Semana 9

Teste de Hipótese - Quem é p valor?

"A formulação de um teste de hipóteses conforme vimos até aqui, parte da fixação do nível de significância α. Pode-se argumentar que esse procedimento pode levar à rejeição da hipótese nula para um valor de α e à não-rejeição para um valor menor.

Outra maneira de analisar consiste em apresentar a probabilidade de significância ou o famoso p valor do teste.

Os passos para seu cálculo são muito parecidos aos já apresentados; a principal diferença está em não construir a região crítica, ao invés disso, o que se faz é indicar a probabilidade de ocorrer valores da estatística mais extremos do que o observado, sob a hipótese de H0 ser verdadeira."

Livro Estatística Básica - Morettin e Bussab - Capitulo 12.

"No teste de hipóteses estatísticas, o valor p ou o valor de probabilidade é, para um determinado modelo estatístico, a probabilidade de que, quando a hipótese nula for verdadeira, o resumo estatístico (como o valor absoluto da diferença média da amostra entre dois grupos comparados) seria maior ou igual aos resultados reais observados."

Wikipedia

Links úteis:

Teste de Hipótese - Khan Academy
https://pt.khanacademy.org/math/statistics-probability/significance-tests-one-sample/idea-of-significance-tests/v/simple-hypothesis-testing

Explicação do que é p-valor sem números. Fantástico!
https://www.youtube.com/watch?v=9jW9G8MO4PQ

Explicação do que é o pvalor com todo o contexto da construção do teste!
https://towardsdatascience.com/p-values-explained-by-data-scientist-f40a746cfc8

Quem é p-valor? Para o que serve? Onde vive e do que se alimenta?

No problema anterior…

Pergunta: Qual o tempo médio da sua casa até o trabalho?

H0: tempo médio é igual 1h30min
H1: tempo médio é diferente de 1h30min

Lembre-se de que amostrados aleatoriamente alguns dias do transporte da casa até o trabalho da Maria e o objetivo é verificar se o tempo médio é igual a 1h30min.

Se a evidência final apoiar nossa hipótese nula, não a rejeitamos. Caso contrário, rejeitamos a hipótese nula.

O trabalho do valor-p aqui é responder a esta pergunta:

Se estou vivendo em um mundo em que o tempo de deslocamento é de 1h30min (ou seja, hipótese nula é verdadeira), quão surpreendente é a minha evidência na vida real?

O p-valor responde a essa pergunta com um número - probabilidade. Quanto menor o valor-p, mais surpreendente é a evidência, mais absurda é a hipótese nula. E quando isso acontece, rejeitamos ela e escolhemos nossa hipótese alternativa.

Se o valor-p for menor que um nível de significância pré determinado, então rejeitamos o valor nulo.

Voltando ao nosso problema de deslocamento da Maria…

Agora que coletamos alguns tempos de deslocamento, realizamos o cálculo e descobrimos que o tempo médio de deslocamento é menor em 30 minutos, com um valor-p de 0,03.

O que isso significa é que, em um mundo em que o tempo de deslocamento da casa até o trabalho da Maria é de 1h30min (hipótese nula é verdadeira), há 3% de chance de vermos que o tempo médio de deslocamento é menor em 30 minutos devido a ruídos aleatórios.

Exploratory analyses House Prices
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

Desafio do Kaggle com base de dados com características de casas a venda. O desafio é realizar uma regressão linear para predizer o preço de venda de casas.

Ao olharmos a base de dados, é possível ver que a distribuição de algumas variáveis são semelhantes, indicando que elas podem ter algum tipo de relação. Será que há alguma interação entre elas? Qual a interação? Podemos ver isso no modelo de regressão criado?

- Por exemplo, distribuição de: GrLivArea x PricesSales

Entendendo Qui-Quadrado: https://math.hws.edu/javamath/ryan/ChiSquare.html

Correlação entre variáveis categóricas:
https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9