

Introdução à Ciência dos Dados

Aula 11 - Fundamentos de Inferência Estatística e Análise de Correlação

Prof. Washington Santos da Silva

15/06/2023

Mestrado Profissional em Administração

Fundamentos de Inferência Estatística

Análise de Correlação

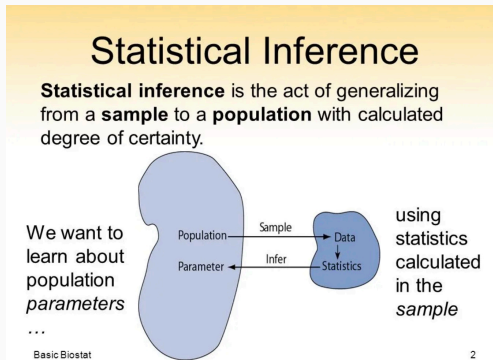
Referências

Fundamentos de Inferência Estatística

Inferência Estatística

Definição

Inferência estatística é o processo que consiste em usar dados de uma amostra para tirar conclusões sobre características (parâmetros) de uma população subjacente da qual a amostra (aleatória) foi retirada.



Conceitos

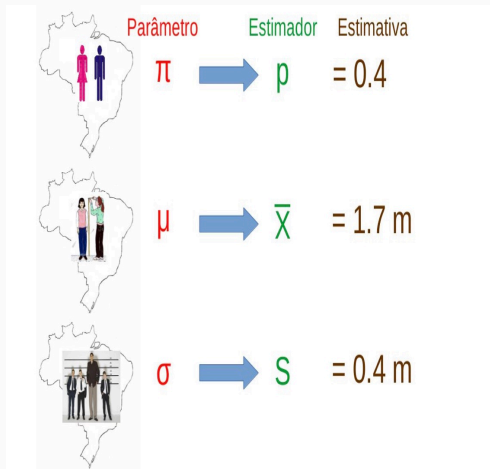
- **população:** conjunto de itens ou eventos com pelo menos uma característica em comum.
- **parâmetro:** Uma característica da população
- **amostra:** é um subconjunto da população
- **estatística:** é qualquer função dos dados da amostra

Procedimentos de Inferência Estatística

- **Estimação:** Pontual e Intervalos de Confiança
- **Testes de Hipóteses**

Estimação

- **parâmetro**: característica da população.
- **estimador**: função (matemática) da amostra (estatística).
- **estimativa**: valor realizado de um estimador.



Estimação

O objetivo dos métodos de **Estimatição** é determinar o valor de um parâmetro populacional com base em uma estatística.

- **Estimativa pontual:** estimativa singular (única).
- **Estimativa por Intervalo de Confiança:** intervalo de valores plausíveis para o parâmetro populacional.

Métodos de Estimação

- Método dos Momentos
- **Método da Máxima Verossimilhança**
- Método dos Mínimos Quadrados Ordinários
- Métodos Bayesianos

Propriedades

Estimador Não Viesado:

$$E(\hat{\theta}) = \theta \rightarrow E(\hat{\theta}) - \theta = 0$$

Estimador Eficiente:

Dizemos que um estimador $\hat{\theta}$ é eficiente, se para um determinado tamanho de amostra n , a $V(\hat{\theta})$ é menor que a variância de qualquer outro estimador não viesado.

Estimador Consistente:

$\hat{\theta}$ é um estimador consistente de θ , se na medida em que o tamanho da amostra aumenta, $\hat{\theta} \rightarrow \theta$

Estimação por Máxima Verossimilhança

- Em estatística, a estimação por máxima verossimilhança é um método de **estimar** os parâmetros de uma **distribuição de probabilidade** assumida, dadas as observações.

Estimação por Máxima Verossimilhança

- Em estatística, a estimação por máxima verossimilhança é um método de **estimar** os parâmetros de uma **distribuição de probabilidade** assumida, dadas as observações.
- A estimativa é obtida **maximizando** uma função de verossimilhança para que, sob o modelo estatístico assumido, os dados observados sejam mais prováveis.

Estimação por Máxima Verossimilhança

- Em estatística, a estimação por máxima verossimilhança é um método de **estimar** os parâmetros de uma **distribuição de probabilidade** assumida, dadas as observações.
- A estimativa é obtida **maximizando** uma função de verossimilhança para que, sob o modelo estatístico assumido, os dados observados sejam mais prováveis.
- O ponto no espaço de parâmetros que maximiza a função de verossimilhança é chamado de **estimativa de máxima verossimilhança**.

