STA 2100 PROBABILITY AND STATISTICS I

PURPOSE

Byt the end of the course the student should be proficient in representing data graphically and handling summary statistics, simple correlation and best fitting line, and handling probability and probability distributions including expectation and variance of a discrete random variable.

DESCRIPTION

Classical and axiomatic approaches to probability. Compound and conditional probability, including Bayes' theorem. Concept of discrete random variable: expectation and variance. Data: sources, collection, classification and processing. Frequency distributions and graphical representation of data, including bar diagrams, histograms and stem-and-leaf diagrams. Measures of central tendency and dispersion. Skewness and kurtosis. Correlation. Fitting data to a best straight line.

Pre-Requisites: STA 2104 Calculus for statistics I, SMA 2104 Mathematics for Science.

COURSE TEXT BOOKS

1. Uppal, S. M., Odhiambo, R. O. & Humphreys, H. M. *Introduction to Probability and Statistics.* JKUAT Press, 2005. ISBN 9966-923-95-0
2. J Crawshaw & J Chambers *A concise course in A-Level statistics, with worked examples*, 3rd ed. Stanley Thornes, 1994 ISBN 0-534- 42362-0.

COURSE JOURNALS

- Journal of Applied Statistics (J. Appl. Stat.) [0266-4763; 1360-0532]
- Statistics (Statistics) [0233-1888]

FURTHER REFERENCE TEXT BOOKS AND JOURNALS

i) GM Clarke & D Cooke *A Basic Course in Statistics.* 5th ed. Arnold, 2004 ISBN13: 978-0-340-81406-2 ISBN10: 0-340-81406-3.
ii) S Ross *A first course in Probability* 4th ed. Prentice Hall, 1994 ISBN-10: 0131856626 ISBN-13: 9780131856622.
iii) P.S. Mann. *Introductory Statistics*. John Wiley & Sons Ltd, 2001 ISBN 13: 9780471395119.
iv) Statistical Science (Stat. Sci.) [0883-4237]
v) Journal of Mathematical Sciences
vi) Journal of Teaching Statistics

Introduction

What is statistics?

The Word statistics has been derived from Latin word "**Status**" or the Italian word "**Statista**", the meaning of these words is "**Political State**" or a Government. Early applications of statistical thinking revolved around the needs of states to base policy on demographic and economic data.

Definition

Statistics: *a branch of science that deals with collection presentation, analysis, and interpretation of data*. The definition points out 4 key aspects of statistics namely

(i) Data collection
(ii) Data presentation,
(iii) Data analysis, and
(iv) Data interpretation

Statistics is divided into 2 broad categories namely descriptive and inferential statistics.

Descriptive Statistics: summary values and presentations which gives some information about the data Eg the mean height of a 1st year student in JKUAT is170cm. 170cm is a statistics which describes the central point of the heights data.

Inferential Statistics: summary values calculated from the sample in order to make conclusions about the target population.

Types of Variables

Qualitative Variables: Variables whose values fall into groups or categories. They are called categorical variables and are further divided into 2 classes namely nominal and ordinal variables

a) Nominal variables: variables whose categories are just names with no natural ordering. Eg gender marital status, skin colour, district of birth etc
b) Ordinal variables: variables whose categories have a natural ordering. Eg education level, performance category, degree classifications etc

Quantitative Variables: these are numeric variables and are further divided into 2 classes namely discrete and continuous variables

a) Discrete variables: can only assume certain values and there are gaps between them. Eg the number of calls one makes in a day, the number of vehicles passing through a certain point etc
b) Continuous variables: can assume any value in a specified range. Eg length of a telephone call, height of a 1st year student in JKUAT etc

