AGA 0505 - Análise de Dados em Astronomia

5. Testes de Hipóteses, Comparação de Distribuições e Análise de Correlação

Laerte Sodré Jr.

1o. semestre, 2023

aula de hoje:

1. teste de hipóteses clássico
2. duas distribuições têm a mesma média?
3. duas distribuições têm a mesma variância?
4. erros nos testes de hipóteses

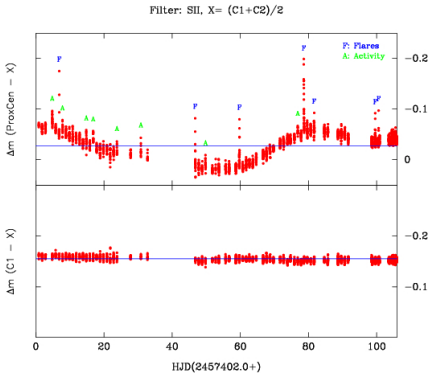
5. comparação de duas distribuições
6. análise de correlação

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

Numerical Recipes- Press, Teukolsky, Vetterling & Flannery

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?
  - 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.

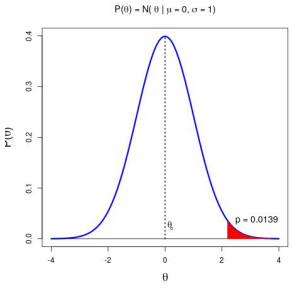
https://reddots.space/differential-photometry-in-practice/

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
- nível de significância α: probabilidade de se errar se H_0 é verdadeira
- nível de confiança: $1 - \alpha$
- valor-p: o teste dá esse valor, que depende do valor da estatística usada e de sua distribuição
- 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 muitas vezes não 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 $\theta = 2.2 \pm 1$
- 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 $\theta = 2.2 \pm 1$ 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!



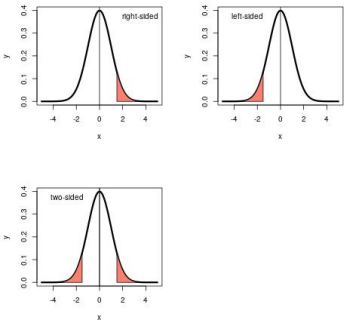
nível de significância, nível de confiança, intervalo de confiança

- 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:
  - é a probabilidade de, se a medida for feita muitas vezes, se obter o valor θ: $1 - \alpha$
  - valores típicos: 95%, 99%
- intervalo de confiança:
  - se a medida for feita muitas vezes, uma fração $1 - \alpha$ cai dentro do intervalo de confiança

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?

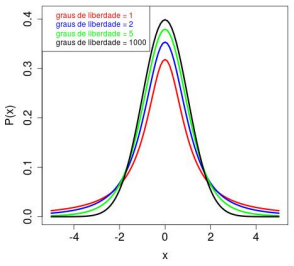
- queremos verificar se a média $\bar{x}$ de N medidas é consistente com um certo valor esperado μ
- para se testar a média de uma amostra em relação à de uma população usa-se a estatística t:

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

σ_s: desvio padrão da amostra

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

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



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

comparação de variâncias

- dados dois conjuntos de dados, $\{x_i\}$ e $\{y_i\}$, com N e M elementos, respectivamente, queremos testar
  H_0: $\sigma_x = \sigma_y$

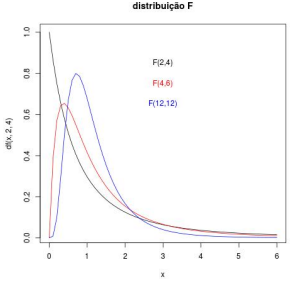
- para distribuições gaussianas a estatística-teste é a F

- a estatística-teste

$$F_* = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2/(N-1)}{\sum_{j=1}^{M}(y_i - \bar{y})^2/(M-1)} = \frac{\sigma_x^2}{\sigma_y^2}$$

segue uma distribuição F com $\nu_1 = N - 1$ e $\nu_2 = M - 1$ graus de liberdade

$$F(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})(\frac{\nu_1}{\nu_2})^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})(1+\frac{\nu_1}{\nu_2}x)^{\frac{\nu_1+\nu_2}{2}}}$$



