Research Methods & Technical Writing

Lesson 7 - Week 7

Sampling Fundamentals

Lecturer: Dr. Msagha J Mbogholi, PhD

Flashback from Lesson 6

- Some of the challenges in data collection are data quality issues, inconsistent data, data downtime, ambiguous data, and duplicate data, among others.
- Primary data collection is the gathering of raw data collected at the source. It is a process of collecting the original data collected by a researcher for a specific research purpose. Primary data collection methods include observations, questionnaires, interviews, online forums, groups, online communities, face to face interviews, online, mail, and phone.
- Other primary data collection methods worth mentioning but which weren't covered in detail in this course include: Warranty cards, distributor or store audits, pantry audits, consumer panels, use of mechanical/electronic devices, projective techniques, depth interviews, and content analysis.
- Secondary data is a type of data that has already been published in books, newspapers, magazines, journals, online portals etc. There is an abundance of data available in these sources about your research area in business studies, almost regardless of the nature of the research area. Therefore, application of appropriate set of criteria to select secondary data to be used in the study plays an important role in terms of increasing the levels of research validity and reliability.

Content

- Introduction
- Definitions
- Sampling distributions
- Standard error
- Estimation
- Sample size determination

Part 1

Introduction

Introduction

- We are now in the second part of this course, so to speak. In this part we shall be describing the analysis of data, in a nutshell.
- What this means is that after collecting data we have to give it meaning. The meaning is in the context of the investigation we are carrying out (when we started off we were investigating a problem to find a solution, remember?)
- Giving data meaning is normally based on some kind of test. Let us illustrate this with a simple example.
- You wake up in the morning feeling ill; your stomach aches, feelings of dizziness, nausea, and so on. Naturally you decide to visit a hospital for a diagnosis. What happens there? Chances are the doctor will want to run some tests. In order to do so they will ask for a sample of your blood, urine and perhaps stool.
- How much blood they take in the lab will depend on the number of tests the doctor has asked to be done (the fear of needles doesn't help here, does it). Further the urine will be collected in a test tube, while the stool (if there's need for it) will be collected in a specimen container.

Introduction (cont'd)

- Notice the term specimen and sample? What do they mean?
- The sample is simply a small portion of all your blood; what this means is that what is drawn by the lab technician can be assumed to represent all the blood in your body. What would happen if they wanted to test all your blood, though; would they draw it all out of your body to test it? Of course, not; you would die if this were to happen.
- The same logic can be applied to the urine and stool; they represent all the urine and stool in your body. Thus simply put a sample or specimen is simply a selected part that represents the whole. With this in mind we can now understand the need for sampling for the purpose of research.
- In this lesson we discuss sampling fundamentals; what you should know before delving into data analysis and proving or disproving your hypothesis.
- The lesson begins by describing some terms used in research related to sampling (and these will appear often in coming lessons). Thereafter, a discussion on the different tests that are used in sampling, and also how to determine the correct sample size to use based on defined parameters.

Introduction (cont'd)

- You may be wondering (apart from our introductory example) why do we need to sample, at all? Kothari (2004) sheds light on this as follows:"
- 1. Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.
- 2. Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.
- 3. Sampling remains the only way when population contains infinitely many members.
- 4. Sampling remains the only choice when a test involves the destruction of the item under study.
- 5. Sampling usually enables to estimate the sampling errors and, thus, assists in obtaining information concerning some characteristic of the population."

Part 2

Definitions

Definitions

- As we delve into the world of sampling there are several new terms that are used which haven't been defined hitherto. It is important to understand the meaning of these terms from the word go, as they tend to confuse learners in their applications in this domain (and more so when used in formulae). This section seeks to define these terms:
- Universe/population: From a statistical point of view, the term 'Universe' refers to the total of the items or units in any field of inquiry, whereas the term 'population' refers to the total of items about which information is desired. (Kothari, 2004). Further, "In statistics, a population is an entire group about which some information is required to be ascertained. A statistical population need not consist only of people. We can have population of heights, weights, BMIs, hemoglobin levels, events, outcomes, so long as the population is well defined with explicit inclusion and exclusion criteria." (Banerjee & Chaudhury, 2010). The population may also be defined as being finite (the number of elements N, can be enumerated/counted) or infinite (the elements can't be counted; for example the number of oxygen molecules in a glass of water).

Definitions (cont'd)

- Sampling frame: A sampling frame is a list of all the items in your population. It's a complete list of everyone or everything you want to study. The difference between a population and a sampling frame is that the population is general and the frame is specific. For example, the population could be "People who live in Jacksonville, Florida." The frame would name *all* of those people, from Adrian Abba to Felicity Zappa. (Glen, n.d.). A frame should always be a good representation of the population.
- Sampling design: A sample design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or the procedure the researcher would adopt in selecting some sampling units from which inferences about the population is drawn. Sampling design is determined before any data are collected. (Kothari, 2004). (the different designs have been discussed in lesson 4 of this course).
- Statistics and parameters: Parameter implies a summary description of the characteristics of the target population. On the other extreme, the statistic is a summary value of a small group of population i.e. sample. (Surbhi S, 2017). To make this clearer let us pick an example or two: you randomly poll voters in an election. You find that 55% of the population plans to vote for candidate A. That is a **statistic**. Why? You only asked a sample—a small percentage— of the population who they are voting for. You calculated what the population was *likely* to do based on the sample. (Glen, n.d.-a).

