Sample selection methods

Why sample selection?
In a census information is collected on all subjects in the population. However, in most epidemiological studies we only study a sample of the population and we hope to generalize the results on the target population. Generalization of the results requires that the sample should represent the population from which the sample is selected. Therefore, selection of the sample i.e. selection of the study subjects should be undertaken in a proper way not according to the likes and dislikes of the researcher.

Since we only study a sample of the population, there are two main potential sources of error which could limit generalization of the results. These are sampling error and selection bias.

Sampling error occurs because we only study a sample of the population and it reflects the variation between the sample estimate and the true population estimate. Sampling error is therefore unavoidable but it is less with larger samples. Sampling error is measured by the standard error of the sample estimate and standard error equals the standard deviation of the sample divided by square root of the sample size ($se=s/\sqrt{n}$).

Selection bias could happen intentionally or unintentionally when the subjects are not selected randomly from a clearly defined population. In a proper sample selection, every subject in the population should have the same chance of being included in the study. Random sampling methods are followed to ensure this chance and avoid selection bias.

Defining the target population and the study population
Selection of study subjects requires that the target population and study population be clearly identified. The target population is the larger population on which you would like to generalize the results of your study and therefore which includes your study population. The study population is part of the target population from which we draw the sample. The study population is more limited in number and more accessible for the research than the target population. We aim to draw the sample form a list of all subjects of the study population. Such a list is called **sampling frame**. For example if we wanted to do a study on job satisfaction of the teachers in Sulaymaniyah city, the target population will be all teachers in Sulaymaniyah city, the study population will be the population of teachers from which we can draw the sample, for example teachers employed by the Department of Education, and the sampling frame will be a list containing the names of all teachers at the DoE.

The **sampling unit** is the subject to be sampled from the study population. In the previous example, a teacher who is going to be interviewed is the

sampling unit. A sampling unit is not necessarily an individual; it could be any thing such as a test, a school or a cluster of households.

So, the sequence of sample selection will be
Target population ‣ Study population ‣ Sampling frame ‣ Sampling unit.

Question: If you wanted to do a study on family size in a Khabat quarter, what would be each of the 4 elements above? And if you wanted to do the same study in a village of 20 families?

Sampling fraction
Sampling fraction is the proportion of the study population sampled. In other words sampling fraction tells us what percentage of the study population will be included in the study.

Sampling fraction= number of subjects studied/ number of subjects on sampling frame

For example if there are 400 pregnant women registered with an antenatal clinic and we decided to interview 80 of them, the sampling fraction will be 80/400= 0.2 i.e. we will sample 20% of the pregnant women.

Sampling methods
Once the sampling frame has been identified, then it is time to select the sampling units i.e. subjects from this list. The best methods of selection aim to ensure that each sampling unit has the same chance of being chosen for the study. These random methods which aim to give equal chance to every one on the sampling frame are called **equal probability selection methods** (epsem). The advantage of these random sampling methods is that the population estimates could be measured from the sample estimates i.e. study results could be extrapolated to the population.

Main methods of random sampling are simple random selection, systematic sampling, stratified sampling, multi-stage sampling and cluster sampling. All these methods use epsem at some stage.

1) Simple random sampling
This is the most straightforward methods of selecting a sample and it is ideal because only chance decides who will be sampled from the study population. It is appropriate when the study population is homogenous in terms of representing the target population and if all subjects could be accessed. Selecting subjects using simple random selection includes to stages:

1) Draw a list including all subjects (sampling frame) and number them.
2) Chose the required number of subjects (sampling units) randomly by:

a. making a card for each subject, shuffling them and drawing the required number of subjects; or better by
b. using random number tables or,
c. using random numbers generated by a computer.

To use a random table such as the table shown below, firstly we have to number the study population from first to last subject and see how many digits the biggest number comprises. Secondly we use random numbers with that number of digits. Thirdly we start arbitrarily from one part of the table and start selecting numbers that exist in the sampling frame. And finally we continue down or across the table until we have selected all subjects.

Example: how do we select 60 children from a sampling frame of 450 children of a school? Firstly number the children from 1 through 450. Secondly was we know that 450 has 3 digits therefore we select numbers with 3 digits from the table. Thirdly we start from the top left of the table i.e. from number 034 (three digits), so child number 34 is the first to be selected. Fourthly we continue to select numbers downwards until we have selected all 60 children. The second number in the table is 977 which is bigger than our study population so we don't need it. The third is 167 which is within range and therefore selected. And so one we select 125, 162, 332, 181, 266, 234, 378, 160, 311, 274, 3, 299 and so on.

