Unit 4: Statistics for Modelling and Simulation

Contents

1.0 Introduction

In this unit we will discuss two ways statistics are computed and applied in modelling and simulations these include: inference and descriptive processes. Statistical inference is generally distinguished from descriptive statistics. In simple terms, descriptive statistics can be thought of as being just a straightforward presentation of facts, in which modelling decisions made by a data analyst have had minimal influence. Statistical inference is the process of drawing conclusions from data that are subject to random variation, for example, observational errors or sampling variation. A complete statistical analysis will nearly always include both descriptive statistics and statistical inference, and will often progress in a series of steps where the emphasis moves gradually from description to inference.

2.0 Intended Learning Outcomes (ILOs)

By the end of this unit you should be able to:

- Differentiate between Descriptive and Inference statistics
- Describe the features of descriptive statistics
- Describe features of Inference statistics
- Compute the essential statistics for simulation

3.0 Main Content

3.1 Descriptive and Inference statistics

Descriptive statistics describe the main features of a collection of data quantitatively. Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to summarize a data set quantitatively without employing a probabilistic formulation, rather than use the data to make inferences about the population that the data are thought to represent. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented. For example in a paper reporting on a study involving human subjects, there typically appears a table giving the overall sample size, sample sizes in important subgroups (e.g., for each treatment or exposure group), and demographic or clinical characteristics such as the average age, the proportion of subjects of each sex, and the proportion of subjects with related co morbidities.

Inferential statistics tries to make inferences about a population from the sample data. We also use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one, or that it might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

3.1 Descriptive Statistics

Descriptive statistics provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of quantitative analysis of data. Descriptive statistics summarize data. For example, the shooting percentage in basketball is a descriptive statistics that summarizes the performance of a player or a team. The percentage is the number of shots made divided by the number of shots taken. A player who shoots 33% is making approximately one shot in every three. One making 25% is hitting once in four. The percentage summarizes or describes multiple discrete events. Or, consider the score of many students, the grade point average. This single number describes the general performance of a student across the range of their course experiences.

One that describes a large set of observations with a single indicator risks distorting the original data or losing important detail. For example, the shooting percentage doesn't tell you whether the shots are three-pointers or lay-ups, and GPA doesn't tell you whether the student was in difficult or easy courses. Despite these limitations, descriptive statistics provide a powerful summary that may enable comparisons across people or other units.

3.1.1 Univariate Analysis

Univariate analysis involves the examination across cases of a single variable, focusing on three characteristics: the distribution; the central tendency; and the dispersion. It is common to compute all three for each study variable.

a. Distribution

The distribution is a summary of the frequency of individual or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of cases that had that value. For instance, computing the distribution of gender in the study population means computing the percentages that are male and female. The gender variable has only two, making it possible and meaningful to list each one. However, this does not work for a variable such as income that has many possible values. Typically, specific values are not particularly meaningful (income of 50,000 is typically not meaningfully different from 51,000). Grouping the raw scores using ranges of values reduces the number of categories to something more meaningful. For instance, we might group incomes into ranges of 0-10,000, 10,001-30,000, etc.

b. Central tendency

The central tendency of a distribution locates the "center" of a distribution of values. The three major types of estimates of central tendency are the *mean*, the *median*, and the *mode*.

The **mean** is the most commonly used method of describing central tendency. To compute the mean, take the sum of the values and divide by the count. For example, the mean quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 36, 15, 25, 15

The sum of these 7 values is 147, so the mean is 147/7 =21.

The mean is computed using the formular: □ X_i / n, where the sum is over i = 1 to n.

The **median** is the score found at the middle of the set of values, i.e., that has as many cases with a larger value as have a smaller value. One way to compute the median is to sort the values in numerical order, and then locate the value in the middle of the list. For example, if there are 500 values, the value in 250th position is the median. Sorting the 8 scores above produces:

15, 15, 15, 20, 21, 25, 36

There are 7 scores and score #4 represents the halfway point. The median is 20. If there is an even number of observations, then the median is the mean of the two middle scores. In the example, if there were an 8th observation, with a value of 25, the median becomes the average of the 4th and 5th scores, in this case 20.5:

15, 15, 15, 20, 21, 25, 25, 36

The **mode** is the most frequently occurring value in the set. To determine the mode, compute the distribution as above. The mode is the value with the greatest frequency. In the example, the modal value 15, occurs three times. In some distributions there is a "tie" for the highest frequency, i.e., there are multiple modal values. These are called **multi- modal** distributions.

Notice that the three measures typically produce different results. The term "average" obscures the difference between them and is better avoided. The three values are equal if the distribution is perfectly "**normal**" (i.e., bell-shaped).

c. Dispersion

Dispersion is the spread of values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 − 15 = 21.

