IA4: Explaining statistics terms to non-statisticians orally and in writing

Submit Part 1 to the IA4 assignment on the LMS by 8:00AM on 18 April. Share your IA4 with your teammates during class on 18 April.
Provide feedback to your teammates by 8:00AM on 23 April (send to your teammates but don't submit on LMS).
Revise Part 1 and submit to a combined team TA4 document on the LMS by on Friday, 26 April.

Explaining statistics to non-statisticians is an essential skill for statistical collaboration meetings with domain experts. It is also essential for work in industry. We all know that statistics and data science is really important and that not many people understand statistics or data science. It is our job to explain statistics and data science to the world.

Part 1

1. Explain **twelve** statistics or data science terms (per individual) in writing, **four** from **List 1**, **four** from **List 2**, and **four not on either list**. At least six of your explanations must include all **five aspects of ADEPT**. All six of the others must include **an example**.

2. One thing missing from the ADEPT method is "Relevance." **How** is this concept or topic **relevant** to the person you are explaining it to? **Add the relevance** to at least **three terms** to make them ADEPTR.

3. Chose at least **seven terms** to **explain orally** to a specific non-statistician (your grandmother, boyfriend, roommate, …). At least four of these terms must use the ADEPT (or ADEPTR) method. These terms can be the same as those you explained in writing.

List 1:
p-value
confidence interval
correlation
independence
reproducibility crisis
cross-validation

List 2:
random sample
standard deviation
hypothesis test
statistically significant
confounding
R-squared
parameter
linear regression
residual

2nd Generation p-value
Bonferroni's correction
nonparametric test
out-of-sample prediction

4. Write a **concise reflective paragraph** about what you learned doing this. Things to consider:
How useful was ADEPT?
How did explaining terms to a real person change your approach?
Did anything unexpected happen?
Do you think ADEPTR is more effective than ADEPT?

- ANSWER -

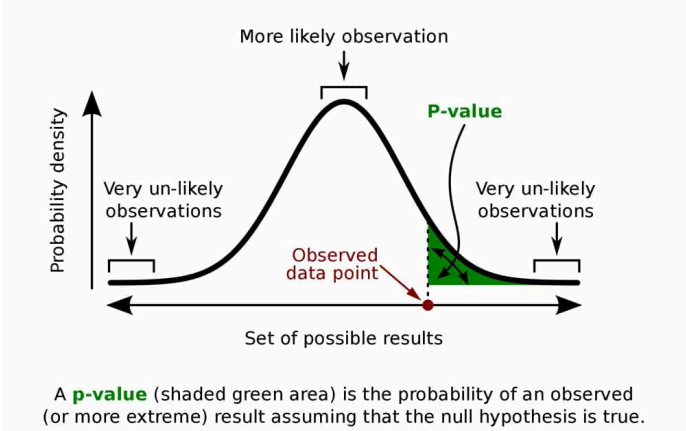
1. Explanations of 12 statistics or data science terms:

List 1 (4 terms):

a. p-value

Analogy: The p-value is like a measure of how surprised we would be if the null hypothesis were true and we still observed the results we did. A small p-value means we would be very surprised.

Diagram:



Example: Imagine you're testing whether a new drug is effective. The p-value tells you how likely it is that the observed difference in outcomes between the treatment and control groups would have occurred by chance if the drug was actually ineffective.

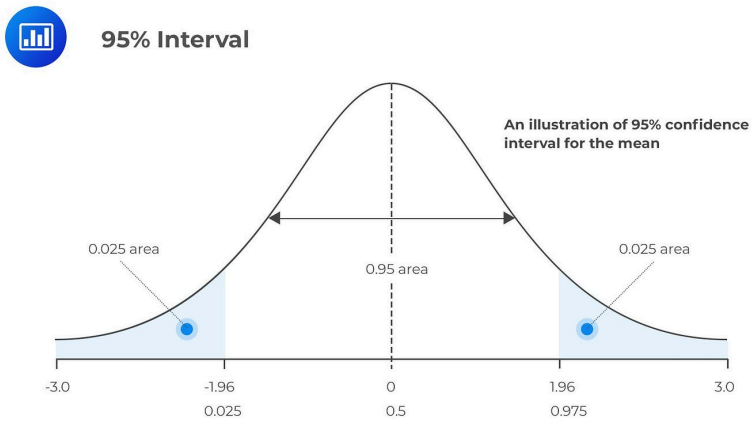
Plain-English Description: The p-value is a number that tells us how likely it is that the results we observed in our data would have happened by chance if the null hypothesis were true. A small p-value means the results are unlikely to have happened by chance, suggesting the null hypothesis is false.