03	47	43	73	86	36	96	47	36	61	46	98	63	71	62	33	26	16	80	45	60	11	14	10	95
97	74	24	67	62	42	81	14	57	20	42	53	32	37	32	27	07	36	07	51	24	51	79	89	73
16	76	62	27	66	56	50	26	71	07	32	90	79	78	53	13	55	38	58	59	88	97	54	14	10
12	56	85	99	26	96	96	68	27	31	05	03	72	93	15	57	12	10	14	21	88	26	49	81	76
55	59	56	35	64	38	54	82	46	22	31	62	43	09	90	06	18	44	32	53	23	83	01	30	30
16	22	77	94	39	49	54	43	54	82	17	37	93	23	78	87	35	20	96	43	84	26	34	91	64
84	42	17	53	31	57	24	55	06	88	77	04	74	47	67	21	76	33	50	25	83	92	12	06	76
63	01	63	78	59	16	95	55	67	19	98	10	50	71	75	12	86	73	58	07	44	39	52	38	79
33	21	12	34	29	78	64	56	07	82	52	42	07	44	38	15	51	00	13	42	99	66	02	79	54
57	60	86	32	44	09	47	27	96	54	49	17	46	09	62	90	52	84	77	27	08	02	73	43	28
18	18	07	92	46	44	17	16	58	09	79	83	86	19	62	06	76	50	03	10	55	23	64	05	05
26	62	38	97	75	84	16	07	44	99	83	11	46	32	24	20	14	85	88	45	10	93	72	88	71
23	42	40	64	74	82	97	77	77	81	07	45	32	14	08	32	98	94	07	72	93	85	79	10	75
52	36	28	19	95	50	92	26	11	97	00	56	76	31	38	80	22	02	53	53	86	60	42	04	53
37	85	94	35	12	83	39	50	08	30	42	34	07	96	88	54	42	06	87	98	35	85	29	48	39
70	29	17	12	13	40	33	20	38	26	13	89	51	03	74	17	76	37	13	04	07	74	21	19	30
56	62	18	37	35	96	83	50	87	75	97	12	25	93	47	70	33	24	03	54	97	77	46	44	80
99	49	57	22	77	88	42	95	45	72	16	64	36	16	00	04	43	18	66	79	94	77	24	21	90
16	08	15	04	72	33	27	14	34	09	45	59	34	68	49	12	72	07	34	45	99	27	72	95	14
31	16	93	32	43	50	27	89	87	19	20	15	37	00	49	52	85	66	60	44	38	68	88	11	80
68	34	30	13	70	55	74	30	77	40	44	22	78	84	26	04	33	46	09	52	68	07	97	06	57
74	57	25	65	76	59	29	97	68	60	71	91	38	67	54	13	58	18	24	76	15	54	55	95	52
27	42	37	86	53	48	55	90	65	72	96	57	69	36	10	96	46	92	42	45	97	60	49	04	91
00	39	68	29	61	66	37	32	20	30	77	84	57	03	29	10	45	65	04	26	11	04	96	67	24
29	94	98	94	24	68	49	69	10	82	53	75	91	93	30	34	25	20	57	27	40	48	73	51	92

Simple random sampling is good for a study population confined to a limited geographical area but if the population is distributed in a wide it may not be logistically possible to do it. For example if we wanted to take a sample from school

children of Sulaymaniyah schools, a random sample means that we might have to go to every possible school in the area. Simple random sampling may also not be appropriate if there are differences in exposure/ disease status of the study population of which we are interested unless we take a very large sample.

2) Systematic sampling

The difference of systematic sampling from simple random sampling is that we don't use random numbers, instead we select the subjects from the sampling frame at regular intervals. The interval is taken from the sampling fraction. For example if the sampling fraction is 10% then we select every 10^{th} subject from the list. Different steps of systematic selection are:

1. Prepare the sampling frame but it should not be ordered in any particular pattern that could affect the selection process. For example it is not right to alternate male and female or other characteristics.
2. Calculate the sampling fraction say it is 10%, so the interval will be 10.
3. Obtain a random number between 1 and 10. Let's assume you get 6, so subject number 6 will be the first subject.
4. Each time you add 10 to the selected subject you will get the next selection i.e. 6, 16. 26, 36, 46 etc. until all subjects are selected.

There is one assumption behind systematic sampling i.e. that the sampling frame is not ordered in any particular pattern. If there is a repeating pattern in ordering the sampling frame list, bias can occur.

Example: how would you select 10 subjects from the sampling list below using systematic sampling?