1. Data Collection:

1.1 Sources of Data

There are 2 sources for data collection namely Primary, and Secondary data

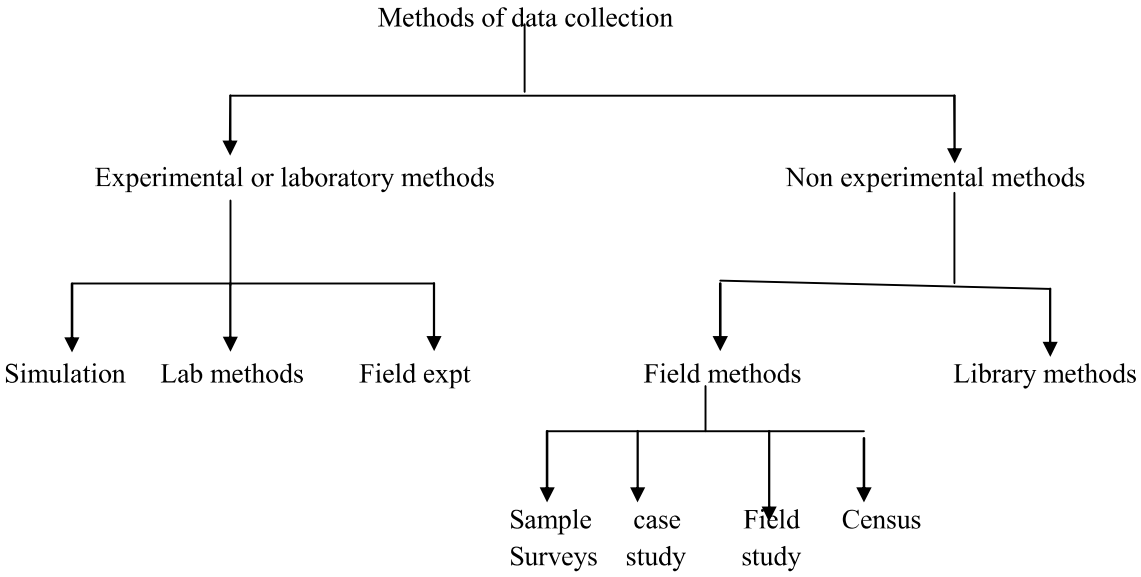
Primary data:- freshly collected ie for the first time. They are original in character ie they are the first hand information collected, compiled and published for some purpose. They haven't undergone any statistical treatment

***Secondary Data*:-** 2nd hand information mainly obtained from published sources such as statistical abstracts books encyclopaedias periodicals, media reports eg census report CD-roms and other electronic devices, internet. They are not original in character and have undergone some statistical treatment at least once.

1.2 Data Collection Methods

The 1st step in any investigation (inquiry) is data collection. Information can either be collected directly or indirectly from the entire population or a sample.

There are many methods of collecting data which includes the ones illustrated in the flow chart below



Experimental methods are so called because in them the investigator in a laboratory tests the hypothesis about the cause and effect relationship by manipulating the independent variables under controlled conditions.

Non-Experimental methods are so called because in them the investigator does not control or change any aspect of the situation under study but simply describes what naturally occurs at a certain point or period of time.

Non-Experimental methods are widely used in social sciences. Some of the Non-Experimental methods used for data collection are outlined below.

a) **Field study**:- aims at testing hypothesis in natural life situations. It differs from field experiment in that the researcher does not control or manipulate the independent variables but both of them are carried out in natural conditions

Merits*:*

(i) The method is realistic as it is carried out in natural conditions
(ii) It's easy to obtain data with large number of variables.

Demerits

(iii) Independent variables are not manipulated.
(iv) Co-operation of the organization is often difficult to obtain.
(v) Data is likely to contain unknown sampling biasness.
(vi) The dross rate (proportion of irrelevant data) may be high in such studies.
(vii) Measurement is not precise as in laboratory because of influence of confounding variables.

b) **Census.** A census is a study that obtains data from every member of a population (totality of individuals /items pertaining to certain characteristics). In most studies, a census is not practical, because of the cost and/or time required.

c) **Sample survey**. A sample survey is a study that obtains data from a subset of a population, in order to estimate population attributes/ characteristics. Surveys of human populations and institutions are common in government, health, social science and marketing research.

d) Case study –It's a method of intensively exploring and analyzing the life of a single social unit be it a family, person, an institution, cultural group or even an entire community. In this method no attempt is made to exercise experimental or statistical control and phenomena related to the unit are studied in natural. The researcher has several discretion in gathering information from a variety of sources such as diaries, letters, autobiographies, records in office, files or personal interviews.

Merits:

(i) The method is less expensive than other methods.
(ii) Very intensive in nature –aims at studying a few units rather than several
(iii) Data collection is flexible since the researcher is free to approach the problem from any angle.
(iv) Data is collected from natural settings.

Demerits

(i) It lacks internal validity which is basic to scientific evidence.
(ii) Only one unit of the defined population is studied. Hence the findings of case study cannot be used as abase for generalization about a large population. They lack external validity.
(iii) Case studies are more time consuming than other methods.

e) **Experiment.** An experiment is a controlled study in which the researcher attempts to understand cause-and-effect relationships. In experiments actual experiment is carried out on certain individuals / units about whom information is drawn. The study is "controlled" in the sense that the researcher controls how subjects are assigned to groups and which treatments each group receives.
f) **Observational study.** Like experiments, observational studies attempt to understand cause-and-effect relationships. However, unlike experiments, the researcher is not able to control how subjects are assigned to groups and/or which treatments each group receives. Under this method information, is sought by direct observation by the investigator.

