ERRORS AND THEIR TYPES

(BIOSTATISTICS)

What is error?

- **Error (statistical error) describes the difference between a value obtained from a data collection process and the 'true' value for the population.**
- The greater the error, the less representative the data are of the population.

Why does error matter?

- **The greater the error, the less reliable are the results of the study.**
- A credible data source will have measures in place throughout the data collection process to minimise the amount of error, and will also be transparent about the size of the expected error so that users can decide whether the data are 'fit for purpose'.

Data can be affected by two types of error:

- Sampling Error
- Non-sampling Error

SAMPLING ERROR

- **Sampling error occurs solely as a result of using a sample from a population, rather than conducting a census (complete enumeration) of the population.**
- It refers to the difference between an estimate for a population based on data from a sample and the 'true' value for that population which would result if a census were taken.
- Sampling errors do not occur in a census, as the census values are based on the entire population.
- Sampling error can be measured and controlled in random samples where each unit has a chance of selection, and that chance can be calculated.
- In general, increasing the sample size will reduce the sample error.

Sampling error can occur when:

- The proportions of different characteristics within the sample are not similar to the proportions of the characteristics for the whole population (i.E. If we are taking a sample of men and women and we know that 51% of the total population are women and 49% are men, then we should aim to have similar proportions in our sample);
- The sample is too small to accurately represent the population; and
- The sampling method is not random.

NON-SAMPLING ERROR

- **Non-sampling error is caused by factors other than those related to sample selection.**
- It refers to the presence of any factor, whether systemic or random, that results in the data values not accurately reflecting the 'true' value for the population.
- Non-sampling error can occur at any stage of a census or sample study, and are not easily identified or quantified.

Non-sampling Error Can Include :

- **Coverage error:** this occurs when a unit in the sample is incorrectly excluded or included, or is duplicated in the sample (e.g. a field interviewer fails to interview a selected household or some people in a household).
- **Non-response error:** this refers to the failure to obtain a response from some unit because of absence, non-contact, refusal, or some other reason. Non-response can be complete non-response (i.e. no data has been obtained at all from a selected unit) or partial non-response (i.e. the answers to some questions have not been provided by a selected unit).
- **Response error:** this refers to a type of error caused by respondents intentionally or accidentally providing inaccurate responses. This occurs when concepts, questions or instructions are not clearly understood by the respondent; when there are high levels of respondent burden and memory recall required; and because some questions can result in a tendency to answer in a socially desirable way (giving a response which they feel is more acceptable rather than being an accurate response).
- **Interviewer error:** this occurs when interviewers incorrectly record information; are not neutral or objective; influence the respondent to answer in a particular way; or assume responses based on appearance or other characteristics.
- **Processing error:** this refers to errors that occur in the process of data collection, data entry, coding, editing and output.

Why do we measure error?

- Error is expected in a data collection process, particularly if the data is obtained from a sample survey. Although non-sampling error is difficult to measure, sampling error can be measured to give an indication of the accuracy of any estimate value for the population. This assists users to make informed decisions about whether the statistics are suited to their needs.

How do we measure error?

- Two common measures of error are: standard error and the relative standard error.

- **Standard Error (SE) is a measure of the variation between any estimated population value that is based on a sample rather than true value for the population.**
- SE of any estimate for a measure of average magnitude of the difference between sample estimate and population parameters taken over the all sample estimate from the population.
- It is important to consider the Standard Error as it affects the accuracy of the estimates and, therefore, the importance that can be placed on the interpretations drawn from the data.

- SE is applied for std. deviation of sampling distribution of any estimate
- The standard error of the mean (SEM) can be expressed as:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

s is the standard deviation of the population.

n is the size (number of observations) of the sample.

- **Relative Standard Error (RSE) is the standard error expressed as a proportion of an estimated value.** It is usually displayed as a percentage. RSEs are a useful measure as they provide an indication of the relative size of the error likely to have occurred due to sampling. A high RSE indicates less confidence that an estimated value is close to the true population value.

