7.1 Sampling + ~~7.2 + 7.7 + 7.3~~

Recall that:

- An element is the entity on which data are collected.
- A population is the collection of all the elements of interest.
- A sample is a subset of the population.

---

* The reason we select a sample is to answer a research questions about population.
* The sampled population is the population from which the sample is drawn.
* A frame is a list of the elements that the sample will be selected from.

Examples: 1 "Finite Population"

A political party in Texas wanted to estimate the proportion of registered voters in the state favoring the candidate.

A sample of 400 registered voters were selected, and 160 of the 400 voters indicated a preference for the candidate.

Hence, the estimated proportion of the population of registered voters favoring the candidate is $\frac{160}{400} = 0.40$

- Sampled Population is all registered voters in Texas.
- Frame is a list of all the registered voters.
  - A sample proportion provides an estimate of population proportion, a sampling error is expected.

2 "Infinite Population" or "Process"

A tire manufacture wanted to estimate the mean useful life of the new tires. The manufacture produced a sample of 120 tires for testing. The test results provided a sample mean of 36,500 miles. Hence, the estimated mean useful life for the population of new tires is 36,500 miles.

- Sampled Population is infinite: since the sample of 120 tires was obtained from a production process at a particular point in time.
- Frame: impossible to construct.
  - A sample mean provides an estimate of population mean, a sampling error is expected.

Example*: suppose that the state wants to develop the portfolio of the company's 2500 managers in order to characterize
* the mean annual salary and
* the proportion of managers who completed the company's training program.

The data that contain this information for all 2500 managers is on the CD. Thus, we can compute the:

- Population mean: $\mu = \$51,800$
- Population stan. deviation: $\sigma = \$4000$
- The data shows that 1500 managers completed the training program
  $\Rightarrow$ Thus, the proportion of the population who completed the training program is $p = \frac{1500}{2500} = 0.60$
- $\mu = \$51,800$, $\sigma = \$4000$, $p = 0.60$, ... are called parameters of the population.
- Parameters: are numerical characteristics of a population.

Now suppose that the data were not saved on CD.
How then the state can estimate the population parameters by using a sample of managers rather than all 2500 managers.

- If a sample of 30 managers (for example) will be selected, then the time and the cost of developing a profile will be less than if we consider the entire population.
- How to select the sample ?!

7.2 Selecting a Sample

1 Selecting a Sample from a finite Population:

- A **simple random sample** of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

- The number of different **simple random samples** of size n that can be selected from a finite population of size N is

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

- For example: $N = 2500$ and $n = 30$. The number of different simple random samples of size 30 is $\binom{2500}{30} = \frac{2500!}{30!\,(2500-30)!} \approx 2.75 \times 10^{69}$

- To select a **simple random** from the finite population of managers:
  1. we construct a **frame** by assigning each manager a number:
     1, 2, 3, ..., 2500
  2. We refere to the table of random numbers: Table 7.1 page 260.
  3. We start the selection of random numbers anywhere in the table and move systematically. We will use the first row and move from left to right.
  4. Since the last number is 2500, has four digits, we select four digits numbers from the table.
  5. If the selection is repeated, we ignor it a second time. : Once an element has been included in the sample, it is removed from the population and cannot be selected again. "sampling without replacement"
     If the selection is greater than 2500, we ignor it. otherwise, : once an element has been included in the sample, it is returned to the population. A previously selected element can be selected again and therefore may appear in the sample more than once. "sampling with replacement"

The four digit random numbers are:

6327 5998 7174 5110 1514 ...

x x x ✓

Until the simple random sample of 30 manager has been obtained.

TABLE 7.1 RANDOM NUMBERS

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

2 Selecting a Sample from a Process "infinite Population"

- If the frame of a population can not be constructed, then we can not use simple random sample. For example:
  - → Sampling from a very large population in which it is impossible to identify all the elements of the population.
  - → Sampling from process in which the sampled population is conceptually infinite. For example:
    - → To select a sample of the elements generated by a production process: The manufacturer produced a sample of 120 new tires to estimate the mean useful life for the population of new tires.
      - → The production of each unit "tire" is independent of the production of the others.
    - → Hence, we select a random sample by selecting any n units produced while the process is operating properly.

Example (Q2 page 262) Assume a finite population has 350 elements. Using the last three digits of the random numbers:

98601 73022 83448 02147 34229 27553 84147 93289 14209

determine the first four elements that will be selected for the simple random sample.

× ✓ × ✓ ✓ × repeated × ✓ it's over ×

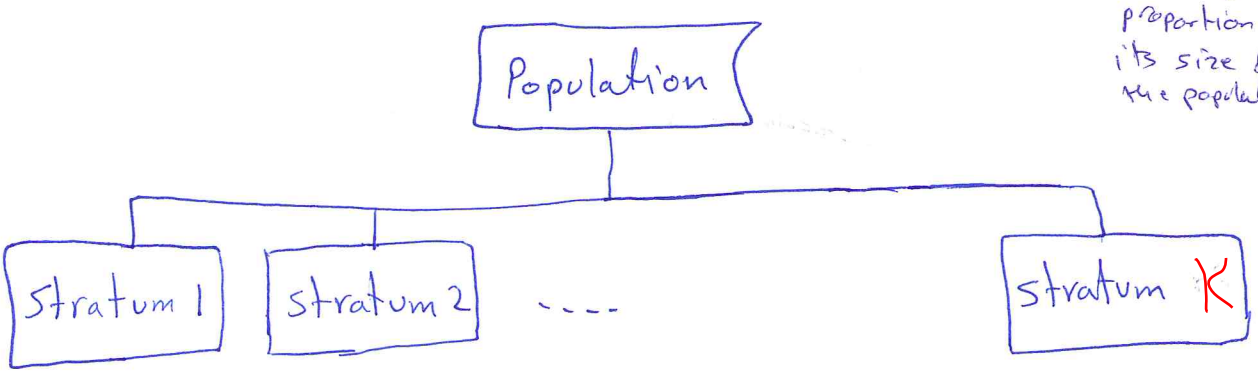
022, 147, 229, 289

7.7 Other Sampling Methods:

In this section we provide an alternative sampling methods other than simple random sampling.

