Non-probability sampling

The commonly used non-probability sampling methods include the following.

Convenience or haphazard sampling

Units are selected in an arbitrary manner with little or no planning involved. Haphazard sampling assumes that the population units are all alike, then any unit may be chosen for the sample. An example of haphazard sampling is the vox pop survey where the interviewer selects any person who happens to walk by. Unfortunately, unless the population units are truly similar, selection is subject to the biases of the interviewer and whoever happened to walk by at the time of sampling.

Volunteer sampling

The respondents are only volunteers in this method. Generally, volunteers must be screened so as to get a set of characteristics suitable for the purposes of the survey (e.g. individuals with a particular disease). This method can be subject to large selection biases, but is sometimes necessary. For example, for ethical reasons, volunteers with particular medical conditions may have to be solicited for some medical experiments.

Another example of volunteer sampling is callers to a radio or television show, when an issue is discussed and listeners are invited to call in to express their opinions. Only the people who care strongly enough about the subject one way or another tend to respond. The silent majority does not typically respond, resulting in a large selection bias. Volunteer sampling is often used to select individuals for focus groups or in-depth interviews (i.e. for qualitative testing, where no attempt is made to generalize to the whole population).

Judgment sampling

With this method, sampling is done based on previous ideas of population composition and behavior. An expert with knowledge of the population decides which units in the population should be sampled. In other words, the expert purposely selects what is considered to be a representative sample. Judgment sampling is subject to the researcher's biases and is perhaps even more biased than haphazard sampling.

Since any preconceptions the researcher has are reflected in the sample, large biases can be introduced if these preconceptions are inaccurate. However, it can be useful in exploratory studies, for example in selecting members for focus groups or in-depth interviews to test specific aspects of a questionnaire.

Quota sampling

This is one of the most common forms of non-probability sampling. Sampling is done until a specific number of units (quotas) for various subpopulations have been selected. Quota sampling is a means for satisfying sample size objectives for the subpopulations.

The quotas may be based on population proportions. For example, if there are 100 men and 100 women in the population and a sample of 20 are to be drawn, 10 men and 10 women may be interviewed. Quota sampling can be considered preferable to other forms of non-probability sampling (e.g. judgment sampling) because it forces the inclusion of members of different subpopulations.

Quota sampling is somewhat similar to stratified sampling, which is probability sampling, in that similar units are grouped together. However, it differs in how the units are selected. In probability sampling, the units are

selected randomly while in quota sampling a non-random method is used—it is usually left up to the interviewer to decide who is sampled. Contacted units that are unwilling to participate are simply replaced by units that are, in effect ignoring nonresponse bias. Market researchers often use quota sampling (particularly for telephone surveys) instead of stratified sampling to survey individuals with particular socio-economic profiles. This is because compared with stratified sampling, quota sampling is relatively inexpensive and easy to administer and has the desirable property of satisfying population proportions. However, it disguises potentially significant selection bias.

As with all other non-probability sample designs, in order to make inferences about the population, it is necessary to assume that persons selected are similar to those not selected. Such strong assumptions are rarely valid.

Snowball or network sampling

Suppose a researcher wishes to find rare individuals in the population, and already knows of the existence of some of these individuals and how to contact them. One approach is to contact those individuals and simply ask them if they know anyone like themselves, then contact those people, etc. The sample grows like a snowball rolling down a hill to hopefully include virtually everybody with that characteristic. Snowball sampling is useful for rare or hard to reach populations such as people with disabilities, homeless people, drug users, or other persons who may not belong to an organised group or such as musicians, painters, or poets, not readily identified on a survey list frame. However, some individuals or subgroups may have no chance of being sampled. In order to be able to generalize the conclusion to the whole population, some assumptions, which are usually not met, are required.

Crowdsourcing

Crowdsourcing has been defined slightly differently by researchers from various areas. Despite the multiplicity of definitions for crowdsourcing, one constant has been the broadcasting of a problem to the public, and an open call for contributions to help solve the problem. Members of the public submit solutions that are then owned by the entity (e.g. individuals, companies, or organizations), which originally broadcast the problem. Crowdsourcing is channelling the experts' desire to solve a problem and then freely sharing the answer with everyone.