The **standard deviation** is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values). The standard deviation shows the relation that set of scores has to the mean of the sample. Again let's take the set of scores:
15, 20, 21, 36, 15, 25, 15
to compute the standard deviation, we first find the distance between each value and the mean. We know from above that the mean is 21. So, the differences from the mean are:
15 − 21 = −6
20 − 21 = −1
21 − 21 = 0
36 − 21 = 15
15 − 21 = −6
25 − 21 = +4
15 − 21 = −6

Notice that values that are below the mean have negative differences and values above it have positive ones. Next, we square each difference:
$(6)^2 = 36$
$(-1)^2 = 1$
$(+0)^2 = 0$
$(15)^2 = 225$
$(-6)^2 = 36$
$(+4)^2 = 16$
$(-6)^2 = 36$
Now, we take these "squares" and sum them to get the **sum of squares** (SS) value. Here, the sum is 350. Next, we divide this sum by the number of scores minus 1. Here, the result is 350 / 6 = 58.3. This value is known as the **variance**. To get the standard deviation, we take the square root of the variance (remember that we squared the deviations earlier). This would be √58.3 = 7.63.

Although this computation may seem intricate, it's actually quite simple. In English, we can describe the standard deviation as:
the square root of the sum of the squared deviations from the mean divided by the number of scores minus one given as: $\sqrt{}(\square\ (x_i – u)^2)/\ n$; where x = observed value and u = the mean

The standard deviation allows us to reach some conclusions about specific scores in our distribution. Assuming that the distribution of scores is close to "normal", the following conclusions can be reached:

a. approximately 68% of the scores in the sample fall within one standard deviation of the mean (u - SD) and (u + SD)
b. approximately 95% of the scores in the sample fall within two standard deviations of the mean (u-2SD) and (u+2SD)
c. approximately 99% of the scores in the sample fall within three standard deviations of the mean (u - 3SD) and (u + 3SD)

For example, since the mean in our example is 21 and the standard deviation is 7.63, we can from the above statement estimate that approximately 95% of the scores will fall in the range of 21 − (2×7.63) to 21 + (2×7.63) or between 5.74 and 36.26. Values beyond two standard deviations from the mean can be considered "outliers". 36 is the only such value in our distribution.

Outliers help identify observations for further analysis or possible problems in the observations. Standard deviations also convert measures on very different scales, such as height and weight, into values that can be compared.

d. Other Statistics

In research involving comparisons between groups, emphasis is often placed on the **significance level** for the **hypothesis** that the groups being compared differ to a degree greater than would be expected by chance. This significance level is often represented as a **p-value**, or sometimes as the standard score of a test statistic. In contrast, an **effect size** conveys the estimated magnitude and direction of the difference between groups, without regard to whether the difference is statistically significant. Reporting significance levels without effect sizes is problematic, since for large sample sizes even small effects of little practical importance can be statistically significant.

3.1.2 Examples of descriptive statistics

Most statistics can be used either as a descriptive statistic, or in an inductive analysis. For example, we can report the average reading test score for the students in each classroom in a school, to give a descriptive sense of the typical scores and their variation. If we perform a formal *hypothesis test* on the scores, we are doing *inductive* rather than descriptive analysis.

The following is a list some statistical common in descriptive analyses:

- Measures of central tendency
- Measures of dispersion
- Measures of association
- Cross-tabulation, contingency table
- Histogram
- Quantile, Q-Q plot
- Scatter plot
- Box plot

3.3 Inference Statistics

The terms **statistical inference**, **statistical induction** and **inferential statistics** are used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems affected by random variation. Initial requirements of such a system of procedures for *inference* and *induction* are that the system should produce reasonable answers when applied to well-defined situations and that it should be general enough to be applied across a range of situations.

The outcome of statistical inference may be an answer to the question "what should be done next?", where this might be a decision about making further experiments or surveys, or about drawing a conclusion before implementing some organizational or governmental policy.

For the most part, statistical inference makes propositions about populations, using data drawn from the population of interest via some form of random sampling. More generally, data about a random process is obtained from its observed behaviour during a finite period of time. Given a parameter or hypothesis about which one wishes to make inference, statistical inference most often uses:

- a statistical model of the random process that is supposed to generate the data, and
- a particular realization of the random process; i.e., a set of data.

The conclusion of a **statistical inference** is a statistical **proposition**.

3.3.1 Some common forms of statistical proposition

- An **estimate** - a particular value that best approximates some parameter of interest,
- A **confidence interval** (or set estimate) - an interval constructed from the data in such a way that, under repeated sampling of datasets, such intervals would contain the true parameter value with the probability at the stated confidence level,
- A **credible interval** - a set of values containing, for example, 95% of posterior belief,
- Rejection of an **hypothesis**
- **Clustering** or classification of data points into groups

3.3.2 Models/Assumptions

Any statistical inference requires some assumptions. A **statistical model** is a set of assumptions concerning the generation of the observed data and similar data. Descriptions of statistical models usually emphasize the role of population quantities of interest, about which we wish to draw inference.