Definitions (cont'd)

- Statistics and parameters (cont'd): further examples of statistics and parameters include (Glen, n.d.-a):
- Parameters:
- 10% of US senators voted for a particular measure. There are only 100 US Senators, you can count what every single one of them voted.
- 40% of 1,211 students at a particular elementary school got below a 3 on a standardized test. You know this because you have each and every students' test score.
- 33% of 120 workers at a particular bike factory were paid less than $20,000 per year. You have the payroll data for all of the workers.

Definitions (cont'd)

- Statistics:
- 60% of US residents agree with the latest health care proposal. It's not possible to actually ask hundreds of millions of people whether they agree. Researchers have to just take samples and calculate the rest.
- 45% of Jacksonville, Florida residents report that they have been to at least one Jaguars game. It's very doubtful that anyone polled in excess of a million people for this data. They took a sample, so they have a statistic.
- 30% of dog owners poop scoop after their dog. It's impossible to survey all dog owners—no one keeps an accurate track of exactly how many people own dogs. This data had to be from a sample, so it's a statistic.
- Table 1 shows the difference between statistics and parameters, including statistical notations for both.

Table 1. Statistics and parameters compared (Surbhi S, 2017)

BASIS FOR COMPARISON	STATISTIC	PARAMETER
Meaning	Statistic is a measure which describes a fraction of population.	Parameter refers to a measure which describes population.
Numerical value	Variable and Known	Fixed and Unknown
Statistical Notation	$\bar{x}$ = Sample Mean	μ = Population Mean
	s = Sample Standard Deviation	σ = Population Standard Deviation
	$\hat{p}$ = Sample Proportion	P = Population Proportion
	x = Data Elements	X = Data Elements
	n = Size of sample	N = Size of Population
	r = Correlation coefficient	ρ = Correlation coefficient

Definitions (cont'd)

- Sampling error: A sampling error occurs when the sample used in the study does not represent the entire population (Fleetwood, 2020). Kothari (2004) refers to this as simply inaccuracies that occur in the information collected, and is inevitable; he also refers to it as "error variance". These errors are normally random (in the case of random sampling). Kothari (2004) demonstrates this in fig 1;
- Sampling error = Frame error + chance error + response error. (If we add measurement error or the non-sampling error to sampling error, we get total error).
- Precision: Precision is the range within which the population average (or other parameter) will lie in accordance with the reliability specified in the confidence level as a percentage of the estimate ± or as a numerical quantity. For example, supposing the estimate is 3000 kgs and the precision required is ± 20%, then the true value will be between {3000 – (20%*3000) = 2400 kgs} and {3000 + (20%*3000) = 3600 kgs}; thus the true value will lie in this range.