As part of Statistics Canada's modernization, crowdsourcing has become an innovative way to collect valuable information for statistical purposes. By using crowdsourcing as the only collection method, surveys can be executed quickly with reduced cost and response burden.

Types of Data Distributions and Density Functions

Distributions

From a practical perspective, we can think of a distribution as a function that describes the relationship between observations in a sample space.

For example, we may be interested in the age of humans, with individual ages representing observations in the domain, and ages 0 to 125 the extent of the sample space. The distribution is a mathematical function that describes the relationship of observations of different heights.

Density Functions

Distributions are often described in terms of their density or density functions.

Density functions are functions that describe how the proportion of data or likelihood of the proportion of observations change over the range of the distribution.

Two types of density functions are probability density functions and cumulative density functions.

- **Probability Density function**: calculates the probability of observing a given value.
- **Cumulative Density function**: calculates the probability of an observation equal or less than a value.

A probability density function, or PDF, can be used to calculate the likelihood of a given observation in a distribution. It can also be used to summarize the likelihood of observations across the distribution's sample space. Plots of the PDF show the familiar shape of a distribution, such as the bell-curve for the Gaussian distribution.

Distributions are often defined in terms of their probability density functions with their associated parameters.

A cumulative density function, or CDF, is a different way of thinking about the likelihood of observed values. Rather than calculating the likelihood of a given observation as with the PDF, the CDF calculates the cumulative likelihood for the observation and all prior observations in the sample space. It allows you to quickly understand and comment on how much of the distribution lies before and after a given value. A CDF is often plotted as a curve from 0 to 1 for the distribution.

Both PDFs and CDFs are continuous functions. The equivalent of a PDF for a discrete distribution is called a probability mass function, or PMF.

Next, let's look at the Gaussian distribution and two other distributions related to the Gaussian that you will encounter when using statistical methods. We will look at each in turn in terms of their parameters, probability, and cumulative density functions.

Gaussian Distribution

The Gaussian distribution, named for Carl Friedrich Gauss, is the focus of much of the field of statistics.

Data from many fields of study surprisingly can be described using a Gaussian distribution, so much so that the distribution is often called the "*normal*" distribution because it is so common.

A Gaussian distribution can be described using two parameters:

- **mean**: Denoted with the Greek lowercase letter mu, is the expected value of the distribution.
- **variance**: Denoted with the Greek lowercase letter sigma raised to the second power (because the units of the variable are squared), describes the spread of observation from the mean.

It is common to use a normalized calculation of the variance called the standard deviation

- **standard deviation**: Denoted with the Greek lowercase letter sigma, describes the normalized spread of observations from the mean.

t-Distribution

It is a distribution that arises when attempting to estimate the mean of a normal distribution with different sized samples. As such, it is a helpful shortcut when describing uncertainty or error related to estimating population statistics for data drawn from Gaussian distributions when the size of the sample must be taken into account.

Although you may not use the Student's t-distribution directly, you may estimate values from the distribution required as parameters in other statistical methods, such as statistical significance tests.

The distribution can be described using a single parameter:

- **number of degrees of freedom**: denoted with the lowercase Greek letter nu (v), denotes the number degrees of freedom.

Key to the use of the t-distribution is knowing the desired number of degrees of freedom.

The number of degrees of freedom describes the number of pieces of information used to describe a population quantity. For example, the mean has *n* degrees of freedom as all *n* observations in the sample are used to calculate the estimate of the population mean. A statistical quantity that makes use of another statistical quantity in its calculation must subtract 1 from the degrees of freedom, such as the use of the mean in the calculation of the sample variance.

Chi-Squared Distribution

The chi-squared distribution is denoted as the lowercase Greek letter chi (X) raised to the second power (X^2).

Like the Student's t-distribution, the chi-squared distribution is also used in statistical methods on data drawn from a Gaussian distribution to quantify the uncertainty. For example, the chi-squared distribution is used in the chi-squared statistical tests for independence. In fact, the chi-squared distribution is used in the derivation of the Student's t-distribution.

The chi-squared distribution has one parameter:

- *degrees of freedom*, denoted k.