Sampling Methods:

1. Simple random sampling: (Finite Population) A sample selected s.t each possible sample of size n has the same prob. of being selected.
2. Stratified Random Sampling: A probability sampling elements selected from population with know prob. of being included in the sample. method in which the population is first divided into strata (small groups) and a simple random sample is then taken from each stratum. ↓ proportion to its size from the population

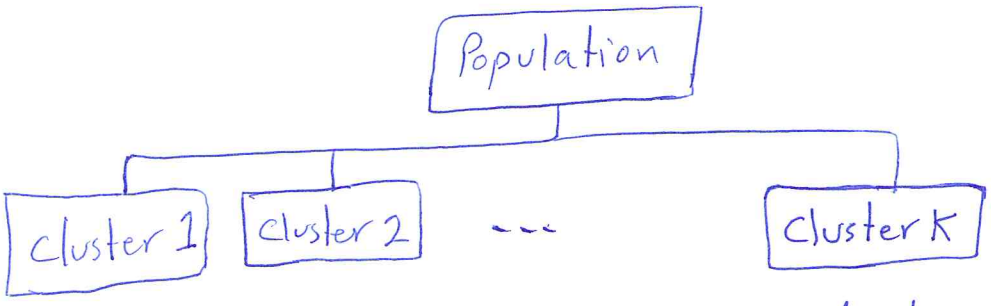


- Each element in the population belongs to one and only one stratum
- Strat can be department, location, age, industry type ...
- Stratified random sampling workes best (gives a good estimat) when the variance among elements in each stratum is small. (the elements of each stratum are as much alike as possible).
- Thus, the stratified random sampling depends on how homogenous the elements are within the strata.

If elements within strata are alike, then strata will have low variance
⇒ Thus, relatively small sample sizes can be used to obtain good estimates

If strata are homogenous, then the stratified random sampling procedure provides results as precise as those of simple random sampling by using a smaller total sample size.

[3] Cluster Sampling: A probability sampling method in which the population is first divided into clusters and then a simple random sample of the clusters is taken.



- The elements in the population are first divided into separate groups called clusters.
- Each element in the population belongs to one and only one cluster.
- Cluster sampling provides the best results when the elements within the clusters are not alike (each cluster is a representative small-scale version of the entire population.
- Thus, cluster sampling depends on how representative each cluster is of the entire population.
  - If all clusters are alike (= they have the same representative), then sampling a small number of clusters will provide good estimate of the population parameters.
- Example of cluster sampling is area sampling where cluster are city blocks.
- Cluster sampling requires a larger total sample size than
  - simple random sampling and
  - stratified random sampling.

  but costs less because the data can be collected in each cluster in a shorter time, and using only one interviewer.

Example: Suppose the Dep. of agriculture wishes to investigate the use of pesticides by farmers in Palestine

مبيدات

A cluster sample could be taken by identifying different ~~countri~~ cities as clusters.

A sample of these cities would then be chosen at random all farmers in these selected cities would be included in the sample.

[4] Systematic Sampling: A probability sampling in which we randomly select one of the first k elements and then select the k^{th} element thereafter.

- When the population is Large, it's time consuming to select a simple random sample by finding random numbers, counting, until the corresponding element. So we use Systematic Sampling:
- For example: A population of 5000 elements.
  Suppose we need a sample of size 50 elements

  $\frac{5000}{50} = 100$

  - A systematic sample selects randomly one elemen in the first 100 elements (65^{th} (suppose it was 65))
  - The systematic sample will be the elements whose orders are:

    65^{th}, $65^{th}+100$, $65^{th}+200$, $65^{th}+300$, ..., $65^{th}+4900$

- That is, we move systematically through the population List, and identifying every 100^{th} elemen after the first element is randomly selected
- Systematic sampling in this way is easier than simple random sample and it has the properties of a simple random sample since the first element is randomly selected.

5 Convenience Sampling: A nonprobability method of sampling whereby elements are selected for the sample on the basis of convenience.

- Elements are included in the sample without know probabilities.
- For example: The university send a sample of selected professors to participate in an international conference, in specific field.
  "Volunteer panels" متطوعين

- Convenience samples are easy to select, but it is impossible to evaluate the "goodness" of the sample in terms of its representative of the population.
- Convenience samples may provide a good results or may not.
- We should be cautious to use the results of convenicence sampling to make inferences about the population, since no statistical methods applied here.

6 Judgment Sampling: A nonprobability method of sampling where by elements are selected for the sample based on the judgment of the person doing the study.

- The person selects elements of the population that he/she feels they are most representative of the population.
- This method is easy way to select a sample.
- Aa example is that a political party select a two or three senators to make an TV interview since they reflect the general opinion of all senators.
- We should be carful to use the results of the judgment sampling to make inference about the population.

Summary:

- We recommend to use prob. sampling methods:
  - simple random sampling
  - stratified random sampling
  - cluster sampling
  - systematic sampling

  → propability sampling

because for these methods, formulas are available for evaluating the "goodness" of the sample results in terms of the closeness of the results to the population parameters, being estimated.

- An evaluation of the goodness cannot be made with
  - convenience sampling and
  - judgment sampling.

  → non propability sampling

Thus, a great care should be used in interpreting the results based on nonprobability sampling methods.

Methods of sampling from a population

It would normally be impractical to study a whole population, for example when doing a questionnaire survey. Sampling is a method that allows researchers to infer information about a population based on results from a subset of the population, without having to investigate every individual. Reducing the number of individuals in a study reduces the cost and workload, and may make it easier to obtain high quality information.

If a sample is to be used, by whatever method it is chosen, it is important that the individuals selected are representative of the whole population. This may involve specifically targeting hard to reach groups.

There are several different sampling techniques available, and they can be subdivided into two groups: probability sampling and non-probability sampling.

In probability (random) sampling, you start with a complete sampling frame of all eligible individuals from which you select your sample. In this way, all eligible individuals have a chance of being chosen for the sample, and you will be more able to generalise the results from your study. Probability sampling methods tend to be more time-consuming and expensive than non-probability sampling.

In non-probability (non-random) sampling, you do not start with a complete sampling frame, so some individuals have no chance of being selected. Consequently, you cannot estimate the effect of sampling error and there is a significant risk of ending up with a non-representative sample which produces non-generalisable results. However, non-probability sampling methods tend to be cheaper and more convenient, and they are useful for exploratory research and hypothesis generation.

Probability Sampling Methods

1. Simple random sampling

In this case each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected.

One way of obtaining a random sample is to give each individual in a population a number, and then use a table of random numbers to decide which individuals to include. For example, if you have a sampling frame of 1000 individuals, labelled 0 to 999, use groups of three digits from the random number table to pick your sample. So, if the first three numbers from the random number table were 094, select the individual labelled “94”, and so on.

As with all probability sampling methods, simple random sampling allows the sampling error to be calculated and reduces selection bias. A specific advantage is that it is the most straightforward method of probability sampling.

2. Systematic sampling

Individuals are selected at regular intervals from the sampling frame. The intervals are chosen to ensure an adequate sample size. If you need a sample size n from a population of size N, you should select every N/n^{th} individual for the sample.

For example, if you wanted a sample size of 100 from a population of 1000, select every 1000/100 = 10^{th} member of the sampling frame.