Technical Description: The p-value is the probability of observing a test statistic at least as extreme as the one observed, assuming the null hypothesis is true. It is used to assess the statistical significance of the results.

b. confidence interval

Analogy: The confidence interval is like a margin of error around your estimate. It gives you a range that you can be reasonably confident contains the true value.

Diagram:



Example: Imagine you want to estimate the average income of a city. If you take a random sample of residents and calculate the sample mean income, the confidence interval tells you the range of values that likely contains the true average income for the entire city.

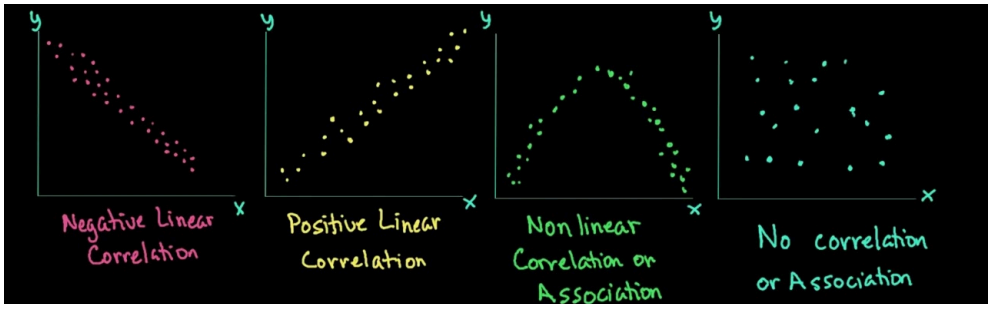
Plain-English Description:A confidence interval is a range of values that is likely to contain the true parameter of interest, based on the observed sample data. It gives a sense of the precision of the estimate.

Technical Description: A confidence interval is constructed using the sample statistic and an appropriate distribution, such that the interval has a specified probability (e.g. 95%) of containing the true parameter value.

c. correlation

Analogy: Correlation is like the strength of a friendship - the more closely two variables are related, the stronger the correlation between them.

Diagram:



Example: The correlation between hours of studying and exam scores is likely to be positive, meaning students who study more tend to score higher on exams.

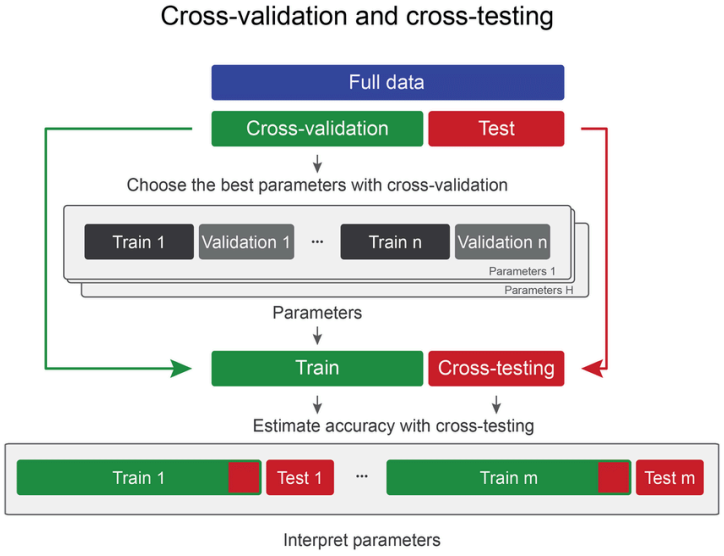
Plain-English Description: Correlation is a statistical measure that describes the strength and direction of the linear relationship between two variables. It tells us how much one variable changes when the other variable changes.