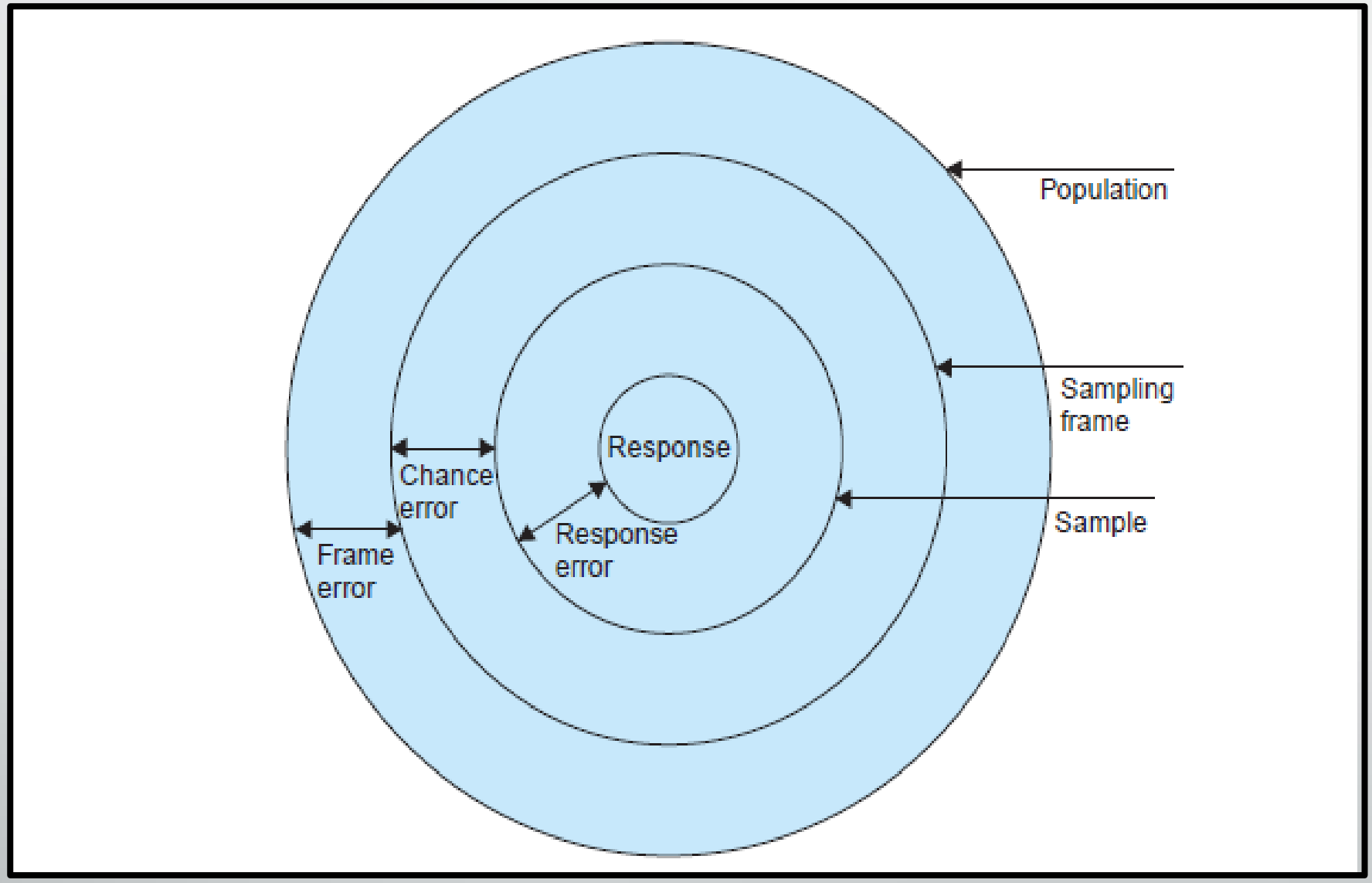

Fig 1. Sampling error (Kothari, 2004)

Definitions (cont'd):

- Confidence level and significance level: "The confidence level or reliability is the expected percentage of times that the actual value will fall within the stated precision limits. Thus, if we take a confidence level of 95%, then we mean that there are 95 chances in 100 (or .95 in 1) that the sample results represent the true condition of the population within a specified precision range against 5 chances in 100 (or .05 in 1) that it does not. Precision is the range within which the answer may vary and still be acceptable; confidence level indicates the likelihood that the answer will fall within that range, and the significance level indicates the likelihood that the answer will fall outside that range. We can always remember that if the confidence level is 95%, then the significance level will be (100 – 95) i.e., 5%; if the confidence level is 99%, the significance level is (100 – 99) i.e., 1%, and so on. We should also remember that the area of normal curve within precision limits for the specified confidence level constitute the acceptance region and the area of the curve outside these limits in either direction constitutes the rejection regions." (Kothari, 2004).

Definitions (cont'd)

- Sampling distribution: Sampling distribution is a statistic that determines the probability of an event based on data from a small group within a large population. Its primary purpose is to establish representative results of small samples of a comparatively larger population. Since the population is too large to analyze, you can select a smaller group and repeatedly sample or analyze them. The gathered data, or statistic, is used to calculate the likely occurrence, or probability, of an event (Burgin, 2023). For example if there is a population N, and we draw a sample of size n, then we will find that each sample will give a different value for statistics such as mean, standard deviation and so on. "All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the particular statistic, say mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure. It may be noted that each item in a sampling distribution is a particular statistic of a sample. The sampling distribution tends quite closer to the normal distribution if the number of samples is large. The significance of sampling distribution follows from the fact that the mean of a sampling distribution is the same as the mean of the universe. Thus, the mean of the sampling distribution can be taken as the mean of the universe." (Kothari, 2004)

Part 3

Sampling distributions

3.1 Sampling distribution of mean

- Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population. If samples are taken from a normal population, $N(\mu, \sigma_p)$ the sampling distribution of mean would also be normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation = $\sigma_p \sqrt{n}$, where μ is the mean of the population, σ_p is the standard deviation of the population and *n* means the number of items in a sample. But when sampling is from a population which is not normal (may be positively or negatively skewed), even then, as per the central limit theorem, the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large i.e., more than 30. In case we want to reduce the sampling distribution of mean to unit normal distribution i.e., $N(0,1)$, we can write the normal variate z. (Kothari, 2004)
- Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is. The formula for the z score is:
- **$z = (x - \mu) / \sigma$;** for example, let's say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be:
- $z = (x - \mu) / \sigma; = (190 - 150) / 25 = 1.6$. your score is 1.6 standard deviations *above* the mean. (Glen, n.d.-c).