Estimação Pontual: Máxima Verossimilhança

- Modelamos um conjunto de observações como uma **amostra aleatória** a partir de uma **distribuição de probabilidade conjunta** desconhecida que é expressa em termos de um conjunto de **parâmetros**.
- O objetivo da estimativa de máxima verossimilhança é determinar os parâmetros para os quais os dados observados têm a maior probabilidade conjunta.
- Avaliando a função de densidade conjunta na amostra de dados observada, $y = (y_1, y_2, \dots, y_n)$, fornece uma função:

$$L_n(\theta; y) = f_n(y, \theta)$$

denominada **função de verossimilhança**.

Estimação Pontual: Máxima Verossimilhança

Para variáveis aleatórias independentes e identicamente distribuídas, $f_n(y; \theta)$ será o produto de funções de densidade univariadas:

$$f_n(y; \theta) = \prod_{k=1}^n f_k(y_k; \theta)$$

Podemos simplificar o algoritmo, tomando o logaritmo natural:

$$\ln(f_n(y; \theta)) = \sum_{k=1}^n \ln(f_k(y_k; \theta))$$

Essa é a **função de log-verossimilhança**, efetivamente utilizada nos algoritmos de estimação.

Estimação Pontual: Máxima Verossimilhança

O objetivo da estimativa de máxima verossimilhança é encontrar os valores dos parâmetros do modelo que maximizam a função de log-verossimilhança no espaço de parâmetros:

$$\hat{\theta} = \max_{\theta \in \Theta} \ln(L_n(\theta; y))$$

Intuitivamente, o método determina estimativas dos parâmetros que tornam mais prováveis os dados observados.

Estimação Pontual: Máxima Verossimilhança

Em uma amostra aleatória de dez relatórios financeiros produzidos pela controladoria de uma determinada empresa, verifica-se que o primeiro, terceiro e décimo relatórios possuem não conformidades graves, enquanto os outros não.

$$x < -c(1, 0, 1, 0, 0, 0, 0, 0, 0, 1)$$

Seja

$$p = P(\text{relatório não conforme})$$

a proporção de relatórios que apresentam uma não conformidade, então a variável aleatória X_i é definida por:

$$X_i = \begin{cases} 1 & \text{se o relatório é não conforme} \\ 0 & \text{se o relatório é conforme} \end{cases}$$

Distribuição de Bernoulli

Uma variável aleatória que segue uma distribuição de Bernoulli pode assumir apenas dois valores: 0 (“fracasso”) ou 1 (“sucesso”), com probabilidades p e $1 - p$, respectivamente.

A função de probabilidade da distribuição de Bernoulli é dada por:

$$P(X = x) = \begin{cases} p, & \text{se } x = 1 \\ 1 - p, & \text{se } x = 0 \end{cases}$$

que pode ser escrita como:

$$f(x; p) = p^x(1 - p)^{1-x}, \quad y \in \{0, 1\}$$

Desejamos um estimador de p que maximize a probabilidade de ocorrência dos dados observados.

Estimação Pontual: Máxima Verossimilhança

Como sabemos que X pode ser modelada por uma distribuição de Bernoulli, a função de verossimilhança é dada por:

$$\begin{aligned}L_n(\theta; y) &= \prod_{k=1}^{10} p^{x_i} (1-p)^{1-x_i}, \\ \ln L_n(\theta; y) &= \sum_{k=1}^n x_i \ln(p) + \sum_{k=1}^n (1-x_i) \ln(1-p), \\ &= n\bar{x} \ln(p) + (n - n\bar{x}) \ln(1-p) \\ \frac{d}{dp} \ln L_n(\theta; y) &= \frac{n\bar{x}}{p} - \frac{n - n\bar{x}}{1-p} = 0, \\ \hat{p} &= \frac{\sum_{k=1}^n x_i}{n} = \frac{3}{10} = 0.3\end{aligned}$$

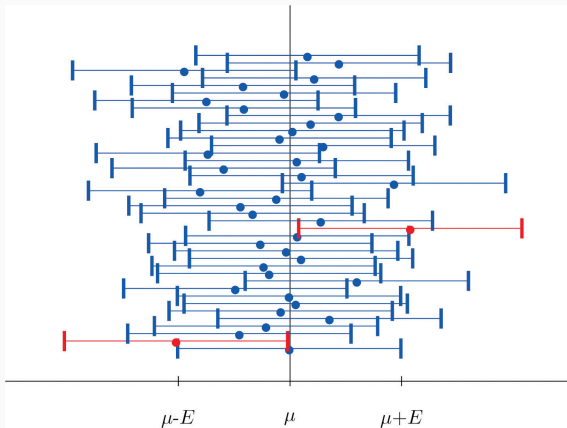
Maximizar $\ln L_n(\theta; y)$ fornece estimativas de p para as quais é mais provável que a amostra observada tenha sido gerada.