Technical Description: Correlation is measured by the correlation coefficient, which ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

d. cross-validation:

Analogy: Cross-validation is like testing a new recipe by making it for different groups of people, to make sure it tastes good to a wide audience, not just your family.

Diagram:



Example: Imagine you are building a machine learning model to predict house prices. Instead of just training and evaluating the model on the entire dataset, you would use cross-validation to repeatedly split the data into training and

testing sets, train the model on the training data, and evaluate it on the testing data. This helps ensure the model generalizes well to new, unseen data.

Plain-English Description: Cross-validation is a technique used to assess the performance of a machine learning model by splitting the available data into training and testing sets, and then iterating through multiple rounds of model training and evaluation. This helps ensure the model can generalize well to new, unseen data.

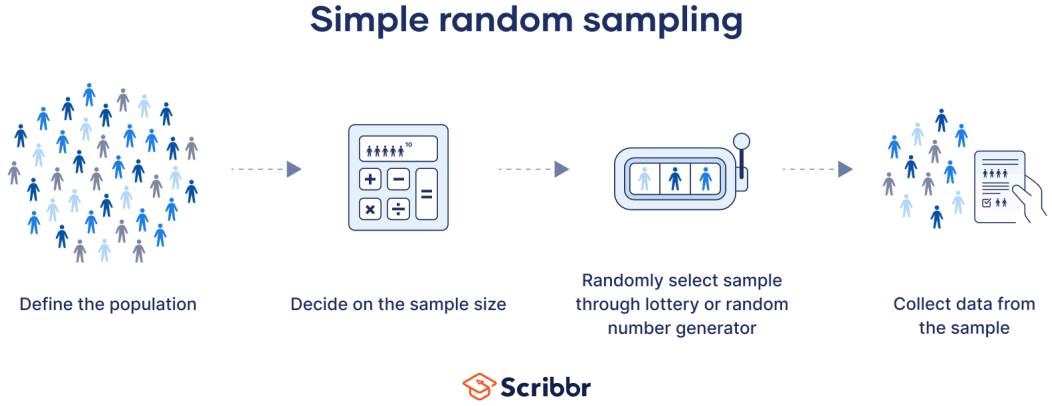
Technical Description: Cross-validation is a model validation technique that involves partitioning the data into complementary subsets, using one subset (the training set) to train the model, and the other subset(s) (the testing set(s)) to evaluate the model's performance. This process is repeated multiple times, with different partitions, to get a more reliable estimate of the model's generalization performance.

List 2 (4 terms):

a. **random sample**

Analogy: A random sample is like picking names out of a hat - everyone has an equal chance of being chosen, so the sample should be representative of the whole population.

Diagram:



Example: Imagine you want to know the average age of all the students at your university. Instead of surveying every single student, you could randomly select 100 students to survey, and use their ages to estimate the average age of all students.

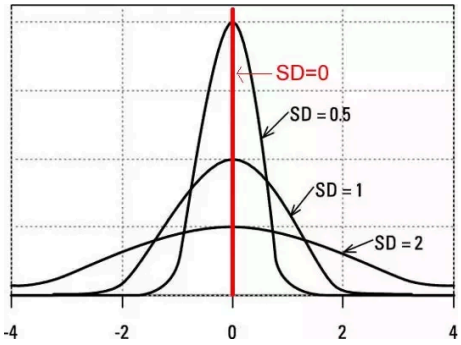
Plain-English Description: A random sample is a subset of a population that is chosen in such a way that every member of the population has an equal chance of being selected. This allows us to draw conclusions about the entire population based on the sample.

Technical Description: A random sample is a subset of a population selected in such a way that every member of the population has an equal probability of being chosen. This allows for unbiased statistical inferences about the population.

b. standard deviation

Analogy: The standard deviation is like measuring how much the individual family members' heights vary from the average height of the family. A low standard deviation means everyone is about the same height, while a high standard deviation means there is more diversity in heights.