1.3 Population and Sample

Population: The entire set of individuals about which findings of a survey refer to.
Sample: A subset of population selected for a study.
Sample Design: The scheme by which items are chosen for the sample.
Sample unit: The element of the sample selected from the population.
Unit of analysis: Unit at which analysis will be done for inferring about the population. Consider that you want to examine the effect of health care facilities in a community on prenatal care. What is the unit of analysis: health facility or the individual woman?.

Sampling Frames

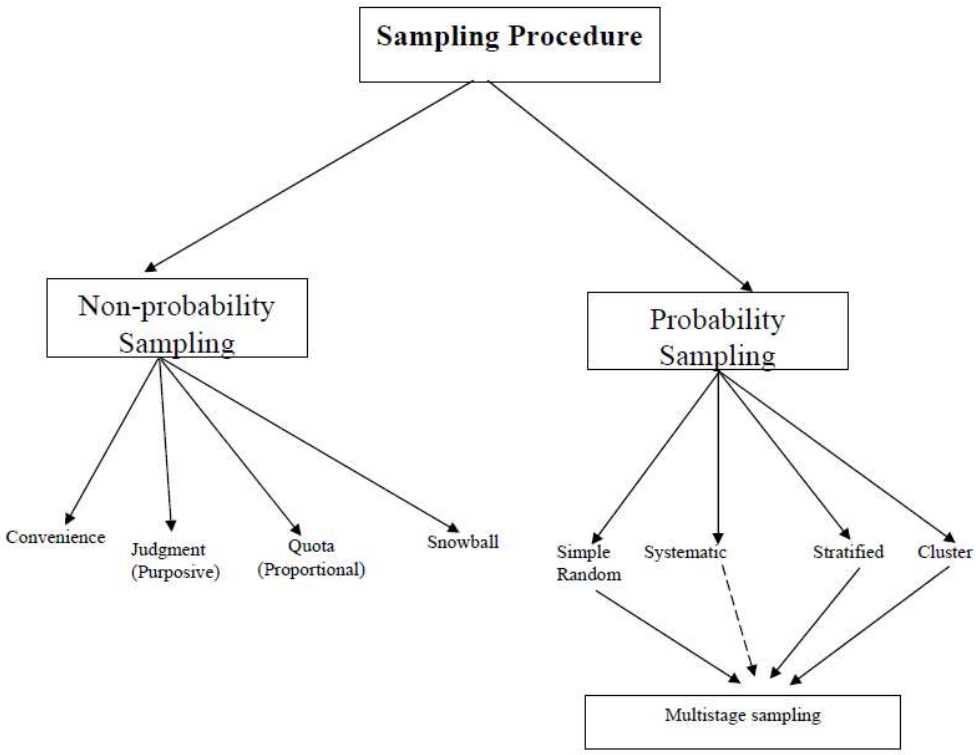
For probability sampling, we must have a list of all the individuals (units) in the population. This list or sampling frame is the basis for the selection process of the sample. “A [sampling] frame is a clear and concise description of the population under study, by virtue of which the population units can be identified unambiguously and contacted, if desired, for the purpose of the survey” - Hedayet and Sinha, 1991

1.4 Sampling

Sampling is a statistical process of selecting a representative sample. We have probability sampling and non-probability sampling **Probability Samples** involves a mathematical chance of selecting the respondent. Every unit in the population has a chance, greater than zero, of being selected in the sample. Thus producing unbiased estimates. They include;

(i) Simple random sampling
(ii) Systematic sampling
(iii) Stratified sampling
(iv) Cluster sampling
(v) multi-stage sampling

Non-probability sampling is any sampling method where some elements of the population have *no* chance of selection (also referred to as “out of coverage”/”undercovered”), or where the probability of selection can't be accurately determined. It yields a non-random sample therefore making it difficult to extrapolate from the sample to the population. They include; Judgement sample, purposive sample, convenience sample: **subjective** Snow-ball sampling: **rare group/disease study**



1.4.1 Sampling Procedure

Sampling involves two tasks

- How to select the elements?
- How to estimate the population characteristics – from the sampling units?

We employ some *randomization* process for sample selection so that there is no preferential treatment in selection which may introduce selectivity bias

1.4.2 Reasons Behind sampling

(i) Cost; the sample can furnish data 0f sufficient accuracy at much lower cost.
(ii) Time; the sample provides information faster than census thus ensuring timely decision making.
(iii) Accuracy; it is easier to control data collection errors in a sample survey as opposed to census.
(iv) Risky or destructive test call for sample survey not census eg testing a new drug.

