arrangement of the sampling units and the number K have the same cyclical effect. For example, in the following two arrangements, the second arrangement will give us an undesirable sample if we take a 1 in K=3 systematic sample:

  (1) 1 1 1 4 4 4 7 7 7

  (2) 1 4 7 1 4 7 1 4 7

However, when variable of interest in the arrangement of the sampling units and the value of K have the same cyclical effect, then we might end up with undesirable sample. In the example, we see that in the second arrangement, when a 1 in 3 systematic sampling plan is used, we will end up with sample that consists of all values of 1, or 4, or 7.

Some Properties of Systematic Sampling

- We can use systematic sampling plan when the number of sampling units in the population is unknown.

  In the United States, this plan is used in exit poll on the election day. A random start from 1 to K is chosen where the value K is determined based on the number of registered voter at the voting station.

In the United States, where exit polls are allowed, media has used systematic sampling plan to select sample for making predictions on the day of election. Voting is not compulsory in the States. It is hard to know how many registered voters would cast their vote at each polling station. Thus, we can't generate simple random sample. Systematic sampling plan is just perfect for this situation. For each selected polling station, we choose the value of K that is directly proportional to the size of registered voters. Interviewer is given a random start chosen from 1 to K, and the interviewer would select and interview voters in the order they step out of the polling station.

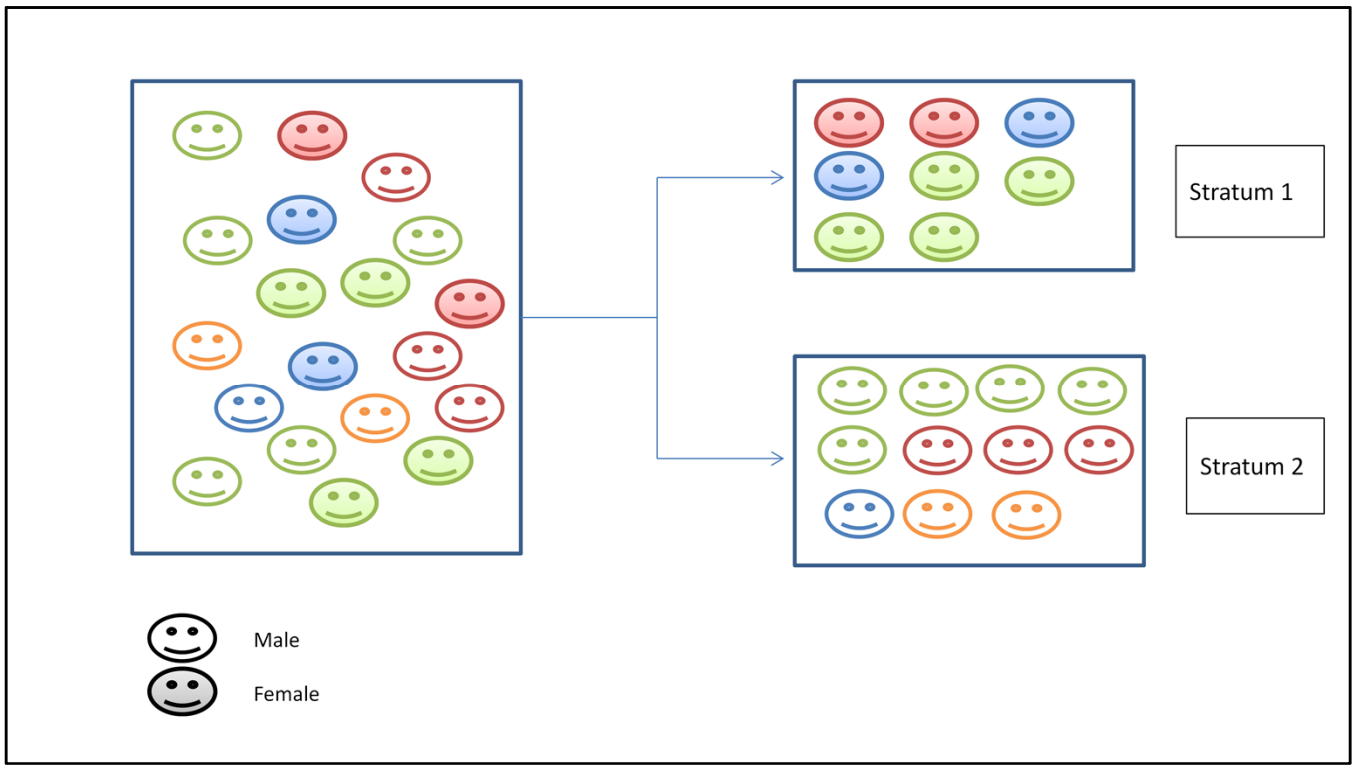
Other Probability Sampling Plans

- Stratified Sampling plan

  We first divide the population of units into groups (strata) and then we take probability sample from each group.

similar to slicing

Sometimes population units fall into natural grouping. It might be more convenient and less costly when we take advantage of this phenomenon. Therefore, instead of taking a probability sample directly from the population, we would take probability sample from each group separately. This is called stratified sampling plan. The group is called stratum. This grouping can also be done artificially. We can group the units that seem to carry similar information together to form stratum. In doing so, we can obtain an overall sample that would produce more efficient results in terms of degree of accuracy when it is extended to the population.



In this example, suppose we have a population of fresh graduates with different majors identified by the colors. The variable of interest is the starting salary. Due to the fact that male graduates fetch higher salary normally because of their experiences in national service than that of female graduates, we can form groups of male graduates and female graduates. Then, random samples are selected from the respective strata.

Example

Drug testing for a particular disease in South East Asia

- The sampling frame is a list of physicians who specialize in the disease and practice in the region.
- For the ease of administering the study, these physicians are grouped according to the countries they are practicing.

In this example, we look at drug testing for a particular disease in South East Asia. It would be more efficient when we engage local organizations to help us in monitoring the process. It is quite natural to take country as stratum and random sample is selected from each country.

Example (cont.)

- We can find the effectiveness of the drug in each country.
- It would be cheaper for the study
- We can obtain a more accurate result on the overall effectiveness of the drug.

