BTS | Barcelona Technology School

Statistical Foundations for DS MBDS 2019

Alex Kumenius – Business Intelligence Data Scientist
Date : October 2019

INTRODUCTION TO DATA

Statistics is the study of how best to **collect**, **analyze**, and **draw conclusions** from `data` .

Statistics is a general *process of investigation* :

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Statistics focuses on making stages 2-4 objective, rigorous, and efficient. Statistics has three primary components :

- How best can we collect data?
- How should it be analyzed?
- And what can we infer from the analysis?

OBSERVATIONS, CASES, VARIABLE AND DATA MATRICES

Effective presentation and description of data is a first step in most analyses.

Data Matrices are a convinient way to record and store data.

Data can be represented as a **data matrix**, which is a common way to organize data.

Each **row** of a data matrix table corresponds to a unique observation or case also called "unit of observation" or an "observational unit". The **columns** represent characteristics, called Variables for each observation.

It is especially important to ask clarifying questions to ensure important aspects of the data are understood, what each variable means and the units of measurement -- ***(Must request for a Codebook)***

In [ ]:
```
# import libraries needed
import os
import pandas as pd
import matplotlib.pyplot as plt
```

In [ ]:
```
# Changing to the Dataset's working directory
os.chdir("D:\\Documents\\EureCat\\Eurecat 2019\\BTS\\Datasets\\")
```

In [ ]:
```
# Check current directory
os.getcwd()
```

In [ ]:
```
# Listing Datasets files
os.listdir()
```

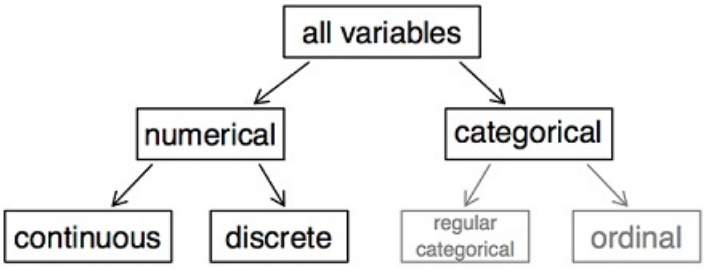
In [ ]:
```
# Open the choosen file
county = pd.read_csv('county.txt', sep='\t', encoding='utf-8')
# show 1st, 5 observations
county.head()
```

In [ ]:
```
county.shape
```

TYPES OF VARIABLES

Generally, in Math and Statistics variables may be Numerical or Categorical Variables.

A Variable is a quantity whose value changes.



NUMERICAL

A Numerical variable can take a wide range of numerical values, and it is sensible to *add, subtract*, or take *averages* with those values. On the other hand, we would **not** classify a variable reporting *"telephone area codes"* as numerical since there is **no** sense to *average, sum*, and *difference*.

Discrete Variables

A **Discrete variable** is a variable whose value is obtained by counting.

Over a particular range of real values ($\mathbb{R}$) is any value in the range that the variable is permitted to take on, there is a positive minimum distance to the nearest other permissible value. The number of permitted values is either ***finite or countably infinite***.

Common examples are variables that must be *integers, non-negative integers, positive integers*, or *only the integers 0 and 1*.

Examples:

- number of students present
- number of red marbles in a jar
- number of heads when flipping three coins
- students' grade level

Continuous Variables

A **Continuous variable** is a **variable whose value is obtained by measuring.**

A Continuous variable is one which can take on infinitely many, uncountable values.
For example, a variable over a non-empty range of the real numbers ($\mathbb{R}$) *a* and *b* is continuous, if it can take on *any value in that range*. The reason is that any range of real numbers between *a* and *b* with $a, b \in \mathbb{R}; \quad a \neq b$ is infinite and uncountable.

Examples:

- **height of students in class**
- **weight of students in class**
- **time it takes to get to school**
- **distance traveled between classes**

In [ ]:
```
# show 1st, 5 observations
county.head()
```

CATEGORICAL

A Categorical Variable takes on a limited, and usually fixed, number of possible values, categories; and the possible values are call the *variable's levels.*

Categorical Variables where their levels have a natural order are "Ordinal Variables".

Categorical Variables without this type of special ordering is called "Nominal Variable".

Examples are *gender, social class, blood type, country affiliation, observation time or rating via Likert scales.*

Categorical data might have an order (e.g. *'strongly agree' vs 'agree'* or *'first observation' vs. 'second observation'*), but numerical operations (additions, divisions, ...) are not possible.

