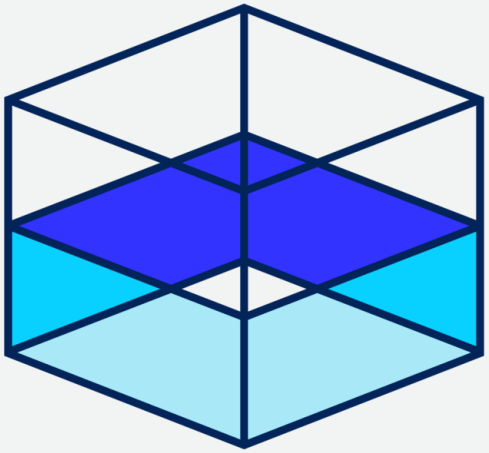


TECH
TALENT
SOUTH

Sources and Types of Data

TECH
TALENT
SOUTH

Updated: 10/09/20

Table of Contents

TECH TALENT SOUTH

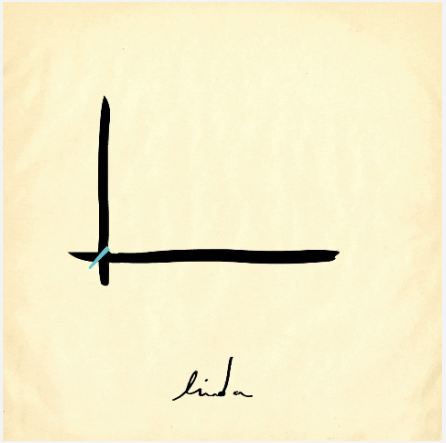
Intro to Data Science

Types of Data

Two types of data:

1. Quantitative
2. Qualitative

Within each types, there are multiple subtypes.



TECH TALENT SOUTH

Intro to Data Science

Quantitative Data

Things that are measured *objectively*.

Numerical values.

Two Types: Discrete and Continuous

TECH TALENT SOUTH

Intro to Data Science

Discrete Data

Discrete data, can only take in certain values. Usually, these are values that can be counted, like the number of students in a class. In this case, we could have any whole number of students. (We couldn't have half a student!)

TECH TALENT SOUTH

Intro to Data Science

Discrete Data

Counts - variables representing frequency of an occurence of an event

- Number of people in a school
- Number of people who voted on a bill

Proportions - also known as bounded counts are the ratios of counts

- Number of students in a school divided by the number of teachers in a school
- Number of people who voted "Yes" on a bill

TECH TALENT SOUTH

Intro to Data Science

Continuous Data

Continuous data is an unfixed number of possible measurements between two realistic points. The data can be any number is not restricted like discrete data is.

Continous data often contain decimal points and can provide great detail. It is also usually contains numbers within an expected range.

TECH TALENT SOUTH

Intro to Data Science

Continuous Data

Continuous data is an unfixed number of possible measurements between two realistic points. The data can be any number is not restricted like discrete data is.

Continous data often contain decimal points and can provide great detail. It is also usually contains numbers within an expected range.

TECH TALENT SOUTH

Intro to Data Science

Continuous Data

Based off of that brief description, what are some examples of continuous data you can think of?

TECH TALENT SOUTH

Intro to Data Science

Continuous Data

Possible examples include:

- A person's height
- A person's weight
- The temperature
- Inches of rain

TECH
TALENT
SOUTH

Intro to Data Science

Interval Data

Ordered units with the same difference. For example, describing the temperature from a list of options such as:

-10, -5, 0, 5, 10

Interval data does not have a "true zero." (In the above example, there is no option for "no temperature.")

We can add and subtract, but cannot multiply or divide to calculate ratios.

TECH TALENT SOUTH

Intro to Data Science

Ratio Values

Ordered units with the same difference, but have a "true zero." For example, the height of a tree.

Interval data does not have a "true zero." (In the above example, there is no option for "no temperature.")

We can add and subtract, but cannot multiply or divide to calculate ratios.

Intro to Data Science

Discrete Data

Discrete data can be measured in different ways: ordered or unordered.

- Nominal Variables (Unordered): gender, location, religion, etc.
- Ordinal (ordered) variables: grade levels, income brackets
- Continuous variables: grouped into a small number of categories (intervals) - income grouped into subsets, blood pressure levels (normal, high-normal etc)

Intro to Data Science

Measurement Hierarchy

nominal < ordinal < interval

Methods applicable to a lower type of variable can be used for a higher one, but not the other way around.

For example, you could use methods designed for nominal data for interval data, but not methods designed for interval data with nominal data.

TECH TALENT SOUTH

Intro to Data Science

Measurement Hierarchy

What types of variables could be used to answer the following question?

Have you studied abroad?

What is your interest level in data science: low, medium, or high?

Based on your test score, what letter grade did you get?

TECH TALENT SOUTH

Intro to Data Science

Measurement Hierarchy

What types of variables could be used to answer the following question?

Have you studied abroad? **binary nominal**

What is your interest level in data science: low, medium, or high? **ordinal**

Based on your test score, what letter grade did you get? **interval**

TECH TALENT SOUTH

Intro to Data Science

Discuss

Why does the measurement hierarchy matter, and how does it affect data analysis?

What are some uses for discrete data that you can think of at this point?

TECH
TALENT
SOUTH

Intro to Data Science

Contingency Tables

Used to summarize discrete data. Contains at least two categories and data to compare the results of the counts.

Preference: Dogs or Cats

	Male	Female	Total
Dog	20	15	35
Cat	12	15	27
Total	32	30	62

Intro to Data Science

Contingency Tables

What can we determine from the table in the previous slide? What other categories could we add to get a deeper analysis?