Systematic sampling is often more convenient than simple random sampling, and it is easy to administer. However, it may also lead to bias, as an example, if a group of students were being selected to get their opinions on college facilities, but the Student Record Department's central list of all students was arranged such that the sex of students alternated between male and female, choosing an even interval (e.g. every 20^{th} student) would result in a sample of all males or all females.

3. Stratified sampling

In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups.

For example, in a study of stroke outcomes, we may stratify the population by sex, to ensure equal representation of men and women. The study sample is then obtained by taking equal sample sizes from each stratum.

In stratified sampling, it may also be appropriate to choose non-equal sample sizes from each stratum. For example, in a study of the health outcomes of nursing staff in a county, if there are three hospitals each with different numbers of nursing staff (hospital A has 500 nurses, hospital B has 1000 and hospital C has 2000), then it would be appropriate to choose the sample numbers from each hospital *proportionally* (e.g. 10 from hospital A, 20 from hospital B and 40 from hospital C).

Stratified sampling improves the accuracy and representativeness of the results by reducing sampling bias. However, it requires knowledge of the appropriate characteristics of the sampling frame (the details of which are not always available), and it can be difficult to decide which characteristic(s) to stratify by.

4. Clustered sampling

In a clustered sample, subgroups of the population are used as the sampling unit, rather than individuals. The population is divided into subgroups, known as clusters, which are randomly selected to be included in the study.

For example individual GP practices or towns could be identified as clusters. In single-stage cluster sampling, all members of the chosen clusters are then included in the study. In two-stage cluster sampling, a selection of individuals from each cluster is then randomly selected for inclusion. Clustering should be taken into account in the analysis. The General Household survey, which is undertaken annually in England, is a good example of a (one-stage) cluster sample. All members of the selected households (clusters) are included in the survey.

Cluster sampling can be more efficient that simple random sampling, especially where a study takes place over a wide geographical region. For instance, it is easier to contact lots of individuals in a few GP practices than a few individuals in many different GP practices. Disadvantages include an increased risk of bias, if the chosen clusters are not representative of the population, resulting in an increased sampling error.

Non-Probability Sampling Methods

1. Convenience sampling

Convenience sampling is an easy method of sampling, because participants are selected based on availability and willingness to take part. Useful results can be obtained, but the results are subject to significant bias, because those who volunteer to take part may be different from those who choose not to (volunteer bias), and the sample may not be representative of other characteristics.

Note: volunteer bias is a risk of all non-probability sampling methods.

2. Judgement (or Purposive) Sampling

Also known as selective, or subjective, sampling. This technique relies on the judgement of the researcher when choosing who to ask to participate. Researchers may implicitly choose a "representative" sample to suit their needs, or specifically approach individuals with certain characteristics. This approach is often used by the media when canvassing the public for opinions and in qualitative research.

Judgement sampling has the advantage of being time-and cost-effective to perform whilst resulting in a range of responses (particularly useful in qualitative research). However, in addition to volunteer bias, it is also subject to errors of judgement by the researcher and the findings will not necessarily be representative.

3. Quota sampling

This method of sampling is often used by market researchers. Interviewers are given a quota of subjects of a specified type in order to recruit. For example, an interviewer might be told to go out and select 20 adult men, 20 adult women, 10 teenage girls and 10 teenage boys so that they could interview them about their television viewing. Ideally the quotas chosen would proportionally represent the characteristics of the underlying population.

This has the advantage of being relatively straightforward and potentially representative, but the selected sample may not be representative of other characteristics that weren't considered (a consequence of the non-random nature of sampling).

4. Snowball sampling

This method is commonly used in social sciences when investigating hard-to-reach groups. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball. For example, when carrying out a survey of risk behaviours amongst drug users, participants may be asked to nominate other users to be interviewed.

Snowball sampling can be effective when a sampling frame is difficult to identify. However, by selecting friends of friends, there is a significant risk of selection bias (choosing a large number of people with similar characteristics or views.

7.3 Point Estimation

- **Sample statistic**: is a sample characteristic such as a sample mean ($\bar{x}$), a sample standard deviation (s), sample proportion ($\bar{p}$), sample variance (s^2), ...

- So, we use sample statistic to estimate the population parameters such as μ, σ, $\hat{p}$ (small p), σ^2, ...

- **Point estimator**:
  - The sample statistic $\bar{x}$ "sample mean" is a point estimator of the population mean μ.
  - The sample standard deviation s is a point estimator of the population standard deviation σ.
  - The sample proportion $\bar{p}$ is a point estimator of the population proportion p.

- **Point estimate**: The numerical value obtained for the point estimators $\bar{x}, s, \bar{p}, \ldots$ as an estimate for the population parameters, $\mu, \sigma, p, \ldots$ are called point estimate.

- Sometimes the point estimates differ from the corresponding population parameters because we are taking a sample and not a census.

- **Target population**: is the population we want to make statistical inference about. It is important for the target population to correspond as closely as possible to the sampled population.

Example (Q12 page 265) A survey question for a sample of 150 individuals, yielded 75 Yes responses, 55 No responses and 20 No opinion.

[a] what is the point estimate of the proportion in the population who respond Yes

[b] " " " " " " " " " " " " " " No

[a] $\bar{p} = \frac{75}{150} = 0.5$ [b] $\bar{p} = \frac{55}{150} = 0.367$

Example (Q 13 page 265) A simple random sample of 5 months of sales data provided the following information

Month	Units Sold (x_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	94	1	1
2	100	7	49
3	85	-8	64
4	94	1	1
5	92	-1	1
	465		116

[a] Develop a point estimate of the population mean number of units sold per month

[b] = = = = = = standard deviation.

[a] $\bar{x} = \frac{\sum x_i}{n} = \frac{465}{5} = 93$

[b] $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{116}{4}} = 5.39$

7.4 Introduction to Sampling Distribution

Recall Example*: Population: 2500 managers of companies

Population parameters: [1] Population mean annual salary μ = \$51,800

[2] Population proportion of managers who completed the company's training program p

[3] Population st. deviation σ = \$4000.

Now • the sample mean $\bar{x}$ is the point estimator of the population mean μ

• the sample proportion $\bar{p}$ is the point estimator of the population proportion p

• If we select a simple random sample of size 30 from the population above 500 times, we obtain:

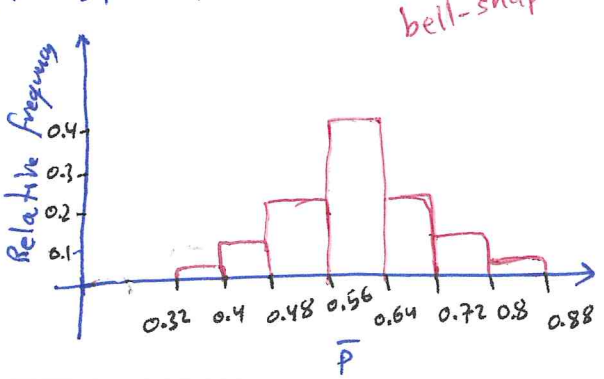
Sample #	Sample mean ($\bar{x}$)	Sample proportion ($\bar{p}$)
1	51,814	0.63
2	52,670	0.70
3	51,780	0.67
⋮	⋮	⋮
500	51,752	0.50

Note that different values of $\bar{x}$ and $\bar{p}$ were obtained.