In [ ]:
```
# show 1st, 5 observations
county.head()
```

The categorical data type is useful in the following cases:

- **A string variable consisting of only a few different values. Converting such a string variable to a categorical variable will save some memory.**
- **The lexical order of a variable is not the same as the logical order ("one", "two", "three"). By converting to a categorical and specifying an order on the categories, sorting and min/max will use the logical order instead of the lexical order.**
- **As a signal to other Python libraries that this column should be treated as a categorical variable (e.g. to use suitable statistical methods or plot types).**

In [ ]:
```
county.dtypes
```

RELATIONSHIPS BETWEEN VARIABLES

Analyses are motivated by *looking* for a *relationship between two or more variables*.

To answer research questions, data must be collected.

Examining Summary Statistics could provide insights for each of the research questions about the study.

A Summary Statistics is a *single number* summarizing a *large amount of data*. In other words, a Summary Statistics is a value computed from the data.

In [ ]:
```
county.head()
```

In [ ]:
```
# Concise Summary of de DataFrame
county.info()
```

In [ ]:
```
# Mean and Standard Deviation
county['fed_spend'].mean(), county['fed_spend'].std()
```

In [ ]:
```
# Summary of Descriptive Statistics - "fed_spend" Variable.
county.fed_spend.describe()
```

In [ ]:
```
# Summary of Descriptive Statistics - "fed_spend" Variable.
county.describe()
```

Additionally, graphs can be used to visually summarize data and are useful for answering such questions as well.

In [ ]:
```
# Create data
x = county.multiunit
y = county.homeownership
colors = 'Blue'

area = np.pi*5
plt.axis([-1, 110, 0, 95])

# Plot
plt.scatter(x, y, s=area, alpha=0.4, c=colors, edgecolor='black')
plt.title('Homeownership vs Multi-Unit')
plt.ylabel('% of Homeownership')
plt.xlabel('% of Units in Multi-Unit Structure')
plt.show()
```

Graphs can be used to study the relationship between two numerical variables (*Scatterplots*). After plotting two supposed related variable, we might brainstorm as to why this relationship exists and investigate each idea to determine which is the most reasonable explanation.

In [ ]:
```
# Create data
x = county.poverty
y = county.fed_spend
colors = 'Blue'

area = np.pi*5
plt.axis([0, 60, 0, 40])

# Plot
plt.scatter(x, y, s=area, alpha=0.4, c=colors, edgecolor='black')
plt.title('Federal Spending Per Capita vs Poverty')
plt.ylabel('Federal Spending Per Capita')
plt.xlabel('Poverty Rate (%)')
plt.show()
```

When two variables show some connection with one another, they are called Associated Variables, better known as Dependent Variables. Variables are said to be *associated* because the plot shows a perceivable pattern. Dependig on the plot pattern, variables may be negatively associated or positive association

If two variables are Not Associated then they are said to be Independent Variables. Two variables are independent if there is no evident relationship between the two.

No pair of variables is both associated and independent

.

Statistical Foundations for DS MBDS 2019

Alex Kumenius – Business Intelligence Data Scientist

Date : October 2019

OVERVIEW OF DATA COLLECTION PRINCIPLES

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider how data are collected so that they are reliable and help achieve the research goals.

Populations and Samples

Consider the following three research questions:

1. **What is the average mercury content in swordfish in the Atlantic Ocean?**
2. **Over the last 5 years, what is the average time to complete a degree for Duke undergraduate students?**
3. **Does a new drug reduce the number of deaths in patients with severe heart disease?**

Each research question refers to a target population, and each observation a case. Generally, it is too expensive to collect data for every case in a population. Instead, a Sample is taken.

A Sample represents a subset of the cases and is often a small fraction of the population. For instance, 60 swordfish in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

Example :

For the three research questions above, identify the target population and what represents an individual case.

- **(1) In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case**
- **(2) Notice that the first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergraduate students who have graduated in the last five years represent cases in the population under consideration. Each such student would represent an individual case.**
- **(3) The population includes all people with severe heart disease. A person with severe heart disease represents a case.**

Anecdotal Evidence

Consider the following possible responses to the three research questions:

1. **A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.**
2. **I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.**
3. **My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.**

Each conclusion is based on data. However, there are two problems :

- **First, the data only represent one or two cases.**
- **Second, and more importantly, it is unclear whether these cases are actually representative of the population.**

Data collected with this lack of criteria are called anecdotal evidence.

Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

Sampling from a population

Sampling is a method that allows researchers to infer information about a population based on results from a subset of the population, without having to investigate every individual.