Intervalos de Confiança

Intervalos de Confiança

- Se você deseja estimar um parâmetro da população, prefere relatar um intervalo de valores em que o parâmetro possa estar ou um único valor?
- Se relatarmos uma estimativa pontual, provavelmente não atingiremos o parâmetro exato da população.
- Se relatarmos um intervalo de valores plausíveis, teremos uma boa chance de capturar o parâmetro.
- Um intervalo de confiança é um intervalo de valores plausíveis para o parâmetro populacional.
- Um intervalo com 95% de confiança, por exemplo, implica que, se o processo de estimação fosse repetido várias vezes, espera-se que 95% dos intervalos calculados contenham o valor verdadeiro do parâmetro.

Intervalos de Confiança



Quantificando a incerteza

- Para construir um intervalo de confiança, precisamos quantificar a variabilidade do estimador utilizado.
- Por exemplo, se desejamos construir um intervalo de confiança para uma proporção populacional (p), precisamos criar uma faixa plausível de valores em torno da estimativa da proporção (\hat{p}).
- Esse intervalo dependerá de quão precisa é a estimativa da proporção.
- Quantificar isso requer uma medição do quanto esperamos que estimativa da proporção varie de amostra para amostra

Métodos

Podemos quantificar a variabilidade de estimadores usando:

- Simulação de Monte Carlo: bootstrappo, ou por
- Teoria: Teorema do Limite Central

Quantificando a incerteza via Teoria

Teorema Central do Limite

Utilizando o Teorema Central do Limite, podemos demonstrar que, para grandes amostras, um estimador por intervalo com $(1 - \alpha)100\%$ de confiança, para uma proporção populacional é:

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

sendo:

α = nível de significância = a probabilidade do intervalo não conter θ .

$(1 - \alpha)$ = é o nível de confiança da estimativa por intervalo.

Quantificando a incerteza via Bootstrapp

- Estimativas de parâmetros obtidas por bootstrapp não-paramétrico oferecem uma alternativa robusta aos métodos clássicos (paramétricos) para inferência estatística.
- Ao contrário dos métodos clássicos de inferência estatística que dependem de suposições paramétricas, ou seja, que os dados sigam uma distribuição de probabilidade, pelo menos aproximadamente, e/ou de aproximações assintóticas, isto é, a validade do estimador requer "grandes" amostras, o bootstrap não paramétrico usa métodos computacionalmente intensivos (Simulação de Monte Carlo) para fornecer resultados inferenciais válidos para diversas situações.

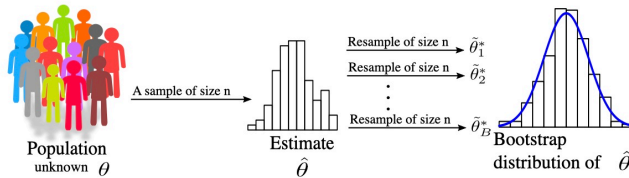
Quantificando a incerteza via Bootstrapp

Bootstrapp: Algoritmo

1. Retire uma amostra aleatória com **reposição** da amostra original, do mesmo tamanho da amostra original.
2. Escolha um estimador de um parâmetro, calcule a estimativa (média, mediana, proporção, etc.) usando a amostra inicial.
3. Repita as etapas (1) e (2) muitas vezes para criar uma distribuição de bootstrap - uma distribuição de estatísticas de bootstrap.
4. Calcule os limites do intervalo com $(1 - \alpha)100\%$ de confiança.