By doing so, we can understand the effectiveness of the drug in each country. The overall cost of study is cheaper because we could set up centre in each country to administer the operation. If culture, diet, …etc have some impact on the effectiveness of the drug and the development of the disease, then by taking countries as strata we can obtain a more accurate result on the overall effectiveness of the drug.

stratum by any potential confounders

Other Probability Sampling Plans

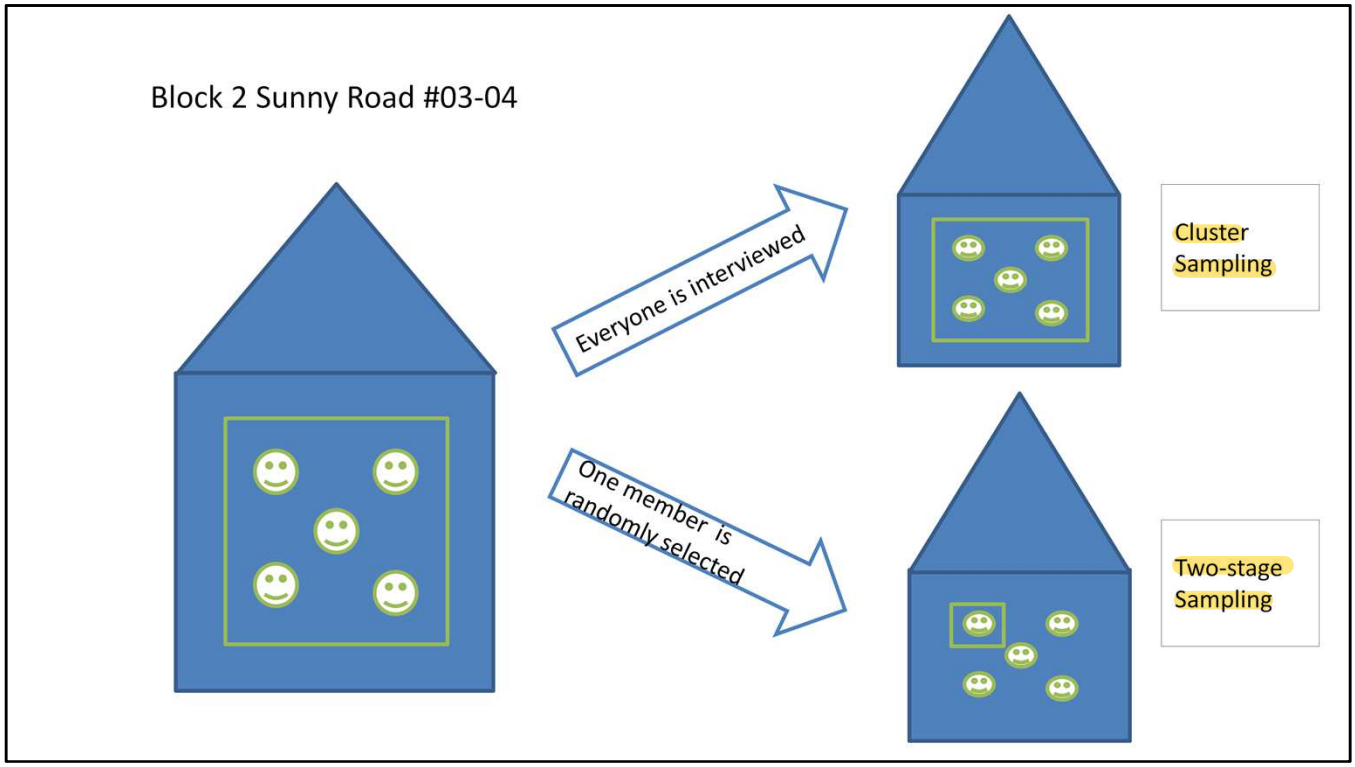
- Multistage Sampling Plan

  In a lot of studies, it might take several stages of selection before reaching the population units. At each stage probability sampling plan is implemented.

In a lot of studies the sampling units might not be the population units. After a sampling unit is selected, we could either collect measurement from every population unit associated with this sampling unit or we could select a few of them using a second probability sampling plan.

List of Addresses
Block 1 Sunny Road #02-01
Block 1 Sunny Road #02-02
.
.
.
Block 2 Sunny Road #02-01
Block 2 Sunny Road #02-02
.
.
.
Block 2 Sunny Road #03-04

In this example, the sampling frame is a list of addresses. The population units are respondents living in these addresses.



After an address is selected, we could collect employment status from everyone living at that address. This is the practice of Ministry of Manpower of Singapore and also most of other countries. The cost is high because more respondents are being interviewed. This sampling scheme is called cluster sampling and in this example each address is taken to be a cluster. In other studies that use the same sampling frame, only one eligible respondent will be selected from each selected address. The second selection will be based on a probability sampling plan as well. It cuts down the cost but the degree of accuracy of the results will suffer. This approach is very popular in market research.

Example

Drug testing of a particular disease

- Sampling frame: a list of physicians specialize in the disease.
- From each of the selected physicians, only some of the patients under the care of these physicians are included in the study.

Returning to the drug testing example, we have to rely on two-stage sampling plan to reach the patients since the sampling frame is a list of physicians.
Suppose there are two sampling frames for us to choose for our study. One is based on a list of the population units and one is not. Which is should we choose? Normally using the one that gives us the direct selection of the population units would provide better results.

Difficulties and Disasters in Sampling

On paper, designing a good sampling plan is simple and direct. But when we want to implement the plan in the real situation we might face some obstacles. Some of these obstacles can be overcome by redesigning our plan; others might have huge impact on the outcomes of the study that leads to misleading and inaccurate conclusion.

Difficulties in Sampling

- Using imperfect sampling frame
  - A perfect sampling frame is the one that consists of exactly all units of the target population.
  - In a lot of situations, a sampling frame might either include unwanted units or exclude desired units.

To find a sampling frame that covers exactly the target population is not easy especially when the target population keep changing over time. For example, if companies or enterprises are the units of interest in a business study, then constantly new companies are born and old companies are out of services. Therefore, at the time when the frame is settled, it would include those that are out of services and exclude those that are born later.

- A frame with too many unwanted units would increase the cost of study.
- For a frame that exclude desired units, we have to
  (1) redefine the target population;
  (2) assess the impact of excluding these units in our study.

Even though it is wise to use a sampling frame that is bigger than the target population, but if it is too big then the cost of the field work would be high. It is because a bigger sample has to be selected for us to screen out the ~~wanted units.~~
For the situation where desired units are excluded we can redefine our target population. For example, in the business study we can define the target population to include only companies that were set up before a certain date. Or we can roughly determine the impact on the outcomes of the study when these desired units are being excluded.

Example

Annual Fresh Graduate Employment Survey

Sampling frame: a list of fresh graduates obtained from the Office of Registrar of the universities.

On the Fresh Graduate Employment study, we can obtain a perfect sampling frame. Office of Registrar of each university keeps records of graduation information. The sampling units are the graduates.

