AGA 0505 - Análise de Dados em Astronomia

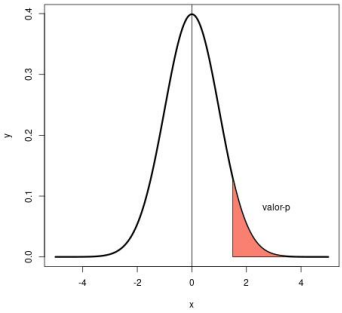
7. Alguns Métodos Frequentistas

Laerte Sodré Jr.

1o. semestre, 2025

aula de hoje:

1. o método da máxima verossimilhança
2. modelos lineares e não-lineares
3. simulações de bootstrap
4. testes de hipóteses
5. comparação de distribuições
6. correlação entre duas variáveis



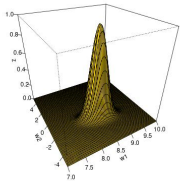
That is the curse of statistics, that it can never prove things, only disprove them!

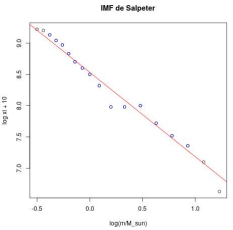
Numerical Recipes- Press, Teukolsky, Vetterling & Flannery

o método da máxima verossimilhança

- R. A. Fisher (1912): *método da máxima verossimilhança* (MV)
- dado um conjunto de dados, D, queremos *ajustar* um *modelo* M que depende de um certo número de parâmetros ajustáveis, w
- ajuste frequentista = inferência dos *valores* dos parâmetros do modelo e de seus erros
- função de verossimilhança: $\mathcal{L}(w)$
  probabilidade dos dados $P(D|w)$, escrita como função dos parâmetros w
- a melhor estimativa para os parâmetros w é a que maximiza a verossimilhança:

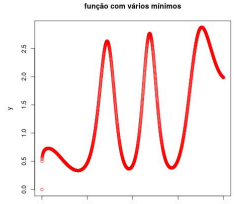
$$\left.\frac{d\mathcal{L}(w)}{dw}\right|_{MV} = 0, \quad \left.\frac{d^2\mathcal{L}(w)}{dw^2}\right|_{MV} < 0$$

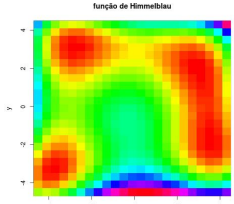




máxima verossimilhança como otimização de funções

- minimização é um problema de *otimização de funções*
- mínimos: podem ser *locais* ou o mínimo *global*
- o mínimo global, no caso geral, pode ser difícil de se achar



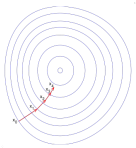


- algoritmo do *"gradiente descendente"* proposto por Cauchy (1847) para derivar os parâmetros orbitais de um planeta

  procedimento iterativo:
  os parâmetros se "movem" na direção oposta à do gradiente da função

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \gamma \nabla \chi^2(\mathbf{w}) \qquad (0 < \gamma < 1)$$

  sempre acha um mínimo (local)
- γ é chamado de *taxa de aprendizado*



exemplo: modelagem de uma função com erros gaussianos

- n dados $D = \{x_i, y_i, \sigma_i\}$, com erros σ_i, que queremos modelar com uma função com m parâmetros w: $y = f(x; w)$
- w: vetor de m parâmetros
- assumindo que as medidas são independentes e os erros gaussianos,

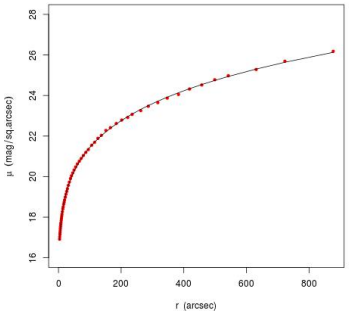
$$\mathcal{L}(w) \propto \prod_{i=1}^{n} P(D_i|w) \propto \exp\left[-\frac{1}{2}\chi^2(w)\right]$$

onde

$$\chi^2(w) = \sum_{i=1}^{n}\left[\frac{(y_i - f(x_i; w))^2}{\sigma_i^2}\right]$$