Quantificando a incerteza via Bootstrapp

Figure Illustration of the bootstrap distribution



Testes Estatísticos de Hipóteses

- Nos EUA, as pessoas que desejam doar um órgão às vezes procuram a ajuda de um "consultor médico" especial. Esses consultores auxiliam o paciente em todos os aspectos da cirurgia, com o objetivo de reduzir a possibilidade de complicações durante o procedimento médico e a recuperação.
- Os pacientes podem escolher um consultor com base na taxa de complicações históricas dos clientes do consultor.
- Uma consultora tentou atrair pacientes observando que a taxa média de complicações para cirurgias de doadores de fígado nos EUA é de cerca de 10% mas seus clientes tiveram apenas 3 complicações nas 62 cirurgias nas quais atuou como consultora.
- Ela afirma que esses dados são uma forte evidência de que seu trabalho contribui significativamente para reduzir as complicações.

Parâmetro em um Teste de Hipóteses

- Um **parâmetro** para um teste de hipótese é o valor populacional "verdadeiro" de interesse.
- Como vimos, estimamos um parâmetro usando uma **estimativa pontual**.

p : = proporção verdadeira de complicações

\hat{p} : = proporção amostral de complicações = $\frac{3}{62} = 0.048$

Correlação versus Causalidade

- É possível avaliar a afirmação causal da médica usando os dados?
- Não. A alegação é que há uma conexão **causal**, mas os dados são observacionais.
- Por exemplo, talvez os pacientes que podem pagar um consultor médica possam pagar melhores cuidados médicos, o que também pode levar a uma menor taxa de complicações.

Duas reivindicações

- **Hipótese Nula:** "Não há nada acontecendo". A taxa de complicações para este consultor não é diferente da média dos EUA de 10
- **Hipótese alternativa:** "Há algo acontecendo". A taxa de complicações para este consultor é **menor** do que a média dos EUA de 10

Teste de hipóteses como um julgamento judicial

- hipótese Nula, H_0 : O réu é inocente
- **Hipótese alternativa**, H_A : O réu é culpado
- **Apresente as evidências**: Coletam-se dados
- **Julgue as evidências**: "Esses dados poderiam plausivelmente ter acontecido, **por acaso, se a hipótese nula fosse verdadeira?**"
- Sim: Não rejeitar H_0
- Não: Rejeitar H_0

Estrutura de um teste de hipóteses

- Comece com uma hipótese nula, H_0 , que representa o status quo.
- Defina uma hipótese alternativa, H_A , que represente a questão da pesquisa, ou seja, o que estamos testando.
- Realize um teste de hipótese sob a suposição de que a hipótese nula é verdadeira e calcule um **valor-p** (probabilidade de se obter um resultado igual ou mais extremo, dado que a hipótese nula é verdadeira).
- se os resultados do teste sugerirem que os dados não fornecem evidências convincentes para a hipótese alternativa, não se rejeita a hipótese nula
- caso contrário, rejeita-se a hipótese nula em favor da alternativa

Estrutura Resumida

- Defina as hipóteses.
- Calcule a estatística de teste com base na amostra
- Calcule o valor p .
- Faça uma conclusão, sobre as hipóteses, no contexto dos dados e da questão da pesquisa.

Erros em TH

- Erro Tipo 1: Rejeitar H_0 quando não deveria rejeitar H_0
- $P(\text{Erro Tipo 1}) = \alpha = \text{nível de significância do teste.}$
- Erro Tipo 2: Não rejeitar H_0 quando você H_0 deveria ser rejeitada.
- $P(\text{Erro Tipo 2})$ é mais difícil de calcular, e aumenta à medida que α diminui.

Qual erro controlar?

Em um tribunal:

- Hipótese nula: O réu é inocente.
- Hipótese alternativa: O réu é culpado.
- O que é "pior": Erro Tipo 1 ou Tipo 2?

Probabilidades

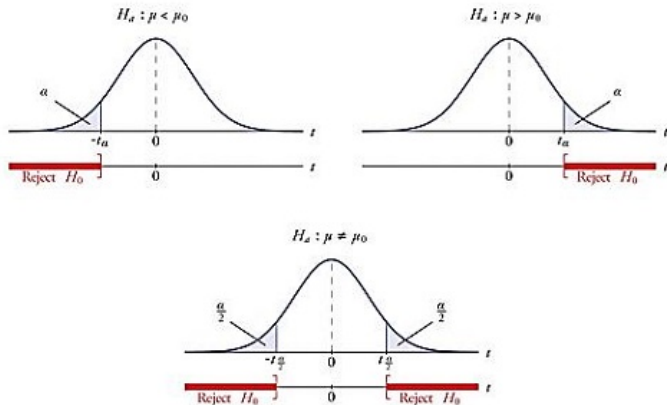
- $P(\text{Erro Tipo 1}) = \alpha$.
- Poder = $P(\text{corretamente rejeitar a hipótese nula})$.
- $P(\text{Erro Tipo 2}) = 1 - \text{Poder}$.