Example

Quarterly unemployment rate of Singapore

- Sampling Frame: a list of home addresses.
  - It includes: (1) people who are under 15 years old; (2) housing units that are not occupied.
  - It excludes housing units that are just completed after the sample is generated.

In Quarterly Unemployment Rate of Singapore Study, we are interested only people who are 15 years old or older. The sampling units which are the home address might include occupants who are under 15 years old. This group of unwanted units can be screened out during the face-to-face interview session. The sampling frame also includes housing units that are not occupied. Actually in order to avoid having too many unoccupied housing units being selected, the current available sampling frame would not include newly completed housing units within a certain period of time even though some owners might have moved in during that period of time. Thus, we have excluded some desired unit. However, if the exclusion of these units does not have high impact on the variable that we are interested in, then the bias is quite small.

Difficulties in Sampling

- Non-response
  - Not all selected units are contactable.
  - Not all selected units are willing to take part in the study.
  - Non-response distorts the results of studies.
  - Usually non-respondents differ from respondents. We need to study the extent of the effect in order to reduce the bias of the collected information.

Another difficulty in sampling is to deal with non-response. Some units that are selected in the sample might not be available or refuse to provide measurements. If the uncontactable or refusal units come from some particular types on the variable of interest, then we have a biased sample. Giving out incentive might increase the response rate. Extra effort made by field work workers can help to convince selected units to take part in the study, and also to retain them for the whole duration of the study.

Example

Drug testing for a particular disease in Singapore

- Sampling frame: a list of physicians specialize in this disease.
- Patients of a selected physician might not be contactable because the contacts are not updated.
- Patients of a selected physician might not be willing to take part in the study.

In medical study, incentive like payment or free medical check-up is offered to offset the inconvenience caused to the selected units. But it is still not easy to convince selected patients to consent to taking part in the study. Also if the duration of the study is long, participated patients might move home without giving notice to the researcher.

Disasters in Sampling

- Getting a Volunteer or Self-selected Sample

  Media like to conduct opinion polls on their websites. It asks readers or viewers who are interested to respond. It mostly would attract those who have strong view on the issues. The results would reflect only the opinion of those who volunteer to respond. It is biased.

The disasters that happen in sampling come from the use of non-probability sampling. It is quite common in practice because the plan can be easily executed and it is also cheaper.But we can expect to obtain biased sample that makes the extension of the results to the population impossible.
News websites like to set up simple survey of opinions to ask viewers and readers to provide their views. It is open to anyone who has internet access and is willing to give their opinion. Those who have responded self select themselves to form the sample. So this is a volunteer or self-selected sample. It is a biased sample because normally only people with strong view on the issue would bother to give their answers.

- Using a Convenience or Haphazard Sample

  These are samples made up of individuals casually met or conveniently available, such as students enrolled in a class or people passing by on a street corner

Sometimes it could be due to the difficulty of constructing or finding a proper sampling frame, a study might use the most convenient group available. A medical researcher might engage a few hospitals to identify suitable patients for a study. Some researchers in tertiary institutes might collect opinions on some issues from students at the beginning of a lecture. A convenient sample is often biased. The information that we can gather from respondents who are easily available is normally different from the hard ones to get.

- Taking a Judgment Sample

  Sample units are chosen from the population by interviewers using their own discretion about which informants are “typical” or “representative”

It is quite obvious that judgment sample is biased because the selection is based on the opinion of experts. The experience and knowledge of these experts would have a lot of influences on how sample is to be constructed. For example, a news reporter might use his judgment to select so called typical people on the street for opinion polls. Of course no one can verify whether the opinions given are of typical as well.

- Selecting a Quota Sample

  This is a process of selection in which the elements are chosen in the field by interviewers using prearranged categories of sample elements to obtain a predetermined number of cases in each category.

In this non-probability sampling plan, each interviewer is assigned a fixed quota of units to interview; and numbers falling into certain categories of variables like sex, age, race, housing type, etc are also fixed. However, interviewers are free to interview anyone they like. We can see that in this plan bias will be introduced by the way interviewers select their respondents. Also the variable of interest might not be strongly associated with these variables. Having the proportions of these categories in the sample similar to those in the population does not make the extension of the results derived from the sample to the population better.

This sampling plan is very popular in market research. To rebuke what marketing consultants claim that quota sampling produces representative sample, we can ask them to produce the proportions of categories of combinations of any two variables. For example, we can choose the two variables Race and Sex and calculate the proportion of Chinese male or Indian female in the sample and compare that figure to the census information. We might get a big surprise.

The Debacle of Non-Probability Sampling

- Literary Digest Poll of U.S. Presidential Election of 1936
  - The candidates were Democratic incumbent Franklin D. Roosevelt and Republican Alf Landon.
  - Wrong sampling frame: 10 million people from magazine subscriber lists, phone directories, car owners received the questionnaires.
  - Low response rate: Only 2.3 million responses for 23% response rate.
  - (Incorrectly) Predicted a 3-to-2 victory for Landon.

We have seen the proper way of generating a good sample and also discussed a few difficulties that we might face. The worst disasters that can happen to us would be the use of non-probability sampling plan. We will be looking at four examples how difficulties and use of non-probability sampling plan could give us terrible outcomes.
Literary Digest Poll of the U.S. Presidential Election of 1936 is a very famous example. In that year, the two candidates were Democratic incumbent Franklin D. Roosevelt and Republican Alf Landon. Literary Digest sent 10 million questionnaires to people taken from magazine subscribers, car owners, telephone directories, club membership directories. So the sampling frame is wrong in the first place. Even though 2.3 million respondents sent back their responses, the response rate is only 23%, which is very low. The magazine predicted that Landon was the winner, but the actually Roosevelt won the election.

The Debacle of Non-Probability Sampling

- Annual Fresh Graduate Employment Survey
  - All fresh graduates receive questionnaires each year.
  - For the study to be recognized, at least 70% overall response rate has to be achieved.
  - Usually the starting salary of current year is compared to that of last year.

In the Annual Fresh Graduate Employment Survey, the sampling frame is perfect. Questionnaires are sent to every graduate in the population. It is stipulated that the survey results is valid only when we reach a response rate of at least 70%.

- Annual Fresh Graduate Employment Survey

Average Starting Salary			
	This year	Last year	% Change
Overall	$3,100	$2,760	**12%**
Professional degree	$3,300	$3,000	**10%**
Non-professional degree	$2,800	$2,600	**8%**

  - The percentage change of the overall average salary should fall between the percentage changes of the two degrees.