- solução de MV: minimização do $\chi^2(w)$
- $\hat{w}$ pode, em alguns casos, ser obtido analiticamente ou, na maioria dos casos, numericamente

- boas soluções têm $\chi^2_{red} = \chi^2/(m-1) \simeq 1$
  χ^2_{red}: "χ^2 reduzido"



erro nos parâmetros de um modelo

- além de "ajustar" os parâmetros, precisamos estimar seus erros a partir da "largura" da verossimilhança
- para um modelo com um único parâmetro, w, e expandindo $\mathcal{L}(w)$ em série de Taylor até segunda ordem em torno do máximo $\hat{w}$:

$$\ln \mathcal{L}(w) = \ln \mathcal{L}(\hat{w}) + \left.\frac{d \ln \mathcal{L}(w)}{dw}\right|_{MV} (w-\hat{w}) + \frac{1}{2} \left.\frac{d^2 \ln \mathcal{L}(w)}{dw^2}\right|_{MV} (w-\hat{w})^2 + \ldots$$

- como o segundo termo se anula na vizinhança de $\hat{w}$:

$$\mathcal{L}(w) \approx \mathcal{L}(\hat{w}) \exp\left[-\frac{(w-\hat{w})^2}{2\sigma_{\hat{w}}^2} \right] \quad \text{onde} \quad \frac{1}{\sigma_{\hat{w}}^2} = -\left.\frac{d^2 \ln \mathcal{L}(w)}{dw^2}\right|_{MV}$$

- a verossimilhança em torno do máximo pode ser aproximada por uma gaussiana de largura σ_w
- no caso de vários parâmetros $\mathbf{w} = \{w_1, w_2, ..., w_m\}$, $\mathcal{L}$ em torno do máximo pode ser aproximada por uma gaussiana multivariada:

$$\mathcal{L}(\mathbf{w}) \approx \mathcal{L}(\mathbf{w}_{MV}) \exp\left[-\frac{1}{2} (\mathbf{w} - \mathbf{w}_{MV}) \mathbf{C}^{-1} (\mathbf{w} - \mathbf{w}_{MV}) T \right] + \ldots$$

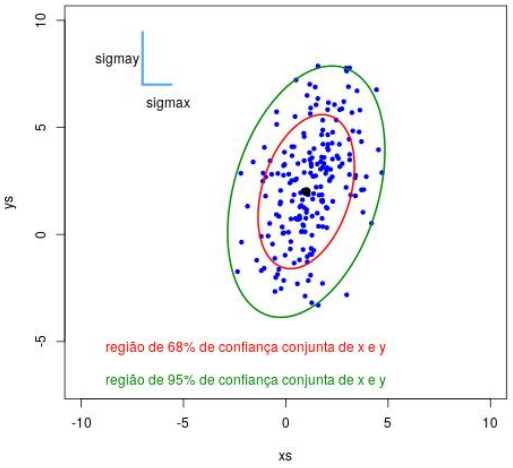
$\mathbf{C}$: matriz de covariância ($m \times m$)

$$C_{ij} = E(w_i - \bar{w}_i) \times E(w_j - \bar{w}_j)$$

- erros nos parâmetros: $\sigma_j^2 = C_{jj}$

erro nos parâmetros de um modelo

- note que nem sempre σ é uma boa forma de representar as incertezas em um ajuste
- quando se tem vários parâmetros, elipses de erro (considerando os parâmetros dois a dois) podem ser mais informativas
- *bootstrap* é um jeito fácil de estimar esse tipo de incerteza



bootstrap: simulando dados com os próprios dados

- *muitas decisões podem ser tomadas usando-se simulações*
- *bootstrap*: técnica baseada em reamostragem dos dados (Efron, 1979)
- suponha que tenhamos um conjunto de dados e queremos determinar a distribuição de alguma estatística w, determinada a partir desses dados
- exemplo: temos um conjunto de pontos (x, y) e queremos ajustar uma reta a eles, $y = a + bx$
- *bootstrap* permite determinar os erros e intervalos de confiança para os parâmetros (a, b)