Diagram:



Example: magine you have a dataset of exam scores. The standard deviation tells you how much the individual scores tend to differ from the average score. A low standard deviation means the scores are clustered close to the average, while a high standard deviation means the scores are more spread out.

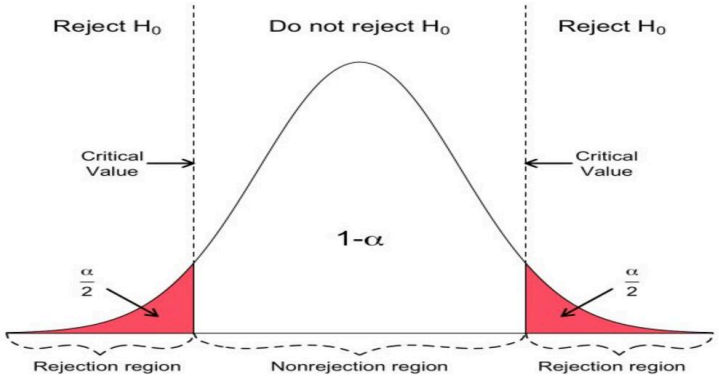
Plain-English Description: The standard deviation is a measure of how spread out the values in a dataset are from the mean or average value. It tells you how much the individual data points tend to deviate from the center.

Technical Description: The standard deviation is the square root of the variance, which is the average squared deviation of each data point from the mean. It quantifies the amount of variation or dispersion in a dataset.

c. hypothesis test

Analogy: A hypothesis test is like a courtroom trial, where the null hypothesis is the "innocent" verdict, and the alternative hypothesis is the "guilty" verdict. The p-value is the evidence that determines which verdict is more likely.

Diagram:



Example: Imagine you want to test if a new teaching method improves student test scores. The null hypothesis would be that there is no difference in scores, and the alternative hypothesis would be that the new method leads to higher scores. The p-value from the test would indicate how likely the observed difference in scores would be if the null hypothesis were true.

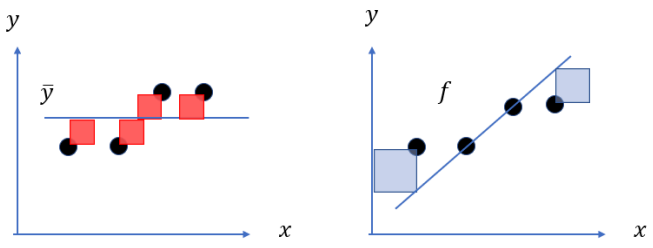
Plain-English Description: A hypothesis test is a statistical procedure used to determine whether there is enough evidence in the sample data to infer that a certain condition (the null hypothesis) is true for the entire population.

Technical Description: Hypothesis tests involve calculating a test statistic, which is then compared to a critical value from a probability distribution. The p-value is used to assess the statistical significance of the results and determine whether to reject or fail to reject the null hypothesis.

d. R-squared

Analogy: R-squared is like a measure of how well a map matches the actual terrain. The higher the R-squared, the better the map (model) represents the real-world (data).

Diagram:



$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Example: Imagine you are building a model to predict housing prices based on factors like square footage, number of bedrooms, and location. The R-squared value would tell you how much of the variation in housing prices is accounted for by the variables included in your model.

Plain-English Description: R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It indicates how well the model fits the observed data.