Because we allow fresh graduates to decide whether they want to take part in the study, the response rates of each degree programme will vary from year to year. That in turn, affects the overall results. In this table, the average salary of graduates who responded to the survey has increased by 12% when it is compared with that of last year. But when we group the graduates into two types of degree programmes, we see that the average salaries of this year for professional degree and non-professional degree have increased by 10% and 8%, respectively, from the year before. This result is not consistent because the percentage increase of the average salary of all graduates should fall between 8% to 10%.

- Annual Fresh Graduate Employment Survey
  - Different proportions of types of degree responded to survey in different years.

The abnormity could come from the following situation. A year before, a larger proportion of non-professional degree holders responded to the survey and that lead to a lower overall average salary because non-professional degree normally fetched lower salary. Then this year, a larger proportion of fresh graduates responded to the survey were professional degree holders. The overall average salary became higher with the help of professional degree holders who fetched higher salary. The phenomenon is acceptable only if the proportion of non-professional degree graduates of last year was large and the proportion of non-professional degree graduates of this year was small. The two samples in this case truly reflected the composition of professional degree and non professional degree graduates of respective years.

The Debacle of Non-Probability Sampling

- Battle of Claiming No 1 High Society Magazine
  - 'Tatler Drops Claim to be No. 1 High Society Magazine' Headline in The Straits Times, April 12, 2006.
  - Tatler commissioned a marketing firm to conduct a survey in an event organized by Tatler.
  - This is a convenient sample.
  - The survey finding showed Tatler to be "the best magazine to advertisers" compared to Prestige and The Peak.

The high society magazine Tatler was sued by another high society magazine Prestige because Tatler claimed that it is the top high society magazine in Singapore. Earlier Tatler had commissioned a marketing research company to conduct a study. The marketing firm conducted a survey at an event organized by Tatler. This is a convenient sample. After the representative of the marketing firm testified in court that the survey might not have met the international market research standard, Tatler agreed to withdraw the claim and pay a lump sum of $40,000 plus legal costs to Prestige.

- Battle of Claiming No 1 High Society Magazine
  - The representative of the marketing firm testified in court that 'the survey had "likely failed to meet the standard promised…. of being in accordance with acceptable international market research standard".'
  - Eventually, Tatler agreed to settle and paid Prestige a lump sum of $40,000 plus legal costs which estimated at $250,000.

The Debacle of Non-Probability Sampling

- The U.S. Presidential Election of 1948
  - The two main candidates: Harry Truman and Thomas Dewey.
  - Three major polls, Grossley, Gallup, and Roper, covered the election campaign declared Dewey the winner.
  - On election day, Truman scored an upset victory.
  - All three polls used quota samples.

In another Presidential Election of the United States in 1948, three polling companies used quota sampling plan to predict the winner of the election between Truman and Dewey. All three companies predicted Dewey as the winner. But Truman scored an upset victory.

- The U.S. Presidential Election of 1948
  - Each interviewer was assigned a fixed quota of units that fall into categories of variables like residence, sex, age, race, and economic status, to interview.
  - The mistakes are: 1. these variables might not link closely to the way they cast their vote; 2. interviewers are free to select anybody they like and thus it leads to selection bias.

As pointed out earlier in the discussion, in quota sampling, the chosen variables might not link closely to how the votes are cast. Also selection bias occurs when interviewers are free to choose the respondents.

Conclusion

"I would trade all your 18,000 case histories for 400 in a probability sample."

Great statistician John Tukey to famous sexologist Alfred Kinsey, who collected data from prisons and bars famous for being gay meeting places to study unusual sexual behaviours.

The above quote is taken from the article 'The problem with sexed-up statistics' by Tim Harford. The article appeared in The Straits Times, June 13, 2015.

From what the great statistician John Tukey has said, a proper sampling frame and the use of probability sampling are ways to extend the results from sample to population.

Estimating a Parameter

Hi! I am Samuel Yeun and I will be bringing you through the second half of this chapter.

Sampling Chapter

1st Half:
How to take a good sample

2nd Half:
What information does our sample give us?

In the earlier few units, we have given you an overview about the sampling process, and how to take a good sample. But, let's stop and think about our experiences in real life. Say, I am interested in studying the average score of all students in university ABC. I managed to get a good sample of 200 students from the university and the average is 3.6/5. Now, if you took a separate census of the entire university, should I guess that your average is EXACTLY the same as my 3.6/5? Most likely not. Why? I'll leave this question as a teaser for now because that's exactly what this second half of the chapter is for. (See the conclusion of this unit for the answer)

If our sample does not tell us the exact value of the census, then what is the use of the sample? What information does our sample give us? Are there things we can do to make our sample better? We will discuss these in more detail now.

What is a parameter?

- Numerical fact about the population
  - Usually unknown to us
- Estimated from a sample

Firstly, let us begin this unit by exploring an important key term: 'Parameter'.
A parameter is a numerical fact about a population, which is usually unknown to us but is of interest to investigators. An example can be "Singapore's quarterly unemployment rate", which was discussed in unit 1. Without a census, it is difficult for us to know the exact value of a parameter. However, we can try to obtain a close estimate of the parameter using a good sample.

What is a parameter?

Purpose of Sampling:

- The results can be extended to the population from which the sample was drawn.

Recall from Unit 2, the purpose of sampling is that the collected results can be extended to the population from which the sample was drawn.

What is a parameter?

Purpose of Sampling:

In other words, our parameter of interest is from a defined population. We can use a sample to give us an estimate about that parameter.

What is a parameter?

E.G.: Unemployment rate

parameter is the main one

Sample

A portion of adults would be selected to provide their employment status.

Population

The collection of every adult (aged 15 or over) who resides in Singapore

Estimate:

Percentage of unemployed adults in sample

Parameter:

Singapore's unemployment rate

Using the example from Unit 1, if we are interested in Singapore's unemployment rate, the sample results are meant to give us an idea about the population's unemployment rate. How do the sample results give us an idea? By providing an estimate of the parameter.

From the population, the parameter of interest is "Singapore's unemployment rate ". We can take a sample of say, 1000 Singapore adults, and find the percentage of unemployed adults in our sample. This percentage will be our estimate of Singapore's unemployment rate.

The Estimation Equation

Estimate = Parameter + Random error

Assumptions:
- *Simple random sample*
- *100% response rate*

However, there is a problem! When we take a sample, it is unlikely that the sample's estimate equals the parameter. The estimate may be higher or lower. What is happening? The answer is: "random error is present".

In the "Unemployment rate" example, we discussed how we can take a sample to estimate the population's parameter of interest – Singapore's unemployment rate. For simplicity, let us assume a simple random sample was conducted. Also for simplicity, assume a 100% response rate. Then, the sample percentage of unemployed adults is a good estimate of the parameter, meaning if we repeat the whole process many times, the sample percentages will fluctuate around the parameter. We say the estimate is unbiased.