- ideia básica do *bootstrap*:
  - seja $\mathcal{D}_0$ o conjunto de dados
  - um novo conjunto de dados $\mathcal{D}_i$ é simulado a partir de $\mathcal{D}_0$ por reamostragem *com substituição*
  - a estatística w_i é determinada a partir desses dados simulados
  - pode-se fazer isso muitas vezes e assim obter-se muitas amostras de w
  - estas amostras podem então ser usadas para estimar os erros e intervalos de confiança de w

ajuste de uma reta aos dados

- suponha que temos n dados $\{x_i, y_i\}$, com erros σ em y_i
- queremos modelar os dados com uma reta: $y = f(x; a, b) = a + bx$
- este problema é denominado *regressão linear ordinária* (OLS)
  - regressão: procedimentos para estimar relações entre variáveis
    y é uma variável contínua
    (se y é discreta: *classificação*)
  - linear: *em relação aos parâmetros* a,b

- note que muitas vezes se pode fazer mudanças de variáveis e usar a regressão linear:
  - $y = A10^{bx} \longrightarrow \log y = \log A + bx$
  - $y = Ax^b \longrightarrow \log y = \log A + b \log x$
  - $f(y) = a + bg(x)$
- f e g podem ser funções arbitrárias!
- o importante é que o **modelo seja linear nos parâmetros** a e b
  (i.e., não dependa de $a^2, ab, ...$)

mínimos quadrados linear geral

- modelos lineares são muito úteis, pois podem envolver funções arbitrárias de x
- modelo linear com m parâmetros a_k:

$$y = \sum_k a_k X_k(x),$$

onde $X_k(x)$ são m funções arbitrárias de x

- exemplo: polinômios de grau n: $y = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$
- máxima verossimilhança = minimização do χ^2:

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{n} \left[\frac{y_i - \sum_k a_k X_k(x_i)}{\sigma_i} \right]^2$$

- solução:

$$a_j = \sum_{k=1}^{m} [\alpha]^{-1}_{jk} \beta_k = \sum_{k=1}^{m} C_{jk} \beta_k,$$

$$\alpha_{kj} = \sum_{i=1}^{n} \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2} \qquad \beta_k = \sum_{i=1}^{n} \frac{y_i X_k(x_i)}{\sigma_i^2}$$

- $\mathbf{C} = \alpha^{-1}$: matriz de covariância
- erros dos parâmetros: $\sigma(a_j)^2 = C_{jj}$

modelos não-lineares

- em geral, em problemas não-lineares, os parâmetros que minimizam o $\chi^2(\mathbf{w})$

$$\chi^2(\mathbf{w}) = \sum_{i=1}^{N} \left[\frac{y_i - f(x_i; \mathbf{w})}{\sigma_i} \right]^2$$

  devem ser obtidos numericamente

- exemplo: ajuste do perfil de brilho da galáxia NGC 4472 (M49) com a lei de Sérsic:

$$\log I(r) = logI_e - b_n[(r/r_e)^{1/n} - 1]$$

$b_n \simeq 0.868n - 0.142$, para $0.5 < n < 16.5$

- 3 parâmetros: $\mathbf{w} = \{I_e, r_e, n\}$
  - r_e: raio efetivo - raio que contém metade da luminosidade da galáxia
  - I_e: brilho superficial em r_e
  - n: expoente de Sérsic
- o perfil de brilho superficial é medido em unidades de mag/arcsec^2:

$$\mu(r) = \mu_0 - 2.5 \log I(r) \text{ mag/arcsec}^2$$

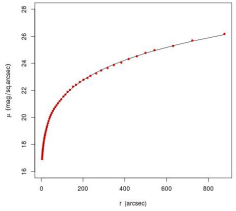
- logo, se $\mu_0' = \mu_0 - 2.5 \log I_e$,

$$\mu(r) = \mu_0' + 2.5 b_n[(r/r_e)^{1/n} - 1]$$

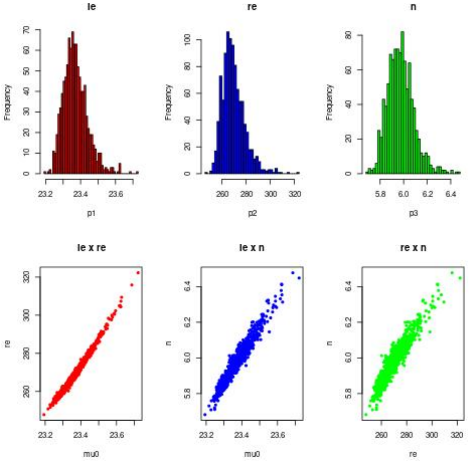
exemplo: perfil de brilho de NGC 4472 (M49)