Thus, the sample mean $\bar{x}$ is a random variable. Hence, $\bar{x}$ has mean or expected value $E(\bar{x}) = \$51,800$, standard deviation $\sigma_{\bar{x}}$, and probability distribution.

• The probability distribution of $\bar{x}$ is called the sampling distribution.

Sampling distribution: A prob. distribution consisting of all possible values of a sample statistic.



7.5 Sampling Distribution of $\bar{x}$

* The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.

• Now we study the properites of the sampling distribution of $\bar{x}$ as other prob. distributions, interms of the following characteristics:
  1. Expected value
  2. standarad deviation
  3. the shap of the distribution

  to determine how close the sample mean $\bar{x}$ is to the population mean μ

1 Expected Value of $\bar{x}$

* The expected value of $\bar{x}$ equals to the mean of the population from which the sample is selected.

$$E(\bar{x}) = \mu \quad \text{---- (1)}$$

* Hence, with a simple random sample, the expected value or mean of the sampling distribution of $\bar{x}$ is equal to the mean of the population.

* Recall that in Example*, the population mean $\mu = \$51,800$ and the expected value of $\bar{x}$ is $E(\bar{x}) = \$51,800$

Hence, $E(\bar{x}) = \mu = \$51,800$ and we say that the point estimator $\bar{x}$ is unbiased.

* Unbiased: A property of a point estimator that is present when the expected value of the point estimator is equal to the population parameter.

* Note that eq. (1) shows that $\bar{x}$ is unbiased estimator of the population mean μ.

2 Standard Deviation of $\bar{x}$

Let: N = Population size

n = sample size

σ = Population standard deviation

$\sigma_{\bar{x}}$ = standard deviation of the sampling distribution of $\bar{x}$.

---

* For finite Population: the standard deviation of $\bar{x}$ is

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{where}$$

$\sqrt{\frac{N-n}{N-1}}$ is the finite population correction factor.

* For infinite population "process" or when N is large and n is small

that is when $\frac{n}{N} \leq 0.05$, then the standard deviation of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

This because the finite population correction factor becomes close to 1, and so has a little effect on the value of $\sigma_{\bar{x}}$. (so, we ignor it).

---

* Note that $\sigma_{\bar{x}}$ is also called the standard error of the point estimator $\bar{x}$

Hence, the standard error is used throughout statistical inference to refere to the standard deviation of a point estimator.

$\sigma_{\bar{p}}$ is the standard error of the proportion.

Recall Example* where the population standard deviation $\sigma = \$4000$ and $N = 2500$ managers, with simple random sample of size $n = 30$, we have $\frac{n}{N} = \frac{30}{2500} = 0.012 \leq 0.05$. Hence the standard error of $\bar{x}$ is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$

[3] Shap (Form) of the sampling Distribution of $\bar{x}$

we consider two cases:

[1] The population has a normal distribution:

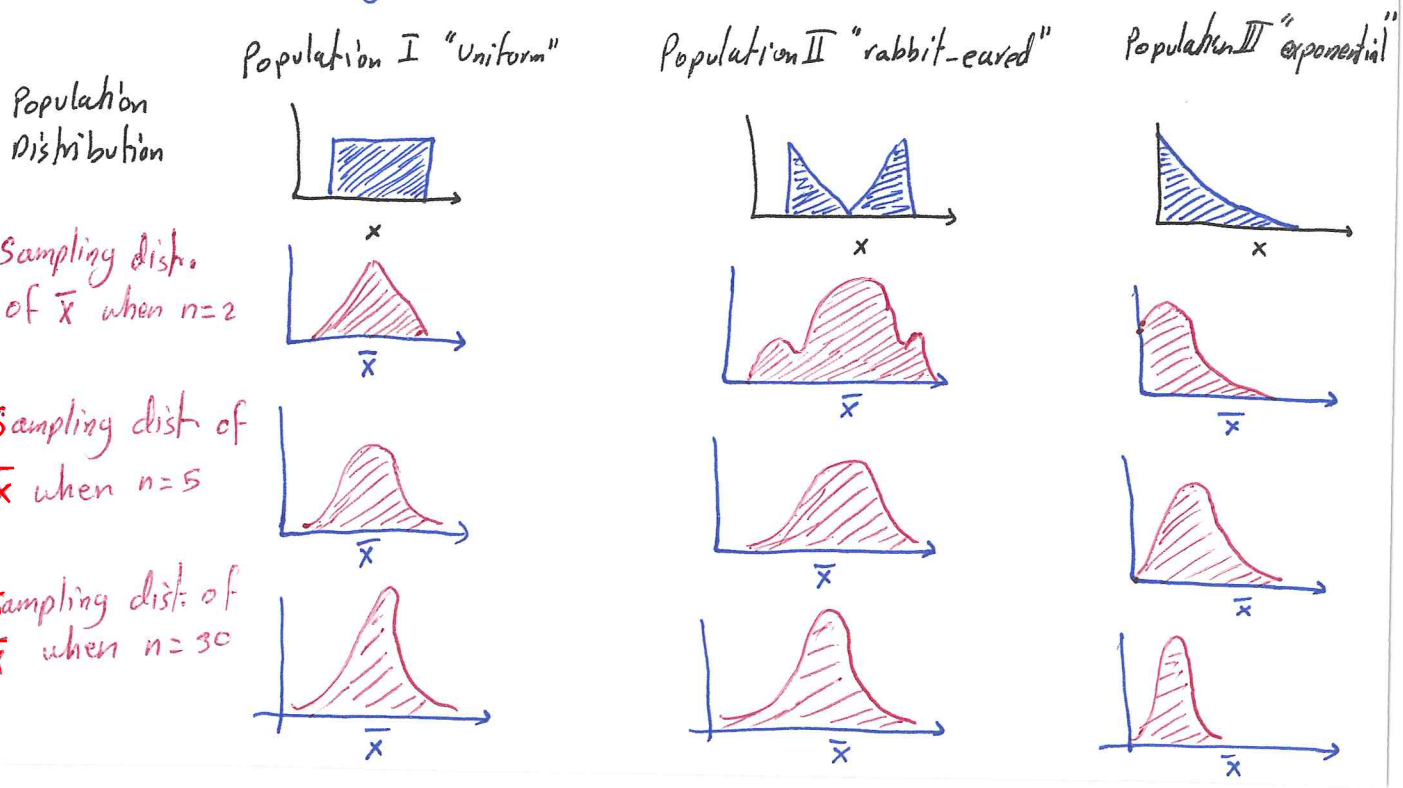
⇒ when the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

[2] The population does not have a normal distribution:

⇒ when the population, from which we select a simple random sample, does not have a normal distribution, we apply the Central limit Theorem to identify the shape of the sampling distribution of $\bar{x}$.

