UNIT I INTRODUCTION TO BIG DATA

Introduction to Big Data Platform – Challenges of Conventional Systems - Intelligent data analysis –Nature of Data - Analytic Processes and Tools - Analysis Vs Reporting - Modern Data Analytic Tools- Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error.

THE NATURE OF DATA

Data is important when it comes time to applying statistical operations and performing model building and testing. There are four types of data:

1. **Nominal**
2. **Ordinal**
3. **Interval**
4. **Ratio**

Each offers a unique set of characteristics, which impacts the type of analysis that can be performed. The distinction between the four types of scales center on three different characteristics:

1. The order of responses – whether it matters or not
2. The distance between observations – whether it matters or is interpretable
3. The presence or inclusion of a true zero

1. Nominal Scales

Nominal data, also known as nominal scale, is a type of qualitative data that's used to label variables without providing numeric values. It's the simplest form of a level of measurement and is the foundation of statistical analysis and other mathematical sciences.

Nominal scales measure categories and have the following characteristics:

- **Order:** The order of the responses or observations does not matter.
- **Distance:** Nominal scales do not hold distance. The distance between a 1 and a 2 is not the same as a 2 and 3.
- **True Zero:** There is no true or real zero. On a nominal scale, zero is uninterpretable.

Consider traffic source as an example in which visitors reach the site through a mutually exclusive channel, or last point of contact. These channels would include:

1. Paid Search
2. Organic Search
3. Email
4. Display

We can count the number of visits from each channel. Those counts can be considered nominal in nature. Suppose the counts looked like this:

Channel	**Count of Visits**

Paid Search	2,143
Organic Search	3,124
Email	1,254
Display	2,077

With nominal data, the order of the four channels would not change or alter the interpretation. Suppose we, instead, viewed the data like this:

Channel	**Count of Visits**
Display	2,077
Paid Search	2,143
Email	1,254
Organic Search	3,124

The order of the categories does not matter. And, the distance between the categories is not relevant. Display is not four times as much as paid search and organic search is not half of organic search. While there is an arithmetic relationship between these counts, that is only relevant if we treat the scales as ratio scales. Finally, zero holds no meaning. We could not interpret a zero because it does not occur on a nominal scale.

Appropriate statistics for nominal scales: mode, count, frequencies

Displays: histograms or bar charts

2. Ordinal Scales

At the risk of providing a tautological definition, ordinal scales measure, well, order. So, our characteristics for ordinal scales are:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ordinal scales do not hold distance. The distance between first and second is unknown as is the distance between first and third along with all observations.
- **True Zero:** There is no true or real zero. An item, observation, or category cannot finish zero.

Let’s work through our traffic source example and rank the channels based on the number of visits to our site, with “1” being the highest number of visits:

Channel	Count of Visits
Organic Search	1

Paid Search	2
Display	3
Email	4

By ranking the channel from most to least number of visitors in terms of last point of contact,we've established an order. However, the distance between the rankings appears unknown. Organic Search could have one more visit compared to Paid Search or one hundred more visitors. The distance between the two items appears unknown. Finally, zero holds no meaning. We could not interpret a zero because it does not occur in an ordinal scale. An item such as Organic Search could not maintain a zero ranking.

Appropriate statistics for ordinal scales: count, frequencies, mode

Displays: histograms or bar charts

3. Interval Scales

Interval scales provide insight into the variability of the observations or data. Classic interval scales are

1. Likert scales (e.g., 1 - strongly agree and 9 - strongly disagree)
2. Semantic Differential scales (e.g., 1 - dark and 9 - light).

The characteristics of interval scales are:

- **Order:** The order of the responses or observations does matter.
- **Distance:** Interval scales do offer distance. That is, the distance from 1 to 2 appears the same as 4 to 5. Also, six is twice as much as three and two is half of four. Hence, we can perform arithmetic operations on the data.
- **True Zero:** There is no zero with interval scales. However, data can be rescaled in a manner that contains zero.

An interval scale measure from 1 to 9 remains the same as 11 to 19 because we added 10 to all values. Similarly, a 1 to 9 interval scale is the same as a -4 to 4 scale because we subtracted 5 from all values. Although the new scale contains zero, zero remains uninterpretable because it only appears in the scale from the transformation.

Unless a web analyst is working with survey data, it is doubtful he or she will encounter data from an interval scale. More likely, a web analyst will deal with ratio scales (next section).

Appropriate statistics for interval scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

An Illustrative Side Note About Temperature

An argument exists about temperature. Is it an interval scale or an ordinal scale? Many researchers argue for temperature as an interval scale. It offers order (e.g., 212°°

F is hotter than 32∘∘ F), distance (e.g., 40∘∘ F to 44∘∘ F is the same as 100∘∘ F to 104∘∘ F), and lacks a true zero (e.g., 0∘∘ F is not the same as 0∘∘ C). However, other researchers argue for temperature as an ordinal scale because of the issue related to distance. 200∘∘ F is not twice as 100 F. The human brain registers both temperatures as equally hot (if standing outside) or mild (if touching a stove). Finally, we would not say that 80 F is twice as warm as 40∘∘ F or that 30∘∘ F is a third colder as 90∘∘ F.

Ratio Scales

Ratio scales appear as nominal scales with a true zero. They have the following characteristics:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ratio scales do have an interpretable distance.
- **True Zero:** There is a true zero.

Income is a classic example of a ratio scale:

Order is established. We would all prefer $100 to $1!

Zero dollars means we have no income (or, in accounting terms, our revenue exactly equals our expenses!)

Distance is interpretable, in that $20 appears as twice $10 and $50 is half of a $100.

In web analytics, the number of visits and the number of goal completions serve as examples of ratio scales. A thousand visits is a third of 3,000 visits, while 400 goal completions are twice as many as 200 goal completions. Zero visitors or zero goal completions should be interpreted as just that: no visits or completed goals. For the web analyst, the statistics for ratio scales are the same as for interval scales.

Appropriate statistics for ratio scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

The table below summarizes the characteristics of all four types of scales.

	Nominal	**Ordinal**	**Interval**
Order matters	No	Yes	Yes
Distance is Interpretable	No	No	Yes
Zero exists	No	No	No