

Introdução à Ciência dos Dados

Aula 11 - Análise de Correlação

Prof. Washington Santos da Silva

15/06/2023

Mestrado Profissional em Administração

Análise de Correlação

Referências

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

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

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.

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.

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

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

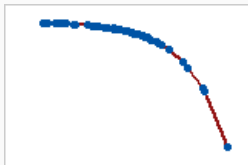
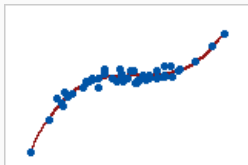
O coeficiente de correlação de Kendall usa pares de observações e determina a força da associação com base no padrão de concordância e discordância entre os pares. Duas VAs X e Y são concordantes se $X_2 - X_1 > 0$ e $Y_2 - Y_1 > 0$, ou $X_2 - X_1 < 0$ e $Y_2 - Y_1 < 0$.

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

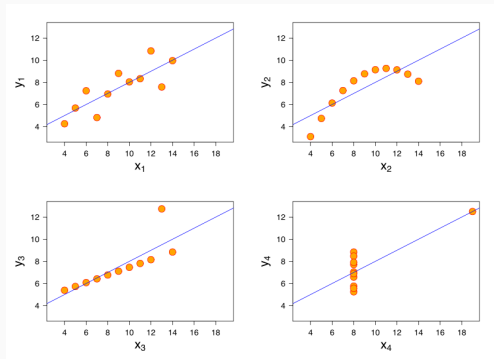
FUNDER, David C.; OZER, Daniel J. Evaluating effect size in psychological research: Sense and nonsense. **Advances in Methods and Practices in Psychological Science**, v. 2, n. 2, p. 156-168, 2019.

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.

Definição

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

Procedimentos Frequentistas de Inferência Estatística

- Intervalos de Confiança
- Testes de Hipóteses

Intervalos de Confiança - Abordagem Frequentista

- Um intervalo de confiança é um **intervalo** construído em torno de um estimador que tem uma determinada **probabilidade (confiança)** de conter o verdadeiro valor do **parâmetro correspondente da população** de interesse.
- 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.

Inferência Estatística: Intervalos de Confiança

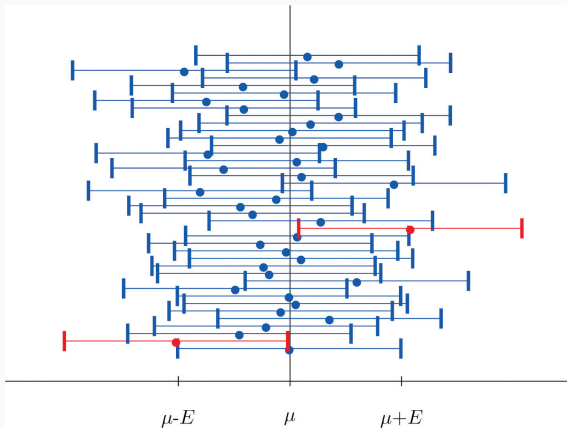
Para construir um intervalo de confiança que contenha o valor verdadeiro do parâmetro populacional θ com uma dada probabilidade/confiança, uma equação da seguinte forma deve ser resolvida:

$$P(L_i \leq \theta \leq L_s) = (1 - \alpha) * 100\%$$

sendo:

- θ o parâmetro a ser estimado
- L_i o limite inferior do intervalo
- L_s o limite superior do intervalo
- α a probabilidade do intervalo não conter θ
- $(1 - \alpha)$ é o nível de confiança da estimativa do intervalo conter θ .

Intervalos de Confiança



Teste de Hipóteses - Abordagem Frequentista

Um teste estatístico de hipóteses é um procedimento que nos permite, com base em certas regras de decisão, confirmar uma hipótese inicial de trabalho, chamada de **hipótese nula**, ou rejeitar esta hipótese nula em favor de uma **hipótese alternativa**.

Inferência Estatística: Teste de Hipóteses

Um teste estatístico de hipóteses baseado em uma amostra e geralmente envolve os seguintes passos:

1. Formular as hipóteses sobre os parâmetros populacionais:
 - A hipótese nula H_0 ,
 - A hipótese alternativa H_1 .
2. Determinar o **nível de significância** α do teste.
3. Determine a distribuição de probabilidade que corresponde à distribuição amostral da estatística de teste.
4. Calcular o valor crítico sob a hipótese nula e definir as regiões de rejeição e de aceitação.
5. Estabelecer as regras de decisão:
 - Se a estatística de teste observada na amostra está localizada na região de aceitação, não rejeitamos a hipótese nula H_0 ;
 - Se a estatística de teste observada na amostra está localizada na região de rejeição, rejeitamos a hipótese nula H_0 em favor da hipótese alternativa H_1 .
 - Tomar a decisão de aceitar ou rejeitar a hipótese nula com base na amostra observada.

Inferência Estatística: 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 $1 - \beta = P(\text{Rejeitar } H_0 / H_0 \text{ é falsa}) = P(H_1 / H_1) = \text{Poder do teste.}$

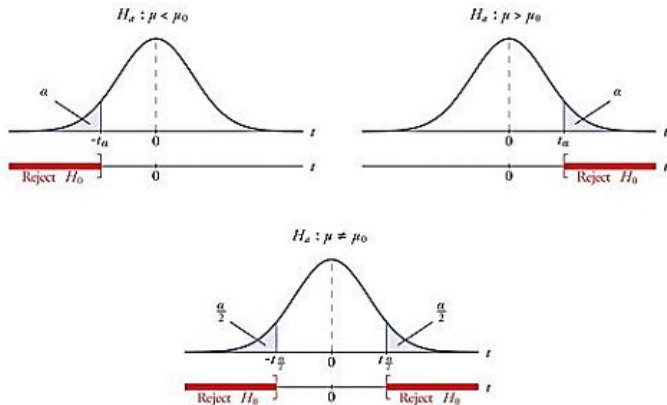
Poder do Teste e Tamanho da Amostra para r

```
# Pacote pwr  
library(pwr)  
  
# funcao que estima o poder do teste ou determine os parametros  
# 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).

Inferência Estatística: Teste de Hipóteses



Conceito

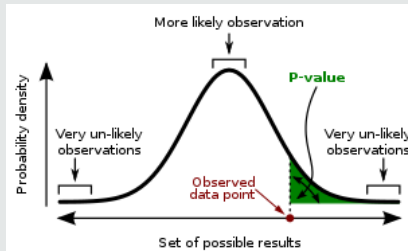
- Em qualquer amostra, por exemplo, a média amostral \bar{x} raramente será exatamente igual ao valor sob a hipótese nula ($H_0 : \mu = \mu_0$, por exemplo). As diferenças entre \bar{x} e μ_0 podem surgir porque a média verdadeira (populacional) (μ), de fato, não é igual a μ_0 (a hipótese nula é falsa) 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.

Inferência Estatística: Teste 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.

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



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

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

Análise de Correlação: Sumário

- 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