M49: galáxia mais luminosa do aglomerado de Virgo

perfil de brilho superficial e modelo

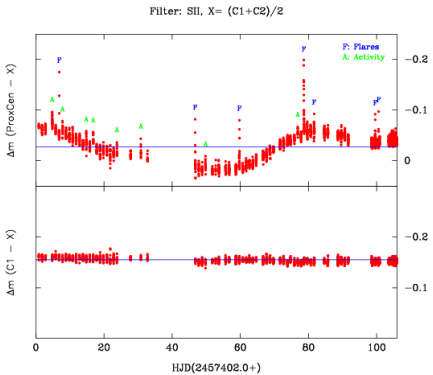


1000 simulações de bootstrap



teste de hipóteses clássico

- tipo de *inferência estatística* frequentista
- é um teste de uma afirmação sobre uma ou mais populações, inferida usualmente de uma *estatística* da distribuição
- exemplos:
  - esta estrela é variável?
  - esta galáxia está mais próxima que 5 Mpc?
  - esta vacina é eficiente?
  - X é o pai da criança?
  - o réu é culpado?



Top: Proxima Centauri differential light curve with SII filter obtained during the PRD 2016 campaign; flares and activity features are clearly visible. The rotation period is 83 days. Bottom: Differential photometry of the C1 comparison star to check its stability.

https://reddots.space/differential-photometry-in-practice/

teste de hipóteses clássico: o que é

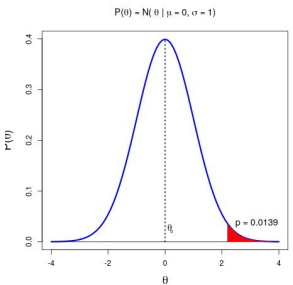
- conceitos importantes: *hipótese nula*, *nível de significância*, *nível de confiança*, valor-p
- duas hipóteses estão envolvidas:
  - a *hipótese nula* H_0: a hipótese que se quer testar (a hipótese *default*)
  - a *hipótese alternativa* H_A: o que acreditamos ser verdadeiro se a hipótese nula é rejeitada
- supõe-se que se conhece a distribuição da estatística de interesse
- o teste dá o valor-p: probabilidade de se obter um valor tão extremo para essa estatística
- o valor-p deve ser *suficientemente* pequeno para se rejeitar H_0, menor que um certo *nível de significância*

- nível de significância α: probabilidade de se errar se H_0 é verdadeira
  valores típicos: 0.05, 0.01
- nível de confiança: $1 - \alpha$
  valores típicos: 95%, 99%
- decisão: se $p < \alpha$ rejeitamos H_0;
  se $p > \alpha$ não rejeitamos H_0
- se H_0 é rejeitada, escolhemos H_A
- note que não rejeitar H_0 não implica que H_0 seja verdadeira!
- em Astronomia raramente se usa α: apenas se avalia o valor-p

um exemplo

- temos uma teoria, onde o valor previsto para uma certa variável é $\theta_0 = 0$; fazemos uma medida e obtemos $\theta = 2.2$
- H_0: $\theta = \theta_0$
- supomos que conhecemos a distribuição de θ, $P(\theta)$, e de sua distribuição cumulativa, $F(\theta)$
  ex.: $P(\theta) \sim N(\mu = 0, \sigma = 1)$
- com $\theta = 2.2$ podemos rejeitar H_0?
- vamos supor um nível de significância α de 1%: $\alpha = 0.01$
- valor p: probabilidade de se obter um valor tão extremo quanto a medida θ, dado H_0
  no caso, $p = 1 - F(\theta) = 0.0139$

- decisão: se $p < \alpha$ rejeita-se H_0
- no caso H_0 não é rejeitada
- para rejeitar H_0, o valor p tem que ser suficientemente pequeno, menor que α!