Central limit Theorem:

In selecting simple random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution as the sample size becomes Large.



* Hence, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution if the sample size $n \geq 30$.

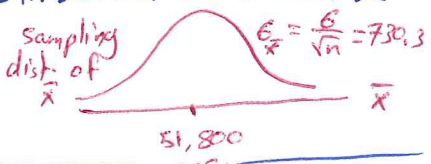
* If the population is highly skewed or If the population has an outliers, then the sampling distribution of $\bar{x}$ can be approximated by a normal distribution if the sample size $n \geq 50$.

Recall Example* where $E(\bar{x}) = \$51,800$ and $\sigma_{\bar{x}} = 730.3$

$\Rightarrow$ we don't know if the population is normally distributed or not.

- So if the population has a normal distribution, then the sampling distribution of $\bar{x}$ is normally distributed.
- If the population does not have normal distribution, then
  1. the simple random sample of size 30 managers and
  2. the central limit Theorem

enable us to conclude that the sampling distribution of $\bar{x}$ can be approximated by a normal distribution.



Example (Q18 page 276) A population has a mean of 200 and standard deviation of 50. A simple random sample of size 100 will be taken and the sample mean $\bar{x}$ will be used to estimat the population mean.

$\mu = 200$
$\sigma = 50$
$n = 100$

[a] what is the expected value of $\bar{x}$? $E(\bar{x}) = \mu = 200$

[b] what is the standard deviation of $\bar{x}$? $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{100}} = \frac{50}{10} = 5$

[c] Show the sampling distribution of $\bar{x}$?

Since $n = 100 \geq 30$, it follows by the Central limit Theorem that the sampling distribution is approximated by a normal distribution with $E(\bar{x}) = 200$ and $\sigma_{\bar{x}} = 5$.

[d] What does the sampling distribution of $\bar{x}$ show?

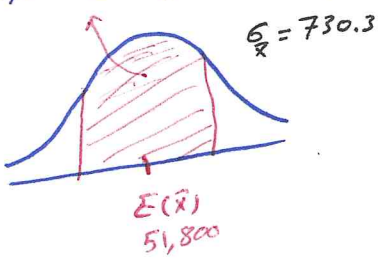
The probability distribution of $\bar{x}$. (see next page)

The sampling distribution of $\bar{x}$ provides probability information about how close the sample mean $\bar{x}$ to the population mean μ.

Recall Example* with $E(\bar{x}) = 51,800$ and $\sigma_{\bar{x}} = 730.3$

what is the Prop. that the sample mean will be within \$500 of the population mean?

$$P(51,800 - 500 \le \bar{x} \le 51,800 + 500) = P(51,300 \le \bar{x} \le 52,300)$$



$$z = \frac{52,300 - 51,800}{730.30} = 0.68$$

$$z = \frac{51,300 - E(\bar{x})}{\sigma_{\bar{x}}} = \frac{51,300 - 51,800}{730.3} = -0.68$$

$$P(51,300 \le x \le 52,300) = P(-0.68 \le z \le 0.68) = P(z \le 0.68) - P(z \le -0.68)$$
$$= 0.7517 - 0.2483 = 0.5034$$

- Hence, a simple random sample of size 30 has roughly 50/100 chance of providing a sample mean within \$500.

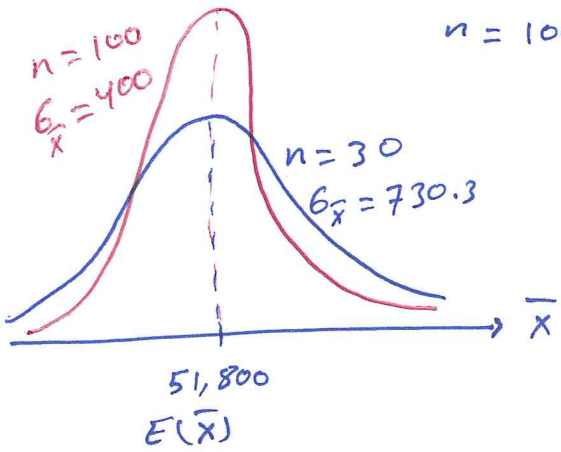
What is the relationship between the sample size n and the sampling distribution of $\bar{x}$?

* Note that $E(\bar{x}) = \mu$ regardless of the sample size n.
* The standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ depends on the sample size n:

As n increases, the standard error $\sigma_{\bar{x}}$ decreases:

For example when $n = 30 \Rightarrow \sigma_{\bar{x}} = 730.3$ "in Example*"

$n = 100 \Rightarrow \sigma_{\bar{x}} = \frac{4000}{\sqrt{100}} = \frac{4000}{10} = 400$



Example (Q 19 page 277) A population has a mean 200 and standard deviation 50. Suppose a simple random sample of size 100 is selected and $\bar{x}$ is used to estimate μ.

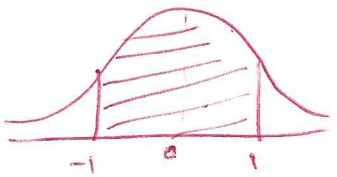
[a] What is the prob. that the sample mean will be within ±5 of the population mean? $E(\bar{x}) = 200$, $\sigma = 50$, $n = 100$

$$P(200-5 \le \bar{x} \le 200+5) = P(195 \le \bar{x} \le 205)$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{10} = 5$$

$$z = \frac{205 - E(\bar{x})}{\sigma_{\bar{x}}} = \frac{205-200}{5} = \frac{5}{5} = 1$$

$$z = \frac{195 - E(\bar{x})}{\sigma_{\bar{x}}} = \frac{195-200}{5} = \frac{-5}{5} = -1$$



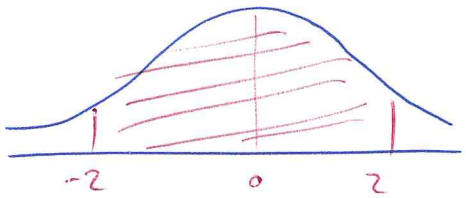
$$P(195 \le \bar{x} \le 205) = P(-1 \le z \le 1) = P(z \le 1) - P(z \le -1) = 0.8413 - 0.1587 = 0.6826$$