3.3.3 Degree of models/assumptions

Statisticians distinguish between three levels of modelling assumptions;

□ **Fully parametric**: The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters. For example, one may assume that the distribution of population values is truly Normal, with unknown mean and variance, and that datasets are generated by 'simple' random sampling. The family of *generalized linear models* is a widely-used and flexible class of parametric models.

- **Non-parametric**: The assumptions made about the process of generating the data are much less than in parametric statistics and may be minimal. For example, every continuous probability distribution has a median, which may be estimated using the sample median or the Hodges-Lehmann-Sen estimator, which has good properties when the data arise from simple random sampling.
- **Semi-parametric**: This term typically implies assumptions 'between' fully and non-parametric approaches. For example, one may assume that a population distribution have a finite mean. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but does not make any parametric assumption describing the variance around that mean. More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components. One component is treated parametrically and the other non-parametrically.

3.3.4 Importance of valid models/assumptions

Whatever level of assumption is made, correctly-calibrated inference in general requires these assumptions to be correct; i.e., that the data-generating mechanisms really has been correctly specified.

- Incorrect assumptions of *'simple' random sampling* can invalidate statistical inference.
- More complex semi- and fully-parametric assumptions are also cause for concern. For example, incorrect assumptions of Normality in the population can invalidates some forms of regression-based inference.
- The use of **any** parametric model is viewed skeptically by most experts in sampling human populations: "most sampling statisticians, when they deal with confidence intervals at all, limit themselves to statements about [estimators] based on very large samples, where the central limit theorem ensures that these [estimators] will have distributions that are nearly normal." Here, the central limit theorem states that the distribution of the sample mean "for very large samples" is approximately normally distributed, if the distribution is not heavy tailed.

3.3.5 Approximate distributions

Given the difficulty in specifying exact distributions of sample statistics, many methods have been developed for approximating these.

With ***finite samples***, approximation results measure how close a limiting distribution approaches the statistic's sample distribution: For example, with 10,000 independent samples the normal distribution approximates (to two digits of accuracy) the distribution of the sample mean for many population distributions. Yet for many practical purposes, the normal approximation provides a good approximation to the sample-mean's distribution when there are 10 (or more) independent samples, according to simulation studies, and statisticians' experience. Following Kolmogorov's work in the 1950s, advanced statistics uses approximation theory and functional analysis to quantify the error of approximation: In this approach, the metric geometry of probability distributions is studied; this approach quantifies approximation error.

With ***infinite samples***, limiting results like the central limit theorem describe the sample statistic's limiting distribution, if one exists. Limiting results are not statements about finite samples, and indeed are logically irrelevant to finite samples. However, the asymptotic theory of limiting distributions is often invoked for work in estimation and testing. For example, limiting results are often invoked to justify the generalized method of moments and the use of generalized estimating equations, which are popular in econometrics and biostatistics. The magnitude of the difference between the limiting distribution and the true distribution (formally, the 'error' of the approximation) can be assessed using simulation. The use of limiting results in this way works well in many applications, especially with low-dimensional models with log-concave likelihoods (such as with one-parameter exponential families).

3.3.6 Randomization-based models

For a given dataset that was produced by a randomization design, the randomization distribution of a statistic (under the null-hypothesis) is defined by evaluating the test statistic for all of the plans that could have been generated by the randomization design.

In frequentist inference, randomization allows inferences to be based on the randomization distribution rather than a subjective model, and this is important especially in survey sampling and design of experiments. Statistical inference from randomized studies is also more straightforward than many other situations.

In Bayesian inference, randomization is also of importance: In survey sampling – *sampling without replacement* ensures the *exchangeability* of the sample with the population; in randomized experiments, randomization warrants a *missing at random* assumption for covariate information.

Objective randomization allows properly inductive procedures. Many statisticians prefer randomization-based analysis of data that was generated by well-defined randomization procedures. However, it is has been observed that in fields of science with developed theoretical knowledge and experimental control, randomized experiments may increase the costs of experimentation without improving the quality of inferences. Similarly, results from randomized experiments are recommended by leading statistical authorities as allowing inferences with greater reliability than do observational studies of the same phenomena. However, a good observational study may be better than a bad randomized experiment.

The statistical analysis of a randomized experiment may be based on the randomization scheme stated in the experimental protocol and does not need a subjective model. However, not all hypotheses can be tested by randomized experiments or random samples, which often require a large budget, a lot of expertise and time, and may have ethical problems.