Even in such a simplified scenario, it is still unlikely that our sample's estimate equals the parameter. Why? Because of random error. The sample is only part of the population. Since the sample was taken randomly, there is the element of random error present.

We can say that in such an example, our Sample's Estimate = Parameter + Random error. Actually, in any probability sample, one can also construct an unbiased estimate for the parameter, so that the above equation holds.

A More Realistic Estimation Equation

Estimate = Parameter + Bias + Random error

Harder to quantify *Easier to Quantify*

Assumptions:

- ~~*Simple random sample*~~
- ~~*100% response rate*~~

In the previous slide, we simplified the situation by assuming a simple random sample was conducted, and there was a 100% response rate. However, when conducting sampling in reality, it is rarely so simple. Due to the way we conduct our sample, problems can arise from a poor sampling frame, non-probability sampling, low response rate, or from other sources. And it is from these problems, that 'Bias' can occur.

Simply put, we can say that in reality, the Sample's estimate = Parameter + Bias + Random error.

Random error can be quantified and detected easily if we know the sample size and standard deviation of the variable that we are measuring. Bias, on the other hand, is the harder issue to quantify and detect. Thus, in the next unit, we will first focus on the issue of bias.

Conclusion

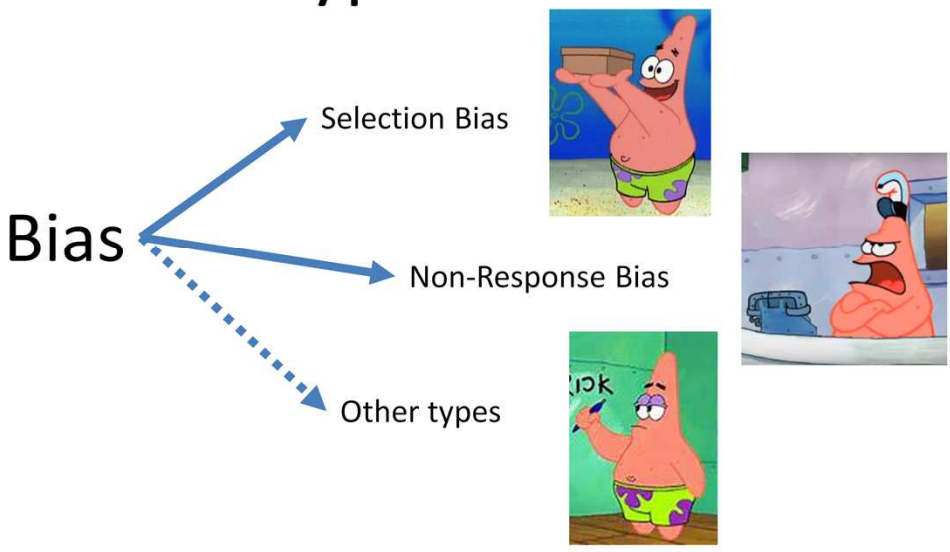
- The estimation equation is a model to help us understand why it is common to see that our sample's estimate differs from the population's parameter.
- Two key portions of the equation are 'Bias' and 'Random error'.

To summarize, the estimation equation is a model to help us understand why it is common to see that our sample's estimate differs from the population's parameter. Two key portions of the equation are 'Bias' and 'Random error'.

Types of Bias

In the previous few units, the word ‘Bias’ has been brought up multiple times. We know that it is bad for the study, but what actually is ‘Bias’? Now that we already have a better idea of the estimation equation, this unit aims to discuss the topic of ‘Bias’ a little bit further.

Types of Bias



Retrieved from: https://www.youtube.com/watch?v=rMog3TXQRds
Retrieved from: https://vignette.wikia.nocookie.net/fantheories/images/d/dd/Stupider.jpg/revision/latest?cb=20180207021402
Retrieved from: https://vignette.wikia.nocookie.net/spongebob/images/8/8b/Its_the_deecewe_box_spongebb.jpg/revision/latest?cb=20120923171331

There are different ways that bias can be introduced into a sample. This Unit will discuss 2 types – selection bias and non-response bias.

Selection Bias

- Systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample
- Caused by:
  - Imperfect Sampling Frame
  - Non-probability Sampling Methods

Selection bias, is the systematic tendency on the part of the sampling procedure to exclude one kind of person or another from the sample. This can happen due to an imperfect sampling frame, or non-probability sampling methods (as seen in Unit 4). For example, an interviewer conducting quota sampling would have unintentional selection bias introduced into the sample when he/she hand-picks the people to be interviewed.

Non-Response Bias

- Systematic tendency from subjects who do not respond to the survey/ questionnaire.
- Caused by:
  - Differences between non-respondents and respondents

  (If non-response rate is high, Non-Response Bias is likely to be significant)

Non-response bias, is the systematic tendency from subjects who do not respond to the survey/questionnaire. Experience shows that non-respondents tend to differ from respondents in important ways (see Chapter 1 for more details). If this is true for the sample and a large number of those selected for the sample do not respond, non-response bias is likely.

Other types of bias

- Phrasing of the questions, tone, or attitude of the interviewers
- When the subjects have a tendency to understate responses about undesirable social habits

There are other ways that bias could enter the study. For example, the answers given by respondents are influenced to some extent by the specific phrasing of the questions, tone, or attitude of the interviewers. In the 1948 United States presidential election survey, changing the order of the candidates' names was found to alter the responses by 5%, giving the higher survey results to the candidate who was named first. Another example could be when the subjects have a tendency to understate responses about undesirable social habits, like smoking.

Conclusion

To minimize bias in a survey, we seek to:

1. Include every population unit in the frame
2. Use a probability sampling method
3. Get 100% response rate

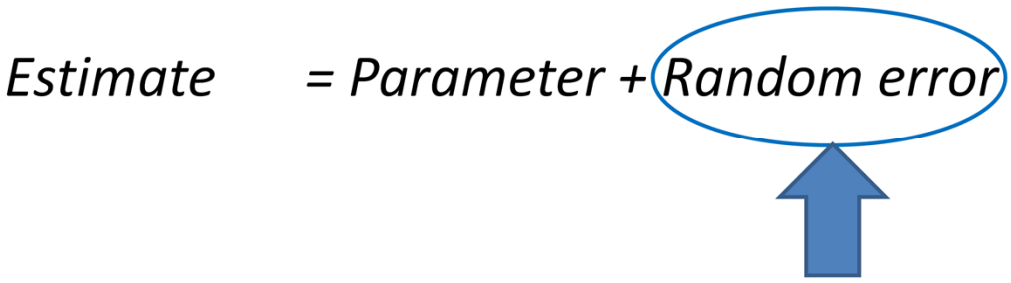
In conclusion, to minimize bias in a survey, we seek to:

Include every population unit in the frame,
use a probability sampling method,
and get a 100% response rate.

