Lecture 02 : Descriptive Statistics

Presented By:

Dr. Vinay Kulkarni

Aegis

SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

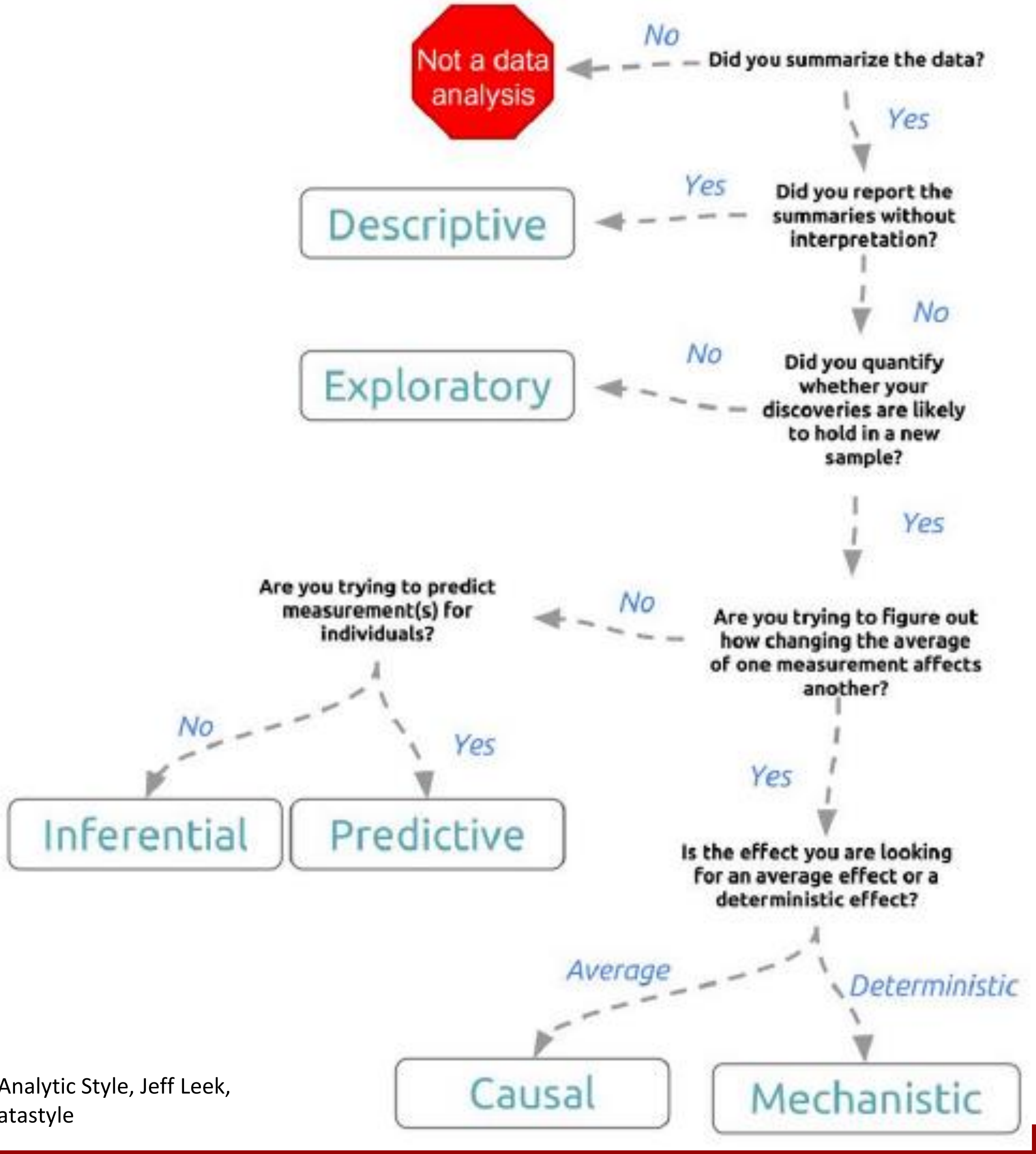
Revision - What is ‘Statistics’

STATISTICS

- The branch of science that deals with
  - Collecting data
  - Organizing and summarizing data
  - Analysis of data
  - Inferring / Predicting / Deciding based on the data and its analysis

SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Revision: Statistics - Branches



Did you summarize the data?