3.2 Sampling distribution of proportion

- Often sampling is done in order to estimate the proportion of a population that has a specific characteristic, such as the proportion of all items coming off an assembly line that are defective or the proportion of all people entering a retail store who make a purchase before leaving. The population proportion is denoted p and the sample proportion is denoted $\hat{p}$.
- Thus if in reality 43% of people entering a store make a purchase before leaving, $p = 0.43$; if in a sample of 200 people entering the store, 78 make a purchase, $\hat{p} = 78/200 = 0.39$
- The sample proportion is a random variable: it varies from sample to sample in a way that cannot be predicted with certainty. Viewed as a random variable it will be written , $\hat{p}$. It has mean $\mu_{\hat{p}}$ and standard deviation $\sigma_{\hat{p}}$ (*The Sample Proportion*, n.d.).

3.2 Sampling distribution of proportion (cont'd)

- Suppose random samples of size *n* are drawn from a population in which the proportion with a characteristic of interest is *p*. The mean $\mu_{\hat{p}}$ and standard deviation $\sigma_{\hat{p}}$ of the sample proportion $\hat{p}$ satisfy:
- $\mu_{\hat{p}} = p$, and $\sigma_{\hat{p}} = \sqrt{p.q/n}$, where $q = (1 - p)$; p represents the proportion of successes while q represents the proportion of failures; thus in a normal distribution $q = (1 - p)$ always.
- Presuming the binomial distribution approximating the normal distribution for large *n*, the normal variate of the sampling distribution of proportion
- $z = (\hat{p} - p) / \sqrt{p.q/n}$, where $\hat{p}$ the sample proportion of successes, can be used for testing of hypotheses.

3.3 Student's t-distribution

- The t-distribution, also known as the Student's t-distribution, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails. It is used for estimating population parameters for small sample sizes or unknown variances. T-distributions have a greater chance for extreme values than normal distributions, and as a result have fatter tails.
- The t-distribution is a continuous probability distribution of the z-score when the estimated standard deviation is used in the denominator rather than the true standard deviation.
- T-tests are used in statistics to estimate significance. (Hayes, 2022)
- Kothari (2004) posits that a sample size of 30 is normally sufficiently large for most tests; this means that the t-distribution is used when the sample n is <= 30. The t variable is calculated as follows:

3.3 Student's t-distribution (cont'd)

$$t = \left(\bar{X} - \mu\right) \Big/ \left(\sigma_s / \sqrt{n}\right)$$

where $\sigma_s = \sqrt{\dfrac{\Sigma \left(X_i - \bar{X}\right)^2}{n} - 1}$

i.e., the sample standard deviation . *t*-distribution is also symmetrical and is very close to the distribution of standard normal variate, *z*, except for small values of *n*. The variable *t* differs from *z* in the sense that we use sample standard deviation (σ_s) in the calculation of *t*, whereas we use standard deviation of population (σ_p) in the calculation of *z*. There is a different *t* distribution for every possible sample size i.e., for different degrees of freedom. The degrees of freedom for a sample of size *n* is $n - 1$. As the sample size gets larger, the shape of the *t* distribution becomes apporximately equal to the normal distribution. In fact for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we can use the normal to approximate the *t*-distribution. But when *n* is small, the *t*-distribution is far from normal but when $n \rightarrow \alpha$, *t*-distribution is identical with normal distribution.

(Kothari, 2004)

3.4 F-distribution

- The F-distribution, also known as the Fisher-Snedecor distribution, is a continuous probability distribution that is often used in hypothesis testing and analysis of variance (ANOVA). It is typically used to compare the variability of two population samples or to determine whether two population variances are equal.
- The F-distribution is a right-skewed distribution with a minimum value of 0 and no maximum value. It is defined by two parameters, known as the degrees of freedom for the numerator (df1) and the degrees of freedom for the denominator (df2). The larger the degrees of freedom, the more the distribution resembles a normal distribution. (Gurus, 2022).
- The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (n-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-n), where n is the number of samples.

3.5 Chi-square distribution: χ^2

- A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between two categorical variables. (Biswal, 2023).
- Further a random variable has a Chi-square distribution if it can be written as a sum of squares of independent standard normal variables.

3.6 Central limit theorem

- The central limit theorem states that irrespective of a random variable's distribution if large enough samples are drawn from the population then the sampling distribution of the mean for that random variable will approximate a normal distribution. This fact holds true for samples that are greater than or equal to 30. In other words, as more large samples are taken, the graph of the sample means starts looking like a normal distribution.
- For example: A set of samples have been collected from a larger sample and the sample mean values are 12.8, 10.9, 11.4, 14.2, 12.5, 13.6, 15, 9, 12.6. Find the population mean.
- Solution: The population mean values are an average of the above sample mean values μ = 112/9 = 12.4; thus the population mean = 12.4 (*Central Limit Theorem - Definition, Formula, Examples*, n.d.)
- The significance of the central limit theorem lies in the fact that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample." (Kothari, 2004)