1.4.3 Probability Sampling Techniques

a)...**Simple Random Sampling** (SRS)

In this design, each element has an equal probability of being selected from a list of all population units (sample of n from N population). Though it's attractive for its simplicity, the design is not usually used in the sample survey in practice for several reasons:

(i) Lack of listing frame: the method requires that a list of population elements be available, which is not the case for many populations.
(ii) Problem of small area estimation or domain analysis: For a small sample from a large population, all the areas may not have enough sample size for making small area estimation or for *domain* analysis by variables of interest.
(iii) Not cost effective: SRS requires covering of whole population which may reside in a large geographic area; interviewing few samples spread sparsely over a large area would be very costly.

Implementation of SRS sampling:

(i) Listing (sampling) Frame
(ii) Random number table (from published table or computer generated)

(iii) Selection of sample

Computer generated random numbers: (STATA output)

832645	573158	467460	838921	171721	152885
708009	285644	727733	343305	539264	907568
305761	995036	740619	054728	746425	713746
536405	504168	750032	367682	626278	855480
217862	782003	409660	155199	129514	484511
844905	296231	103727	053603	562252	219726
670523	707073	049209	830572	337034	716264
334920	023934	808901	740693	170372	095017
885588	384435	129958	303040	264636	858065
458268	058670	888935	064613	661404	411861
277649	076177	482951	876389	898190	927367
977683	759956	553916	983998	331578	981306

b)..Systematic Sampling

Systematic sampling, either by itself or in combination with some other method, may be the most widely used method of sampling." In systematic sampling we select samples "evenly" from the list (sampling frame): First, let us consider that we are dividing the list evenly into some "blocks". Then, we select a sample element from each block.

In systematic sampling, only the first unit is selected at random, the rest being selected according to a predetermined pattern. To select a systematic sample of n units, the first unit is selected with a random start r from 1 to k sample, where $k=N/n$ sample intervals, and after the selection of first sample, every k^{th} unit is included where $1 \leq r \leq k$.

An example:
Let N=100, n=10, then k=100/10. Then the random start r is selected between 1 and 10 (say, r=7). So, the sample will be selected from the population with serial indexes of: 7, 17, 27, ...,97. i.e., r, r+k, r+2k,....., r+(n-1)k

What could be done if k=N/n is not an integer?

Selection of systematic sampling when sampling interval (k) is not an integer
Consider, n=175 and N=1000. So, k=1000/175 = 5.71
One of the solution is to make k rounded to an integer, i.e., k=5 or k=6.Now, if k=5, then n=1000/5=200; or, If k=6, then n=1000/6 = 166.67 ~ 167. Which n should be chosen?

Solution
if k=5 is considered, stop the selection of samples when n=175 achieved.
if k=6 is considered, treat the sampling frame as a circular list and continue the selection of samples from the beginning of the list after exhausting the list during the first cycle.

An alternative procedure is to keep k non-integer and continue the sample selection as follows: Let us consider, k=5.71, and r=4. So, the first sample is 4th in the list. The second = (4+5.71) =9.71 ~9th in the list, the third =(4+2*5.71) =15.42 ~ 15th in the list, and so on. (The last sample is: 4+5.71*(175-1) = 997.54 ~ 997th in the list). Note that, k is switching between 5 and 6

Advantages:

Systematic sampling has many attractiveness:

(i) Provides a better random distribution than SRS
(ii) Simple to implement
(iii) May be started without a complete listing frame (say, interview of every 9th patient coming to a clinic).
(iv) With ordered list, the variance may be smaller than SRS (see below for exceptions

Disadvantages:

(i) Periodicity (cyclic variation)
(ii) linear trend

i.

When to use systematic sampling?

i) Even preferred over SRS
ii) When no list of population exists
iii) When the list is roughly of random order
iv) Small area/population

c)..Stratified Sampling

In stratified sampling the population is partitioned into groups, called *strata*, and sampling is performed separately within each *stratum*.

This sampling technique is used when;

i) Population groups may have different values for the responses of interest.
ii) we want to improve our estimation for each group separately.
iii) To ensure adequate sample size for each group.

In stratified sampling designs:

i) Stratum variables are mutually exclusive (no over lapping), e.g., urban/rural areas, economic categories, geographic regions, race, sex, etc. The principal objective of stratification is to reduce sampling errors.
ii) The population (elements) should be *homogenous* within-stratum, and the population (elements) should be *heterogeneous* between the strata.

Advantages

(i) Provides opportunity to study the stratum; variations - estimation could be made for each stratum
(ii) Disproportionate sample may be selected from each stratum
(iii) The precision is likely to increase as variance may be smaller than simple random case with same sample size
(iv) Field works can be organized using the strata (e.g., by geographical areas or regions)
(v) Reduce survey costs.