Reducing the number of individuals in a study reduces the cost and workload, and may make it easier to obtain high quality information, but this has to be balanced against having a large enough sample size with enough power to detect a true association.

Bias in sampling

It is important that the individuals selected are representative of the whole population. This may involve specifically targeting hard to reach groups.

For example, if the electoral roll for a town was used to identify participants, some people, such as the homeless, would not be registered and therefore excluded from the study by default.

Why pick a sample randomly? ***Why not just pick a sample by hand?***

Example 1

Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study.

- ***What kind of students do you think she might collect?***
- ***Do you think her sample would be representative of all graduates?</i>***

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be well-representative of the population.

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional.

When selecting samples by hand, we run the risk of picking a biased sample - bias into a sample, even if that bias is unintentional or difficult to discern.

Sampling randomly helps resolve this problem.

There are five important potential sources of bias that should be considered when selecting a sample, irrespective of the method used.

Sampling bias may be introduced when:

1. ***Any pre-agreed sampling rules are deviated from***
2. ***People in hard-to-reach groups are omitted***
3. ***Selected individuals are replaced with others, for example if they are difficult to contact***
4. ***There are low response rates***
5. ***An out-of-date list is used as the sample frame (for example, if it excludes people who have recently moved to an area).***

> ***SIMPLE RANDOM SAMPLE*** *The most **basic random sample** is called a Simple Random Sample, each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample. Taking Simple Random Sample helps minimize bias.*

Although Simple Random Sample helps minimize bias, bias can crop up in other ways. Even when cases are picked up at random, caution must be exercised if the non-response bias is high, and can skew results. Another common downfall is a convenience sample, where individuals who are easily accessible are more likely to be included in the sample.

It is often difficult to discern what sub-population a convenience sample represents

.

Explanatory and response variables

Consider the following question from $county$ ***dataset :***

- ***Is federal spending, on average, higher or lower in counties with high rates of porverty ?***

If we suspect poverty might affect spending in a county, then poverty is the explanatory variable and federal spending is the response variable in the relationship. If there are many variables, it may be possible to consider a number of them as explanatory variables.

Sometimes the explanatory variable is called the independent variable and the response variable is called the dependent variable. However, this becomes confusing since a pair of variables might be independent or dependent, so we avoid this language.

To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropiate analysis.

```
explanatory     might affect          response
 variable    ------------------->     variable
```

> ***ASSOCIATION DOES NOT IMPLY CAUSATIOM*** *Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other. Even in some cases, there is no explanatory or response variable.*

In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

Introducing Observational Studies and Experiments

There are two primary types of data collection : Observational Studies and Experiments.

- *Observational Studies when data is collected in a way that does not directly interfere with how the data arise. We merely observe the data that arise.*
- *Observational Studies can provide evidence of naturally ocurring association between variables, but they cannot by themselves show a causal connection.*

When we want to investigate the possibility of a causal connection, we will conduct an Experiment. And ususally there will be both an explanatory and a response variable.

When individuals are randomly assigned to a group, the experiment is called a randomized experiment.

OBSERVATIONAL STUDIES

Generally, data in Observational Studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned to each subject by the researchers.

Making causal conclusions based on Experiment is often reasonable</u>. However, making the same causal conclusions based on observational data can be misleading and is not recommended. Thus, observational studies are generally only sufficient to show associations.

Confounding variable is a variable that is correlated with both the explanatory and response variable. However, there is no guarantee that all confounding variables can be examined or measured.

The $county$ dataset is an observational study with confounding variables, and its data cannot easily be used to make causal conclusions.

Observational studies come in two forms:

1. *Prospective studies identifies individuals and collects information as events unfold.*
2. *Retrospective studies collect data after events have taken place.*

Some datasets may contain both prospectively and retrospectively-collected variable, such as county datasets, Local governments prospectively collect some variables as events unfolded (e.g. retails sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

EXPERIMENTS

Studies where the researchers assign treatments to cases are called experiments.

When this assignment includes randomization, e.g. using a coin flip to decide which treatment a patient receives, it is called a Randomized experiment. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

Principles of Experimental Design

Randomized experiments are generally built on four principles :

1. ***Controlling. Researchers assign treatments to cases, and control any other differences in the groups.***

For example:

When patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

1. ***Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled.***

For example :

Some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such di↵erences, and it also prevents accidental bias from entering the study.

1. ***Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response.***

In a single study, we replicate by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

1. ***Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influence the response.***

Under these circumstances, they may first group individuals based on this variable into blocks and then randomize cases within each block to the treatment groups. This strategy is often referred to as blocking.

For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

It is important to incorporate the first three experimental design principles into any study