3.7 Sandler's A test

- This test was developed by Joseph Sandler and is based on a simplification of the t-test described earlier in this lesson.
- It is mostly used by psychologists use this test in case of two groups that are matched with respect to some extraneous variable(s). While using *A*-test, we work out *A*-statistic that yields exactly the same results as the student's t-test.
- The A-statistic is found as: (sum of squares of the differences)/ (the squares of the sum of the differences).; the number of degrees of freedom (df) applied is the same as for the t-test, i.e. d.f. = (n – 1), n being equal to the number of pairs.
- The critical value of *A*, at a given level of significance for given d.f., can be obtained from the table of *A*-statistic. One has to compare the computed value of *A* with its corresponding table value for drawing inference concerning acceptance or rejection of null hypothesis. If the calculated value of *A* is equal to or less than the table value, in that case *A*-statistic is considered significant where upon we reject *H*o and accept *Ha*. But if the calculated value of *A* is more than its table value, then *A*-statistic is taken as insignificant and accordingly we accept *H*o. This is so because the two test statistics viz., *t* and A are inversely related.

Part 4

Standard error

Standard error

- The standard deviation of sampling distribution of a statistic is known as its standard error (S.E). Kothari (2004) posits that "the utility of the concept of standard error in statistical induction arises on account of the following reasons:
- The standard error helps in testing whether the difference between observed and expected frequencies could arise due to chance. The criterion usually adopted is that if a difference is less than 3 times the S.E., the difference is supposed to exist as a matter of chance and if the difference is equal to or more than 3 times the S.E., chance fails to account for it, and we conclude the difference as significant difference.
- The standard error helps in testing whether the difference between observed and expected frequencies could arise due to chance. The criterion usually adopted is that if a difference is less than 3 times the S.E., the difference is supposed to exist as a matter of chance and if the difference is equal to or more than 3 times the S.E., chance fails to account for it, and we conclude the difference as significant difference…..(cont'd)

Standard error (cont'd)

- …(cont'd)
- The standard error enables us to specify the limits within which the parameters of the population are expected to lie with a specified degree of confidence. Such an interval is usually known as confidence interval. There are several formulae for computing the standard errors concerning various measures based on samples.." These shall not be discussed in detail in this lesson; however, their application shall be covered as we cover relevant content later in the course.
- However, the standard deviation of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size :
- $\sigma_{\bar{x}} = \sigma / \sqrt{n}$
- The standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of $\sqrt{n}$.

Part 5

Estimation

5.1 Introduction

- Before understanding how to calculate estimates in this part let us introduce a few terms and properties related to them. These terms are all explained by (Glen, 2016):"
- An estimator is a statistic that estimates some fact about the population. You can also think of an estimator as the rule that creates an estimate. For example, the sample mean($\bar{x}$) is an estimator for the population mean, μ.
- The quantity that is being estimated (i.e. the one you want to know) is called the **estimand**. For example, let's say you wanted to know the average height of children in a certain school with a population of 1000 students. You take a sample of 30 children, measure them and find that the mean height is 56 inches. This is your sample mean, the **estimator**. You use the sample mean to **estimate** that the population mean (your **estimand**) is about 56 inches.
- Estimators can be a **range of values** (like a confidence interval) or a **single value** (like the standard deviation). When an estimator is a range of values, it's called an **interval estimate**. For the height example above, you might add on a confidence interval of a couple of inches either way, say 54 to 58 inches. When it is a single value — like 56 inches — it's called a **point estimate."**

5.1 Introduction (cont'd)

- Estimators can be described in the following ways (Glen, 2016):"
- **Biased**: a statistic that is either an overestimate or an underestimate.
- **Efficient**: a statistic with small variances (the one with the smallest possible variance is also called the "best"). *Inefficient* estimators can give you good results as well, but they usually requires much larger samples.
- **Invariant**: statistics that are not easily changed by transformations, like simple data shifts.
- **Shrinkage**: a raw estimate that's improved by combining it with other information. See also: The James-Stein estimator.
- **Sufficient**: a statistic that estimates the population parameter as well as if you knew all of the data in all possible samples.
- **Unbiased**: an accurate statistic that neither underestimates nor overestimates."
- Consequently, Kothari (2004) posits that a good estimator should be unbiased, efficient, sufficient and consistent (should approach the value of population parameter as the sample size becomes larger and larger)

5.2 Estimating population mean (μ)

- The sample mean $\bar{x}$ is the best estimator of the population mean, μ , and its sampling distribution, so long as the sample is sufficiently large, approximates the normal distribution. If we know the sampling distribution of $\bar{x}$ we can make statements about any estimate that we may make from the sampling information. As an example suppose we wish to find the average height of soccer players; we take a sample of 40 players and we find the average (arithmetic mean) is 1.85 meters; that is to say, $\bar{x} = 1.85$; we repeat the exercise several times drawing different random samples. We get different values of $\bar{x}$, say, 1.79, 1.90, 1.82, 1.80, and so on. Each of these means is a separate point estimate of the population.
- This is a characteristic of a distribution of sample means (and also of other sample statistics). Even if the population is not normal, the sample means drawn from that population are dispersed around the parameter in a distribution that is generally close to normal; the mean of the distribution of sample means is equal to the population mean. (Kothari, 2004)

5.2 Estimating population mean (μ)

- The relationship between the dispersion of a population distribution and that of the sample mean can be stated as under:

$$\sigma_{\bar{X}} = \frac{\sigma_p}{\sqrt{n}}$$

where $\sigma_{\bar{X}}$ = standard error of mean of a given sample size

σ_p = standard deviation of the population

n = size of the sample.

