Part 3: Inferential Statistics

Sampling and Sampling Distributions

Sampling is widely used in business as a means of gathering information about a population.

Reasons for Sampling

There are several good reasons for taking a sample instead of conducting a census:

1. The sample can save money.
2. The sample can save time.
3. For given resources, the sample can broaden the scope of the study.
4. Because the research process is sometimes destructive, the sample can save product.
5. If accessing the population is impossible, the sample is the only option.

Reasons for Taking a Census

1. To eliminate the possibility that by chance a randomly selected sample might not be representative of the population.
2. The client does not have an appreciation for random sampling and feels more comfortable with conducting a census.

Sampling frame – a list, map or directory that is being used to represent the population in the process of sampling.

Types of Sampling

1. **Random Sampling** – Every unit of the population has the same probability of being selected into the sample; also called **probability sampling**.
2. **Nonrandom sampling** – Not every unit of the population has the same probability of being selected to the sample.

Random Sampling Techniques

1. Simple random sampling – the most elementary random sampling technique

   With simple random sampling, each unit of the frame is numbered from 1 to N (where N is the size of the population). Next, a table of random numbers or a random number generator is used to select *n* items to the sample. A random number generator is usually a computer program that allows computer or calculator output to yield random numbers.

Simple random sampling is easier to perform on small than on large populations. The process of numbering all the members of the population and selecting items is cumbersome for large populations.

Sample design – a definite plan, determined completely before any data are actually collected, for obtaining a sample from a given population.

2. **Stratified random sampling** – a procedure that consists of stratifying (or dividing) the population into a number of overlapping subpopulations (or strata) and taking a sample from each stratum.
    a. **Proportionate stratified random sampling** – occurs when the percentage of the sample taken from each stratum is proportionate to the percentage that each stratum is within the whole population.
    b. **Disproportionate stratified random sampling** – occurs whenever the proportions of the strata in the sample are different than the proportions of the strata in the population.
3. **Systematic sampling** – is sued because of its convenience and relative ease of administration. With systematic sampling, every kth item is selected to produce a sample of size *n* from a population of size N.
   Characteristics:
    a. The elements of the population are treated as an ordered sequence.
    b. Elements selected to the sample are selected at constant interval from the ordered sequence frame.
4. **Cluster or Area sampling** – involves dividing the population into nonoverlapping areas or clusters. However, in contrast to stratified random sampling where strata are homogeneous, cluster sampling identifies clusters that are internally heterogeneous.

   Two of the foremost advantages of cluster sampling are convenience and cost. Clusters are usually convenient to obtain and the cost of sampling from the entire population is reduced because the scope of the study is reduced to the cluster.

Nonrandom sampling techniques

1. **Convenience sampling** – Elements are selected for the sample for the convenience of the researcher. The researcher typically chooses items that are readily available, nearby, and/or willing to participate.

2. **Judgment sampling** – occurs when elements are selected for the sample by the judgment of the researcher. Researchers often believe that they can obtain a representative sample by using sound judgment.
3. **Quota sampling** – a nonrandom sampling technique similar to stratified random sampling where the population is stratified on some characteristic. Elements eventually selected for the sample are chosen by nonrandom processes.

Sampling error – occurs when the sample is not representative of the population
Nonsampling errors include:

a. missing data
b. recording errors
c. input processing errors
d. analysis errors
e. response errors
   e.1 telescoping error – occurs when a respondent attributes an event to a wrong time period.
   e.2 omission error - occurs when a respondent fails to mention past events.
   e.3 detail error - occurs when a respondent remembers an event incorrectly.

Sampling Distributions

To illustrate the concept of a sampling distribution, let us construct the one for the mean of a random sample of size *n* = 2 drawn without replacement from the finite population of size N = 5 whose elements are 3, 5, 7, 9 and 11. The mean of this population is

And the standard deviation is

Now, if we take a random sample of size *n* = 2 from this population, there are ___ possibilities, namely,

And their means are:
Since each sample has probability ___, we thus get the following sampling distribution of the mean of a random sample of size *n* = 2 from the given population:

$\bar{x}$	Probability, f(x)

Calculate its mean $\mu_{\bar{X}}$ and its standard deviation $\delta_{\bar{X}}$, where the subscripts serve to distinguish between these parameters and those of the original population

Observations:

1. The mean of the sampling distribution of the mean is ____ to the population mean.
2. The standard deviation of the sampling distribution of the mean is _______ the population standard deviation.

Stating formally, for random samples of size n taken from a population with mean μ and standard deviation δ, the sampling distribution of $\overline{\mathrm{x}}$ has the mean

$\boxed{\mu_{\overline{X}} = \mu}$ **mean of sampling distribution of** $\overline{\mathrm{x}}$

and the standard deviation

$\boxed{\delta_{\overline{X}} = \frac{\delta}{\sqrt{n}}}$ **standard deviation of the mean or standard error of the mean**

The role of $\delta_{\overline{X}}$ in statistics is fundamental, as it measures the extent to which the sample means can be expected to fluctuate, or vary, due to chance. If $\delta_{\overline{X}}$ is small, the chances are good that the mean of a sample will be close to the mean of the population; if $\delta_{\overline{X}}$ is large, we are more likely to get a sample mean which differs considerably from the mean of the population.

Problems:

1. When we sample an infinite population, what happens to the standard error of the mean if the sample size is (a) increased from 60 to 240; (b) decreased from 640 to 40?
2. If the standard deviation of the mean for the sampling distribution of random samples of size 36 from a large or infinite population is 2, how large must be the size of the sample if the standard deviation is to be reduced to 1.2?

Central Limit Theorem

For large samples, the sampling distribution of the mean can be approximated closely with a normal distribution.

Combining this theorem with $\mu_{\overline{X}} = \mu$ and $\delta_{\overline{X}} = \frac{\delta}{\sqrt{n}}$ for random samples from infinite populations, we find that $\overline{x}$ is the mean of a random sample of size *n* from an infinite population with mean μ and standard deviation δ and the sample size *n* is large, then

$$\boxed{z = \frac{\overline{x} - \mu}{\delta / \sqrt{n}}}$$

is a value of a random variable having approximately the standard normal distribution.

The central limit theorem is of fundamental importance in statistics because it justifies the use of normal-curve methods for a wide range of problems. This theorem applies automatically to **sampling from infinite populations or from finite populations whose sampling size *n* is large**.

Problems:

1. The amount of time that a drive-through bank teller spends on a customer is a random variable with mean μ = 3.2 minutes and a standard deviation δ=1.6 minutes. If a random sample of 64 customers is observed, find the probability that their mean time at the teller's counter is (a) at most 2.7 minutes; (b) more than 3.5 minutes; (c) at least 3.2 minutes but less than 3.4 minutes? Ans: 0.0062, 0.0668, 0.3413
2. The average life of a bread-making machine is 7 years with a standard deviation of 1 year. Assuming that the lives of these machines follow approximately a normal distribution, find (a) the probability that the mean life of a random sample of 9 such machines falls between 6.4 and 7.2 years; (b) the value of $\overline{x}$ to the right of which 15% of the means computed from random samples of size 9 would fall. Ans: 0.6898; 5.35 years
3. If a certain machine makes electrical resistors having a mean resistance of 40Ω and a standard deviation of 2Ω, what is the probability that a random sample of 36 of these resistors will have a combined resistance o f more than 1458Ω? Ans: 0.0668