- o valor p é calculado a partir dessa distribuição

erros no teste de hipótese

- hipótese nula H_0; H_A: uma hipótese alternativa
  exemplo:
  - H_0: a estrela não é variável
  - H_A: a estrela é variável
- note que:
  - a rejeição de H_0 não implica que H_A seja verdadeira
  - a não rejeição de H_0 não implica que H_0 seja verdadeira
- tipos de erro:
  - erro tipo I: H_0 é rejeitada quando é verdadeira (**falso positivo**)
  - erro tipo II: H_0 não é rejeitada quando é falsa (**falso negativo**)
- note que esses erros dependem apenas de H_0, não de H_A

	variabilidade	**verdadeira**
decisão	não variável	variável
não variável	ok!	**erro tipo II**
variável	**erro tipo I**	ok!

	verdade	
decisão	H_0	H_A
H_0	ok!	**erro tipo II**
H_A	**erro tipo I**	ok!

tipos de variáveis

- existem testes para diferentes tipos de variáveis e distribuições
- convém distinguir entre 3 tipos de variáveis:
  - nominal: os valores são membros de um conjunto não-ordenado:
    Ex.: tipos morfológicos de galáxias; os nomes dos estados do Brasil
  - ordinal: os valores são membros de um conjunto ordenado discreto:
    Ex.: a ordem dos planetas, a sequência de pré-requisitos de uma disciplina
    o que importa é que os valores estão intrinsecamente ordenados
  - contínua: os valores são números reais
    Ex.: distâncias, tempo
- uma variável contínua pode ser transformada em ordinal por binagem e em nominal se a ordem dos bins for desconsiderada
- *muitas decisões podem ser tomadas usando-se simulações*

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

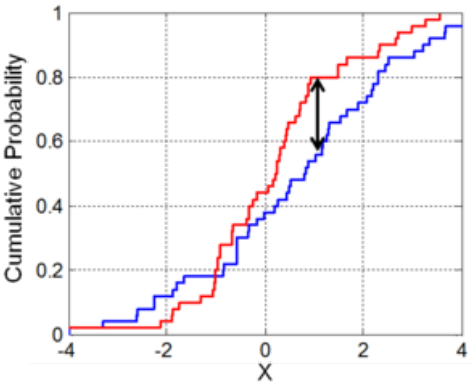
- teste do χ^2 para tabelas de contingência de variáveis nominais
  - exemplo: x: tipo morfológico de uma galáxia (E, S0, Sa...); y: tipo de atividade nuclear (passiva, formação estelar, Sey I, Sey II, LINER)
  - H_0: as duas variáveis não estão associadas

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

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

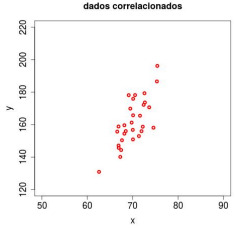
comparação de duas distribuições com o teste KS

- 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”



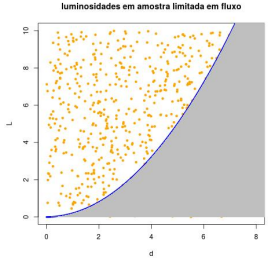
Correlação entre duas variáveis

- 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

exemplo: efeitos de seleção!



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

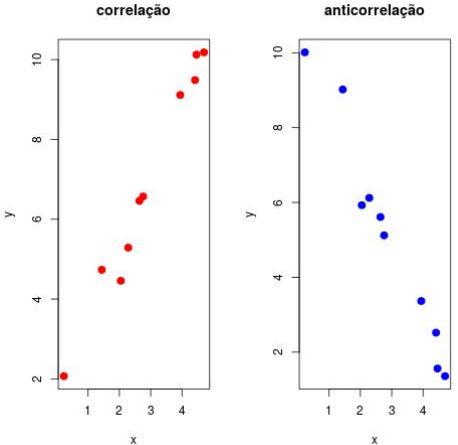
$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (-1 \leq \rho \leq 1)$$

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)