TASK:

Explain the difference between correlation and causation. Why is it important to understand the distinction between the two when interpreting statistical results? Provide at least two examples of a correlation that does not necessarily imply causation. How can hypothesis testing be used to determine whether a correlation is statistically significant and can be used to make causal inferences?

definition

1. Correlation occurs when two variables change at the same time.
2. Causation: A causation is a relationship in which the change in one variable causes the other variable to change.

Difference between Correlation and Causation:

Correlation describes an association between types of variables: when one variable changes, so does the other. A correlation is a statistical indicator of the relationship between variables. These variables change together: they covary. But this covariation isn't necessarily due to a direct or indirect causal link.

Causation means that changes in one variable brings about changes in the other; there is a cause-and-effect relationship between variables. The two variables are correlated with each other and there is also a causal link between them.

- Why Correlation Is Not Causation

A correlation is a relationship or connection between two variables where whenever one changes, the other is likely to also change. But a change in one variable doesn't cause the other to change. That would be causation. Your growth from a child to an adult is an example. When your height increased, your mass increased, too. Getting taller didn't also make you get wider. Instead, maturing to adulthood caused both variables to increase — that's causation.

Examples of Correlation without Causation

1. Children and Music Lessons

After a study of human brain development, researchers concluded that kids between 4 and 6 years old who took music lessons showed evidence of boosted brain development in areas related to memory and attention. Based on this study, our biased brain might connect the dots quickly and conclude that music lessons improve brain development. But there are other variables to consider. The fact that the children took music lessons is an indicator of wealth. So they probably had access to other resources that are known to boost brain development like good nutrition.

The point of this example is that researchers can't assume from only this data that music lessons affect brain development. Yes, there's clearly a correlation, but there's no actual evidence of causation. We need more data to get a true causal explanation.

1.
   - Cancer and Mobile Phones

If you study a chart that shows both the number of cancer cases and the number of mobile phones, you'll notice that both numbers went up in the last 20 years. If your brain processes this information with cause-relation cognitive bias, you might decide that mobile phones cause cancer. But that's ridiculous. There's no proof of that other than the fact that both data points happen to increase. A lot of other things have also increased in the past 20 years, and they can't all cause cancer or be caused by mobile phone use.

hypothesis testing

we'll show how hypothesis testing can be used to determine whether a correlation is statistically significant and can be used to make causal inferences.

Hypothesis testing is used to determine whether a correlation is statistically significant and whether a causal link can be inferred. One approach is to carry out controlled experiments, in which one variable is manipulated to observe its effect on another variable. The following Python example shows how hypothesis testing can be used with the statsmodels and

Average Treatment Effect (ATE): The ATE of 1.354 indicates that, on average, individuals with a higher income (treatment group) have a happiness score that is approximately 1.354 points higher than individuals with a lower income (control group). This suggests a positive relationship between income and happiness.

Correlation Coefficient: The correlation coefficient of 0.760 indicates a strong positive correlation between income and happiness scores. This suggests that as income increases, happiness scores tend to increase as well.

P-value: The very small p-value (3.736097155334566e-22) indicates that the observed correlation coefficient is statistically significant. In other words, there is strong evidence to reject the null hypothesis of no correlation between income and happiness.

In summary, based on these findings, we can conclude that there is a significant positive correlation between income and happiness scores. Additionally, the estimated ATE suggests that higher income is associated with higher levels of happiness on average.

data set:

https://www.kaggle.com/datasets/levyedgar44/income-and-happiness-correction

References:

Marco, P. (2024). Introduction to Causal Inference with Machine Learning in Python. Towardsdatascience. https://towardsdatascience.com/introduction-to-causal-inference-with-machine-learning-in-python-1a42f897c6ad

Cornellius, Y.(2022). 4 Python Packages to Learn Causal Analysis. Towardsdatascience. https://towardsdatascience.com/4-python-packages-to-learn-causal-analysis-9a8eaab9fdab

Akansha, K. Everything you need to know about Hypothesis Testing in Machine Learning. Analyticsvidhya. https://www.analyticsvidhya.com/blog/2021/09/hypothesis-testing-in-machine-learning-everything-you-need-to-know/

Anthony, F.(2022). Correlation Is Not Causation. Builtin. https://builtin.com/data-science/correlation-is-not-causation

Pritha, B.(2023). Correlation vs. Causation | Difference, Designs & Examples. Scribbr. https://www.scribbr.com/methodology/correlation-vs-causation/