Unit 8: Inferential Statistics

Welcome to week 8 where you will be introduced to the different analyses available for the different types of quantitative data. This will mean looking at ways to describe your data and how to create and perform hypothesis tests. This is called inference, as we are making inferences about a population from a sample of data. Inference is the process of extracting meaningful and useful business information from raw data. This process is known as ***statistical inference*** because it involves using the data to make valid inferences about the underlying population. As data is inherently variable, all such inferences will necessarily be subject to uncertainty. This uncertainty is quantified using probability.

Data consist of the observed values of one or more variables of interest. They are usually organized into **datasets**. A dataset can be thought of as a table, whose columns represent the different **variables**, and whose rows represent the individual **observations**. Each individual cell in the table contains the value that the variable takes for the given observation.

In order to extract meaningful and useful business intelligence from data, it is important that the methodology to be employed is valid for the variable or variables of interest. The first important distinction concerns the **level of measurement** of each variable.

When exploring data, it is helpful to try to summarise the data in some meaningful way. One approach is to construct a graphical summary. In this section, we (very briefly) explore some of the more useful graphical summaries.

It is often convenient to summarise numerical data using a few simple summary measures. Most commonly, two such measures are used. The first of these is a measure of **location**, and represents the value taken by “a typical observation” – that is, by an observation that falls “right in the middle of the data”. As well as knowing the magnitude of a typical observation, some idea of the “variability” or “spread” of the data is also useful. This is provided by a measure of **dispersion**.

Instead of *estimating* some population value of interest that underlies the data, an alternative form of inference is to use the data to provide evidence about whether some assumption of interest regarding that population value is likely to be true. Such a form of inference is known as **hypothesis testing** and is usually preferred to the estimation approach when interest lies in comparing the relevant values underlying two or more different populations.

These are all techniques we can employ to find patterns and meaning from our data.

In this unit we shall:

- Define the different levels of quantitative data.

- Define measures of location and spread.
- Introduce the concept of inference and hypothesis testing.

On completion of this unit you will be able to:

- Identify the different levels of measurement.
- Produce measures of location and spread.
- Perform appropriate hypothesis tests.

Reflection:

Levels of measurement, also known as scales of measurement, provide information on the precision with which data are collected. In the context of scientific study, a variable is any data point that can carry on a variety of values (e.g., height or test scores).

The level of measurement is a statistical classification that links the values of the variables. So, the information contained inside the values is described by the scale of measurement (Team, 2022).

There are four possible data types: nominal, ordinal, interval, and ratio.

- Nominal data can only be categorized,
- Ordinal data can be categorized and ranked,
- Interval data can be categorized, ranked, and uniformly spaced, and
- Ratio data has a natural zero.

What you can do with your data analysis may be constrained by the precision with which you measure the variable of interest. From simple (nominal) to sophisticated and precise (high), there is a scale of measuring levels (ratio).

Importance of levels of measurement:

How you can examine your data is dependent on the granularity of your measurements. The different levels restrict the types of inferential statistics that may be performed on your data to either confirm or deny your hypothesis, as well as the types of descriptive statistics that can be used to provide an overall overview of your data.

Many of your variables will have multiple possible levels of measurement, and you will need to decide on that level of measurement before you can begin collecting data (BHandari, 2020).

The average of a group of numbers is a single number that essentially informs us where the set of numbers is located, hence averages are what we mean when we talk about location measures. Dispersion in a data set can be quantified by looking at how widely different values lie within the set. For any data set summary, these two figures are the most crucial.

We have prior knowledge of three averages and one measure of dispersion:

1. **The mode**: In other words, this is the most common piece of information or numerical value.

2. **The median**: If we align together all the information in sequence, this is the data point or value that lies exactly in the middle. If there is no "middle" number, we use the number that falls exactly in the "middle" of the two

3. **The (arithmetic) mean**: If we add all the numbers together and divide that total by the total number of numbers, we get this value.

4. **The range**: The median is a measure of dispersion defined as the difference between the smallest and greatest values (Maths with David, 2020).

Hypotheses Test:

Testing hypotheses is a scientific method for examining our preconceived notions about the universe by employing statistical methods. Most commonly, scientists employ this method to examine the hypothesized results of their theoretical analyses.

When testing a hypothesis, you do the following five things:

1. Provide a formal statement of your research's null and alternative hypotheses (Ho) and (Ha or H1).
2. Gather information that can be used to verify the theory.
3. Carry out the correct statistical analysis.
4. Decide whether to reject or fail to reject your null hypothesis.
5. Present the findings in your results and discussion section (Bevans, 2019).

References:

Bevans, R. (2019). *Hypothesis Testing | A Step-by-Step Guide with Easy Examples*. [online] Scribbr. Available at: https://www.scribbr.com/statistics/hypothesis-testing/.

BHandari, P. (2020). *Levels of Measurement | Nominal, Ordinal, Interval and Ratio*. [online] Scribbr. Available at: https://www.scribbr.com/statistics/levels-of-measurement/.

Maths with David. (2020). *KS4. Statistics 5. Measures of Location and Spread*. [online] Available at: https://mathswithdavid.com/ks4-statistics-5-measures-of-location-and-spread/ [Accessed 5 Oct. 2022].

Team, C. (2022). *Level of Measurement - Overview, Types of Scales, Examples*. [online] Corporate Finance Institute. Available at: https://corporatefinanceinstitute.com/resources/knowledge/other/level-of-measurement/.