[b] What is the prob. that the sample mean will be within ±10 of the population mean?

$$P(200-10 \le \bar{x} \le 200+10) = P(190 \le \bar{x} \le 210)$$

$$z = \frac{210 - E(\bar{x})}{\sigma_{\bar{x}}} = \frac{210-200}{5} = \frac{10}{5} = 2$$

$$z = \frac{190 - E(\bar{x})}{\sigma_{\bar{x}}} = \frac{190-200}{5} = \frac{-10}{5} = -2$$



$$P(190 \le \bar{x} \le 210) = P(-2 \le z \le 2) = P(z \le 2) - P(z \le -2) = 0.9772 - 0.0228 = 0.9544$$

Example (Q 21 page 277) Suppose a simple random sample of size 50 is selected from a population with $\sigma = 10$. Find the standard error if

[a] The population is infinite: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} = \frac{10}{7.07} = 1.4$

[b] The population size is $N = 50{,}000$ since $\frac{n}{N} = 0.001 \le 0.05$, it follows that $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 1.4$

[c] The population size is $N = 5000$ since $\frac{n}{N} = 0.01 \le 0.05$ $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 1.4$

[d] The population size is $N = 500$ since $\frac{n}{N} = 0.1 > 0.05$ $\sigma_{\bar{x}} = \sqrt{\frac{450}{499}}(1.4) = 0.9496(1.4) = 1.33$

7.6 Sampling Distribution of $\bar{p}$

- Recall that the sample proportion $\bar{p}$ is the point estimator of the population proportion p.
- we compute $\bar{p}$ by the formula $\bar{p} = \frac{x}{n}$ where

  x = number of elements of intrest in the sample

  n = sample size
- Note that the sample proportion $\bar{p}$ is a random variable and it's prob. distribution is called the sampling distribution of $\bar{p}$.
- The sampling distribution of $\bar{p}$ is the probability distribution of all possible values of the sample proportion $\bar{p}$.

- To determine how close the sample proportion $\bar{p}$ is to the population proportion p, we study the properties of the sampling distribution of $\bar{p}$ in terms of the following characteristics:
  1. Expected value of $\bar{p}$
  2. Standard deviation of $\bar{p}$
  3. the shap (Form) of the sampling distribution of $\bar{p}$.

1 Expected value of $\bar{p}$

$E(\bar{p}) = p$ Hence, $\bar{p}$ is an unbiased estimator of p.

2 Standard deviation of $\bar{p}$

* For Finite Population: the standard deviation of $\bar{p}$ is

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}} \quad \text{where}$$

$\sqrt{\frac{N-n}{N-1}}$ is the finite population correction factor.

* For infinite population "process" or when $\frac{n}{N} \leq 0.05$ the standard deviation of $\bar{p}$ is

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

* Note that $\sigma_{\bar{p}}$ is also called the standard error of proportion. Thus, $\sigma_{\bar{x}}$ and $\sigma_{\bar{p}}$ are called standard errors.

Recall Example* where $p = 0.6$, $N = 2500$, $n = 30$

$\frac{n}{N} = \frac{30}{2500} = 0.012 \leq 0.05$ so we ignore the finite population correction factor $\sqrt{\frac{N-n}{N-1}}$

Hence, the standard error of $\bar{p}$ is $\Rightarrow$ $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.6(1-0.6)}{30}} = \sqrt{\frac{(0.6)(0.4)}{30}} = 0.0894$

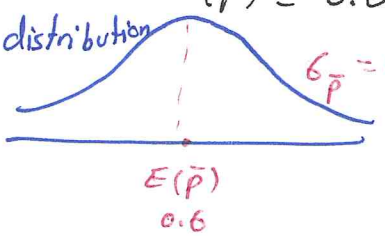
---

3 Shap (Form) of the Sampling Distribution of $\bar{p}$

$\bar{p} = \frac{x}{n}$ where x is binomial random variable that indicates the number of elements of intres in the sample n which is taken from Large population. "n is constant"

- Since n is constant, it follows that $P(\frac{x}{n}) = f(x) =$ binomial prob. of x. Hence $\bar{p}$ is a discrete prob. distribution.
- But, as we showed in ch6, a binomial distribution can be approximated by a normal distribution if $np \geq 5$ and $n(1-p) \geq 5$.
- The sampling distribution of $\bar{p}$ can be approximated by a normal distribution if $np \geq 5$ and $n(1-p) \geq 5$.

Recall the Example* where $E(\bar{p}) = 0.6$ and $\sigma_{\bar{p}} = 0.0894$



$np = 30(0.6) = 18 \geq 5$

$n(1-p) = 30(0.4) = 12 \geq 5$

Thus, the sampling distribution of $\bar{p}$ can be approximated by normal Dist.

Example (Q 31 page 282) A simple random sample of size 100 is selected from a population with $p = 0.4$.

[a] What is the expected value of $\bar{p}$? $E(\bar{p}) = p = 0.4$

[b] What is the standard error of $\bar{p}$? $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.4)(0.6)}{100}} = 0.049$

[c] Show the sampling distribution of $\bar{p}$?

$np = 100(0.4) = 40 \geq 5$ and

$n(1-p) = 100(0.6) = 60 \geq 5$

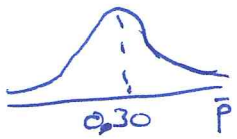
Hence, the sampling distribution is approximated by a normal distribution with $E(\bar{p}) = 0.4$ and $\sigma_{\bar{p}} = 0.049$

[d] What does the sampling distribution of $\bar{p}$ show?

The probability distribution of $\bar{p}$.

Example (Q 35 page 283) $p = 30\%$ and $n = 100$.

[a] Assume $p = 0.30$. what is the sampling distribution of $\bar{p}$?



$np = 100(0.30) = 30 \geq 5$ and

$n(1-p) = 100(0.70) = 70 \geq 5$.

$\Rightarrow$ The normal distribution is appropriate with $E(\bar{p}) = 0.30$

and $\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.3(0.7)}{100}} = 0.0458$

[b] What is the prob. that the sample proportion $\bar{p}$ will be between 0.20 and 0.40?

$P(0.20 \leq \bar{p} \leq 0.40) = P(-2.18 \leq z \leq 2.18) = 0.9854 - 0.0146 = 0.9708$

$z = \frac{0.20 - E(\bar{p})}{\sigma_{\bar{p}}} = \frac{0.20 - 0.30}{0.0458} = -2.18$

$z = \frac{0.40 - E(\bar{p})}{\sigma_{\bar{p}}} = \frac{0.40 - 0.30}{0.0458} = 2.18$

[c] What is the prob. that the sample proportion $\bar{p}$ will be between 0.25 and 0.35?