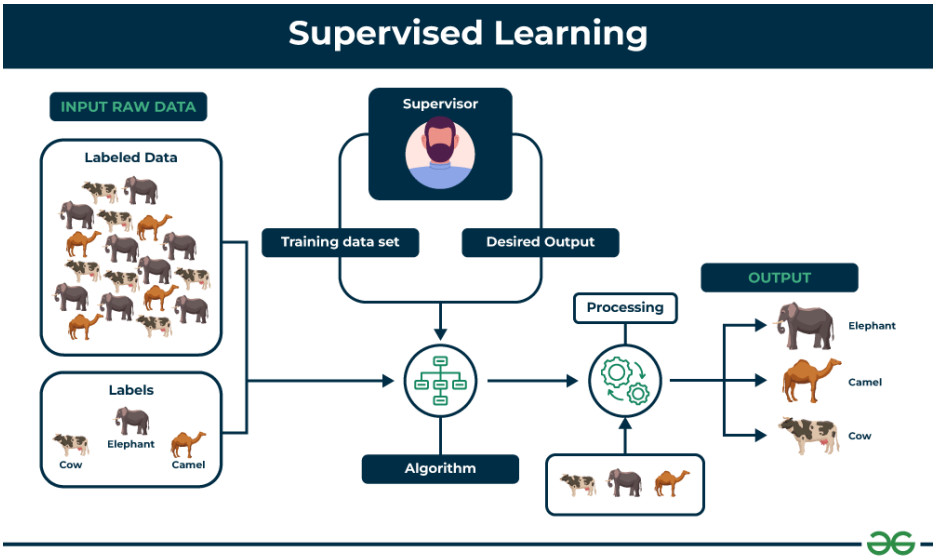
Technical Description: R-squared, also known as the coefficient of determination, is a value between 0 and 1 that represents the proportion of the variance in the dependent variable that is predictable from the independent variables in a linear regression model. It is a measure of how well the regression line approximates the real data points.

Outside of Lists 1 and 2:

a. Supervised Learning

Analogy: Supervised learning is like having a teacher who provides you with labeled examples, and you use those examples to learn how to make predictions.

Diagram:



Example: Imagine you want to build a model to predict whether an email is spam or not. You would provide the algorithm with a dataset of emails that have already been labeled as spam or not spam, and the algorithm would learn the patterns in the data to make predictions on new, unlabeled emails.

Plain-English Description: Supervised learning is a type of machine learning where the algorithm is trained on a dataset of labeled examples, and then used to make predictions on new, unseen data.

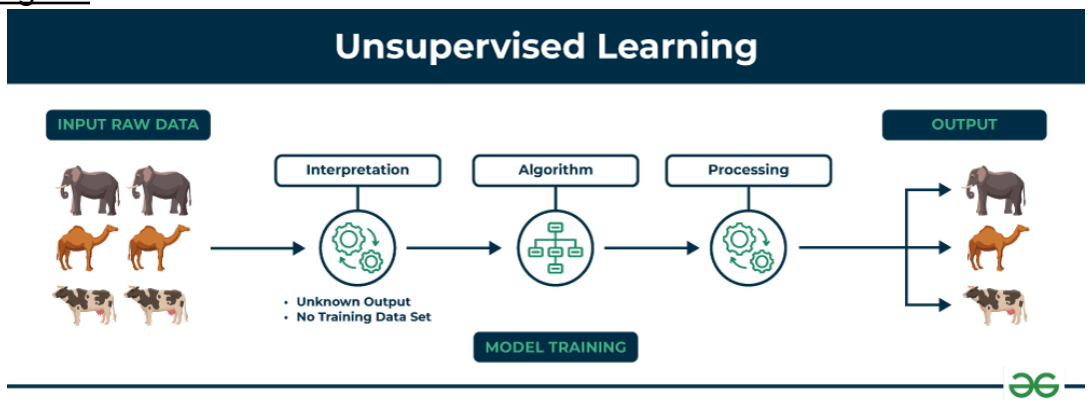
Technical Description: In supervised learning, the goal is to learn a function that maps input features to output labels, using a training dataset with known

input-output pairs. Common supervised learning tasks include classification and regression.

b. Unsupervised Learning

Analogy: Unsupervised learning is like exploring a new city without a map - you have to discover the patterns and structures on your own, without any guidance.

Diagram:



Example: Imagine you have a dataset of customer purchase histories, and you want to find natural groupings or segments of customers with similar buying patterns. Unsupervised learning algorithms like clustering can be used to discover these groups without any prior knowledge about the customer segments.