Did you report the summaries without interpretation?

Did you quantify whether your discoveries are likely to hold in a new sample?

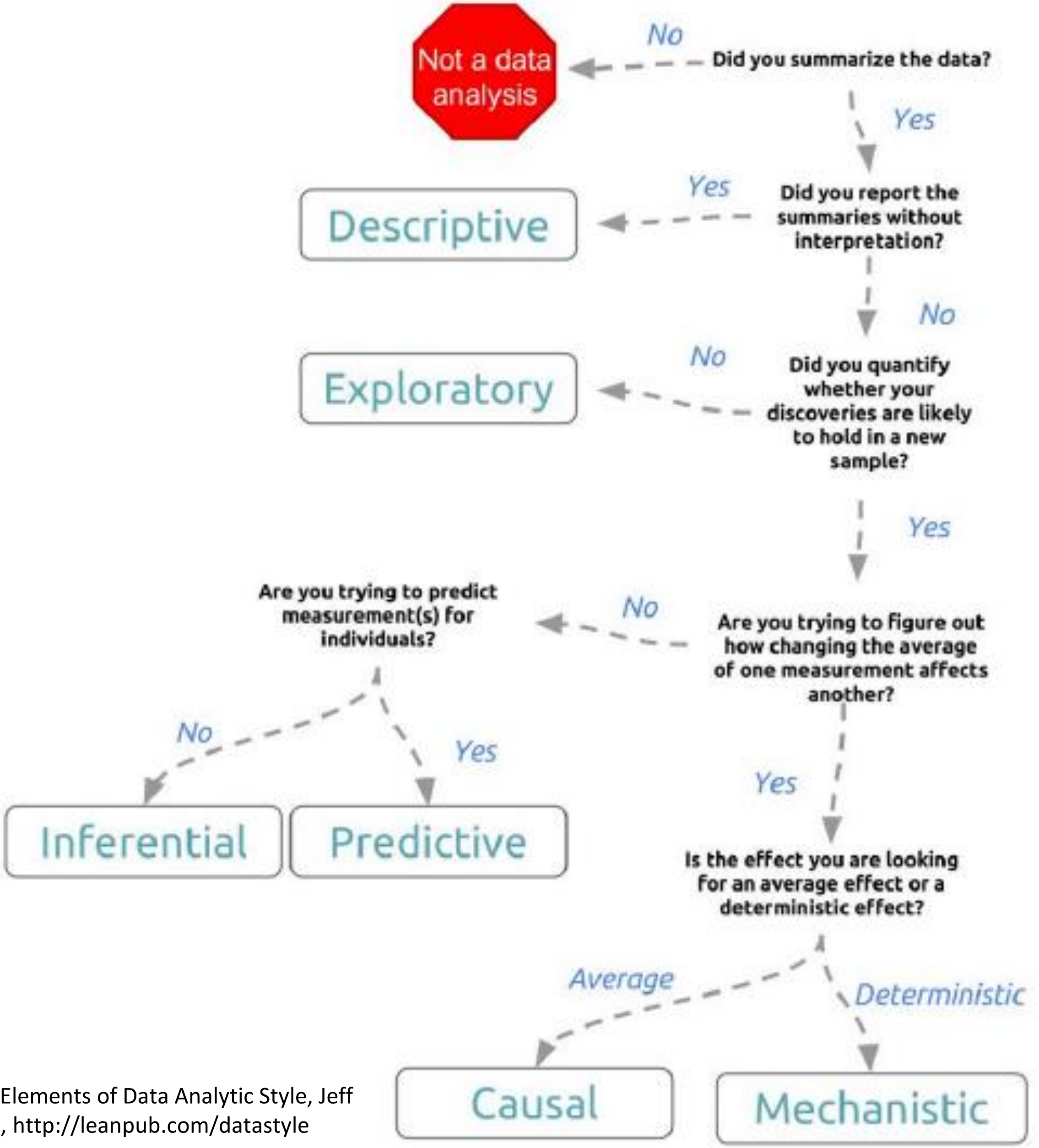
Are you trying to figure out how changing the average of one measurement affects another?