- The best estimate for the standard deviation of the population is the standard deviation of the sample, and this is what we use. We can now swap the two in the formula and get:

$$\sigma_{\bar{X}} = \frac{\sigma_s}{\sqrt{n}}$$

where

$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}}$$

5.2 Estimating population mean (μ)

- Consider the following example drawn from Kothari (2004):
- Suppose we take one sample of 36 items and work out its mean to be equal to 6.20 and its standard deviation to be equal to 3.8, then the best point estimate of population mean is 6.20. The standard error of mean would be $3.8 / \sqrt{36} = 3.8 / 6 = 0.663$. If we take the interval estimate of μ to be
- $\bar{x} \pm 1.96 (\sigma_{\bar{x}}) \Rightarrow 6.20 \pm 1.24$ or from 4.96 to 7.44, it means that there is a 95 per cent chance that the population mean is within 4.96 to 7.44 interval.
- Let us work out one simple example to demonstrate this concept, also adapted from Kothari (2004):
- From a random sample of 36 Malindi civil service personnel, the mean age and the sample standard deviation were found to be 40 years and 4.5 years respectively. Construct a 95 per cent confidence interval for the mean age of civil servants in Malindi.

5.2 Estimating population mean (μ)

- Solution: we see that $n = 36$, $\bar{x} = 40$ years, and $\sigma_s = 4\ 5.$ years;
- Further, the standard variate, z, for 95 per cent confidence is 1.96 (as per the normal curve area table). Thus, 95 per cent confidence interval for the mean age of population is:

$$\bar{X} \pm z \frac{\sigma_s}{\sqrt{n}}$$

or $$40 \pm 1.96 \frac{4.5}{\sqrt{36}}$$

or $$40 \pm (1.96)\ (0.75)$$

or $$40 \pm 1.47 \text{ years}$$

5.3 Estimating population proportion ($\hat{p}$)

- In the real world, you usually don't know facts about the entire population and so you use sample data to estimate p. This sample proportion is written as $\hat{p}$, pronounced *p-hat*. It's calculated in the same way, except you use data from a sample: just divide the total number of items in the sample by the number of items you're interested in.
- For example, in a survey of 3121 people, 412 are under-vaccinated. What is the proportion of under-vaccinated people in the local population?
- Answer: You don't know population data for the local area, so use the sample data:
- $\hat{p} = x / n$
  $= 412/3121$
  $= 0.132$
- (Glen, 2016a)

5.3 Estimating population proportion ($\hat{p}$)

- Kothari (2004) also describes a different approach to estimating population proportion:"
- … if we take a random sample of 50 items and find that 10 per cent of these are defective i.e., p = .10, we can use this sample proportion (p = .10) as best estimator of the population proportion ($\hat{p}$ = p = .10). In case we want to construct confidence interval to estimate a population proportion, we should use a normal distribution (ideally, as sample size increases the distribution, even if initially binomial, will always tend to a normal distribution).
- The mean of the sampling distribution of the proportion of successes (μ_p) is taken as equal to p and the standard deviation for the proportion of successes, also known as the standard error of proportion, is taken as equal to $\sqrt{pq}/ n$. But when population proportion is unknown, then we can estimate the population parameters by substituting the corresponding sample statistics p and q in the formula for the standard error of proportion to obtain the estimated standard error of the proportion as shown below:

5.3 Estimating population proportion ($\hat{p}$)

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

Using the above estimated standard error of proportion, we can work out the confidence interval for population proportion thus:

$$p \pm z \cdot \sqrt{\frac{pq}{n}}$$

where

p = sample proportion of successes;

$q = 1 - p$;

n = number of trials (size of the sample);

z = standard variate for given confidence level (as per normal curve area table).

5.3 Estimating population proportion ($\hat{p}$)

- Let us demonstrate this concept using a simple example, also from Kothari (2004):
- A market research survey in which 64 consumers were contacted states that 64 per cent of all consumers of a certain product were motivated by the product's advertising. Find the confidence limits for the proportion of consumers motivated by advertising in the population, given a confidence level equal to 0.95.
- The solution is provided in the next slide:

5.3 Estimating population proportion ($\hat{p}$)

- $n = 64$, $p = 64\%$ or $.64$, $q = 1 - p = 1 - .64 = .36$
- and the standard variate (z) for 95 per cent confidence is 1.96 (as per the normal curve area table). Thus, 95 per cent confidence interval for the proportion of consumers motivated by advertising in the population is:

$$p \pm z \cdot \sqrt{\frac{pq}{n}}$$

$$= .64 \pm 1.96 \sqrt{\frac{(0.64)(0.36)}{64}}$$

$$= .64 \pm (1.96)(.06)$$

$$= .64 \pm .1176$$

Thus, lower confidence limit is 52.24%

upper confidence limit is 75.76%

Table 2(a). Estimation formulae (Kothari, 2004)