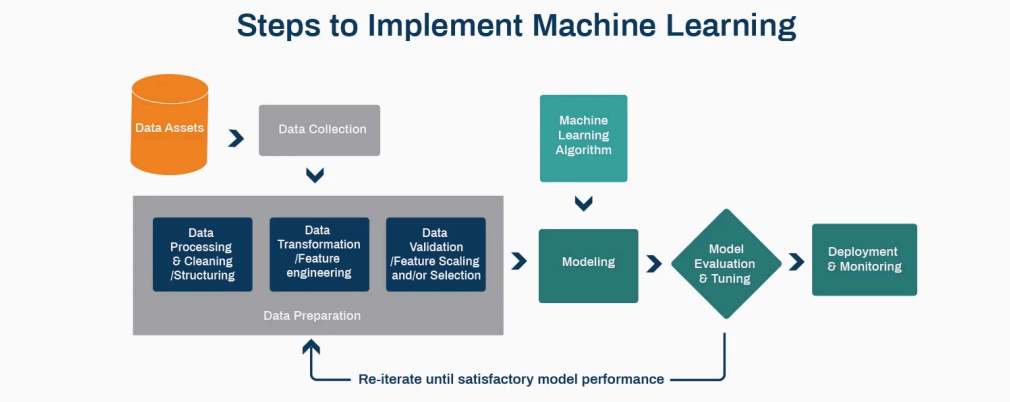
Plain-English Description: Unsupervised learning is a type of machine learning where the algorithm is given unlabeled data and tasked with finding inherent patterns or structures in the data, without any specific output variable or label.

Technical Description: In unsupervised learning, the goal is to discover hidden structures or patterns in data, without any predetermined output variables. Common unsupervised learning tasks include clustering, dimensionality reduction, and anomaly detection.

c. Machine Learning

Analogy: Machine learning is like a child learning to recognize different types of animals - the more examples they see, the better they get at identifying new ones.

Diagram:



Example: Imagine you want to build a model that can recognize handwritten digits. You would provide the machine learning algorithm with a large dataset of images of handwritten digits, along with their true labels. The algorithm would then learn the patterns in the data and be able to accurately classify new, unseen digit images.

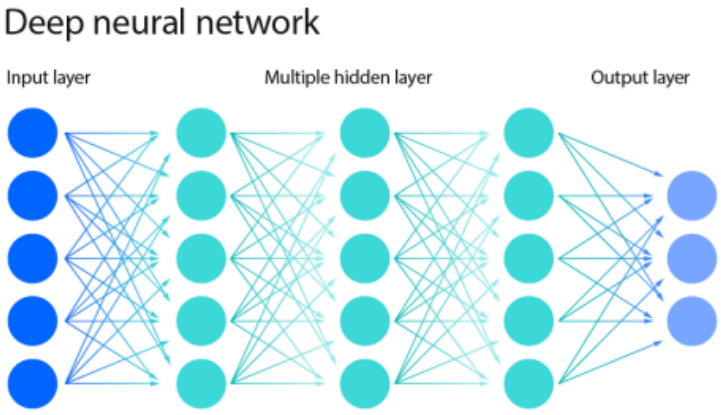
Plain-English Description: Machine learning is a field of artificial intelligence that involves the development of algorithms and statistical models that allow computers to perform specific tasks effectively without using explicit instructions, relying instead on patterns and inference in data.

Technical Description: Machine learning is the study of algorithms and statistical models that computer systems use to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is a subfield of artificial intelligence.

d. Deep Learning

Analogy: Deep learning is like the human brain, with multiple layers of interconnected "neurons" that can learn to recognize complex patterns in data.

Diagram:



Example: Imagine you want to build a model that can recognize faces in images. Deep learning algorithms, such as convolutional neural networks, can be trained on a large dataset of face images to learn the relevant features and patterns that distinguish different faces.

Plain-English Description: Deep learning is a specialized form of machine learning that uses artificial neural networks with multiple hidden layers to learn and make predictions from complex, high-dimensional data.

Technical Description: Deep learning is a subset of machine learning that uses artificial neural networks with multiple hidden layers to learn representations of data. Deep learning models are able to learn hierarchical features and discover patterns in unstructured data, such as images, text, and audio.