Are you trying to predict measurement(s) for individuals?

Is the effect you are looking for an average effect or a deterministic effect?

Ref: Elements of Data Analytic Style, Jeff Leek,
http://leanpub.com/datastyle

Revision: Statistics - Branches



Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Inferential
Predictive
Yes
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

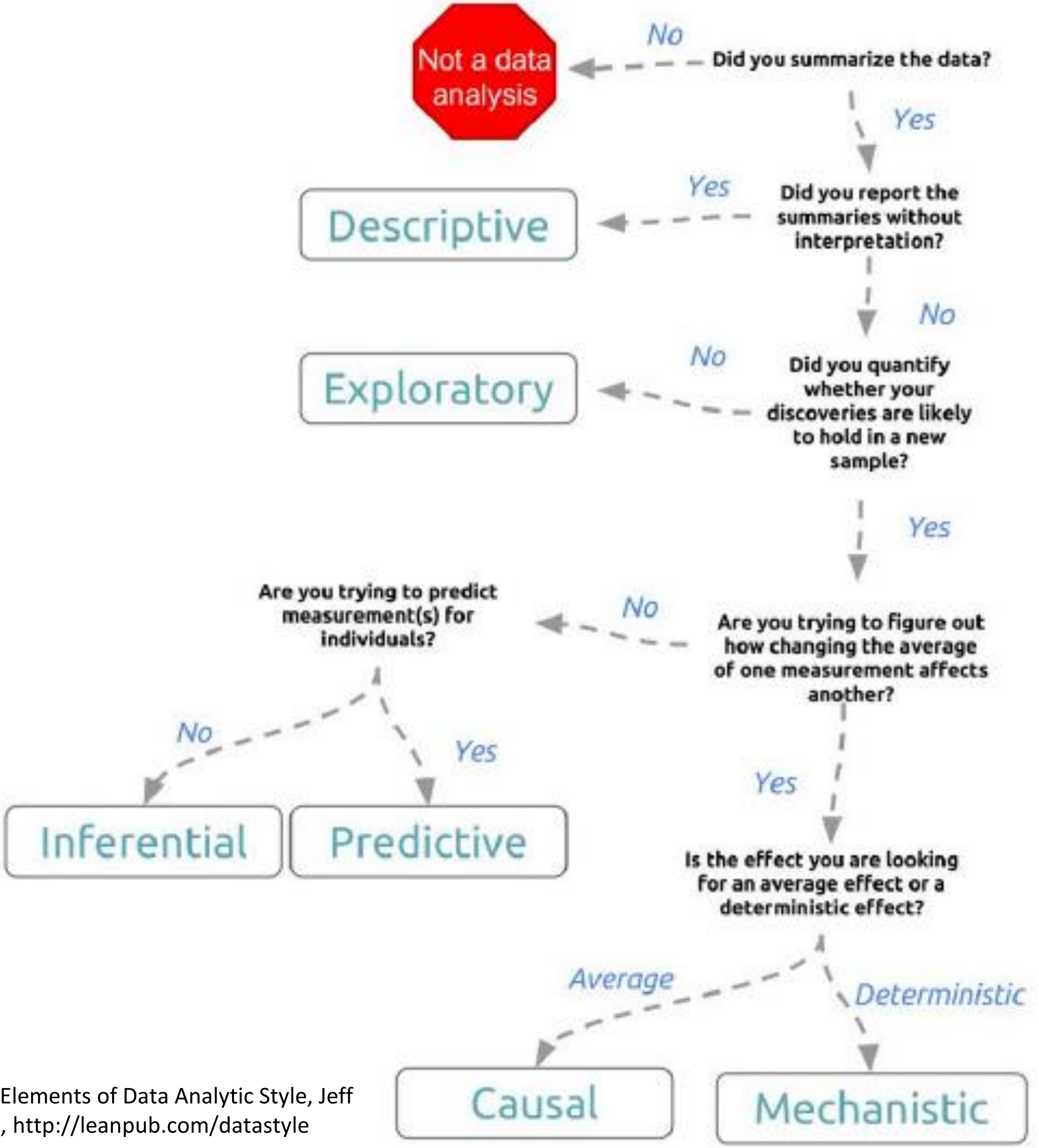
Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

