# 05 - Executando o teste de normalidade

[0:00] Vamos começar realmente a trabalhar com testes de hipóteses.

[0:05] No curso anterior, começamos a mexer com inferência estatística, teremos agora uma continuidade desse assunto.

[0:11] No curso anterior, nós estudamos estimação, estimação pontual, estimação intervalar. Aprendemos a calcular o intervalo de confiança, tamanho de amostra.

[0:20] Aprendemos a obter probabilidades de uma distribuição normal, o que era nível de confiança, nível de significância. Todas essas coisas que a gente vem aprendendo lá, vamos aplicar agora dentro de testes de hipóteses.

[0:32] Compreenderemos a aplicação prática de tudo isso, em conjunto, dentro de teste de hipóteses.

[0:38] O que é teste de hipóteses? É uma regra de decisão que ajuda a avaliar hipóteses feitas sobre os parâmetros populacionais e aceitá-las ou rejeitá-las como provavelmente verdadeiras ou falsas, tudo isso tendo como base uma amostra.

[0:56] É o que já começamos a estudar. A partir de uma amostra, temos uma estimativa de um parâmetro da população e desocobre se ele é representativo ou não.

[1:03] Algo interessante é que podemos testar várias coisas. O primeiro teste com o qual vou dar um susto em vocês, assim como eu prometi no vídeo anterior, é o teste de normalidade.

[1:10] Antes analisávamos se uma variável é normal ou não verificando se nela aparecia um "sininho", só visualmente.

[1:19] Mas existe uma forma de fazer isso mais formalmente, um teste mais robusto, que tenha um valor estatístico, diga-se, com tanta probabilidade que você pode aceitar ou rejeitar a hipótese dessa distribuição, se seguiu a normal.

[1:36] Outro exemplo de porque a gente está usando teste de hipóteses é um bem clássico, da diferença de renda entre sexos, homens ganham mais, mulheres ganham menos. Será que isso realmente é verdade, com base em um dado concreto?

[1:55] Por exemplo, o nosso Dataset tem essa informação, a gente vai testar isso no nosso treinamento. Outra coisa, por exemplo, também é um fabricante te passar uma informação e você tem que tomar aquilo como verdade. Podemos testar se essa afirmação do fabricante é verdadeira ou não usando o teste de hipóteses.

[2:13] Vamos começar com o teste de normalidade. Como eu disse, eu vou dar um susto em vocês, alguns conceitos você não vai entender mesmo.

[2:21] Mas a ideia realmente é essa: entender como faz, aplicar um primeiro teste e nos próximos vídeos, vamos fazer esse passo a passo, e eu vou mostrando para você como calcular essas estatísticas na mão, com o lápis. Ok?

[2:35] E no final eu ensinarei como você executa um teste, assim como eu vou fazer agora, pegando uma funcionalidade do Scipy, do Statsmodels, do Python, para executar isso rapidamente.

[2:46] Então, vamos lá. Primeira coisa, no nosso Colab, localizaremos o item "2.1 Teste de normalidade". Neste item, encontraremos "Importando bibliotecas" e, em seguida, um link. Com este link, vou importar o Normaltest. É o mais simples, um teste de normalidade.

[2:53] Algo que eu quero chamar a atenção é que é sempre bom ler a documentação desses testes que estamos utilizando.

[2:59] A documentação diz qual é a hipótese nula do teste. A gente já vai entender o que é isso, hipótese nula. O que ele está testando exatamente? Ele vai dizer para você.

[3:09] No nosso caso, ele está dizendo que a função normaltest testa a hipótese nula, que é o H0, vamos conhecer, de que a amostra é proveniente de uma distribuição normal. Essa é a hipótese nula. É isso que eu tenho que rejeitar ou aceitar de acordo com o que vamos obter de resultado do teste.

[3:30] Vamos importar.

from scipy.stats import normaltestCOPIAR CÓDIGO

[3:44] O que eu quero fazer aqui, eu vou definir um nível de significância. Eu vou colocar isso dentro de uma variável. A gente já conhece o que é significância, o que é nível de confiança.

[3:56] O nível de confiança, trocando em miúdos, é a probabilidade de dar certo, do meu estimador estar correto; e o de significância é dele estar errado. Perfeito?