1. Calculate the interval (30/10=3)
2. Get a random number between 1 and 3, assume we get 2
3. So first subject: is number 2 (B)
4. then add 3 to each selected subject and get the rest .

sn	name	sex	sn	name	sex
1	A	m	16	P	m
2	**B**	m	**17**	**Q**	f
3	C	f	18	R	m
4	D	m	19	S	m
5	**E**	f	**20**	**T**	f
6	F	f	21	U	m
7	G	f	22	V	m
8	**H**	m	**23**	**W**	m
9	I	m	24	X	f
10	J	f	25	Y	m
11	**K**	m	**26**	**Z**	f
12	L	f	27	AA	f
13	M	f	28	AB	f
14	**N**	f	**29**	**AC**	m
15	O	f	30	AD	m

More complex sampling methods

It is not always possible to apply the previous two sampling methods and we need to use more complex sampling methods. Simple random sampling and systematic sampling are not appropriate in the following situations:

1. When there is no existing sampling frame available and a sampling frame could not be easily created.
2. When the study population is distributed in a wide area which makes it logistically difficult to select all subjects using simple methods.
3. When the population is composed of subgroups with distinct characteristics which may be interested to be sufficiently represented in the sample.

When the above situations exist we use more complex sampling methods which includes stratified sampling, multi-stage sampling and cluster sampling. These methods could be used alone or in combination. Like simple methods, these methods aim to give every subject an equal chance to be selected using epsem.

3) Stratified sampling

When the population consists of distinct subgroups (strata) that may differ in relation to the exposure/outcome under study and we are interested to include these subgroups in the study, stratified sampling is more appropriate than simpler methods. We stratify the sample by the groups and usually apply epsem by sampling the same proportion from each group i.e. using the same sampling fraction. But if this leads to a very small sample size in small strata, the sample size of such strata has to be increased in order to give them adequate representation.

For example if we want to do a survey of malnutrition in a city. and we know there are 3 different ethnic groups in the city where malnutrition rates may differ between them, it would be more appropriate to stratify the sample by ethnic groups in order to give the chance to all ethnic groups be adequately represented. If some ethnic groups have a small population they may not be represented by simple or systematic random sampling.

4) Multi-stage sampling

Multi-stage sampling means that the sample is selected in two or more stages. At the first stage we select what we call first-stage units and at the next stage we select what we call second-stage units and so on. The advantage of this method is that we don't need to have a sampling frame of all sampling units; we only need that for the second stage. The disadvantage of multi-stage sampling is that its estimation of the population characteristic is not as precise as simple random sampling with the same sample size. A larger sample size is required to increase precision of the population estimate.

For example if want to do a survey of smoking amongst school children in Sulaymaniyah city, we can use a two-stage sampling method. At the first stage we take a random sample of all schools in the city and at the second stage we take a sample of children from selected schools. To do this we need a list of first-stage units (all schools) and a list of the second-stage units (the children) of selected schools. Selection of the second-stage units is by simple random sampling. Selection of the first-stage units is by simple random selection if units (schools in the example) contain equal number of pupils. But if they differ considerably, we have to use what is called probability proportionate to size (pps) in order to ensure that every pupil has an equal chance to be selected (epsem). This means that if a school consists of twice as many pupils as another, it should have twice the chance of the first school to be selected.

If there are more than 2 stages in the sampling process e.g. provinces, districts and households, the method is called multi-stage sampling.

5) Cluster sampling

Cluster sampling is like a two-stage sampling with the difference that in the second stage, all sampling units of each selected second-stage unit (i.e. all pupils of selected schools in the previous example) are included. In this case the second-stage units are called clusters. Selection of the clusters is by simple random sampling. A cluster could be a school, a group of households, a village etc.

Central Organization for Statistics & Information Technology and UNICEF undertake a multiple indicator cluster survey in Iraq every 5 years or so. These health and demographic surveys were undertaken in 1995, 2000 and 2006. This is a stratified two-stage cluster survey in which the sample is stratified to provinces and urban & rural areas. Quarters/ villages are selected by systematic sampling (first stage SU). Then clusters are selected from these areas (second stage SU) and all families of each selected cluster (about 20 households) are sampled.

Design factor and design effect

The more complex sampling methods involve more than one strata or stage and therefore there is additional variability in the sample estimate compared to random sampling methods. In random sampling the source of variability is only between the subjects selected (sampling units) but in a 2-stage sampling method, in addition to this (in which case they are called second-stage sampling units) there will be variability in the first-stage sampling units.

The effect of the sampling design on variability of the sample estimate is

measured by **design factor**. Design factor is calculated by comparing the standard error of the sample estimate under the complex design compare to the standard error using random sampling and the same sample size.

Design factor= se under complex design/ se under simple random sampling

The square of the design factor is called **design effect**. Design effect is used to adjust the sample size and the confidence interval of the sample estimate. This is a way to account to the additional variability of we encounter using more complex sampling methods.

Further reading

Essentials of medical statistics, by Kirkwood
Field trials of health interventions in developing countries, edited by P Smith and R marrow