Erros Tipo I (α), Tipo II (β) e Poder do Teste.

Possibilidades envolvidas em um teste de hipóteses

Realidade	Decisão	Aceitar H_0	Rejeitar H_0
	H_0 é verdadeira	Decisão correta $1 - \alpha = P(\text{Aceitar } H_0 / H_0 \text{ é V}) = P(H_0 / H_0)$	Erro do Tipo I $\alpha = P(\text{Erro do tipo I}) = P(\text{Rejeitar } H_0 / H_0 \text{ é V}) = \text{Nível de significância do teste} = P(H_1 / H_0)$
	H_0 é falsa	Erro do Tipo II $\beta = P(\text{Erro do tipo II}) = P(\text{Aceitar } H_0 / H_0 \text{ é falsa}) = P(\text{Aceitar } H_0 / H_1 \text{ é V}) = P(H_0 / H_1)$	Decisão correta $1 - \beta = P(\text{Rejeitar } H_0 / H_0 \text{ é falsa}) = P(H_1 / H_1) = \text{Poder do teste.}$

Testes Estatísticos de Hipóteses



Poder do Teste e Tamanho da Amostra para r

```
# Pacote pwr  
library(pwr)  
  
# funcao que estima o poder do teste ou determina os parâmetros  
# para obter o poder desejado  
pwr.r.test(n = NULL, r = NULL, sig.level = 0.05, power = NULL,  
           alternative = c("two.sided", "less", "greater"))
```

sendo:

- n = tamanho da amostra
- r = tamanho do efeito - Cohen (1988):
 - $r = 0.1$ (fraca)
 - $r = 0.3$ (moderada)
 - $r = 0.5$ (forte)
- sig.level = nível de significância = erro tipo I = α
- $\text{power} = (1 - \beta)$ = poder do teste. (0.8 ou 80% é um poder adequado em geral).

Conceito

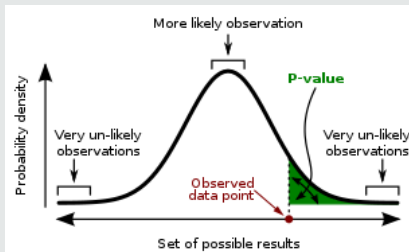
- Em qualquer amostra, por exemplo, a média amostral \bar{x} raramente será exatamente igual ao valor sob a hipótese nula ($H_0 : \mu = \mu_0$, por exemplo). As diferenças entre \bar{x} e μ_0 podem surgir porque a média verdadeira (populacional) (μ), de fato, não é igual a μ_0 (a hipótese nula é falsa) ou porque a média verdadeira é igual μ_0 (a hipótese nula é verdadeira), mas \bar{x} difere de μ devido à amostragem aleatória.
- É impossível distinguir entre essas duas possibilidades com certeza. Embora uma amostra de dados não possa fornecer evidências conclusivas sobre a hipótese nula, é possível fazer um **cálculo probabilístico** que permite testar a hipótese nula de forma que leva-se em conta a incerteza da amostragem.
- Esse cálculo envolve o uso dos dados amostrais para calcular o **valor-p** da hipótese nula.

Testes de Hipóteses usando valor-p

valor-p

valor-P é a probabilidade, calculada assumindo-se que H_0 seja "verdadeira", de se obter um valor da estatística de teste igual ou superior à estatística de teste observada. Um valor-p muito pequeno significa que um resultado tão extremo quanto o observado seria muito improvável sob a hipótese nula.

$$\text{valor} - p = P(|T| \geq |t| | H_0)$$



O que você deseja fazer?

Procedimentos de Inferência Estatística

- **Estimação** -> Intervalo de Confiança
- **Decisão** -> Teste de Hipóteses

Análise de Correlação

Análise e Coeficiente de Correlação

Objetivo

A Análise de Correlação mede, **principalmente**, a força da relação linear entre variáveis numéricas. A medida baseia-se no coeficiente de correlação entre duas variáveis aleatórias.

