|  |  |
| --- | --- |
| **Data Scientist** | **Data Engineers** |
| Optimize data processing | Optimize data flow |
| Define metrics |  |
| Establish collection method |  |
| Work with enterprise system |  |

|  |  |
| --- | --- |
| **Data Scientist** | **Statisticians** |
| Large scale data collection and cleaning | Traditional data collection; surveys, polls, and experiments |
| Create ML models |  |
| Implement algorithm |  |

|  |  |
| --- | --- |
| **Data Scientist** | **Business Analyst** |
| Help BA on automate reports and data extraction | Focus on Database Design, ROI assessment |
|  | Some work on finance planning, optimization, and risk management |

Data Science Life Cycle:

1. Define problem
2. Clean the data
3. Exploratory data analysis
4. Predict/infer solutions from the data

# Data Design

1. Reducing bias with probability sampling
   1. Simple random sampler (SRS)

A simple random sample is taken by sampling at random without replacement. To take a simple random sample of Size 2 from the population of eight people, you can write each letter from A through H on a separate index card, place all the cards into a hat, mix the cards as well, and draw two cards from the hat without looking. Here are all possible. simple random samples of Size 2 from the population of 8 people.

A screenshot of a computer

Description automatically generated

So when selecting a sample, you could get any of these samples with equal chance.

* 1. Cluster Sampling

A cluster sample is taken by dividing the population into clusters and then using SRS to select clusters at random. To take a cluster sample of Size 2 from the population of 8 people, you can pair up each person to form four clusters of 2 people per cluster and then use SRS to select one cluster. The selected cluster is a sample of Size 2. Suppose you pair up the individuals like this, those are your clusters.

A group of people in circle frames

Description automatically generated

So when selecting a cluster, you could get either AB, CD, EF, or GH with equal chance.

A white background with black text

Description automatically generated

* 1. Stratified sampling

A stratified sample is taken by dividing the population into strata and then producing one simple random sample per stratum. To take a stratified sample of Size 2 from the population of 8 people, you would first divide the population into 2 Strata. Strata 1 would contain A, B, C, D, and E, and Strata 2 would contain F, G, and H.



Then you would use SRS to select one individual from each strata. This results in a sample of Size 2. Here are the possible samples you could get.

A group of people in different poses

Description automatically generated

Now, you might have noticed that the strata in this example are different sizes. In stratified sampling, the strata do not have to be the same size, and that can be an advantage. For example, you could stratify the US by occupation, then take the samples from each strata of size proportional to the distribution of occupations in the US. This would ensure that occupations that are not as common would still show up in a stratified sample. A simple random sample might completely miss people in such occupations. So stratified sampling helps you ensure that subgroups of the population are well represented in the sample, which can lessen variation in estimation. However, stratified sampling can sometimes be more difficult to perform because sometimes you don't know how large each strata is.

1. Using non-probability sampling

Sometimes, data collection can be very expensive and time-consuming, and you may not have access to all the members of a population. In that case, you would not be able to take a random sample. Instead, you could use a different technique called non-probability sampling, which is a subjective, in other words, non-random method. With this method, there is no guarantee that every individual in the population has a chance of being included in the sample. So this method has higher risks of sampling bias, but it's also easier and less expensive to implement. It's often used in qualitative or exploratory research, where the goal is to gain an initial understanding of a specific phenomenon. In this lesson, I will introduce key types of non- probability sampling with examples of each.

* 1. Volunteer sampling

With this approach, volunteers make up the sample. For example, if a researcher wants to learn more about the experience of having a particular disease, they could circulate a survey where the eligibility requirement is that the participant has the disease. Those who meet that criteria and want to participate voluntarily share their responses to the survey questions.

* 1. Purposive sampling

With this approach, a researcher uses their best judgment to select a sample that they think is a best fit. For example, if a researcher wants to learn more about the experiences of employees with disabilities, the sample could be intentionally selected to suit that goal.

* 1. Quota sampling

Another method is quota sampling, which is one of the most common forms of non-probability sampling. With this approach, sampling is done until a specific number of units, in other words, quotas, for various subpopulations have been selected. This helps in achieving sample size objectives and ensuring that members of different subpopulations are included. For example, if a researcher wants to learn more about consumer preferences of individuals across different socioeconomic groups, they could use quota sampling.

* 1. Snowball sampling

In snowball sampling, participants recruit other participants to join the sample. This is used when it's hard to directly reach members of a population with the required characteristics. For example, if a researcher wants to learn more about the musical techniques of artists in niche genre, they may consider using snowball sampling. If you need to collect data with limited resources for a time sensitive project and probability sampling is not an option, you can consider using one of these non-probability sampling methods.

