Correlation and Causality

Correlation

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.

A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

Misalnya, kita ingin mengetahui ukuran/besaran hubungan antara variabel x dengan variabel y. Selain itu, korelasi juga mengukur arah hubungan, apakah hubungannya positif (variabel x meningkat, variabel y juga meningkat) atau hubungannya negatif (variabel x meningkat, variabel y menurun). Nilai korelasi antara -1 sampai 1.

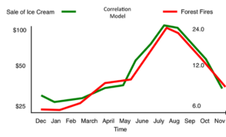
Perlu diingat, ketika kedua variabel saling berkorelasi, belum tentu kedua variabel tersebut saling berkausalitas atau sebab akibat. Misalnya variabel x berkorelasi dengan variabel y, belum tentu perubahan variabel x dipengaruhi oleh variabel y.

Causation

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

Korelasi dapat digunakan sebagai langkah awal dari pengeriksaan hubungan sebab akibat.

Correlation vs Causation



Why are correlation and causation important?

to identify the extent to which one variable relates to another variable. For example:

1. Is there a relationship between a person's education level and their health?
2. Is pet ownership associated with living longer?
3. Did a company's marketing campaign increase their product sales?

> if there is a correlation, then this may guide further research into investigating whether one action causes the other.
>
> By understanding correlation and causality, it allows for policies and programs that aim to bring about a desired outcome to be better targeted.

For two variables, a statistical correlation is measured by the use of a Correlation Coefficient, represented by the symbol (r)

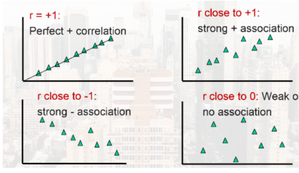
The coefficient's numerical value ranges from +1.0 to −1.0, which provides an indication of the strength and direction of the relationship.

If the correlation coefficient has a negative value (below 0) it indicates a negative relationship between the variables.

If the correlation coefficient has a positive value (above 0) it indicates a positive relationship between the variables

How is correlation measured?

Dalam mengukur korelasi juga dapat menggunakan scatter plot

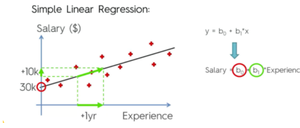


How can causation be established?

Causality is the area of statistics that is commonly misunderstood and misused by people in the mistaken belief that because the data shows a correlation that there is necessarily an underlying causal relationship .

The use of a experiment is the most effective way of establishing causality between variables.

Contohnya menggunakan simple linear regression



.

Jika outputnya berupa kategori seperti yes or no, kita dapat menggunakan classification. Misalnya kita ingin mengetahui apakah customer akan churn atau tidak? maka kita cari faktor-faktor yang memengaruhi customer akan churn atau tidak, kemudian kita lihat besaran dan arah korelasi dari masing-masing variabel. Selanjutnya kita menggunakan model classification untuk mengetahui kausalitas dan outputnya.

Experimental Design (A/B Testing)

A/B Testing merupakan pengaplikasian statistika inferensial di Industri. Misalnya, di sebuah industri memiliki fitur baru, untuk menguji apakah fitur baru tersebut sesuai dengan objektif bisnis maka menggunakan A/B Testing.

Sebagai contoh: Website kampanye Barack Obama, tim kampanye memiliki ide untuk mengubah tampilan website untuk meningkatkan conversion rate (relawan kampanye obama). Sebelum diaplikasikan perubahan tersebut dilakukan A/B Testing, mana tampilan website yang mampu meningkatkan conversion rate.

Steps to do an Experiment

Define an Experiment

What's the name of the experiment

> Ex : AB Test New Design for Registration Webview

Define Hypothesis

> Ex : New design will increase the conversion rate (registration rate) (H1)

Who is the participant

> Ex : The user that visit registration webview

What variable that will be tested

> Ex : Existing Design & New Design

Define Metrics

Macroconversions

Metrics most closely align with hypothesis

> e.g. completed registration (registration rate)

Microconversions

The other actions that users take on our side,

> e.g. watching a video

Vanity Metrics

> e.g. page views

Define the Durations

> Sample Size: Menentukan jumlah sampel yang diperlukan.
>
> Seasonal Effect: Memerperhatikan efek musiman, berdasarkan data historis di waktu tertentu (dalam kasus ini hari minggu) banyak orang yang melakukan registrasi. Maka dari itu, kita hanya perlu melakukan eksperimen di hari minggu. Durasi waktu tergantung dengan jumlah sampel yang dipilih atau diperlukan.

Do Interim Analysis

Hypothesis Testing

Melakukan analisis sementara, misalnya dalam durasi tertentu (misal seminggu secara berulang) kita melakukan analisis untuk melihat hasil sementara. Contohnya, berdasarkan analisis sementara diketahui bahwa tampilan website baru mampu meningkatkan conversion rate lebih tinggi

> Proportion Test: Jika untuk mengukur probabilitas
>
> *T* test: Jika untuk mengukur page view atau banyaknya orang yang mengunjungi website
>
> Anova: jika lebih dari dua sampel, atau dalam kasus ini lebih dari dua tampilan website

Post Analysis

Analisis secara keseluruhan

Hypothesis Testing

> Proportion Test: Jika untuk mengukur probabilitas
>
> *T* test: Jika untuk mengukur page view atau banyaknya orang yang mengunjungi website
>
> Anova: jika lebih dari dua sampel, atau dalam kasus ini lebih dari dua tampilan website