3.3.7 Modes of inference

Different schools of statistical inference have become established. These schools (or 'paradigms') are not mutually-exclusive, and methods which work well under one paradigm often have attractive interpretations under other paradigms. The two main paradigms in use are **frequentist** and **Bayesian** inference, which are both summarized below.

a. Frequentist inference

This paradigm regulates the production of propositions by considering (notional) repeated sampling of datasets similar to the one at hand. By considering its characteristics under repeated sample, the frequentist properties of any statistical inference procedure can be described - although in practice this quantification may be challenging. Examples of frequentist inference are: P-value and Confidence interval

The frequentist calibration of procedures can be done without regard to utility functions. However, some elements of frequentist statistics, such as statistical decision theory, do incorporate utility functions. Loss functions must be explicitly stated for statistical theorists to prove that a statistical procedure has an optimality property. For example, median-unbiased estimators are optimal under absolute value loss functions, and least squares estimators are optimal under squared error loss functions.

While statisticians using frequentist inference must choose for themselves the parameters of interest, and the estimators/test statistic to be used, the absence of obviously-explicit utilities and prior distributions has helped frequentist procedures to become widely- viewed as 'objective'.

b. Bayesian inference

The Bayesian calculus describes degrees of belief using the 'language' of probability; beliefs are positive, integrate to one, and obey probability axioms. Bayesian inference uses the available *posterior beliefs* as the basis for making statistical propositions. There are several different justifications for using the Bayesian approach. Examples of Bayesian inference are: *Credible intervals* for interval estimation and *Bayes factors* for model comparison

Many informal Bayesian inferences are based on "intuitively reasonable" summaries of the posterior. For example, the posterior mean, median and mode, highest posterior density intervals, and Bayes Factors can all be motivated in this way. While a user's utility function need not be stated for this sort of inference, these summaries do all depend (to some extent) on stated earlier beliefs, and are generally viewed as subjective conclusions.

Formally, Bayesian inference is calibrated with reference to an explicitly stated utility, or loss function; the 'Bayes rule' is the one which maximizes expected utility, averaged over the subsequent uncertainty. Formal Bayesian inference therefore automatically provides optimal decisions in a decision theoretic sense. Given assumptions, data and utility, Bayesian inference can be made for essentially any problem, although not every statistical inference need have a Bayesian interpretation. Some advocates of Bayesian inference assert that inference *must* take place in this decision-theoretic framework, and that Bayesian inference should not conclude with the evaluation and summarization of posterior beliefs.

3.4 Other Essential Statistics for Simulations

3.4.1 Sample Size Determination

A common goal of survey research is to collect data representative of a population. The researcher uses information gathered from the survey to generalize findings from a drawn sample back to a population, within the limits of random error. However, when critiquing business education research, Wunsch (1986) stated that “two of the most consistent flaws included:

1. disregard for sampling error when determining sample size, and
2. disregard for response and non-response bias”.

Within a quantitative survey design, determining sample size and dealing with no response bias is essential. “One of the real advantages of quantitative methods is their ability to use smaller groups of people to make inferences about larger groups that would be prohibitively expensive to study”. The question then is, how large of a sample is required to infer research findings back to a population?

Standard textbook authors and researchers offer tested methods that allow studies to take full advantage of statistical measurements, which in turn give researchers the upper hand in determining the correct sample size. Sample size is one of the four inter-related features of a study design that can influence the detection of significant differences, relationships or interactions (Peers, 1996). Generally, these survey designs try to minimize both alpha error (finding a difference that does not actually exist in the population) and beta error (failing to find a difference that actually exists in the population) (Peers, 1996).

However, improvement is needed. Researchers are learning experimental statistics from highly competent statisticians and then doing their best to apply the formulas and approaches

Foundations for Sample Size Determination

Primary Variables of Measurement

The researcher must make decisions as to which variables will be incorporated into formula calculations. For example, if the researcher plans to use a seven-point scale to measure a continuous variable, e.g., job satisfaction, and also plans to determine if the respondents differ by certain categorical variables, e.g., gender, tenured, educational level, etc., which variable(s) should be used as the basis for sample size? This is important because the use of gender as the primary variable will result in a substantially larger sample size than if one used the seven-point scale as the primary variable of measure.

Cochran (1977) addressed this issue by stating that “One method of determining sample size is to specify margins of error for the items that are regarded as most vital to the survey. An estimation of the sample size needed is first made separately for each of these important items”. When these calculations are completed, researchers will have a range of n’s, usually ranging from smaller n’s for scaled, continuous variables, to larger n’s for dichotomous or categorical variables.

The researcher should make sampling decisions based on these data. If the n’s for the variables of interest are relatively close, the researcher can simply use the largest n as the sample size and be confident that the sample size will provide the desired results.

More commonly, there is a sufficient variation among the n's so that we are reluctant to choose the largest, either from budgetary considerations or because this will give an over- all standard of precision substantially higher than originally contemplated. In this event, the desired standard of precision may be relaxed for certain of the items, in order to permit the use of a smaller value of n. The researcher may also decide to use this information in deciding whether to keep all of the variables identified in the study. "In some cases, the n's are so discordant that certain of them must be dropped from the inquiry; . . .".