- Covariância

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)], \\ &= E[XY] - \mu_X\mu_Y \end{aligned}$$

- Correlação

$$\text{Cor}(X, Y) = \rho_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} \quad (-1 \leq \rho_{xy} \leq +1)$$

Coeficiente de Correlação de Pearson

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Sendo x_i e y_i os dados ou observações.

Coeficiente de Correlação de Pearson

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Sendo x_i e y_i os dados ou observações. **Pressuposto:** Distribuição aproximadamente normal.

Coeficientes de Correlação: Estimadores

Coeficiente de Correlação de Pearson

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Sendo x_i e y_i os dados ou observações. **Pressuposto:** Distribuição aproximadamente normal. Sensível a valores extremos.

Coeficientes de Correlação: Estimadores

Coeficiente de Correlação de Pearson

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Sendo x_i e y_i os dados ou observações. **Pressuposto:** Distribuição aproximadamente normal. Sensível a valores extremos.

Coeficiente de Correlação de Spearman

$$r_s = 1 - \frac{6 \sum_{i=1}^n (x_i - y_i)^2}{n(n^2 - 1)}$$

Sendo x_i e y_i os postos (*rank*) das observações.

Coeficientes de Correlação: Estimadores

Coeficiente de Correlação de Pearson

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Sendo x_i e y_i os dados ou observações. **Pressuposto:** Distribuição aproximadamente normal. Sensível a valores extremos.

Coeficiente de Correlação de Spearman

$$r_s = 1 - \frac{6 \sum_{i=1}^n (x_i - y_i)^2}{n(n^2 - 1)}$$

Sendo x_i e y_i os postos (*rank*) das observações. **Método não paramétrico:** robusto a outliers. menor poder do teste.

Coeficiente de Correlação de Kendall

$$r_k = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \text{sgn}(x_j - x_i) \text{sgn}(y_j - y_i)$$

O coeficiente de correlação de Kendall usa pares de observações e determina a força da associação com base no padrão de concordância e discordância entre os pares. Duas VAs X e Y são concordantes se $X_2 - X_1 > 0$ e $Y_2 - Y_1 > 0$, ou $X_2 - X_1 < 0$ e $Y_2 - Y_1 < 0$.

Coeficiente de Correlação de Kendall

$$r_k = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n \text{sgn}(x_j - x_i) \text{sgn}(y_j - y_i)$$

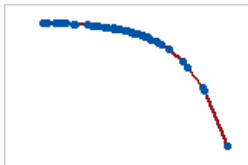
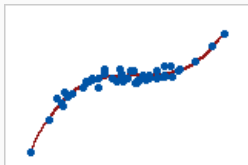
O coeficiente de correlação de Kendall usa pares de observações e determina a força da associação com base no padrão de concordância e discordância entre os pares. Duas VAs X e Y são concordantes se $X_2 - X_1 > 0$ e $Y_2 - Y_1 > 0$, ou $X_2 - X_1 < 0$ e $Y_2 - Y_1 < 0$.

Método não paramétrico: robusto a outliers. menor poder do teste.

Relações Monotônicas

Importante!

Algumas relações monotônicas podem ser capturadas pelo coeficiente de correlação de Spearman (e de Kendall).



Níveis de Correlação: Interpretação

r	Magnitude
$r \geq 0.5$	correlação forte/alta
$0.3 \leq r \leq 0.5$	correlação moderada
$0.1 \leq r \leq 0.3$	correlação fraca/pequena
$r < 0.1$	correlação muito fraca/pequena

COHEN, Jacob. **Statistical power analysis for the behavioral sciences**. Routledge, 1988.

Níveis de Correlação: Interpretação

r	Magnitude
$r \geq 0.4$	muito forte/alta
$0.3 \leq r < 0.4$	forte/alta
$0.2 \leq r < 0.3$	moderada/média
$0.1 \leq r < 0.2$	pequena
$0.05 \leq r < 0.1$	correlação muito fraca/pequena
$r < 0.05$	correlação extremamente fraca/pequena

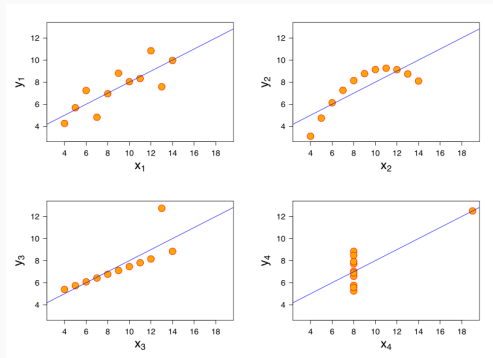
FUNDER, David C.; OZER, Daniel J. Evaluating effect size in psychological research: Sense and nonsense. **Advances in Methods and Practices in Psychological Science**, v. 2, n. 2, p. 156-168, 2019.