Random Error

In our previous unit, we discussed the issue of bias. For this unit, we will discuss what random error means in more detail.

Recall: The Estimation Equation



Let us reuse the Unemployment example again (from 'The Estimation Equation' Unit), where we are interested to estimate the unemployment rate of Singaporeans. Specifically, refer to the slide titled: The Estimation Equation. Recall that for simplicity and without losing any generality, we assumed that we have taken a simple random sample from Singapore's population. Thus, it was also safe to assume there is no bias in the estimation. We discussed that even for such a simplified case, it was unlikely that the sample's estimate equals the parameter. This is due to random error.

What is Random Error

Estimate:
Percentage of unemployed adults in sample

Parameter:
Singapore's unemployment rate

Would we expect all their "sample percentage of unemployed adults" to be the same?

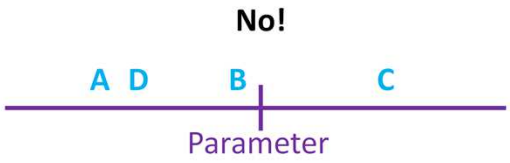
Retrieved from: https://cdn.vox-cdn.com/thumbor/HIHZdlKkUfA3Q7RdoNILafPYe2Y=/0x0:283x341/1200x0/filters:focal(0x0:283x341)/cdn.vox-cdn.com/uploads/chorus_asset/file/6759169/Screen_Shot_2016-07-07_at_3.30.18_PM.0.png

Let's take it a step further. Imagine that many other researchers are also interested in the unemployment rate of Singapore. These researchers all get a different simple random sample with the size of 1000 Singapore adults, and calculate their own "sample percentage of unemployed adults". Again for simplicity, we will assume their estimates are unbiased. Still, would we expect all their "sample percentage of unemployed adults" to be the same? I'll give you a few seconds to think about your answer…

What is Random Error

Estimate:
Percentage of unemployed adults in sample

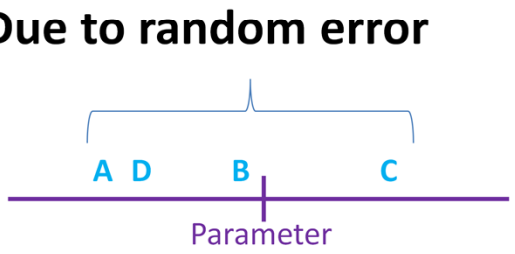
Parameter:
Singapore's unemployment rate



Retrieved from: https://cdn.vox-cdn.com/thumbor/HIHZdlKkUfA3Q7RdoNILafPYe2Y=/0x0:283x341/1200x0/filters:focal(0x0:283x341)/cdn.vox-cdn.com/uploads/chorus_asset/file/6759169/Screen_Shot_2016-07-07_at_3.30.18_PM.0.png

The answer is: No! It is expected that their sample percentages will fluctuate around the parameter.

What is Random Error



These researchers have different sample percentages, all clustered around the parameter. It is not that some researchers have the wrong answer. It is just due to random error. To recap, the sample is only part of the population. Since the sample was taken randomly, there is the element of random error present.

Sample Size and Random Error

- Suppose that Ben and Jerry are interested in finding the unemployment rate of Singapore.
- Ben gets a Simple Random Sample of 600 Singapore adults
- Jerry gets a Simple Random Sample of 5000 Singapore adults

Now that we have discussed what random error is, let's look into its relation to the sample size.
Here's a scenario:
Suppose that Ben and Jerry are interested in finding the unemployment rate of a very large population, for example: Singapore.
Ben gets a simple random sample of 600 Singapore adults
Jerry gets a simple random sample of 5000 Singapore adults

Sample Size and Random Error

Ben's Estimate = Parameter + Ben's Random error

Jerry's Estimate = Parameter + Jerry's Random error

Whose estimate is likely to have a smaller random error?

Apart from sample size, everything else that Ben and Jerry do are exactly the same. We will assume there is no bias in their estimates. Yet again, it is not likely that their sample percentage will be exactly the population's parameter. Both of their estimates are likely to be a little bit off, due to random error. But hang on, make a guess! Whose estimate is likely to have a smaller random error?

Sample Size and Random Error

Ben's Estimate = Parameter + Ben's Random error

Jerry's Estimate = Parameter + Jerry's Random error

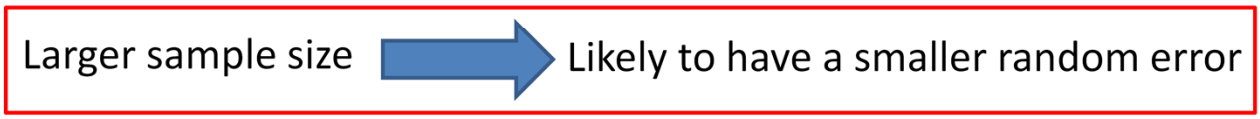
Whose estimate is likely to have a smaller random error?

Jerry

Jerry's estimate is more likely to have a smaller random error!

Sample Size and Random Error

When estimating the population's unemployment **rate**



When estimating a population's rate (or average), a sample with larger size is likely to have a smaller random error. Jerry has a larger sample size of 5000, compared to Ben's 600 sample size. Thus Jerry is more likely to have a smaller random error.

Conclusion

- When finding an estimate for the Parameter, it is difficult to be 100% sure.
- There will always be uncertainty.
- When estimating the rate from a population, a sample with larger size is likely to have a smaller random error.

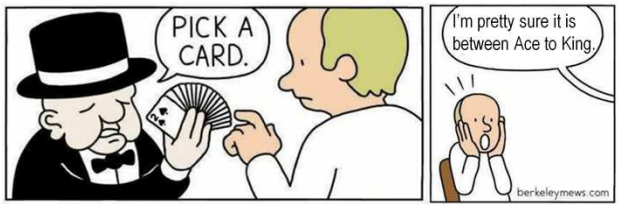
When finding an estimate for the Parameter, it is difficult to be 100% sure. As seen in the example with many researchers, there will always be fluctuations even when the same sampling techniques are employed among the researchers. Likewise, even if you were to repeat the study many times by yourself, your sample percentages are likely to fluctuate. There will always be uncertainty. However, we also concluded that when estimating the rate from a population, a sample with larger size is likely to have a smaller random error.