Error Estimation

Cochran's (1977) formula uses two key factors:

(1) the risk the researcher is willing to accept in the study, commonly called the margin of error, or the error the researcher is willing to accept, and

(2) the alpha level, the level of acceptable risk the researcher is willing to accept that the true margin Alpha Level.

The alpha level used in determining sample size in most educational research studies is either .05 or .01 (Ary, Jacobs, & Razavieh, 1996). In Cochran's formula, the alpha level is incorporated into the formula by utilizing the t-value for the alpha level selected (e.g., t-value for alpha level of .05 is 1.96 for sample sizes above 120). Researchers should ensure they use the correct t- value when their research involves smaller populations, e.g., t-value for alpha of .05 and a population of 60 is 2.00.

In general, an alpha level of .05 is acceptable for most research. An alpha level of .10 or lower may be used if the researcher is more interested in identifying marginal relationships, differences or other statistical phenomena as a precursor to further studies.

An alpha level of .01 may be used in those cases where decisions based on the research are critical and errors may cause substantial financial or personal harm, e.g., major programmatic changes.

Acceptable Margin of Error

The general rule relative to acceptable margins of error in educational and social research is as follows: For categorical data, 5% margin of error is acceptable, and, for continuous data, 3% margin of error is acceptable (Krejcie & Morgan, 1970). For example, a 3% margin of error would result in the researcher being confident that the true mean of a seven point scale is within ±.21 (.03 times seven points on the scale) of the mean calculated from the research sample. For a dichotomous variable, a 5% margin of error would result in the researcher being confident that the proportion of respondents who were male was within ±5% of the proportion calculated from the research sample. Researchers may increase these values when a higher margin of error is acceptable or may decrease these values when a higher degree of precision is needed.

Variance Estimation

A critical component of sample size formulas is the estimation of variance in the primary variables of interest in the study. The researcher does not have direct control over variance and must incorporate variance estimates into research design. Cochran (1977) listed four ways of estimating population variances for sample size determinations:

(1) take the sample in two steps, and use the results of the first step to determine how many additional responses are needed to attain an appropriate sample size based on the variance observed in the first step data;

(2) use pilot study results;

(3) use data from previous studies of the same or a similar population; or

(4) estimate or guess the structure of the population assisted by some logical mathematical results.

The first three ways are logical and produce valid estimates of variance; therefore, they do not need to be discussed further. However, in many educational and social research studies, it is not feasible to use any of the first three ways and the researcher must estimate variance using the fourth method.

A researcher typically needs to estimate the variance of scaled and categorical variables. To estimate the variance of a scaled variable, one must determine the inclusive range of the scale, and then divide by the number of standard deviations that would include all possible values in the range, and then square this number. For example, if a researcher used a seven-point scale and given that six standard deviations (three to each side of the mean) would capture 98% of all responses, the calculations would be as follows:

7 (number of points on the scale)
S = ---
6 (number of standard deviations)

When estimating the variance of a dichotomous (proportional) variable such as gender, Krejcie and Morgan (1970) recommended that researchers should use .50 as an estimate of the population proportion. This proportion will result in the maximization of variance, which will also produce the maximum sample size. This proportion can be used to estimate variance in the population. For example, squaring .50 will result in a population variance estimate of .25 for a dichotomous variable.

Basic Sample Size Determination

a. Continuous Data

Before proceeding with sample size calculations, assuming continuous data, the researcher should determine if a categorical variable will play a primary role in data analysis. If so, the categorical sample size formulas should be used. If this is not the case, the sample size formulas for continuous data described in this section are appropriate.

Assume that a researcher has set the alpha level a priori at .05, plans to use a seven point scale, has set the level of acceptable error at 3%, and has estimated the standard deviation of the scale as 1.167. Cochran's sample size formula for continuous data and an example of its use is presented here along with the explanations as to how these decisions were made.

$$n_0 = \frac{(t)^2 * (s)^2}{(d)^2} = \frac{(1.96)^2(1.167)^2}{(7*.03)^2} = 118$$

Where t = value for selected alpha level of .025 in each tail = 1.96 (the alpha level of .05

indicates the level of risk the researcher is willing to take that true margin of error may exceed the acceptable margin of error.)
s = estimate of standard deviation in the population = 1.167 (estimate of variance deviation for 7 point scale calculated by using 7 [inclusive range of scale] divided by 6 [number of standard deviations that include almost all (approximately 98%) of the possible values in the range]).
d = acceptable margin of error for mean being estimated = .21 (number of points on primary scale * acceptable margin of error; points on primary scale = 7; acceptable margin of error = .03 [error researcher is willing to except]).

