AGA 0505 - Análise de Dados em Astronomia

2. Probabilidades

Laerte Sodré Jr.

1o. semestre, 2023

aula de hoje:

1. o que são probabilidades
2. distribuições de probabilidades
3. a distribuição normal ou gaussiana
4. probabilidades condicionais e conjuntas
5. as regras fundamentais das probabilidades e o teorema de Bayes
6. combinação de distribuições
7. exercícios sobre probabilidades

A teoria das probabilidades é o senso comum reduzido ao cálculo

Pierre-Simon Laplace

o que são probabilidades?

- teoria das probabilidades: provê um meio de modelar dados/observações e quantificar as incertezas

- fontes de incertezas:
  - incerteza intrínseca ao fenômeno, como na Mecânica Quântica ou em fenômenos caóticos e/ou estocásticos
  - incerteza nas medidas ou observações
  - incerteza nos modelos

  tudo isso leva a uma incerteza nas predições e/ou na interpretação dos dados

- há uma disputa dentro da Estatística, tendo como base a natureza das probabilidades:

  métodos bayesianos x
  métodos frequentistas

o que são probabilidades

- probabilidade frequentista:
  - medida da frequência de eventos (em vários experimentos ou ensemble de sistemas estatisticamente equivalentes)
  - $P(x)$: o número de vezes que um evento ocorre dividido pelo número total de tentativas, no limite de um número infinito de tentativas
  - $P(x)$: número entre 0 e 1 que mede a frequência com que a proposição x aparece em uma população ou amostra
- problemas:
  - não comporta eventos únicos ou não repetíveis
  - lida com propriedades assintóticas (“no limite...”)

- probabilidade bayesiana:
  medida da plausibilidade de um evento
  - $P(x)$: número entre 0 e 1 que mede o grau com que a proposição x é verdadeira
    (com 0 falsa e 1 verdadeira)

- problemas:
  - nem sempre se consegue definir um modelo probabilístico adequado
  - em particular, pode-se ter uma certa ambiguidade devido à escolha de diferentes priores
    (veremos isso mais tarde)

o que são probabilidades?

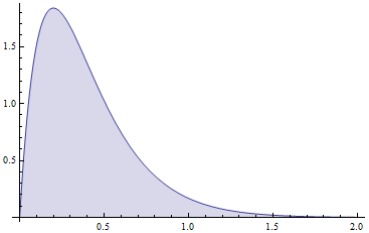
- probabilidade como medida da plausibilidade (chatGPT):
  - a probabilidade pode ser vista como uma medida da incerteza associada a uma determinada proposição ou evento, em que valores mais altos de probabilidade indicam que é mais plausível ou mais provável que a proposição ou evento seja verdadeiro, enquanto valores mais baixos indicam que é menos provável
  - exemplo: um meteorologista diz que há 70% de chance de chuva amanhã: com base na análise das condições climáticas atuais, é plausível que haja chuva, mas ainda há uma certa incerteza sobre isso; a probabilidade de 70% sugere que há uma maior chance de chuva do que de não chuva, mas ainda existe uma possibilidade significativa de que não chova

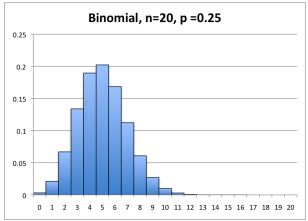
distribuições de probabilidades

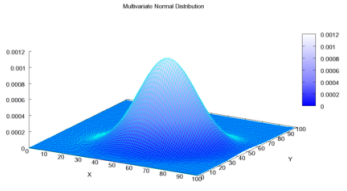
- x: variável aleatória contínua ou discreta
- *variável aleatória*: variável cujos valores são resultados de um processo aleatório ou estocástico, obedecendo a uma certa distribuição de probabilidades, $P(x)$
- notação: $x \sim P(x)$
  x é uma variável aleatória com distribuição que obedece a $P(x)$
- $P(x)$: número entre 0 e 1 que mede a incidência da variável x ou o grau com que uma proposição x é verdadeira
- $P(x)$ pode ser uma função discreta ou contínua
- $P(x)$ contínua: função de densidade de probabilidades (FDP ou PDF):
  $P(x)dx$: número entre 0 e 1 que mede o grau de plausibilidade de que uma certa variável x esteja entre x e $x + dx$
- no caso de variáveis discretas a distribuição de probabilidades é chamada de função de massa de probabilidades (FMPs)
- as FDPs/FMPs podem ser multivariadas, i.e., funções de várias variáveis:
  $P(x, y, z)$