Disadvantages

(i) Sampling frame is needed for each stratum
(ii) Analysis method is complex
(iii) Correct variance estimation
(iv) Data analysis should take sampling "weight" into account for disproportionate sampling of strata
(v) Sample size estimation is difficult in practice

Allocation of Stratified Sampling

The major task of stratified sampling design is the appropriate allocation of samples to different strata.

Types of allocation methods:
(i) Equal allocation
(ii) Proportional to stratum size
(iii) Cost based sample allocation

Equal Allocation

Divide the number of sample units *n* equally among the *K* strata. ie $n_i = \frac{n}{k}$ Example: n = 100 and k= 4 strata $n_i = \frac{100}{4} = 25$ *units* in each stratum.

Disadvantages of equal allocation:
May need to use weighting to have unbiased estimates

Proportional allocation

Make the proportion of each stratum sampled identical to the proportion of the population. Ie Let the sample fraction be f= n/N. So, $n_i = fN_i = n\frac{N_i}{N}$, Where $\frac{N_i}{N}$ is the stratum weight.

Example: N = 1000, n = 100 $f = \frac{100}{1000} = 0.1$ now suppose $N_1 = 700$ and $N_2 = 300$ then $n_1 = 700 * 0.1 = 70$ and $n_2 = 300 * 0.3 = 30$

Disadvantage of proportional allocation:
Sample size in a stratum may be low thus providing unreliable stratum-specific results.

d)..Cluster Sampling

In many practical situations the population elements are grouped into a number of clusters. A list of clusters can be constructed as the sampling frame but a complete list of elements is often unavailable, or too expensive to construct. In this case it is necessary to use cluster sampling where a random sample of clusters is taken and some or all elements in the selected clusters are observed. Cluster sampling is also preferable in terms of cost, because it is much cheaper, easier and quicker to collect data from adjoining elements than elements chosen at random. On the other hand, cluster sampling is less informative and less efficient per elements in the sample, due to similarities of elements within the same cluster. The loss of efficiency, however, can often be compensated by increasing the overall sample size. Thus, in terms of unit cost, the cluster sampling plan is efficient.

e)..Multi-Stage Samples

Here the respondents are chosen through a process of defined stages. Eg residents within Kibera (Nairobi) may have been chosen for a survey through the following process:
Throughout the country (Kenya) the Nairobi may have been selected at random, (stage 1), within Nairobi, Langata (constituency) is selected again at random (stage 2), Kibera is then selected within Langata (stage 3), then polling stations from Kibera (stage 4) and then individuals from the electoral voters' register (stage 5)! As demonstrated five stages were gone through before the final selection of respondents were selected from the electoral voters' register.

Advantages of probability sample

(i) Provides a quantitative measure of the extent of variation due to random effects
(ii) Provides data of known quality
(iii) Provides data in timely fashion
(iv) Provides acceptable data at minimum cost
(v) Better control over nonsampling sources of errors
(vi) Mathematical statistics and probability can be applied to analyze and interpret the data

1.4.4 Non-probability Sampling

Social research is often conducted in situations where a researcher cannot select the kinds of probability samples used in large-scale social surveys. For example, say you wanted to study homelessness - there is no list of homeless individuals nor are you likely to create such a list. However, you need to get some kind of a sample of respondents in order to conduct your research. To gather such a sample, you would likely use some form of non-probability sampling.
There are four primary types of non-probability sampling methods:

a)..Convinience Sampling

It's a method of choosing subjects who are available or easy to find. This method is also sometimes referred to as haphazard, accidental, or availability sampling. The primary advantage of the method is that it is very easy to carry out, relative to other methods.

Demerit

- One can never be certain what population the participants in the study represent. The population is unknown.
- The method is haphazard, and the cases studied probably don't represent any population you could come up with. However, it's very useful for pilot studies

Advantages of convenience sample

(i) It's *very easy* to carry out with few rules governing how the *sample* should be collected.
(ii) The *relative cost* and *time* required to carry out a convenience sample are *small* in comparison to probability sampling techniques. This enables you to achieve the *sample size* you want in a *relatively fast* and *inexpensive* way.
(iii) The convenience sample may help you gather useful data and information that would not have been possible using *probability sampling techniques*, which require more formal access to *lists of populations* [see, for example, the article on simple random sampling].

For example, imagine you were interested in understanding more about employee satisfaction in a single, large organisation in the United States. You intended to collect your data using a questionnaire. The manager who has kindly given you access to conduct your research is unable to get permission to get a *list* of all employees in the organisation, which you would need to use a *probability sampling technique* such as simple random sampling or systematic random sampling. However, the manager has managed to secure permission for you to spend two days in the organisation to collect as many questionnaire responses as possible. You decide to spend the two days at the entrance of the organisation where all employees have to pass through to get to their desks. Whilst a *probability sampling technique* would have been preferred, the convenience sample was the only sampling technique that you could use to collect data. Irrespective of the disadvantages of convenience sampling, discussed below, without the use of this sampling technique, you may not have been able to get access to any data on employee satisfaction in the organisation.