$P(0.25 \leq \bar{p} \leq 0.35) = P(-1.09 \leq z \leq 1.09) = 0.8621 - 0.1379 = 0.7242$

$z = \frac{0.25 - E(\bar{p})}{\sigma_{\bar{p}}} = \frac{0.25 - 0.30}{0.0458} = -1.09$

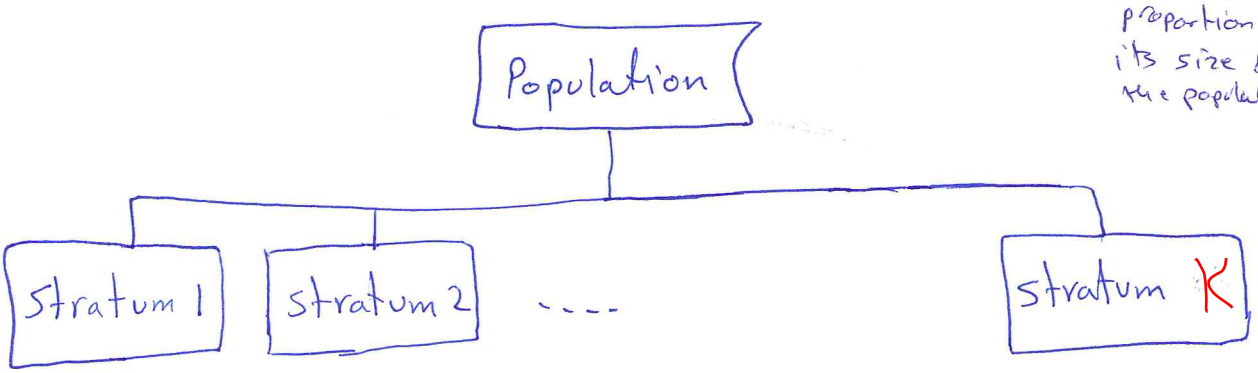
$z = \frac{0.35 - E(\bar{p})}{\sigma_{\bar{p}}} = \frac{0.35 - 0.30}{0.0458} = 1.09$

7.7 Other Sampling Methods:

In this section we provide an alternative sampling methods other than simple random sampling.

Sampling Methods:

1. Simple random sampling: (Finite Population) A sample selected s.t each possible sample of size n has the same prob. of being selected.
2. Stratified Random Sampling: A probability sampling elements selected from population with know prob. of being included in the sample. method in which the population is first divided into strata (small groups) and a simple random sample is then taken from each stratum. ↓ proportion to its size from the population

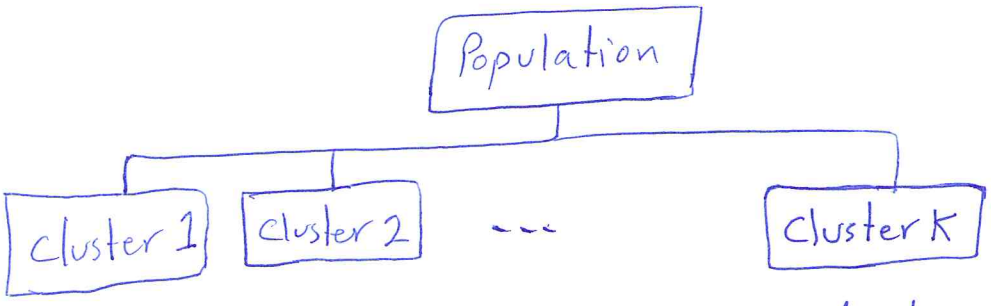


- Each element in the population belongs to one and only one stratum
- Strat can be department, location, age, industry type ...
- Stratified random sampling workes best (gives a good estimat) when the variance among elements in each stratum is small. (the elements of each stratum are as much alike as possible).
- Thus, the stratified random sampling depends on how homogenous the elements are within the strata.

If elements within strata are alike, then strata will have low variance
⇒ Thus, relatively small sample sizes can be used to obtain good estimates

If strata are homogenous, then the stratified random sampling procedure provides results as precise as those of simple random sampling by using a smaller total sample size.

[3] Cluster Sampling: A probability sampling method in which the population is first divided into clusters and then a simple random sample of the clusters is taken.



- The elements in the population are first divided into separate groups called clusters.
- Each element in the population belongs to one and only one cluster.
- Cluster sampling provides the best results when the elements within the clusters are not alike (each cluster is a representative small-scale version of the entire population.
- Thus, cluster sampling depends on how representative each cluster is of the entire population.
  - If all clusters are alike (= they have the same representative), then sampling a small number of clusters will provide good estimate of the population parameters.
- Example of cluster sampling is area sampling where cluster are city blocks.
- Cluster sampling requires a larger total sample size than
  - simple random sampling and
  - stratified random sampling.

  but costs less because the data can be collected in each cluster in a shorter time, and using only one interviewer.

Example: Suppose the Dep. of agriculture wishes to investigate the use of pesticides by farmers in Palestine

مبيدات

A cluster sample could be taken by identifying different ~~countri~~ cities as clusters.

A sample of these cities would then be chosen at random all farmers in these selected cities would be included in the sample.

[4] Systematic Sampling: A probability sampling in which we randomly select one of the first k elements and then select the k^{th} element thereafter.

- When the population is Large, it's time consuming to select a simple random sample by finding random numbers, counting, until the corresponding element. So we use Systematic Sampling:
- For example: A population of 5000 elements.
  Suppose we need a sample of size 50 elements

  $\frac{5000}{50} = 100$

  - A systematic sample selects randomly one elemen in the first 100 elements (65^{th} (suppose it was 65))
  - The systematic sample will be the elements whose orders are:

    65^{th}, $65^{th}+100$, $65^{th}+200$, $65^{th}+300$, ..., $65^{th}+4900$

- That is, we move systematically through the population List, and identifying every 100^{th} elemen after the first element is randomly selected
- Systematic sampling in this way is easier than simple random sample and it has the properties of a simple random sample since the first element is randomly selected.

5 Convenience Sampling: A nonprobability method of sampling whereby elements are selected for the sample on the basis of convenience.

- Elements are included in the sample without know probabilities.
- For example: The university send a sample of selected professors to participate in an international conference, in specific field.
  "Volunteer panels" متطوعين

- Convenience samples are easy to select, but it is impossible to evaluate the "goodness" of the sample in terms of its representative of the population.
- Convenience samples may provide a good results or may not.
- We should be cautious to use the results of convenicence sampling to make inferences about the population, since no statistical methods applied here.

6 Judgment Sampling: A nonprobability method of sampling where by elements are selected for the sample based on the judgment of the person doing the study.