Confidence Intervals

When we look at results from a sample, a common terminology that pops up is "Confidence Interval". What is this all about?

Confidence Intervals

- Range of values that we are reasonably certain our unknown parameter lies in.

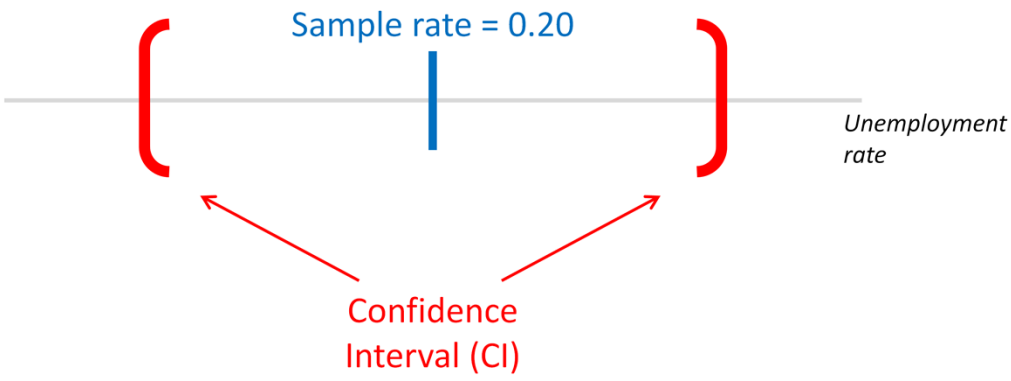


Retrieved from: http://www.berkeleymews.com/?p=888

Simply put, the confidence interval is the range of values that we are reasonably certain our parameter lies in. The purpose of sampling is to estimate the unknown value of the parameter – see 'Estimating a Parameter' unit.

Confidence Intervals

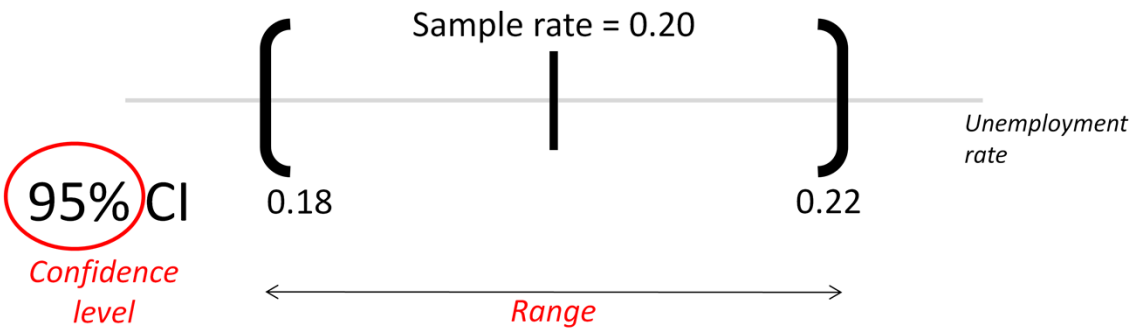
From a sample of 1000 Singapore adults:



To make reading easier, let's use the unemployment example again. I obtained a simple random sample of 1000 Singapore adults and my sample's unemployment rate is 0.20. Should I say that the population's unemployment rate is 0.20? It is a good guess, but it is not likely to be 0.20 due to random error. So along with the rate, I wish to report more about the random error in the study. How can I do that? By reporting the confidence interval (Also abbreviated as: CI).

Reporting Confidence Intervals

From a sample of 1000 Singapore adults:

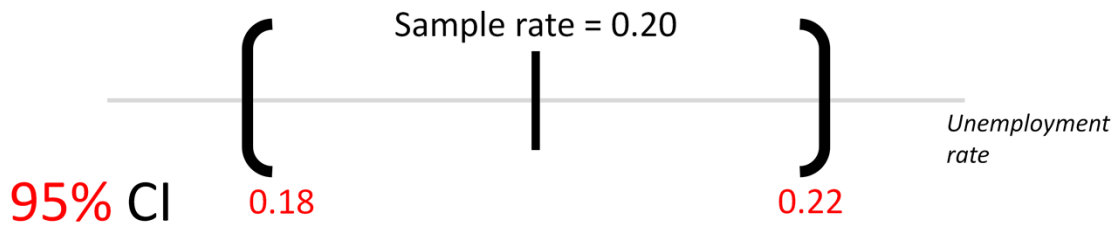


When we read sample data from computers or journals, it is common to see the confidence interval already calculated and shown together with the estimate. For this unit, we will not cover how to calculate the confidence interval, but instead focus on how to interpret it.

A confidence interval is reported with two parts – A 'confidence level', and a 'range'.

Confidence Intervals Interpretation (1)

From a sample of 1000 Singapore adults:

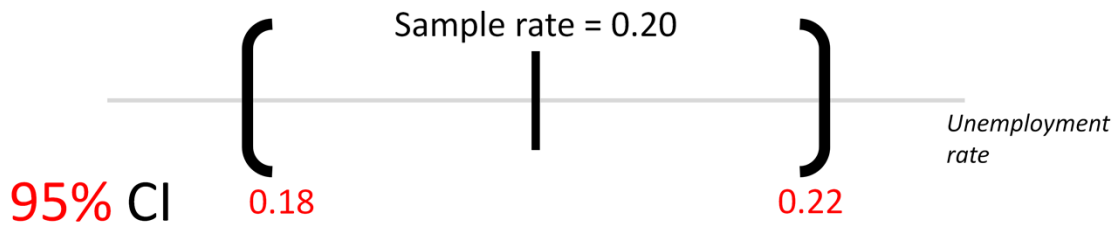


"We are 95% confident that the range from 0.18 and 0.22 contains the population parameter."

In this example, we can report the confidence interval by saying: "We are 95% confident that the range from 0.18 and 0.22 contains the population parameter.".

Reporting Confidence Intervals (2)

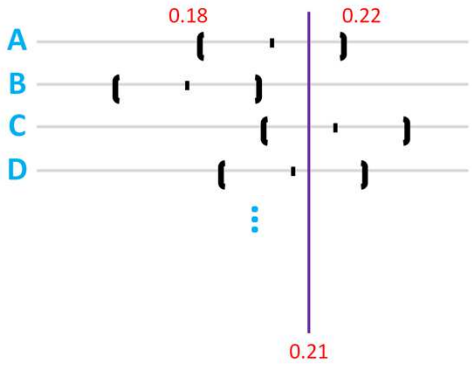
From a sample of 1000 Singapore adults:



95% CI: 0.20 ± 0.02

Another way to report the confidence interval is to say “95% CI: 0.20 ± 0.02”.