Disadvantages of convenience sampling

- The convenience sample often suffers from a number of *biases*. This can be seen in both of our examples, whether the 10,000 students we were studying, or the employees at the large organisation. In both cases, a convenience sample can lead to the *under-representation* or *over-representation* of particular *groups* within the *sample*. If we take the large organisation:

It may be that the organisation has multiple sites, with employee satisfaction varying considerably between these sites. By conducting the survey at the headquarters of the organisation, we may have missed the differences in employee satisfaction amongst those at different sites, including non-office workers. We also do not know why some employees agreed to take part in the survey, whilst others did

not. Was it because some employees were simply too busy? Did they not trust the intentions of the survey? Did others take part out of kindness or because they had a particular grievance with the organisation? These types of *biases* are quite typical in convenience sampling.

- Since the *sampling frame* is *not know*, and the *sample* is *not chosen at random*, the *inherent bias* in convenience sampling means that the sample is *unlikely* to be *representative* of the *population* being studied. This undermines your ability to make *generalisations* from your *sample* to the *population* you are studying.

If you are an undergraduate or master's level dissertation student considering using *convenience sampling*, you may also want to read more about how to put together your *sampling strategy* [see the section: Sampling Strategy

b)..Quota Sampling

Quota sampling is designed to overcome the most obvious flaw of availability sampling. Rather than taking just anyone, you set quotas to ensure that the sample you get represents certain characteristics in proportion to their prevalence in the population. Note that for this method, you have to know something about the characteristics of the population ahead of time. Say you want to make sure you have a sample proportional to the population in terms of gender - you have to know what percentage of the population is male and female, then collect sample until yours matches. Marketing studies are particularly fond of this form of research design.

The primary problem with this form of sampling is that even when we know that a quota sample is representative of the particular characteristics for which quotas have been set, we have no way of knowing if sample is representative in terms of any other characteristics. If we set quotas for gender and age, we are likely to attain a sample with good representativeness on age and gender, but one that may not be very representative in terms of income and education or other factors.

Moreover, because researchers can set quotas for only a small fraction of the characteristics relevant to a study quota sampling is really not much better than availability sampling. To reiterate, you must know the characteristics of the entire population to set quotas; otherwise there's not much point to setting up quotas. Finally, interviewers often introduce bias when allowed to self-select respondents, which is usually the case in this form of research. In choosing males 18-25, interviewers are more likely to choose those that are better-dressed, seem more approachable or less threatening. That may be understandable from a practical point of view, but it introduces bias into research findings.

Imagine that a researcher wants to understand more about the career goals of students at a single university. Let's say that the university has roughly 10,000 students. suppose we were interested in *comparing the differences* in career goals between *male* and *female* students at the single university. If this was the case, we would want to ensure that the *sample* we selected had a *proportional* number of *male* and *female* students relative to the *population*.

To create a quota sample, there are three steps:

Choose the relevant grouping chsr and divide the population accordingly *gender*

Calculate a quota (number of *units* that should be included in *each* for group

Continue to invite units until the quota for each group is met

Advantages of quota sampling

i) It particularly useful when you are unable to obtain a probability sample, but you are still trying to create a sample that is as representative as possible of the population being studied. In this respect, it is the non-probability based equivalent of the stratified random sample.

ii) Unlike probability sampling techniques, especially stratified random sampling, quota sampling is much *quicker* and *easier* to carry out because it does not require a *sampling frame* and the strict use of random sampling techniques.
iii) The quota sample improves the *representation* of particular *strata* (*groups*) within the *population*, as well as ensuring that these *strata* are *not over-represented*. For example, it would ensure that we have sufficient male students taking part in the research (60% of our *sample size* of 100; hence, 60 male students). It would also make sure we did not have more than 60 male students, which would result in an *over-representation* of male students in our research.
iv) It allows *comparison* of *groups*.

Disadvantages of quota sampling

i) In quota sampling, the *sample* has not been chosen using *random selection*, which makes it impossible to determine the possible *sampling error*.
ii) this *sampling bias*. Thus no*statistical inferences* from the *sample* to the *population*. This can lead to problems of *external validity*.
iii) Also, with quota sampling it must be possible to clearly divide the *population* into *strata*; that is, *each unit* from the population must only belong to *one stratum*. In our example, this would be fairly simple, since our *strata* are *male* and *female* students. Clearly, a student could only be classified as either male or female. No student could fit into both categories (ignoring transgender issues).

c)..Purposive Sampling

Purposive sampling is a sampling method in which elements are chosen based on purpose of the study. Purposive sampling may involve studying the entire population of some limited group or a subset of a population. As with other non-probability sampling methods, purposive sampling does not produce a sample that is representative of a larger population, but it can be exactly what is needed in some cases - study of organization, community, or some other clearly defined and relatively limited group.