# Python vs R

|  |  |
| --- | --- |
| **Python** | **R** |
| Data analysis integrated with web applications | Data analysis tasks require standalone computing or analysis on individual servers |
| Statistical code needs to be incorporated into a production database. | Better user-friendly data analysis statistics and graphical models |
| Productivity and code readability. | Used primarily in academics and research |
| Used by programmers for data analysis or statistic and by developers that turn to data science | Complex syntax and steep learning curve |
| Easy syntax and easy to learn |  |

# Structuring tabular data

A data frame is a tabular data structure where each column is labeled and each row is labeled.

# Exploratory Data Analysis

Data types:

1. Nominal data:

No inherent order

Example: political party, computer operating system such as Windows, macOS, Linux, gender, languages.

1. Ordinal data:

Data with ordered categories

Example: clothing size such as small, medium, large, extra large or highest level of education, such as high school, undergrad, master's, Ph.D, or even a Yelp rating on a scale from 0-5 stars.

1. Numerical data:

Data consist of amounts or quantities

Example: height, weight, price, distance, blood pressure.

Main properties of data:

1. Granularity

What each record in the data represents. Consider asking how fine or coarse is the data.

1. Scope

The scope of the data set refers to the coverage of the data set in relation to what you're interested in analyzing. What does the data describe? Does the data cover the topic you're interested in?

1. Temporality

This refers to how the data is situated in time and specifically to the date and time fields in the data set. When was the data collected?

1. Faithfulness

How accurately does a data describe the real world? Should you trust this data?

# Data Cleaning

Data cleaning detecting and fixing corrupt or inaccurate records from a record set

Data cleaning addresses:

1. Missing values
2. Formatting of values
3. Structure of the data overall
4. Extracting information from complex values
5. Unit conversion
6. Interpretation of magnitudes

Data cleaning example:

1. Missing data
2. Misspellings
3. Duplicated rows
4. Inconsistent formats
5. Outliers or extreme values
6. Unspecified units

# Inference and Statistical Analysis

Statistical inference: the process of using data analysis to deduce properties of an underlying probability distribution.

Statistical inference methods:

1. Hypothesis tests
2. Confidence intervals

**Hypothesis tests**:

“Let's say I have a fictional data set that represents a sample of avocado trees. I have a new fertilizer, and I want to test if the new fertilizer has an effect on how long the avocado takes to grow. I have data on a sample of avocado trees, and for each tree, the data set contains the amount of time it took for the avocado to grow and whether the new fertilizer was used.”

Null hypothesis:

1. Usually no associations between variables
2. Usually attributes observed trends to random chance

Avocado null hypothesis:

“In the population, the distribution of growth duration of avocado trees that receive fertilizer is the same as that of avocado trees that did not receive fertilizer. If the two distributions are different in the sample, it's due to random chance.”

Alternative hypothesis:

Usually attributes trends observed in the data to associations between variables.

Avocado alternative hypothesis:

“In the population, the distribution of growth duration of avocado trees that receive fertilizer is different from that of avocado trees that did not receive fertilizer due to an association between growth duration and whether fertilizer was used. And on average, avocado trees that received fertilizer took a shorter period of time to grow than those that did not receive fertilizer.”

**Test Statistic**: helps decide between the two hypotheses.

Null hypothesis: on average, the growing time is the same for avocado trees that receive fertilizer and for those that did not.

Alternative hypothesis: on average, avocado trees that receive fertilizer took a shorter period of time to grow than those that did not receive fertilizer.

Test statistic:

The average growing time among avocado trees that receive fertilizer - the average growing time among avocado trees that did not receive fertilizer.

Smaller values of this test statistic would indicate that the alternative hypothesis is better supported, while larger values of this test statistic would indicate that the null hypothesis is better supported.

**Create a permutation**

Randomly permute or rearrange the data. Code on 08\_03.ipynb

**Conducting a permutation test**

Code on 08\_04.ipynb

P value represents the approximate probability that the observed value of the test statistic or a more extreme value shows up among the simulated values of the test statistic under the assumption that the null hypothesis is true.

Draw a conclusion and use a P value threshold of significance, also known as a P value cutoff. The most commonly used P value cutoffs are 5% and 1%

If p value < p value cut off, then reject null hypothesis

**Bootstraping**

* Simulates new random samples by resampling
* Consists of sampling at random

Code on 08\_05.ipynb

A diagram of a diagram

Description automatically generated