Therefore, for a population of 1,679, the required sample size is 118. However, since this sample size exceeds 5% of the population (1,679*.05=84), Cochran's (1977) correction formula should be used to calculate the final sample size. These calculations are as follows:

$$n = \frac{no}{(1 + no / Population)} = \frac{(118)}{(1 + 118/1679)} = 111$$

Where population size = 1,679.
n0 = required return sample size according to Cochran's formula= 118. n1 = required return sample size because sample > 5% of population.
These procedures result in the minimum returned sample size. If a researcher has a captive audience, this sample size may be attained easily.
However, since many educational and social research studies often use data collection methods such as surveys and other voluntary participation methods, the response rates are typically well below 100%. Salkind (1997) recommended over-sampling when he stated that "If you are mailing out surveys or questionnaires . . . count on increasing your sample size by 40%-50% to account for lost mail and uncooperative subjects". But Over- sampling can add costs to the survey but is often necessary. A second consequence is, of course, that the variances of estimates are increased because the sample actually obtained is smaller than the target sample.
However, many researchers criticize the use of over-sampling to ensure that this minimum sample size is achieved and suggestions on how to secure the minimal sample size are scarce. If the researcher decides to use over-sampling, four methods may be used to determine the anticipated response rate:
(1) take the sample in two steps, and use the results of the first step to estimate how many additional responses may be expected from the second step;
(2) use pilot study results;
(3) use responses rates from previous studies of the same or a similar population; or
(4) estimate the response rate. The first three ways are logical and will produce valid estimates of response.

b. Categorical Data

The sample size formulas and procedures used for categorical data are very similar, but some variations do exist. Assume a researcher has set the alpha level a priori at .05, plans to use a proportional variable, has set the level of acceptable error at 5%, and has estimated the standard deviation of the scale as .5. Cochran's sample size formula for categorical data and an example of its use is presented here along with explanations as to how these decisions were made.

$$no= \frac{(t)2 * (p)(q)}{(d)2}$$

$$no= \frac{(1.96)2(.5)(.5)}{(.05)2} = 384$$

Where t = value for selected alpha level of .025 in each tail = 1.96 (the alpha level of .05 indicates the level of risk the researcher is willing to take that true margin of error may exceed the acceptable margin of error).
Where (p)(q) = estimate of variance = .25 (maximum possible proportion (.5) * 1- maximum possible proportion (.5) produces maximum possible sample size).
Where d = acceptable margin of error for proportion being estimated = .05 (error researcher is willing to except).

Therefore, for a population of 1,679, the required sample size is 384. However, since this sample size exceeds 5% of the population (1,679*.05=84), Cochran's (1977) correction formula should be used to calculate the final sample size. These calculations are as follows:

$$n_1= \frac{n_0}{(1 + n_o / \text{Population})}$$

$$n_1= \frac{(384)}{(1 + 384/1679)} = 313$$

Where population size = 1,679,
n_0 = required return sample size according to Cochran's formula= 384,
n_1 = required return sample size because sample > 5% of population

These procedures result in a minimum returned sample size of 313. Using the same oversampling procedures as cited in the continuous data example, and again assuming a response rate of 65%, a minimum drawn sample size of 482 should be used. These calculations were based on the following:

Where anticipated return rate = 65%.
Where n2 = sample size adjusted for response rate. Where minimum sample size (corrected) = 313.

Therefore, n2 = 313/.65 = 482.

3.4.2 The Central Limit Theorem

The main idea of the central limit theorem (CLT) is that the average of a sample of observations drawn from some population with any shape-distribution is approximately distributed as a normal distribution if certain conditions are met. In theoretical statistics there are several versions of the central limit theorem depending on how these conditions are specified. These are concerned with the types of assumptions made about the distribution of the parent population (population from which the sample is drawn) and the actual sampling procedure.

One of the simplest versions of the theorem says that if is a random sample of size n (say, n larger than 30) from an infinite population, finite standard deviation, then the standardized sample mean converges to a standard normal distribution or, equivalently, the sample mean approaches a normal distribution with mean equal to the population mean and standard deviation equal to standard deviation of the population divided by the square root of sample size n. In applications of the central limit theorem to practical problems in statistical inference, however, statisticians are more interested in how closely the approximate distribution of the sample mean follows a normal distribution for finite sample sizes, than the limiting distribution itself. Sufficiently close agreement with a normal distribution allows statisticians to use normal theory for making inferences about population parameters (such as the mean) using the sample mean, irrespective of the actual form of the parent population.

It is well known that whatever the parent population is, the standardized variable will have a distribution with a mean 0 and standard deviation 1 under random sampling. Moreover, if the parent population is normal, then it is distributed exactly as a standard normal variable for any positive integer n. The central limit theorem states the remarkable result that, even when the parent population is non-normal, the standardized variable is approximately normal if the sample size is large enough (say > 30). It is generally not possible to state conditions under which the approximation given by the central limit theorem works and what sample sizes are needed before the approximation becomes good enough. As a general guideline, statisticians have used the prescription that if the *parent distribution is symmetric and relatively short-tailed*, then the sample mean reaches approximate normality for smaller samples than if the parent population is skewed or long-tailed.