Advantages of purposive sampling

i) There are a wide range of *qualitative research designs* that researchers can draw on. Achieving the goals of such qualitative research designs requires different types of *sampling strategy* and *sampling technique*. One of the major benefits of purposive sampling is the wide range of sampling techniques that can be used across such qualitative research designs; purposive sampling techniques that range from *homogeneous sampling* through to *critical case sampling*, *expert sampling*, and more.
ii) Whilst the various purposive sampling techniques each have different goals, they can provide researchers with the justification to make *generalisations* from the sample that is being studied, whether such generalisations are *theoretical*, *analytic* and/or *logical* in nature. However, since each of these types of purposive sampling differs in terms of the nature and ability to make generalisations, you should read the articles on each of these purposive sampling techniques to understand their relative advantages.
iii) Qualitative research designs can involve multiple phases, with each phase building on the previous one. In such instances, different types of sampling technique may be required at each phase. Purposive sampling is useful in these instances because it provides a wide range of non-probability sampling techniques for the researcher to draw on. For example, *critical case sampling* may be used to investigate whether a phenomenon is worth investigating further, before adopting an *expert sampling* approach to examine specific issues further.

Disadvantages of purposive sampling

i) Purposive samples, irrespective of the type of purposive sampling used, *can be* highly prone to *researcher bias*. The idea that a purposive sample has been created based on the *judgement* of the researcher is not a good defence when it comes to alleviating possible researcher biases,

ii) specially when compared with *probability sampling* techniques that are designed to reduce such biases. However, this judgemental, subjective component of purpose sampling is only a major disadvantage when such judgements are *ill-conceived* or *poorly considered*; that is, where judgements have not been based on clear criteria, whether a theoretical framework, expert elicitation, or some other accepted criteria.
iii) The subjectivity and non-probability based nature of *unit* selection (i.e. selecting people, cases/organisations, etc.) in purposive sampling means that it can be difficult to defend the representativeness of the sample. In other words, it can be difficult to convince the reader that the judgement you used to select units to study was appropriate. For this reason, it can also be difficult to convince the reader that research using purposive sampling achieved *theoretical*/*analytic*/*logical generalisation*. After all, if different units had been selected, would the results and any generalisations have been the same?

d)..Snowball Sampling

Snowball sampling is a method in which a researcher identifies one member of some population of interest, speaks to him/her, and then asks that person to identify others in the population that the researcher might speak to. This person is then asked to refer the researcher to yet another person, and so on.
Snowball sampling is very good for cases where members of a special population are difficult to locate. For example,.*populations* that are subject to social stigma and marginalisation, such as suffers of AIDS/HIV, as well as individuals engaged in illicit or illegal activities, including prostitution and drug use. Snowball sampling is useful in such scenarios because:
The method creates a sample with questionable representativeness. A researcher is not sure who is in the sample. In effect snowball sampling often leads the researcher into a realm he/she knows little about. It can be difficult to determine how a sample compares to a larger population. Also, there's an issue of who respondents refer you to - friends refer to friends, less likely to refer to ones they don't like, fear, etc.
Snowball sampling is a useful choice of *sampling strategy* when the *population* you are interested in studying is *hidden* or *hard-to-reach*.

Advantages of Snowball Sampling

(i) The chain referral process allows the researcher to reach populations that are difficult to sample when using other sampling methods.
(ii) The process is cheap, simple and cost-efficient.
(iii) This sampling technique needs little planning and fewer workforce compared to other sampling techniques.

Disadvantages of Snowball Sampling

(i) The researcher has little control over the sampling method. The subjects that the researcher can obtain rely mainly on the previous subjects that were observed.
(ii) Representativeness of the sample is not guaranteed. The researcher has no idea of the true distribution of the population and of the sample.
(iii) Sampling bias is also a fear of researchers when using this sampling technique. Initial subjects tend to nominate people that they know well. Because of this, it is highly possible that the subjects share the same traits and characteristics, thus, it is possible that the sample that the researcher will obtain is only a small subgroup of the entire populatio

1.4.5 Limitations of Sampling

a) Sampling frame: may need complete enumeration
b) Errors of sampling may be high in small areas
c) May not be appropriate for the study objectives/questions
d) Representativeness may be vague, controversial

1.5 Survey Administration

a)..Self-Administered Surveys

Self-administered surveys have special strengths and weaknesses.