[4:10] Eu tenho uma célula aqui vazia, depois eu apago. Significância, vou colocar dentro desta variável, eu vou dar 5%, 0,05. É aquela significância padrão, nível de confiança 95%.

significancia = 0.05COPIAR CÓDIGO

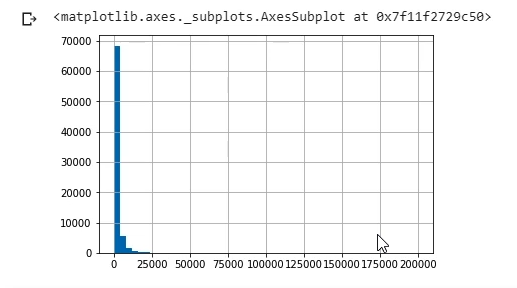
[4:27] Testando a variável renda do nosso Dataset, que a gente abriu no vídeo anterior. Vamos pegar a renda e visualizar primeiro como é que essa renda está se comportando.

[4:36] Então dados.renda.hist(). O Pandas tem a função que já é herdada do Matplotlib para a gente plotar um histograma.

[4:46] Eu vou colocar dentro um parâmetro que se chama bins e eu vou falar que é igual a 50. Ele está dizendo que eu quero 50 "barrinhas" no meu histograma. Vai ficar um pouco mais fácil de visualizar.

dados.renda.hist(bins = 50)COPIAR CÓDIGO

[4:56] Então está aqui. Visualmente, aquele nosso padrão de visualizar teste e visualizar uma normal, eu esperaria que a normal tivesse uma forma de sino, e eu estou vendo aqui que não tem esse padrão.



[5:13] Parece que é uma coisa assim: sobe, desce e vem para o lado direito. É uma coisa meio assimétrica. A gente já viu isso nos outros cursos.

[5:27] Visualmente, eu consigo perceber que isso não segue uma normal, mas vamos testar isso usando um teste formal, para termos uma estatística e dizer se, realmente, essa variável não segue uma distribuição normal. Com certeza. Sem precisar visualizar.

[5:46] Vamos começar. Vou fazer direto o teste. É só chamar, normaltest(), que a gente já importou lá em cima, e passar o dado para ele por dados.Renda. Roda isso. Ele já rodou o teste, está aqui.

normaltest(dados.renda)

NormaltestResult(statistic=152380.75803960307, pvalue=0.0)

COPIAR CÓDIGO

[6:03] O Output desse teste são dois valores: um é o Estatística de teste e o outro é o P valor. Tudo isso a gente vai conhecer ao longo do nosso curso.

[6:16] Como ele tem dois Outputs, é uma tupla, eu posso passar isso para duas variáveis. Vou chamar de stat\_test, vírgula, p\_valor. Perfeito. E eu posso printá-los embaixo.

stat\_test, p\_valor = normaltest(dados.Renda)

print(stat\_test)

print(p\_valor)

152380.75803960307

0.0COPIAR CÓDIGO

[6:41] Vamos separá-los, porque eu posso precisar usar isso para alguma outra coisa. Assim conseguimos visualizá-los separadamente.

[6:59] Aqui em cima eu já deixei uma regrinha de decisão: Rejeitar H0 se o valor de p for maior ou igual a 0,05. Quando eu rejeito o H0 e o que ele é? O H0 é a afirmação de que a amostra é proveniente de uma distribuição normal.

[7:15] A regra de rejeição é simples: a minha variável de p\_valor é menor ou igual a 0,05? O Alfa? Que é o quê? A minha significância, aquela variavel que eu criei mais acima no código.

p\_valor <= significância

TrueCOPIAR CÓDIGO

[7:41] O que ele está dizendo? Que é verdadeira essa afirmação. Então, o que eu faço? Eu rejeito o H0. Que é o quê, novamente? É a hipótese de que a amostra é proveniente de uma distribuição normal.

[7:54] O que ele está dizendo aqui é que não, não é proveniente de uma distribuição normal, eu rejeito essa hipótese aqui totalmente. Visualmente a gente confirma isso e, agora, com uma estatística mais formal, a gente também confirma isso.