Under certain conditions, in large samples, the sampling distribution of the sample mean can be approximated by a normal distribution. The sample size needed for the approximation to be adequate depends strongly on the shape of the parent distribution. Symmetry (or lack thereof) is particularly important. For a symmetric parent distribution, even if very different from the shape of a normal distribution, an adequate approximation can be obtained with small samples (e.g., 10 or 12 for the uniform distribution). For symmetric short-tailed parent distributions, the sample mean reaches approximate normality for smaller samples than if the parent population is skewed and long-tailed. In some extreme cases (e.g. binomial) samples sizes far exceeding the typical guidelines (e.g., 30) are needed for an adequate approximation.

For some distributions without first and second moments (e.g., Cauchy), the central limit theorem does not hold.

3.4.3 The Least Squares Model

Many problems in analyzing data involve describing how variables are related. The simplest of all models describing the relationship between two variables is a linear, or straight-line, model. The simplest method of fitting a linear model is to "eye-ball" a line through the data on a plot. A more elegant, and conventional method is that of "least squares", which finds the line minimizing the sum of distances between observed points and the fitted line. With this you will:

- Realize that fitting the "best" line by eye is difficult, especially when there is a lot of residual variability in the data.
- Know that there is a simple connection between the numerical coefficients in the regression equation and the slope and intercept of regression line.
- Know that a single summary statistic like a correlation coefficient does not tell the whole story. A scatter plot is an essential complement to examining the relationship between the two variables.

3.4.4 ANOVA: Analysis of Variance

Analysis of Variance or ANOVA enables us to test the difference between 2 or more means. ANOVA does this by examining the ratio of variability between two conditions and variability within each condition. For example, if we give a drug that we believe will improve memory to a group of people and give a placebo to another group of people, we might measure memory performance by the number of words recalled from a list we ask everyone to memorize. A **t-test** would compare the likelihood of observing the difference in the mean number of words recalled for each group. An ANOVA test, on the other hand, would compare the variability that we observe between the two conditions to the variability observed within each condition. We measure variability as the sum of the difference of each score from the mean. When we actually calculate an ANOVA we use a short-cut formula. Thus, when the variability that we predict (between the two groups) is much greater than the variability we don't predict (within each group) then we will conclude that our treatments produce different results.

3.4.5 Exponential Density Function (EDF)

EDF is use to take important class of decision problems under uncertainty such as the chance between events. For example, the chance of the length of time to next breakdown of a machine not exceeding a certain time, such as the photocopying machine in your office not to break during this week.

Exponential distribution gives distribution of time between independent events occurring at a constant rate. Its density function is:
f(t) = □ exp(-□t),

where □ is the average number of events per unit of time, which is a positive number. The mean and the variance of the random variable t (time between events) are 1/ □, and 1/$□^2$, respectively

Applications include probabilistic assessment of the time between arrival of patients to the emergency room of a hospital, and arrival of ships to a particular port.

3.4.6 Poisson Process

An important class of decision problems under uncertainty is characterized by the small chance of the occurrence of a particular event, such as an accident. Poisson gives probability of exactly x independent occurrences during a given period of time if events take place independently and at a constant rate. It may also represent number of occurrences over constant areas or volumes. The following statements describe the *Poisson Process*:

1. The occurrences of the events are independent.
2. The occurrence of events from a set of assumptions in an interval of space or time has no effect on the probability of a second occurrence of the event in the same, or any other, interval.
3. Theoretically, an infinite number of occurrences of the event must be possible in the interval.
4. The probability of the single occurrence of the event in a given interval is proportional to the length of the interval.
5. In any infinitesimally small portion of the interval, the probability of more than one occurrence of the event is negligible.

Poisson processes are often used, for example in quality control, reliability, insurance claim, incoming number of telephone calls, and queuing theory.

An Application: One of the most useful applications of the Poisson Process is in the field of queuing theory. In many situations where queues occur it has been shown that the number of people joining the queue in a given time period follows the Poisson model. For example, if the rate of arrivals to an emergency room is □ per unit of time period (say 1 hr), then:

P (n arrivals) = $□^{n}$ $e^{-□□}$/ n!

The mean and variance of random variable n are both □□. However if the mean and variance of a random variable having equal numerical values, then it is not necessary that its distribution is a Poisson.

Applications:

P (0 arrival) = $e^{-□}$

P (1 arrival) = □ $e^{-□□}$/ 1!P (2 arrival) = $□^{2}$

$e^{-□□}$/ 2!

and so on. In general:

P (n+1 arrivals) = □□Pr (n arrivals) / n.

3.4.7 Uniform Density Function (UDF)

This function gives the probability that observation will occur within a particular interval when probability of occurrence within that interval is directly proportional to interval length.

For example, it is used to generate random numbers in sampling and Monte Carlo simulation.