	In case of infinite population	*In case of finite population**
Estimating population mean (μ) when we know σ_p	$\bar{X} \pm z \cdot \frac{\sigma_p}{\sqrt{n}}$	$\bar{X} \pm z \cdot \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$
Estimating population mean (μ) when we do not know σ_p	$\bar{X} \pm z \cdot \frac{\sigma_s}{\sqrt{n}}$	$\bar{X} \pm z \cdot \frac{\sigma_s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$
	In case of infinite population	*In case of finite population**

Table 2(b). Estimation formulae (Kothari, 2004)

and use σ_s as the best estimate of σ_p and sample is large (i.e., $n > 30$)		
Estimating population mean (μ) when we do not know σ_p and use σ_s as the best estimate of σ_p and sample is small (i.e., $n \leq 30$)	$\bar{X} \pm t \cdot \frac{\sigma_s}{\sqrt{n}}$	$\bar{X} \pm t \cdot \frac{\sigma_s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$
Estimating the population proportion $(\hat{p})$ when p is not known but the sample is large.	$p \pm z \cdot \sqrt{\frac{pq}{n}}$	$p \pm z \cdot \sqrt{\frac{pq}{n}} \times \sqrt{\frac{N-n}{N-1}}$

* In case of finite population, the standard error has to be multiplied by the finite population multiplier viz., $\sqrt{(N-n)/(N-1)}$.

5.3 Estimating population proportion ($\hat{p}$)

- Table 2 (a and b) present a summary of the important formulae that are used in statistics regarding estimation.
- The table is important due to the fact that it differentiates instances where the population is finite, and when it is infinite.
- In most literature the assumption is that the population is finite, meaning the value of N is always known; thus the learner is usually at a loss as to how to proceed when confronted with situations where the population is infinite.
- Hopefully table 2 solves this dilemma.

Part 6

Sample size determination

6.1 Introduction

- One of the major dilemmas that a novice researcher is faced with is the sample size (n). How big or how small should it be? If it is too small then it might not be representative of the population; too big and it will be mostly a waste of resources. Kothari (2004) advises researchers to keep in mind the following when determining sample size:
- Nature of the universe - Universe may be either homogenous or heterogeneous in nature. If the items of the universe are homogenous, a small sample can serve the purpose. But if the items are heterogeneous, a large sample would be required. Technically, this can be termed as the dispersion factor.
- *Number of classes proposed:* If many class-groups (groups and sub-groups) are to be formed, a large sample would be required because a small sample might not be able to give a reasonable number of items in each class-group.
- *Nature of study:* If items are to be intensively and continuously studied, the sample should be small. For a general survey the size of the sample should be large, but a small sample is considered appropriate in technical surveys.
- *Type of sampling:* Sampling technique plays an important part in determining the size of the sample. A small random sample is apt to be much superior to a larger but badly selected sample.

6.1 Introduction (cont'd)

- …(cont'd)
- *Standard of accuracy and acceptable confidence level:* If the standard of accuracy or the level of precision is to be kept high, we shall require relatively larger sample. For doubling the accuracy for a fixed significance level, the sample size has to be increased fourfold.
- *Availability of finance:* In practice, size of the sample depends upon the amount of money available for the study purposes. This factor should be kept in view while determining the size of sample for large samples result in increasing the cost of sampling estimates.
- *Other considerations:* Nature of units, size of the population, size of questionnaire, availability of trained investigators, the conditions under which the sample is being conducted, the time available for completion of the study are a few other considerations to which a researcher must pay attention while selecting the size of the sample.
- There are generally two ways by which we can determine the size of the sample: by using standard deviation together with confidence level, and based on Bayesian statistics.

6.2 Using standard deviation and confidence level

- This is done using Andrew Fisher's formula. Kibuacha (2021) describes the 6 steps as follows:
- Determine the population size (if known).
- Determine the confidence interval.
- Determine the confidence level.
- Determine the standard deviation (a standard deviation of 0.5 is a safe choice where the figure is unknown)
- Convert the confidence level into a Z-Score. This table for z scores (versus confidence levels) is available in most literature therefore we shall not share it here.
- Put these figures into the sample size formula to get your sample size. The formula is as follows:

GeoPoll

$$\text{Sample Size} = \frac{(\text{Z-score})^2 \times \text{StdDev} \times (1-\text{StdDev})}{(\text{confidence interval})^2}$$

6.2 Using standard deviation and confidence level

- Let us demonstrate this using an example from Kibuacha (2021) :
- Say you choose to work with a 95% confidence level, a standard deviation of 0.5, and a confidence interval (margin of error) of ± 5%, you just need to substitute the values in the formula:
- ***((1.96)^2 x 0.5(1 - 0.5)) / (.05)^2***
- ***= (3.8416 x .25) / .0025***
- ***= .9604 / .0025***
- ***= 384.16***
- Your sample size should be 385.
- Fortunately there are many online calculators that you can use to calculate sample size (how cool is that?) such as the one at easycalculation.com