nível de significância, nível de confiança, intervalo de confiança

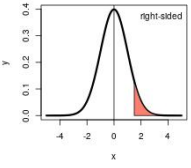
- nível de significância α:
  - probabilidade de se errar quando H_0 é verdadeira
  - valores típicos: 0.05, 0.01
  - usado em testes de hipótese
- nível de confiança: $1 - \alpha$
  - é a probabilidade de, se a medida for feita muitas vezes, se obter o valor θ
  - valores típicos: 95%, 99%
  - note que α é arbitrário!
- intervalo de confiança:
  - se a medida for feita muitas vezes, uma fração $1 - \alpha$ cai dentro do intervalo de confiança

- intervalo de credibilidade bayesiano: dá um intervalo ligado à probabilidade do valor de um parâmetro ou hipótese

testes unilaterais e bilaterais

- em inglês: testes *right-sided*, *left-sided*, *two-sided*
- exemplo: queremos saber se
  - uma medida é maior que a média? teste *right-sided*: área a direita
  - uma medida é menor que a média? teste *left-sided*: área a esquerda
  - uma medida é diferente da média? teste *two-sided*: área dos dois lados
  - p: área sob a distribuição correspondente à estatística testada- área na parte direita, esquerda, ou nas duas partes da distribuição

- para uma hipótese ser rejeitada, a área, p, deve ser *pequena*, menor que o nível de significância escolhido!



duas distribuições têm a mesma média?

- o teste/estatística depende do que se quer testar!
- exemplo: duas distribuições têm a mesma média?
- aplica-se o teste t de Student

- Student era o pseudônimo de William Sealy Gosset, um químico trabalhando para a cervejaria Guinness na Irlanda
- ele desenvolveu a estatística t para monitorar a qualidade do *stout*!

duas distribuições têm a mesma média?

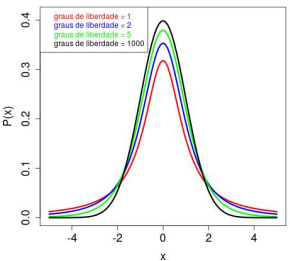
- queremos verificar se a média $\bar{x}$ de N medidas é consistente com um certo valor esperado μ
- para se testar a média de uma amostra em relação à de uma população usa-se a estatística t:

$$t = \frac{\bar{x} - \mu}{\sigma_s / \sqrt{N}}$$

σ_s: desvio padrão da amostra

- a estatística t distribui-se como uma *distribuição t com $\nu = N - 1$ graus de liberdade*:

$$P(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$



- note que, conforme ν aumenta, $P(t)$ se aproxima de uma gaussiana
- o valor p é calculado a partir dessa distribuição

comparação de duas distribuições com o χ^2

- testes do χ^2 para comparação de um histograma com um modelo ou entre dois histogramas
- estatística χ^2 para comparação de dados e modelo

$$\chi^2 = \sum_{i=1}^{N_{bin}} \frac{(N_i - n_i)^2}{n_i}$$

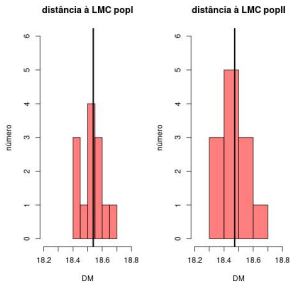
sendo N_i e n_i as contagens observadas e do modelo ou

- estatística χ^2 para comparação de dois conjuntos, N e M

$$\chi^2 = \sum_{i=1}^{N_{bin}} \frac{(N_i - M_i)^2}{N_i + M_i}$$

binados e com o mesmo número de bins

- comparação entre os módulos de distâncias de estrelas de população I e população II da Grande Nuvem de Magalhães



H0: as duas populações têm a mesma média

valor p = 0.092

comparação de duas distribuições com o χ^2

- teste do χ^2 para tabelas de contingência de variáveis nominais:
  os valores são membros de um conjunto não-ordenado (tipos morfológicos de galáxias; nomes dos estados do Brasil)
  - exemplo: será que o tipo morfológico de uma galáxia se correlaciona com sua atividade nuclear?
    x: tipo morfológico de uma galáxia (E, S0, Sa...);
    y: tipo de atividade nuclear (passiva, formação estelar, Sey I, Sey II, LINER)
  - H_0: as duas variáveis não estão associadas