2. Relevance of the concepts:

p-value: The p-value is highly relevant to non-statisticians because it is a key component in determining whether the results of a study are statistically significant and worth acting upon. Understanding the p-value can help non-statisticians critically evaluate the claims made in scientific studies and news reports, which often rely on p-values to support their findings.

confidence interval: Confidence intervals are relevant to non-statisticians because they provide a range of plausible values for a parameter of interest, rather than just a single point estimate. This gives a sense of the precision and uncertainty around a measurement, which is important for making informed decisions in various contexts, such as personal finance, medical treatment, or policy-making.

random sample: The concept of a random sample is relevant to non-statisticians because it is the foundation for making valid inferences about a larger population based on a smaller subset. Understanding random sampling helps non-statisticians appreciate the importance of proper study design and the limitations of convenience samples, which can lead to biased conclusions.

3. Oral explanations of 7 terms using ADEPT/ADEPTR:

1. p-value (ADEPTR)
2. confidence interval (ADEPTR)
3. random sample (ADEPTR)
4. standard deviation (ADEPT)
5. correlation (ADEPT)
6. supervised learning (ADEPT)
7. hypothesis test (ADEPT)

4. Reflective paragraph:

The ADEPT method proved to be a very effective framework for explaining statistical and data science concepts to non-statisticians. By breaking down the explanations into plain-English descriptions, illustrative examples, diagrams, analogies, and technical details, I was able to create comprehensive and accessible explanations that could be tailored to the specific audience.

Incorporating the "Relevance" component (ADEPTR) was particularly valuable, as it helped me bridge the gap between the abstract statistical concepts and the non-statistician's real-world experiences and interests. By explaining how the concepts were relevant to the person's own context, I was able to make the information more meaningful and engaging.

Explaining the terms orally, rather than just in writing, forced me to be more interactive and responsive to the person's level of understanding. I had to closely monitor their body language and receptiveness, and adjust my approach accordingly. This made the explanations more dynamic and effective, as I could adapt the pace, examples, and level of detail based on the person's needs.

The addition of the cross-validation concept further strengthened the ADEPTR method. Cross-validation is a crucial technique for evaluating the performance and generalization of machine learning models, which is highly relevant to many real-world applications. Incorporating this term, along with the others, allowed me to create a more comprehensive and robust set of explanations that covered a wide range of important statistical and data science principles.

Overall, I found that the ADEPTR method was highly effective for breaking down complex statistical and data science topics in a way that is both informative and engaging for non-technical audiences. The combination of plain-English descriptions, illustrative examples, diagrams, analogies, technical details, and relevance to the audience's context proved to be a powerful approach for facilitating deeper understanding and retention of the concepts.

Part 2
Provide feedback on your teammates' assignments. You must provide feedback on the assignment in general and feedback specifically on at least two of their written explanations with the goal of helping them improve their capacity to explain technical concepts.

Part 3
Revise your written explanations based on the feedback you received. Write a short description of what you changed from Part 1 to Part 2 and why. Indicate which explanation you think is your best. "Elevate" this explanation to the top of your team's page. Your part 3 submission should have each teammate's best explanation at the top as well as the "redone first round loser" from class on 25 April. We may start collecting the very best explanations to post on a PUSAKA webpage.

Some guidelines for how to explain statistical terms to non-statisticians:

* Make sure you understand the term. Explain it to yourself first.
* The best two words in an explanation are "for example".
* Use examples!
* Explain the main points. Be concise. The non-statistician (domain expert) doesn't need to know all the technical details.
* Use examples!
* Be positive :) Say what it is, not what it's not.
* Use examples!
* Use words not formulas.
* Use analogies and examples that your collaborator understands.
* Use plain English. Avoid jargon.