DESCRIPTIVE STATISITCS
Seeks to summarize the measurements in a single data set without further interpretation.

EXPLORATORY ANALYSIS
Builds on descriptive data analysis by searching for discoveries, trends, correlations or relationships between the measurement of multiple variables to generate ideas or hypotheses.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Revision: Statistics - Branches


Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Inferential
Predictive
Yes
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

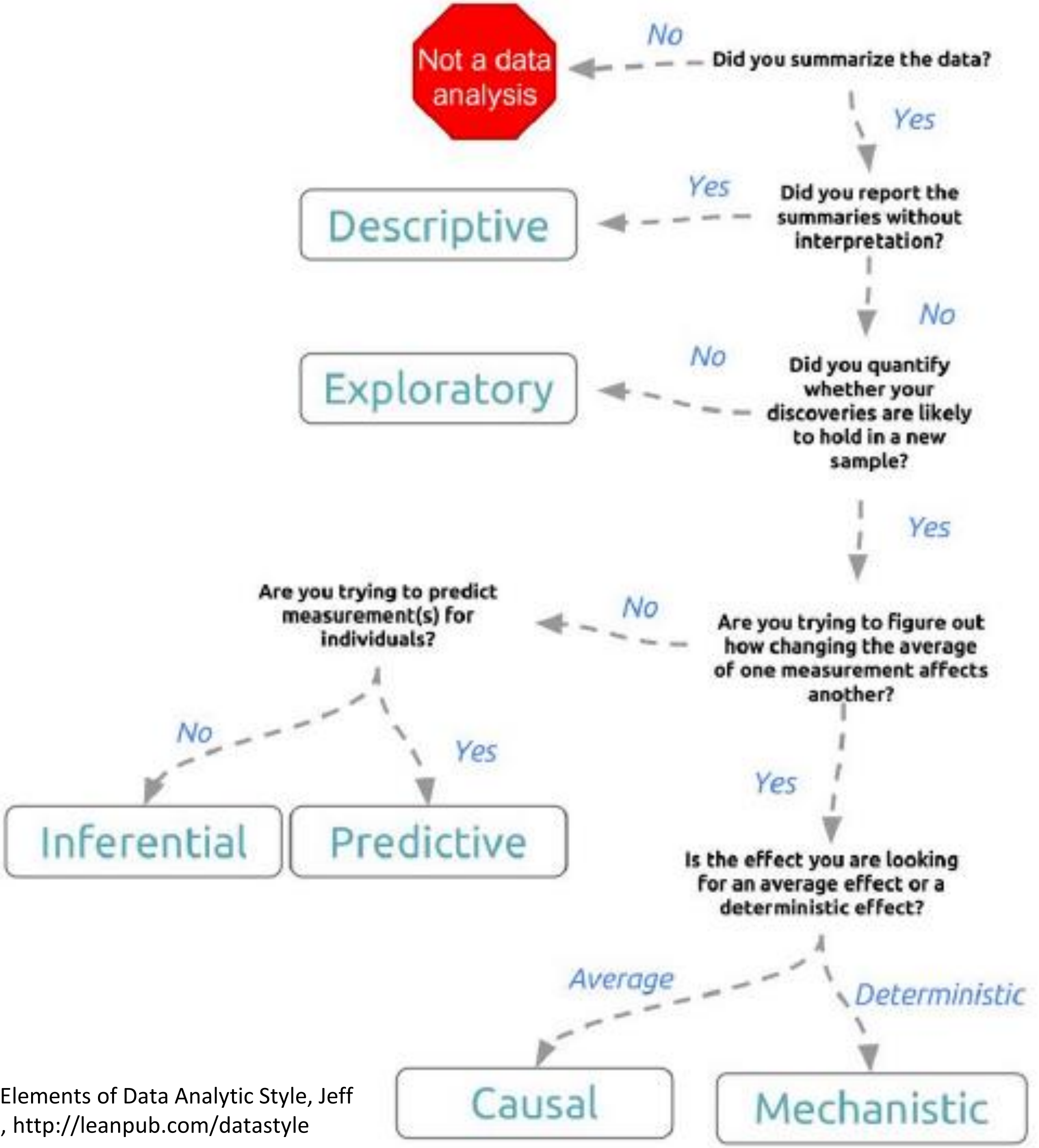
Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