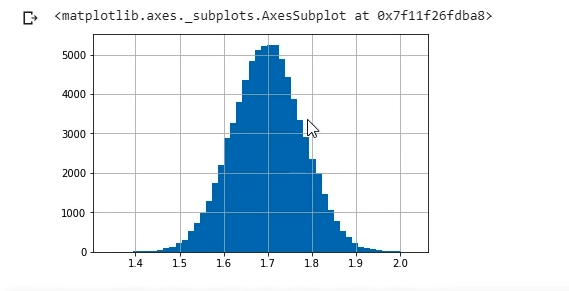
[8:08] Agora, vamos descobrir uma variável que realmente siga uma distribuição normal e saber o que esse teste mostra para a gente.

[8:14] A variável "Altura", como eu disse no vídeo anterior e venho falando isso nos outros cursos, fui eu que gerei ela a partir de, justamente, uma distribuição normal, que é uma variável aleatória proveniente de uma normal.

[8:28] Vamos visualizar de novo, só para termos certeza de tudo que eu estou falando, Altura.hist. E vou colocar o mesmo bins igual a 50 para ficar igualzinho.

dados.Altura.hist(bins = 50)COPIAR CÓDIGO

[8:42] Olha lá que beleza, o "sininho". Com média de 1,70 m (Um metro e setenta). Eu criei assim mesmo, de propósito.



[8:52] Aqui, novamente, a regra de rejeição. Vamos copiar esse trecho de código do stat\_test, para adiantarmos o nosso trabalho. E já vamos copiar esse p\_valor <= significância também.

[9:05] Nós vamos substituir o p\_valor e o stats aqui. A única coisa que eu tenho que fazer é mudar de Renda para Altura e rodar o teste.

stat\_test, p\_valor = normaltest(dados.Altura)

print(stat\_test)

print(p\_valor)

0.19973093957002253

0.9049591541967501COPIAR CÓDIGO

[9:19] É possível notar os números da estatística no trecho acima. Depois a gente vai entender o que é estatística. Não estamos usando, mas vamos precisar dela para tirar algumas decisões. E o nosso P valor, que é o principal. Todo teste que você vai rodar, vai plotar o P valor para você, não se preocupe.

[9:33] E a regra de decisão é sempre essa. Já é um valor alto, quase 1. Vou rodar a minha regra de decisão, que eu criei aqui.

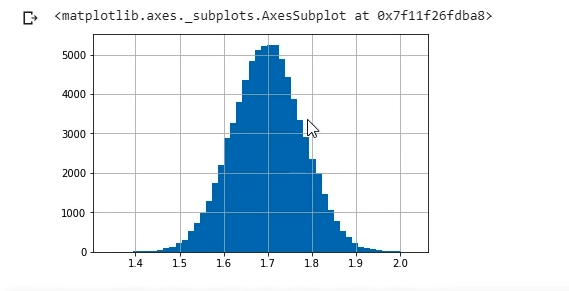
p\_valor <= significancia

FalseCOPIAR CÓDIGO

[9:44] E o que ele está dizendo? Falso. Eu rejeito H0 se valor de P for menor ou igual 0,5. O que ele está falando é que não, que é falso.

[9:54] O que eu faço? Não posso rejeitar. Eu não posso rejeitar a hipótese nula, H0, de que a amostra é proveniente de uma distribuição normal. E realmente é, fui eu que criei.

[10:07] E está aqui. Visualmente, realmente parece uma normal. Bonitinho o sino.



[10:12] Pessoal, é isso que eu queria mostrar nesse vídeo, esse susto inicial. Não se preocupem, a gente vai ver o que é P valor, a gente vai aprender a calcular um P valor manualmente e também H0, hipótese nula, rejeição, aceitar hipótese alternativa. Os níveis de significância a gente já conhece. Mas todas essas coisinhas que eu fui falando aqui, a gente vai ver passo a passo.

[10:33] No próximo vídeo, já vamos entender os passos. E no outro a gente vai começar a colocar a mão na massa. Vamos fazer testes, passo a passo, e, no final, eu mostro como fazer isso, de uma forma bem simples e prática, utilizando funcionalidades do Python.

[10:46] No próximo vídeo a gente vê as etapas básicas de um teste. Até lá.