Confidence Intervals Interpretation



About 95% of the researchers will have intervals that contain the population parameter.

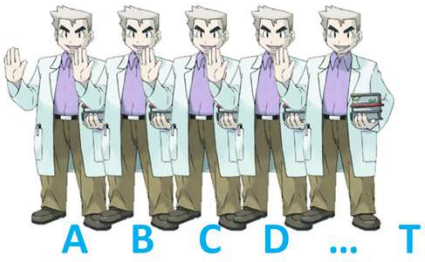
Retrieved from: https://cdn.vox-cdn.com/thumbor/HIHZdlKkUfA3Q7RdoNILafPYe2Y=/0x0:283x341/1200x0/filters:focal(0x0:283x341)/cdn.vox-cdn.com/uploads/chorus_asset/file/6759169/Screen_Shot_2016-07-07_at_3.30.18_PM.0.png

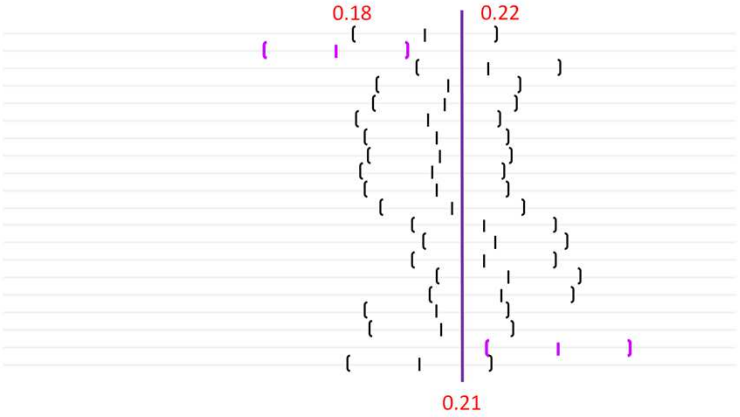
Perhaps the term "confident" may seem new to some of us. How do we interpret this term? Recall from the previous unit the hypothetical scenario where other researchers also got a different simple random sample with the size of 1000 Singapore adults. The confidence interval depends on the sample, so each of the researchers report their own different estimate and different confidence intervals (The different intervals do have different range lengths but the differences are too small to be seen in this diagram). From our previous slide, the confidence interval of 0.18 to 0.22 is just one among the many other intervals reported.

Now, let us assume we managed to get our hands on Singapore's census information and discovered that the population parameter = 0.21. Most of the researchers have intervals that contain 0.21. That's good!

The 95% in the phrase '95% confidence interval', means that about 95% of the researchers will have intervals that contain the population parameter!

Confidence Intervals Interpretation





About 95% of the researchers will have intervals that contain the population parameter.

Retrieved from: https://cdn.vox-cdn.com/thumbor/HIHZdlKkUfA3Q7RdoNILafPYe2Y=/0x0:283x341/1200x0/filters:focal(0x0:283x341)/cdn.vox-cdn.com/uploads/chorus_asset/file/6759169/Screen_Shot_2016-07-07_at_3.30.18_PM.0.png

Extending the interpretation further into a more concrete example, suppose there are 20 researchers, each with their own estimate and 95% confidence interval. The 95% confidence level means we are expecting about 19 out of 20 of them (95%) to have the population parameter in their intervals. Note that in this example, only 18 of the researchers have the parameter in the interval. This is not an error. It does not have to be exactly 19 researchers (we are expecting ABOUT 19 researchers).

Warning

Is it possible to report a confidence interval with confidence level of 100%?

No!

Assuming we have conducted simple random sampling (or other probability sampling methods), and no other problems when collecting the sample, it is still sensible to use our sample to estimate the parameter.

To wrap up this part on interpretation of intervals, here is a question for you. “When a researcher takes a sample, is it possible to report a confidence interval with confidence level of 100%?”. In general, no!

In real life, when doing estimation, we do not know the population parameter. Any researcher will not know if his/ her interval contains the parameter. Of course, to get a 100% confidence level, we have to include every possible value, in this case 0% to 100%. But that would not be useful at all.

Commonly, researchers choose to report a confidence level of 95%. Some choose other values, like 90% or 99%. But to report with 100% confidence is too big a claim to make!

This may sound worrying. Some of us may wonder: If there are no reports with 100% confidence level, then does this mean all samples are useless? No! Building on the previous unit, our sample is meant to give us information so that we can make informed decisions. Assuming we have conducted simple random sampling (or other probability sampling methods), and no other problems when collecting the sample, it is still sensible to use our sample to estimate the parameter.

Confidence Intervals ↔ Random Error

Now that we have understood how to interpret the confidence interval, we can move on to the next portion of the unit. Earlier on, we briefly mentioned that the 'confidence interval' is a method to report more about the 'random error' in a study, but how are the two concepts actually related?

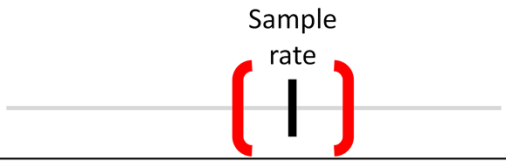
Random Error and Range

Researcher's name	Sample size	Confidence Interval
Ben	600	**95** % CI
Jerry	5000	**95** % CI

At 95% confidence level, who is more likely to have a smaller confidence interval range?

Let us re-use the scenario of Ben and Jerry attempting to find Singapore's unemployment rate (from the 'Random Error' unit).
Ben uses a simple random sample of 600 Singapore adults.
Jerry uses a simple random sample of 5000 Singapore adults.
When reporting 95% confidence level, do you think one of them is more likely to have a smaller confidence interval range?

Random Error and Range

Researcher's name	Sample size	Confidence Interval
Ben	600	**95** % CI 
Jerry	5000	**95** % CI 

Jerry's estimate is more likely to have a smaller random error
➔ more likely to have a smaller range

Yes! Actually, from our earlier unit, we know that Jerry's estimate is more likely to have a smaller random error, and hence is more likely to have a smaller range. In other words, he can be more certain of where the parameter is.

Conclusion

- If we use probability sampling, the confidence interval is helpful in providing information about the error in the estimate.
- At 95% confidence level, a smaller random error in the estimate implies a smaller confidence interval range.

In summary, if we conduct a probability sampling method, the confidence interval is helpful in giving readers more information, to better guess what the parameter may be. Conversely, if the sample is not obtained by a probability method, for example a convenience sample, then a derived confidence interval may not be reliable. Also, we can always work towards having more certainty of where our parameter lies. At 95% confidence level, a smaller random error in the estimate implies a smaller confidence interval range.