- The person selects elements of the population that he/she feels they are most representative of the population.
- This method is easy way to select a sample.
- Aa example is that a political party select a two or three senators to make an TV interview since they reflect the general opinion of all senators.
- We should be carful to use the results of the judgment sampling to make inference about the population.

Summary:

- We recommend to use prob. sampling methods:
  - simple random sampling
  - stratified random sampling
  - cluster sampling
  - systematic sampling

  → propability sampling

because for these methods, formulas are available for evaluating the "goodness" of the sample results in terms of the closeness of the results to the population parameters, being estimated.

- An evaluation of the goodness cannot be made with
  - convenience sampling and
  - judgment sampling.

  → non propability sampling

Thus, a great care should be used in interpreting the results based on nonprobability sampling methods.

Methods of sampling from a population

It would normally be impractical to study a whole population, for example when doing a questionnaire survey. Sampling is a method that allows researchers to infer information about a population based on results from a subset of the population, without having to investigate every individual. Reducing the number of individuals in a study reduces the cost and workload, and may make it easier to obtain high quality information.

If a sample is to be used, by whatever method it is chosen, it is important that the individuals selected are representative of the whole population. This may involve specifically targeting hard to reach groups.

There are several different sampling techniques available, and they can be subdivided into two groups: probability sampling and non-probability sampling.

In probability (random) sampling, you start with a complete sampling frame of all eligible individuals from which you select your sample. In this way, all eligible individuals have a chance of being chosen for the sample, and you will be more able to generalise the results from your study. Probability sampling methods tend to be more time-consuming and expensive than non-probability sampling.

In non-probability (non-random) sampling, you do not start with a complete sampling frame, so some individuals have no chance of being selected. Consequently, you cannot estimate the effect of sampling error and there is a significant risk of ending up with a non-representative sample which produces non-generalisable results. However, non-probability sampling methods tend to be cheaper and more convenient, and they are useful for exploratory research and hypothesis generation.

Probability Sampling Methods

1. Simple random sampling

In this case each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected.

One way of obtaining a random sample is to give each individual in a population a number, and then use a table of random numbers to decide which individuals to include. For example, if you have a sampling frame of 1000 individuals, labelled 0 to 999, use groups of three digits from the random number table to pick your sample. So, if the first three numbers from the random number table were 094, select the individual labelled “94”, and so on.

As with all probability sampling methods, simple random sampling allows the sampling error to be calculated and reduces selection bias. A specific advantage is that it is the most straightforward method of probability sampling.

2. Systematic sampling

Individuals are selected at regular intervals from the sampling frame. The intervals are chosen to ensure an adequate sample size. If you need a sample size n from a population of size N, you should select every N/n^{th} individual for the sample.

For example, if you wanted a sample size of 100 from a population of 1000, select every 1000/100 = 10^{th} member of the sampling frame.

Systematic sampling is often more convenient than simple random sampling, and it is easy to administer. However, it may also lead to bias, as an example, if a group of students were being selected to get their opinions on college facilities, but the Student Record Department's central list of all students was arranged such that the sex of students alternated between male and female, choosing an even interval (e.g. every 20^{th} student) would result in a sample of all males or all females.

3. Stratified sampling

In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups.

For example, in a study of stroke outcomes, we may stratify the population by sex, to ensure equal representation of men and women. The study sample is then obtained by taking equal sample sizes from each stratum.

In stratified sampling, it may also be appropriate to choose non-equal sample sizes from each stratum. For example, in a study of the health outcomes of nursing staff in a county, if there are three hospitals each with different numbers of nursing staff (hospital A has 500 nurses, hospital B has 1000 and hospital C has 2000), then it would be appropriate to choose the sample numbers from each hospital *proportionally* (e.g. 10 from hospital A, 20 from hospital B and 40 from hospital C).

Stratified sampling improves the accuracy and representativeness of the results by reducing sampling bias. However, it requires knowledge of the appropriate characteristics of the sampling frame (the details of which are not always available), and it can be difficult to decide which characteristic(s) to stratify by.

4. Clustered sampling

In a clustered sample, subgroups of the population are used as the sampling unit, rather than individuals. The population is divided into subgroups, known as clusters, which are randomly selected to be included in the study.

For example individual GP practices or towns could be identified as clusters. In single-stage cluster sampling, all members of the chosen clusters are then included in the study. In two-stage cluster sampling, a selection of individuals from each cluster is then randomly selected for inclusion. Clustering should be taken into account in the analysis. The General Household survey, which is undertaken annually in England, is a good example of a (one-stage) cluster sample. All members of the selected households (clusters) are included in the survey.

Cluster sampling can be more efficient that simple random sampling, especially where a study takes place over a wide geographical region. For instance, it is easier to contact lots of individuals in a few GP practices than a few individuals in many different GP practices. Disadvantages include an increased risk of bias, if the chosen clusters are not representative of the population, resulting in an increased sampling error.

Non-Probability Sampling Methods

1. Convenience sampling

Convenience sampling is an easy method of sampling, because participants are selected based on availability and willingness to take part. Useful results can be obtained, but the results are subject to significant bias, because those who volunteer to take part may be different from those who choose not to (volunteer bias), and the sample may not be representative of other characteristics.

Note: volunteer bias is a risk of all non-probability sampling methods.

2. Judgement (or Purposive) Sampling

Also known as selective, or subjective, sampling. This technique relies on the judgement of the researcher when choosing who to ask to participate. Researchers may implicitly choose a "representative" sample to suit their needs, or specifically approach individuals with certain characteristics. This approach is often used by the media when canvassing the public for opinions and in qualitative research.

Judgement sampling has the advantage of being time-and cost-effective to perform whilst resulting in a range of responses (particularly useful in qualitative research). However, in addition to volunteer bias, it is also subject to errors of judgement by the researcher and the findings will not necessarily be representative.

3. Quota sampling

This method of sampling is often used by market researchers. Interviewers are given a quota of subjects of a specified type in order to recruit. For example, an interviewer might be told to go out and select 20 adult men, 20 adult women, 10 teenage girls and 10 teenage boys so that they could interview them about their television viewing. Ideally the quotas chosen would proportionally represent the characteristics of the underlying population.

This has the advantage of being relatively straightforward and potentially representative, but the selected sample may not be representative of other characteristics that weren't considered (a consequence of the non-random nature of sampling).

4. Snowball sampling

This method is commonly used in social sciences when investigating hard-to-reach groups. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball. For example, when carrying out a survey of risk behaviours amongst drug users, participants may be asked to nominate other users to be interviewed.

Snowball sampling can be effective when a sampling frame is difficult to identify. However, by selecting friends of friends, there is a significant risk of selection bias (choosing a large number of people with similar characteristics or views.