INFERENTIAL ANALYSIS
Goes beyond exploratory analysis by quantifying whether an observed pattern will hold beyond the data set in hand – relationships among measurements at population scale. This is the most common form of data analysis.

PREDICTIVE ANALYSIS
This uses a subset of measurements (features) to predict another measurement (outcome) for a person or a unit. There is however no attempt to explain why the prediction works.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Revision: Statistics - Branches



Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Inferential
Predictive
Yes
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Ref: Elements of Data Analytic Style, Jeff Leek, http://leanpub.com/datastyle

CAUSAL ANALYSIS

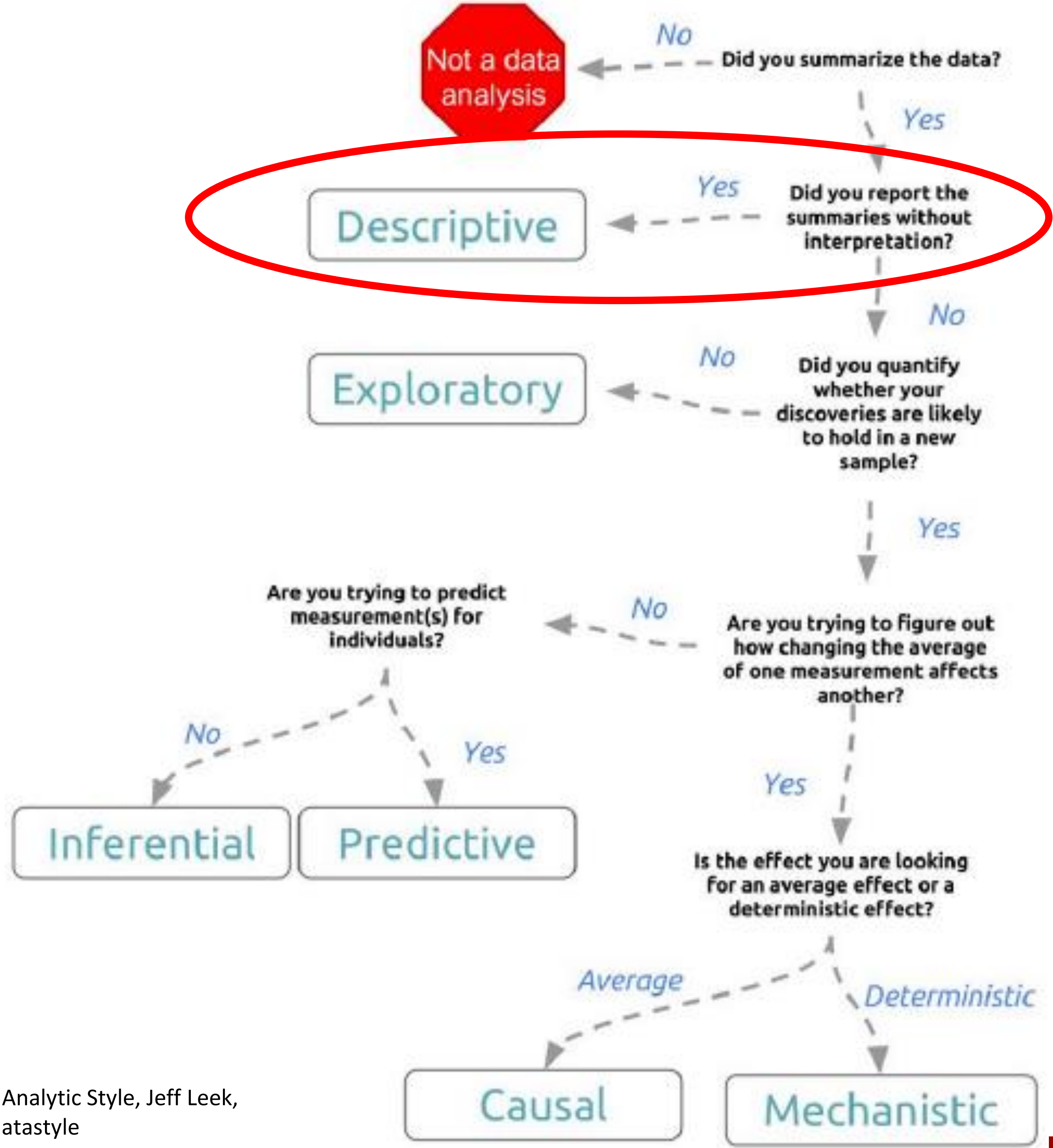
Seeks to reliably find out what happens to one measurement if you make changes to another measurement. Unlike predictive or inferential data analysis, causal analysis identifies both – magnitude and direction of relationships between variables.