Níveis de Correlação: Pacote effectsize

Pacote effectsize

Vale a pena conhecer o pacote [effectsize](#). Veja como implementar as possibilidades de "regras de bolso" para a interpretação de estimativas de r em [Automated Interpretation of Indices of Effect Size](#)

Cuidado! Quarteto de Anscombe



[Leia Wikipedia](#)

Cuidado!

uma medida estatística que sumariza a informação dos dados, não pode substituir o exame visual dos dados.

Correlação: Inferência Estatística - Teste de Hipóteses

Teste para r - Pearson

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2,\alpha)}$$

Usando um software estatístico, é fácil calcular a probabilidade de observar um valor igual o maior que $|t|$. Essa probabilidade recebe o nome de **valor-p**.

Se valor-p < 0.05 A estimativa é estatisticamente diferente de zero.

Se valor-p > 0.05 A estimativa **não** é estatisticamente diferente de zero.

Melhor: Intervalo de Confiança para ρ

Correlação: Inferência - Intervalo de Confiança

IC para r - Pearson

1. Aplica-se a transformada Z de Fisher a r

$$z = \ln\left(\frac{1+r}{1-r}\right)$$

2. Estima-se o IC com o valor z com:

$$IC = [z - z_{crtico} * ep; z + z_{crtico} * ep],$$
$$ep = \frac{1}{\sqrt{n-3}}$$

3. Converte-se a estimativa baseada em z para valores r

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} = \tanh(z)$$

- Se as variáveis aleatórias forem numéricas, contínuas e normalmente distribuídas, use o coeficiente de Pearson

Análise de Correlação: Sumário

- Se as variáveis aleatórias forem numéricas, contínuas e normalmente distribuídas, use o coeficiente de Pearson
- Se as variáveis forem medidas usando uma escala ordinal ou tiverem distribuições assimétricas e/ou outliers (i.e, não normalidade), e determinados tipos de relações não lineares, utilize o coeficiente de Spearman.

Análise de Correlação: Sumário

- Se as variáveis aleatórias forem numéricas, contínuas e normalmente distribuídas, use o coeficiente de Pearson
- Se as variáveis forem medidas usando uma escala ordinal ou tiverem distribuições assimétricas e/ou outliers (i.e, não normalidade), e determinados tipos de relações não lineares, utilize o coeficiente de Spearman.
- Relações identificadas utilizando coeficientes de correlação devem ser interpretadas por aquilo que são: **associações, não como relações de causa e efeito**. Sob determinadas condições, que precisam ser verificadas, é possível inferir relações de causalidade.

Análise de Correlação: Sumário

- Se as variáveis aleatórias forem numéricas, contínuas e normalmente distribuídas, use o coeficiente de Pearson
- Se as variáveis forem medidas usando uma escala ordinal ou tiverem distribuições assimétricas e/ou outliers (i.e, não normalidade), e determinados tipos de relações não lineares, utilize o coeficiente de Spearman.
- Relações identificadas utilizando coeficientes de correlação devem ser interpretadas por aquilo que são: **associações, não como relações de causa e efeito**. Sob determinadas condições, que precisam ser verificadas, é possível inferir relações de causalidade.
- Não é apropriado concluir que mudanças em uma variável **causam** mudanças em outra com base apenas na correlação.

Análise de Correlação: Sumário

- O coeficiente de correlação de Pearson é muito sensível a valores extremos. Um único valor que é muito diferente dos outros valores em um conjunto de dados pode alterar muito o valor do coeficiente. Você deve tentar identificar a causa de qualquer valor extremo. Corrija qualquer entrada de dados ou erros de medição. Considere a remoção de valores de dados associados a eventos anormais e únicos (causas especiais). Em seguida, repita a análise.

Análise de Correlação: Sumário

- O coeficiente de correlação de Pearson é muito sensível a valores extremos. Um único valor que é muito diferente dos outros valores em um conjunto de dados pode alterar muito o valor do coeficiente. Você deve tentar identificar a causa de qualquer valor extremo. Corrija qualquer entrada de dados ou erros de medição. Considere a remoção de valores de dados associados a eventos anormais e únicos (causas especiais). Em seguida, repita a análise.
- Um baixo coeficiente de correlação de Pearson não significa que não exista relação entre as variáveis. As variáveis podem ter uma relação não linear. Para verificar relações não lineares graficamente, crie um gráfico de dispersão ou use regressão simples.