6.3 Using Bayesian approach

- The Bayesian approach is an advanced approach and we shall only describe the steps to follow as shared by Kothari (2004):
- (i) Find the expected value of the sample information (EVSI)* for every possible *n*;
- (ii) Also workout reasonably approximated cost of taking a sample of every possible *n*;
- (iii) Compare the EVSI and the cost of the sample for every possible *n*. In other words, workout the expected net gain (ENG) for every possible *n* as stated below:
- For a given sample size (*n*): (EVSI) – (Cost of sample) = (ENG)
- (iv) From (iii) above the optimal sample size, that value of *n* which maximizes the difference between the EVSI and the cost of the sample, can be determined.
- The learner can find more literature regarding the Bayesian approach from the following links:
- https://www.journals.uchicago.edu/doi/full/10.1086/693433
- https://www.jstor.org/stable/2349003

https://doi.org/10.3390/ijerph192114245

Summary (1 of 2)

- Parameter implies a summary description of the characteristics of the target population. On the other extreme, the statistic is a summary value of a small group of population i.e. sample.
- Precision is the range within which the population average (or other parameter) will lie in accordance with the reliability specified in the confidence level as a percentage of the estimate ± or as a numerical quantity.
- The confidence level or reliability is the expected percentage of times that the actual value will fall within the stated precision limits.
- Sampling distribution is a statistic that determines the probability of an event based on data from a small group within a large population: there is sampling distribution of the mean, sampling distribution of proportion, student's t-distribution, F distribution, and chi-square distribution. There is also Sandler's A-test which borrows from t-distribution.

Summary (2 of 2)

- The central limit theorem states that irrespective of a random variable's distribution if large enough samples are drawn from the population then the sampling distribution of the mean for that random variable will approximate a normal distribution. This fact holds true for samples that are greater than or equal to 30.
- The standard deviation of sampling distribution of a statistic is known as its standard error (S.E).
- The sample mean $\bar{x}$ is the best estimator of the population mean, μ , and its sampling distribution, so long as the sample is sufficiently large, approximates the normal distribution.
- There are generally two ways by which we can determine the size of the sample: by using standard deviation together with confidence level, and based on Bayesian statistics

References

- Biswal, A. (2023, February 17). *What is a Chi-Square Test? Formula, Examples & Uses | Simplilearn*. Simplilearn.com. https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test
- Burgin, E. (2023, March 10). *Sampling Distribution: Definition, Factors and Types*. Indeed.com; Indeed. https://www.indeed.com/career-advice/career-development/what-is-sampling-distribution#:~:text=Understanding%20sampling%20distribution&text=Each%20random%20sample%20selected%20may,fall%20somewhere%20along%20the%20graph.
- *Central Limit Theorem - Definition, Formula, Examples*. (n.d.). Cuemath. Retrieved May 13, 2023, from https://www.cuemath.com/data/central-limit-theorem/
- Glen, S. (n.d.-a). *Difference Between a Statistic and a Parameter*. Statistics How To. https://www.statisticshowto.com/statistics-basics/how-to-tell-the-difference-between-a-statistic-and-a-parameter/

References

- Glen, S. (n.d.-b). *Sampling Frame: Definition, Examples" From StatisticsHowTo.com: Elementary Statistics for the rest of us!* StatisticsHowTo.com. Retrieved April 10, 2023, from https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/sampling-frame/
- Glen, S. (n.d.-c). *Z-Score: Definition, Formula and Calculation*. Statistics How To. Retrieved May 13, 2023, from https://www.statisticshowto.com/probability-and-statistics/z-score/
- Glen, S. (2016a, May 5). *Population Proportion*. Statistics How To. https://www.statisticshowto.com/population-proportion/
- Glen, S. (2016b, September 8). *Estimator: Simple Definition and Examples*. Statistics How To. https://www.statisticshowto.com/estimator/

References

- Hayes, A. (2022, October 24). *What Is T-Distribution in Probability? How Do You Use It?* Investopedia. https://www.investopedia.com/terms/t/tdistribution.asp#:~:text=The%20t%2Ddistribution%2C%20also%20known
- Kibuacha, F. (2021, April 7). *How to determine sample size for a research study*. GeoPoll; GeoPoll. https://www.geopoll.com/blog/sample-size-research/
- Kothari, C. R. (2004). *Research methodology : methods & techniques* (2nd ed.). New Age International (P) Ltd., Publishers, Cop.
- Surbhi S. (2017, September 1). *Difference Between Statistic and Parameter (with Comparison Chart and Illustration) - Key Differences*. Key Differences. https://keydifferences.com/difference-between-statistic-and-parameter.html
- *The Sample Proportion*. (n.d.). Saylordotorg.github.io. Retrieved May 13, 2023, from https://saylordotorg.github.io/text_introductory-statistics/s10-03-the-sample-proportion.html