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}},$$

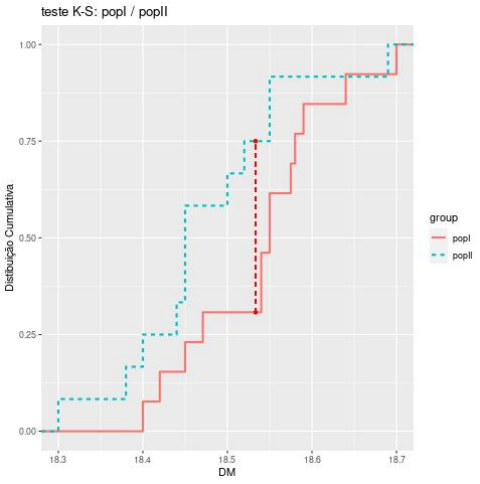
onde N_{ij} é o número de objetos de tipo x_i com atividade nuclear y_j e n_{ij} é o número esperado sob H_0

- se as variáveis forem independentes, pode-se mostrar que:

$$n_{ij} = \frac{(\sum_j N_{ij})(\sum_i N_{ij})}{N} \qquad N = \sum_{ij} N_{ij}$$

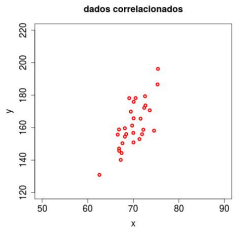
comparação de duas distribuições com o teste KS

- teste de Kolmogorov-Smirnov (KS):
  - compara a distribuição cumulativa dos dados de diferentes amostras
  - se aplica tanto a conjuntos de dados binados como não binados
  - teste *não-paramétrico*: não assume uma forma específica para cada distribuição
  - teste “robusto”



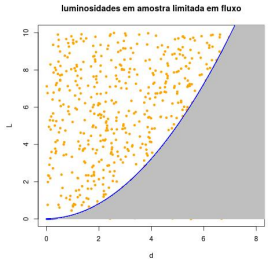
correlação entre duas variáveis

- vamos considerar agora *medidas de associação* entre duas variáveis
- uma variável está correlacionada ou depende da outra?
- o conhecimento de uma variável ajuda a saber o valor da outra?



- cuidado: correlações podem não significar que as variáveis estejam intrinsecamente correlacionadas ou que uma seja causa da outra

exemplo: efeitos de seleção!



a luminosidade mínima observada em uma amostra limitada em fluxo depende da distância

coeficiente de correlação de Pearson

- o coeficiente de correlação de Pearson mede a força de uma correlação **LINEAR**
- sejam $\{x_i\}$ e $\{y_i\}$, $i = 1, ..., N$ duas variáveis aleatórias
- o coeficiente de correlação de Pearson é definido como

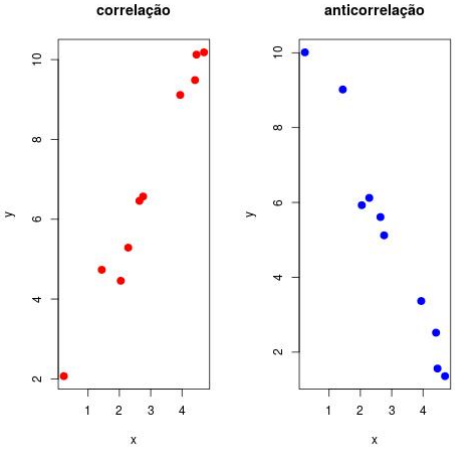
$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (-1 \leq \rho \leq 1)$$

onde

$$\sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2, \qquad \sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

$$\sigma_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

- se $\rho > 0$: correlação; se $\rho < 0$: anticorrelação



coeficiente de correlação de Spearman

- testes não-paramétricos: não pressupõem uma forma para a distribuição dos dados
- coeficiente de correlação de Spearman: mede o *ordenamento relativo* de duas variáveis
- suponha que se ordene os dados $\{x_i\}$ e $\{y_i\}$ tal que $\{X_i\}$ e $\{Y_i\}$ representem a posição dos dados na sequência ordenada: $1 < X_i < N$ e $1 < Y_i < N$
- coeficiente de correlação de ordem de Spearman:

$$\rho_S = 1 - 6\frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{N^3 - N}$$

- Spearman é mais robusto que correlação linear: se a correlação é detectada, provavelmente é real
- Spearman mede monotonicidade: r_s é o mesmo para a relação entre x e y e para $\log x$ e $\log y$ (para x e y positivos)