They are useful in describing the characteristics of a large population and make large samples feasible.

Advantages:

i) ***Low cost.*** Extensive training is not required to administer the survey. Processing and analysis are usually simpler and cheaper than for other methods.
ii) ***Reduction in biasing error.*** The questionnaire reduces the bias that might result from personal characteristics of interviewers and/or their interviewing skills.
iii) ***Greater anonymity***. Absence of an interviewer provides greater anonymity for the respondent. This is especially helpful when the survey deals with sensitive issues.
iv) Convenience to the respondents (may complete any time at his/her own convenient time)
v) Accessibility (greater coverage, even in the remote areas)
vi) May provide more reliable information (e.g., may consult with others or check records to avoid recall bias)

Disadvantages:

i) ***Requires simple questions.*** The questions must be straightforward enough to be comprehended solely on the basis of printed instructions and definitions.
ii) ***No opportunity for probing.*** The answers must be accepted as final. Researchers have no opportunity to clarify ambiguous answers.
iii) ***Low response rate;*** respondents may not respond to all questions and/or may not return questionnaire
iv) The respondent must be literate to read and understand the questionnaire
v) Introduce self selection bias
vi) Not suitable for complex questionnaire

b). . .Interview Surveys

Unlike questionnaires interviewers ask questions orally and record respondents' answers. This type of survey generally decreases the number of —"do not know" and —"no answer" responses, compared with self-administered surveys. Interviewers also provide a guard against confusing items. If a respondent has misunderstood a question, the interviewer can clarify, thereby obtaining relevant responses.

Interviewer selection: background characteristics (race, sex, education, culture) listening skill recording skill experience unbiased observation/recording

Interviewer training: be familiar with the study objectives and significance thorough familiarity with the questionnaire contextual and cultural issues privacy and confidentiality informed consent and ethical issues unbiased view mock interview session

Supervision of the interviewer: Spot check Questionnaire check Reinterview (reliability check)

Advantages

i) ***Flexibility.*** Allows flexibility in the questioning process and allows the interviewer to clarify terms that are unclear.
ii) ***Control of the interview situation.*** Can ensure that the interview is conducted in private, and respondents do not have the opportunity to consult one another before giving their answers.
iii) ***High response rate.*** Respondents who would not normally respond to a mail questionnaire will often respond to a request for a personal interview.
iv) May record non-verbal behaviour, activities, facilities, contexts
v) Complex questionnaire may be used
vi) Illiterate respondents may participate

Disadvantages

i) ***Higher cost.*** Costs are involved in selecting, training, and supervising interviewers; perhaps in paying them; and in the travel and time required to conduct interviews.

ii) ***Interviewer bias*.** The advantage of flexibility leaves room for the interviewer's personal influence and bias, making an interview subject to interviewer bias.
iii) ***Lack of anonymity*.** Often the interviewer knows all or many of the respondents. Respondents may feel threatened or intimidated by the interviewer, especially if a respondent is sensitive to the topic or to some of the questions.
iv) Less accessibility
v) Inconvenience
vi) Often no opportunity to consult records, families, relatives

c).. Telephone Interview

Advantages:

(i) Less expensive
(ii) Shorter data collection period than personal interviews
(iii) Better response than mail surveys

Disadvantages:

(i) Biased against households without telephone, unlisted number
(ii) Nonresponse
(iii) Difficult for sensitive issues or complex topics
(iv) Limited to verbal responses

d)… Focus Groups

A focus group typically can be defined as a group of people who possess certain characteristics and provide information of a qualitative nature in a focused discussion. They ars generally composed of six to twelve people. Size is conditioned by two factors: the group must be small enough for everyone to participate, yet large enough to provide diversity. This group is special in terms of purpose, size, composition, and procedures. Participants are selected because they have certain characteristics in common that relate to the topic at hand, such as parents of gang members, and, generally, the participants are unfamiliar with each other. Typically, more than one focus group should be convened, since a group of seven to twelve people could be too atypical to offer any general insights on the gang problem.

A trained moderator probes for different perceptions and points of view, without pressure to reach consensus. Focus groups have been found helpful in assessing needs, developing plans, testing new ideas, or improving existing programs

Advantages:

i) Flexibility allows the moderator to probe for more in-depth analysis and ask participants to elaborate on their responses.
ii) Outcomes are quickly known.
iii) They may cost less in terms of planning and conducting than large surveys and personal interviews.

Limitations

i) A skilled moderator is essential.
ii) Differences between groups can be troublesome to analyze because of the qualitative nature of the data.
iii) Groups are difficult to assemble. People must take the time to come to a designated place at a particular time.
iv) Participants may be less candid in their responses in front of peers.