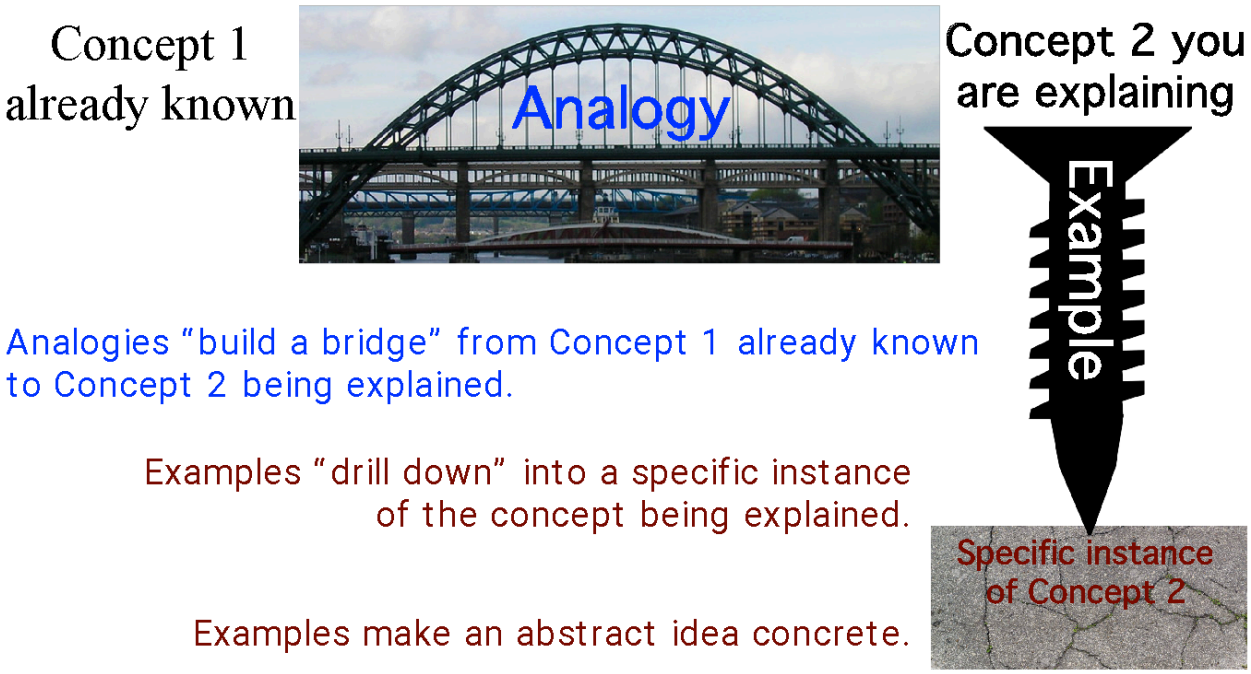
Here is a *definition* of a p-value:
a **p-value** is the proportion of times you would observe data AS EXTREME OR MORE EXTREME if you were to repeat your experiment many times, ASSUMING THE NULL HYPOTHESIS IS TRUE.

The definition above is correct, but it is not an *explanation*. Your assignment is to **explain** what a p-value is (and the other terms above) to a non-statistician. Explanations often include definitions, but they are more than just "defining what the term is".

- A definition conveys a fact
- An explanation conveys insight, i.e. how and why facts fit together.

Here's a relevant analogy: A dictionary is full of facts ("this means that"), but a famous quote arranges those facts into a deeper truth. Similarly, an explanation conveys understanding of a statistical concept much better than a definition will.

Here is a diagram of an analogy (with plain language) attempting to explain the difference between an analogy and an example. In this framework, an example of an analogy is that London is an analogy for New York City (i.e., London is a related concept, if someone understands London, they have a great start at understanding New York City). Central Park is an example of New York City (i.e., it's a specific area of New York City).



Analogies "build a bridge" from Concept 1 already known to Concept 2 being explained.

Examples "drill down" into a specific instance of the concept being explained.

Examples make an abstract idea concrete.

Learning Objectives for this assignment include:

- Practice using the ADEPT method for explanations both in person and in writing
- Understanding the difference between an analogy and an example and using them correctly in your explanations
- Giving effective and helpful feedback to your peers
- Gaining a better understanding of statistics and data science terms
- Testing to see if we should change ADEPT to ADEPTR