The mass function of geometric mean of n independent uniforms [0,1] is: $P(X = x) = n\ x^{(n - 1)}\ (\mathrm{Log}[1/x^{n}])^{(n - 1)} / (n - 1)!$.
$z_L = [U^L\text{-}(1\text{-}U)^L] / L$ is said to have Tukey's symmetrical distribution.
You may like to use *Uniform Applet* to perform your computations, then visit also:
http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/pvalues.htm

3.4.8 Test for Randomness

We need to test for both randomness as well as uniformity. The tests can be classified in 2 categories: Empirical or statistical tests, and theoretical tests.
Theoretical tests deal with the properties of the generator used to create the realization with desired distribution, and do not look at the number generated at all. For example, we would not use a generator with poor qualities to generate random numbers.
Statistical tests are based solely on the random observations produced.

A. Test for independence:

Plot the x_i realization vs x_{i+1}. If there is independence, the graph will not show any distinctive patterns at all, but will be perfectly scattered.

B. Runs tests.(run-ups, run-downs):

This is a direct test of the independence assumption. There are two test statistics to consider: one based on a normal approximation and another using numerical approximations.
Test based on Normal approximation:
Suppose you have N random realizations. Let K be the total number of runs in a sequence. If the number of positive and negative runs are greater than say 20, the distribution of K is reasonably approximated by a Normal distribution with mean (2N - 1) /3 and (16N - 29) / 90. Reject the hypothesis of independence or existence of runs if | Zo| < Z(1-alpha/2) where Zo is the Z score.

C. Correlation tests:

Do the random numbers exhibit discernible correlation? Compute the sample Autcorrelation Function.

Frequency or Uniform Distribution Test:
Use Kolmogorov-Smirimov test to determine if the realizations follow a U(0,1).

3.4.9 Some Useful SPSS Commands

a. Test for Binomial:

```
NPAR TEST BINOMIAL(p)=GENDER(0, 1)
```

b. Gooness-of-fit for discrete r.v.:

```
NPAR TEST CHISQUARE=X (1,3)/EXPECTED=20 30 50
```

C. Two population t-test

```
T-TEST GROUPS=GENDER(1,2)/VARIABLES=X
```

4.0 Self-Assessment Exercise(s)

Answer the following questions:

1. State the Cochran's sample size formula for continuous and categorical data
2. Assume a researcher has set the alpha level a priori at 10%, plans to use a proportional variable, has set the level of acceptable error at 5%, and has estimated the standard deviation of the scale as .5. Find the sample size for a population of 2500
3. What are the Cochran's key factors for error estimation?
4. State the essential feature of Poisson process
5. List four ways of estimating population variances for sample size determinations according to Cochran.
6. What are the objectives of randomization, and state the importance of randomization in frequentist and Bayesian inferences

5.0 Conclusion

Statistics is the basis of simulation. In this unit we have simply introduced some basic statistics in modelling and simulations. We hope that the reader will broaden his/her understanding by consulting the referenced texts or othe statistics books.

6.0 Summary

In this unit we were able to

- Differentiate between the two broad components of statistics: descriptive and inference statistics
- Have concise discussions of descriptive statistics on; Univarite statistics measures: the distribution, central tendency, dispersion, etc. and gave some examples.
- Discuss Inference statistics under the following subheads: definition, Model/assumptions, approximate distributions, random-based models and modes of inference.
- Introduce some essential statistical measures in simulation such as;
    - sample size determination
    - central limit theorem

    - least square model
    - Analysis of variance
    - Exponential distribution function
    - Poisson distribution
    - Uniform distribution
    - Test for randomness
    - Some commands of Special package for statistical analysis (SPSS)

7.0 Further Readings

- Gordon, S. I., & Guilfoos, B. (2017). *Introduction to Modeling and Simulation with MATLAB® and Python*. Milton: CRC Press.
- Zeigler, B. P., Muzy, A., & Kofman, E. (2019). *Theory of modeling and simulation: Discrete event and iterative system computational foundations*. San Diego (Calif.): Academic Press.
- Kluever, C. A. (2020). *Dynamic systems modeling, simulation, and control*. Hoboken, N.J: John Wiley & Sons.
- Law, A. M. (2015). *Simulation modeling and analysis*. New York: McGraw-Hill.
- Verschuuren, G. M., & Travise, S. (2016). *100 Excel Simulations: Using Excel to Model Risk, Investments, Genetics, Growth, Gambling and Monte Carlo Analysis*. Holy Macro! Books.
- Grigoryev, I. (2015). *AnyLogic 6 in three days: A quick course in simulation modeling*. Hampton, NJ: AnyLogic North America.
- Dimotikalis, I., Skiadas, C., & Skiadas, C. H. (2011). *Chaos theory: Modeling, simulation and applications: Selected papers from the 3rd Cghaotic Modeling and Simulation International Conference (CHAOS2010), Chania, Crete, Greece, 1-4 June, 2010*. Singapore: World Scientific.
- Velten, K. (2010). *Mathematical modeling and simulation: Introduction for scientists and engineers*. Weinheim: Wiley-VCH.