Análise de Correlação: Sumário

- O coeficiente de correlação de Pearson é muito sensível a valores extremos. Um único valor que é muito diferente dos outros valores em um conjunto de dados pode alterar muito o valor do coeficiente. Você deve tentar identificar a causa de qualquer valor extremo. Corrija qualquer entrada de dados ou erros de medição. Considere a remoção de valores de dados associados a eventos anormais e únicos (causas especiais). Em seguida, repita a análise.
- Um baixo coeficiente de correlação de Pearson não significa que não exista relação entre as variáveis. As variáveis podem ter uma relação não linear. Para verificar relações não lineares graficamente, crie um gráfico de dispersão ou use regressão simples.
- Use o coeficiente de correlação de Spearman para examinar a força e a direção da relação monotônica entre duas variáveis contínuas ou ordinais. Em uma relação monotônica, as variáveis tendem a se mover na mesma direção relativa, mas não necessariamente a uma taxa constante.

- É necessário muito cuidado com **correlações espúrias**!:
 - Leitura: [Leaders: Stop Confusing Correlation with Causation](#)
 - Leitura: [Correlation vs. Causation](#)
 - Leitura: [An introduction to Causal inference](#)
 - Leitura: [Correlation vs. Trends: A Common Misinterpretation](#)
 - Site: [Spurious Correlation](#)

Reportando uma análise de correlação

Modelo American Psychological Association (APA)

Um coeficiente de correlação de Pearson foi estimado para avaliar a relação linear entre [variável 1] e [variável 2].

Houve uma correlação [negativa ou positiva] [significativa ou não significativa], entre a **variável 1** e a **variável 2**, t = valor calculado, p = [valor – p do teste], r (**graus de liberdade**) = [estimativa de r], n = tamanho da amostra. IC (95%) [inferior, superior].

Exemplo Hipotético

Um coeficiente de correlação de Pearson foi estimado para avaliar a relação linear entre X e Y .

Os resultados indicam uma correlação positiva significativa entre X e Y : $t = 8.01$, $\text{valor} - p = .002$, $r(149) = .55$, $n = 151$, IC 95% [0.43, 0.65].



Questões

1. Usando o pacote 'BatchGetSymbols' importe dados dos últimos 1000 dias para as seguintes ações: 'AAPL', 'WEGE3.SA', 'AMZN', 'GOOG'.
2. Converta os dados diários dos preços de fechamento das ações para o formato 'xts', usando o pacote 'xts', este é um formato específico para o armazenamento de séries temporais em R. Além disso, elabore gráficos das séries temporais de preços para cada uma das ações.
3. Calcule os retornos compostos continuamente a partir dos preços no formato 'xts' usando a função 'Return.calculate' do pacote 'PerformanceAnalytics', elabore gráficos das séries temporais dos retornos para cada uma das ações.
4. Faça a fusão da séries de retornos das ações em um único objeto usando a função 'merge.xts()' do pacote 'xts' e nomeie o objeto como 'retornos'.
5. Faça uma análise gráfica da correlação entre os retornos das ações usando a função 'chart.Correlation()' do pacote 'PerformanceAnalytics' aplicada sobre o objeto 'retornos'.

Questões

6. Encontre a matriz de correlações de Pearson entre os retornos.
7. Obtenha estimativas pontuais do r de Pearson entre os retornos das ações e teste a hipótese nula de que $r = 0$.
8. Dadas as regularidades empíricas dos retornos de ações, qual estimador de ρ , você considera mais apropriado para verificar quais séries de retornos possuem correlações estatisticamente significativas? Reporte os resultados do estimador escolhido, conforme o padrão da APA.

Referências

-  STOCK, James H.; Watson, Mark W. Econometria: uma abordagem moderna. Pearson Universidades, 2004. Capítulo 3. Disponível na Biblioteca Virtual Pearson:
<https://plataforma.bvirtual.com.br/Account/Login>
-  WOOLDRIDGE, Jeffrey M. Introdução à econometria: uma abordagem moderna. São Paulo: Thomson, 2006. Disponível na Biblioteca do Campus