MECHANISTIC ANALYSIS

Mechanistic analysis seeks to demonstrate that changing one measurement always and exclusively leads to a specific deterministic behaviour in another.

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Today's Focus


Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Inferential
Predictive
Yes
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Ref: Elements of Data Analytic Style, Jeff Leek,
http://leanpub.com/datastyle

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Agenda

- Descriptive Statistics: Introduction
- Methods of Descriptive Statistics
- Central Tendency
- Measures of Central Tendency
- Measures of Position
- Measures of Dispersion
- Measures of Quality and Outliers

Descriptive Statistics

- **DESCRIPTIVE STATISITICS**
  - Seeks to summarize the measurements in a single data set without further interpretation.

- Goals of Descriptive Statistics
  - Summarize data
  - Understand and communicate
  - Ground work prior to Inferential statistics

Descriptive Statistics

Two methods used to describe data

- Graphical
- Numerical

- Graphical descriptions
  - Categorical variables
    - Bar graph
    - Pie Chart
    - Pareto Chart
  - Quantitative variables
    - Dot plot
    - Stem and leaf display
    - Histogram

Summarizing Quantitative Data

Measures used to summarize quantitative data

- Measures of Central Tendency
- Measures of Variation / Dispersion
- Measures of Position
- Measures of Quality and Outliers

Summarizing Quantitative Data

Measures used to summarize quantitative data

- **Measures of Central Tendency**
  - The Mean
  - The Mode
  - The Median
  - The Mid-range
- Measures of Variation / Dispersion
- Measures of Position
- Measures of Quality and Outliers

The Mean

- For the Population

$$\mu = (x_1 + x_2 + \ldots + x_N) / N = \sum x_i / N$$

- For the Sample

$$\bar{x} = (x_1 + x_2 + \ldots . + x_n) / n = \sum x_i / n$$

SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

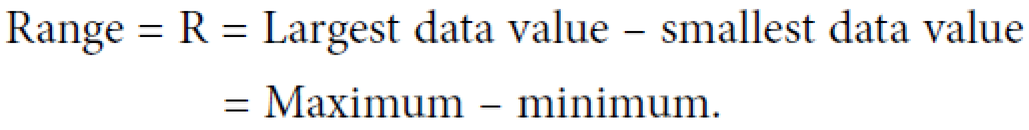
Summarizing Quantitative Data

Measures used to summarize quantitative data

- Measures of Central Tendency
- **Measures of Variation / Dispersion**
  - The Range
  - The Variance
  - The Standard Deviation
- Measures of Position
- Measures of Quality and Outliers