Preference: Dogs or Cats

	Male	Female	Total
Dog	20	15	35
Cat	12	15	27
Total	32	30	62

Intro to Data Science

Qualitative Data

Qualitative data is used to categorize. It is not numerical in nature.

This includes data from interviews, focus groups, and observational studies.

Even though it is does not provide concrete numerical information, it can still be very useful.

Intro to Data Science

Qualitative Data

Qualitative data is used to categorize. It is not numerical in nature.

This includes data from interviews, focus groups, and observational studies.

Even though it is does not provide concrete numerical information, it can still be very useful.

Intro to Data Science

Qualitative Data

What are some examples of qualitative data you can think of?

TECH TALENT SOUTH

Intro to Data Science

Qualitative Data

There are many ways to gather qualitative data, including:

1. Interviews: Researchers ask questions and keep track of the results

2. Focus Groups: Groups are picked out by a researcher, often within a similar demographic, and are asked questions while reactions and feedback are recorded.

3. Observation: Researchers observes settings where respondents are and records relevant information.

Intro to Data Science

Qualitative Data

4. Longitudinal Studies: Data collection from the same source over an extended period of time.

5. Case Studies: An individual occurence or event is studied in depth.

TECH TALENT SOUTH

Intro to Data Science

Qualitative Data

Deductive Approach

- Based on predetermined structures to analyze the data. It is usually used when the researcher has a generally knowledge of the expected results of the study.

Inductive Approach

- Is not based on any predetermined structures or prior knowledge. It is used when the researcher has little knowledge of the subject and its expected outcome.

Intro to Data Science

Qualitative Data

Advantages

- **Helps with in-depth analysis** - subjects can be asked questions so very specific data can be obtained
- **Rich data** - because the data is not restricted to numbers, the results can cover a wide-range of topics, making them useful for future studies

Intro to Data Science

Qualitative Data

Disadvantages

- **Time consuming** - it takes much longer and is more expensive to perform a qualitative test/often a smaller sample size must be used
- **Hard to Generalize** - smaller sample sizes make it harder to draw broad conclusions
- **Skill-Dependen** - quality of gathering data depends on researchers ability to interview and observe

Intro to Data Science

Bias

Regardless of type, it is important to avoid bias when sourcing data. Having a sample size that does not accurately reflect the target population can skew data and produce incorrect results.

Let's discuss a few common types of bias.

Intro to Data Science

Bias

Regardless of type, it is important to avoid bias when sourcing data. Having a sample size that does not accurately reflect the target population can skew data and produce incorrect results.

Let's discuss a few common types of bias.

Intro to Data Science

Selection Bias

This occurs when a sample population does not reflect the true population.

For example, say you want to research the effects of a new heart medication. When selecting your candidates, you select those who already have other preexisting conditions.

If your research shows the heart medications causes complications, you will now be unsure whether it was a result of the medication or the preexisting condition.

TECH TALENT SOUTH

Intro to Data Science

Selection Bias

Another common example is when candidates are able to self volunteer.

For example, say you are testing the efficacy of a new diet shake and offer it for free to test subjects who volunteer to try it to lose weight.

The subjects who volunteer are most likely already concerned about their health, and are likely to already be exercising regularly or doing other activities to care for their health.

Those who volunteered are a specific group of people who are not representative of the population as a whole.

TECH TALENT SOUTH

Intro to Data Science

Selection Bias

To avoid selection bias, researchers attempt to cast a wide net, testing a large group of people from various backgrounds or communities, seeking to elminiate bias from the start.

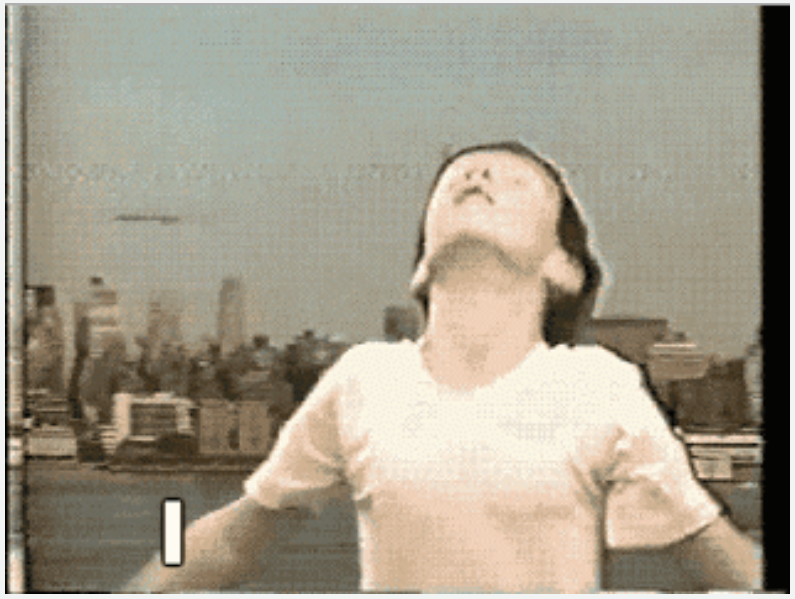
They also will randomize the experimental and control groups.

Still, some bias is often unavoidable and can be difficult to avoid.

Intro to Data Science

Non-Response Bias

If you attempt to poll a large number of people about a topic, many people will opt to not respond. Only those passionate about a topic one way or the other will respond, leading to a loud minority dictating the results.



TECH TALENT SOUTH

Intro to Data Science

Social Desirability Bias

Subjects may be prone to answer what is considered socially acceptable, but not what they truly believe.

Indirect and non-personal questions can help to avoid this. People will be more likely to answer truthfully.

Intro to Data Science

Bias

What are some examples of bias that you can think of?

How could you avoid them?

TECH
TALENT
SOUTH