Standard Error v/s Relative Standard Error

- The **Standard Error measure indicates the extent to which a survey estimate is likely to deviate from the true population** and is **expressed as a number.**
- The **Relative Standard Error (RSE) is the standard error expressed as a fraction of the estimate** and is usually **expressed as a percentage.**
- Estimates with a RSE of 25% or greater are subject to high sampling error and should be used with caution.

PROBABLE ERROR

- In statistics, **probable error** defines the half-range of an interval about a central point for the distribution, such that half of the values from the distribution will lie within the interval and half outside.
- Measure of the error of estimate for a sample from a normal distribution, it is computed by multiplying the standard error with 0.6745
- Thus for a symmetric distribution, it is equivalent to half the interquartile range, or the median Absolute deviation.

$$PE = 0.67449\ (SE)$$

PROBABLE ERROR OF COEFFICIENT OF CORRELATION

- It is an measure of testing reliability of an observed value of coefficient of correlation. it depends on the condition of random sampling
- It is represented by "r"

$$P.E.r = 0.6745\ (1-r^2)/\sqrt{n}$$

r = coefficient of correlation.
n = number of pairs of observation.

What can measures of error tell us?

- The standard error can be used to construct a confidence interval.
  A confidence interval is a range in which it is estimated the true population value lies.
- Confidence intervals of different sizes can be created to represent different levels of confidence that the true population value will lie within a particular range.
- A common confidence interval used in statistics is the 95% confidence interval. In a 'normal distribution', the 95% confidence interval is measured by two standard errors either side of the estimate.

SIGNIFICANCE OF PROBABLE ERROR

- Can be used of determining limits within which coefficient of correlation of population is expected to be located
- It is used to test if an observed value of sample correlation coefficient is significant of any correlation in population
- If $r < PE$, then correlation=insignificant
- If $r > 6PE$ then r= significant
- If $r < 6PE$ then sample size is too small for any estimation

Type I And Type II Errors

- In statistical hypothesis testing, a **type I error** is the incorrect rejection of a true null hypothesis (H_0) (also known as a "false positive" finding), while a **type II error** is incorrectly retaining a false null hypothesis (also known as a "false negative" finding).
- More simply stated, a type I error is to falsely infer the existence of something that is not there, while a type II error is to falsely infer the absence of something that is.

- A **type I error** (or **error of the first kind**) is the incorrect rejection of a true null hypothesis.
- Usually a type I error leads one to conclude that a supposed effect or relationship exists when in fact it doesn't.
- (H_0)=true but is rejected
- Let the probability of making type I error by rejecting H_0 = a
- Then probability of accepting $H_{0 = 1-a}$
- **Examples of type I errors-**

a. a test that shows a patient to have a disease when in fact the patient does not have the disease,

b. a fire alarm going on indicating a fire when in fact there is no fire, or

c. an experiment indicating that a medical treatment should cure a disease when in fact it does not.

- A **type II error** (or **error of the second kind**) is the failure to reject a false null hypothesis.
- Similarly, probability of making type II error= b
- **Examples of type II errors –**

a. a blood test failing to detect the disease it was designed to detect, in a patient who really has the disease;

b. a fire breaking out and the fire alarm does not ring; or

c. a clinical trial of a medical treatment failing to show that the treatment works when really it does

LEVEL OF SIGNIFICANCE

- Statistical tests fix the probability of committing type I error at certain level, called the level of significance.
- If the calculative probability is less than LOS, then null hypothesis is rejected or accepted otherwise
- 2 commonly used LOS are-
- 1% LOS and 5% LOS
- Simply, LOS means chances of making error
- If we chose 5% LOS , it implies that 5 out of 100 we are likely to reject the correct H_0
- Example: if a=0.05 the probability of making error is 5% and when a=0.01 the probability of making error is 1%


THANK
YOU