The Range

- The Range



$$\begin{aligned}\text{Range} = R &= \text{Largest data value} - \text{smallest data value}\\ &= \text{Maximum} - \text{minimum.}\end{aligned}$$

- Range: Measure of distance between the extremes in the data
- It does not tell us how the observations are distributed between the smallest and the largest data values

The Variance and Standard Deviation (Population)

- Variance (Population)

$$\sigma^2 = \frac{\sum_{i=1}^{N}(Y_i - \mu)^2}{N}$$

- Standard Deviation (Population)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \mu)^2}{N}}$$

The Variance and Standard Deviation (Sample)

- Variance (Sample)

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

- Standard Deviation (Sample)

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Rule: Sample v/s Population

Any set of data should be considered as a **Sample** until it is clearly specified that data is the whole **Population**

Alternative Formulae

- Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}$$

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}$$

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Alternative Formulae

- Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}}$$

Summarizing Quantitative Data

Measures used to summarize quantitative data

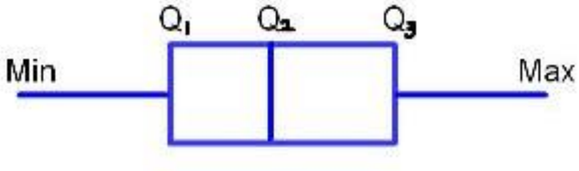
- Measures of Central Tendency
- Measures of Variation / Dispersion
- **Measures of Position**
- **Measures of Quality and Outliers**
  - The Percentiles
  - The Deciles
  - The Quartiles
  - The z-score

The Percentiles, Deciles and Quartiles

- Percentiles
  - Divide the data set, in order of magnitude, into 100 parts
  - Hence 99 percentiles can be determined
- Deciles
  - Divide the data set, in order of magnitude, into 10 parts
- Quartiles
  - Divide the data set, in order of magnitude, into 4 equal parts, each a quartile

Characteristics of Quartiles

- Quartiles help us to identify the following
  - Min, 25th Percentile, Median, 75th Percentile, Max



- Inter Quartile Range : Q3 – Q1
  - Range of the middle 50% of the data set
- IQR is resistant to extreme values
  - Variance and Standard Deviation are not
- **Quartiles can help identify ‘outliers’ by defining the ‘fences’**
  - Lower Fence = Q1 – 1.5 * IQR
  - Upper Fence = Q3 + 1.5 * IQR

Exercises

- Using “R” load the file tempdata.csv into the R variable “tempdata”
- Using tempdata do the following:
  - Create:
    - dotchart, stem plot, histogram, boxplot and interpret the results
  - Use the following R functions and interpret the results
    - Quantiles, IQR, var, sd, range, summary

Population v/s Sample: Exercise

- Create a dataset with 10000 observations

Population v/s Sample: Exercise

- Create a dataset with 10000 observations
- By the method of “Random Sampling” create three sample sets of 100 observations each

Population v/s Sample: Exercise

- Create a dataset with 10000 observations
- By the method of “Random Sampling” create three sample sets of 100 observations each
- For each sample:
  - Calculate the measures of “Descriptive Statistics”
  - Tabulate your observations for each set

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Population v/s Sample: Exercise

- Create a dataset with 10000 observations
- By the method of “Random Sampling” create three sample sets of 100 observations each
- For each sample:
  - Calculate the measures of “Descriptive Statistics”
  - Tabulate your observations for each set
- Calculate the measures of “Descriptive Statistics” for the population

Population v/s Sample: Exercise

- Create a dataset with 10000 observations
- By the method of “Random Sampling” create three sample sets of 100 observations each
- For each sample:
  - Calculate the measures of “Descriptive Statistics”
  - Tabulate your observations for each set
- Calculate the measures of “Descriptive Statistics” for the population
- Compare the descriptive measures calculated for each sample set with those of the population
  - Record and explain your observations

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Descriptive Statistics: Exercise

- Repeat the experiment by employing “Systematic Sampling” instead of random sampling
  - What are your observations?