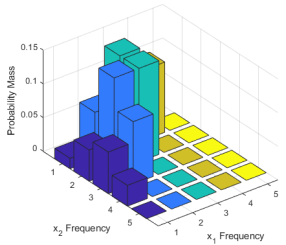
distribuições de probabilidades

exemplos:









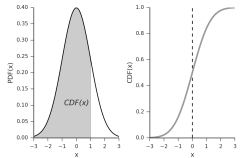
distribuições cumulativas

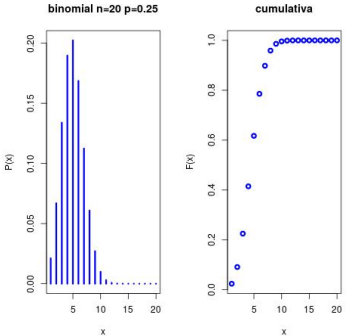
- $P(x)$: função de distribuição de probabilidades (FDP)
- FDP cumulativa:

$$F(x) = \int_{-\infty}^{x} P(x')dx'$$

$$F(\infty) = \int_{-\infty}^{\infty} P(x')dx' = 1$$

probabilidades são normalizadas!





probabilidades condicionais e conjuntas

- probabilidades são, normalmente, condicionais, isto é, dependem ou podem depender de outras proposições:
  $P(x|y)$: lê-se probabilidade de x dado y
  $P(x|y)$ mede a plausibilidade de x se a proposição y é admitida como verdadeira
  $P(x|y,z,w)$: a probabilidade de x é condicional a tudo o que estiver à direita da barra "|"

  exemplo: gaussiana

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[-\frac{(x-\mu)^2}{2\sigma^2}\Big]$$

- probabilidade conjunta: $P(x,y)$
  probabilidade conjunta de duas proposições x e y
- $P(x,y|z)$: probabilidade conjunta de x e y dado z

- $P(x) = P(x|H)$: toda probabilidade depende de hipóteses (H), implícitas ou explícitas
  por exemplo: a distribuição é gaussiana

distribuição normal ou gaussiana

- distribuição gaussiana (univariada):

$$P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

2 parâmetros:

- média μ
- desvio padrão σ (ou variância σ^2)

- notação: $x \sim N(\mu, \sigma)$

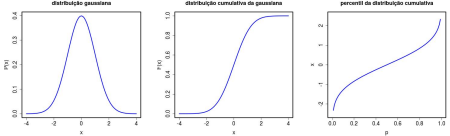
- distribuição cumulativa:

$$F(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$$

onde

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

é a *função erro* (esse nome tem razões históricas)



as regras fundamentais das probabilidades

- as regras fundamentais das probabilidades:
  - $P(x) \geq 0$
  - regra da soma: $P(x) + P(\bar{x}) = 1$
    onde $\bar{x}$ representa a probabilidade de x ser falso
  - regra do produto (ou da cadeia): $P(x, y) = P(x|y) \times P(y) = P(y|x) \times P(x)$
- teorema de Bayes ($\sim$1740)
  - da regra do produto vem que:

$$P(x, y) = P(x|y) \times P(y) \qquad P(y, x) = P(y|x) \times P(x)$$

  - como $P(x, y) = P(y, x)$, temos o teorema de Bayes:

$$P(x|y) = \frac{P(y|x) \times P(x)}{P(y)}$$

as regras fundamentais das probabilidades: algumas consequências

- regra da soma:

$$P(x) + P(\bar{x}) = 1$$

- generalização para um conjunto de proposições discretas ou contínuas, “mutuamente exclusivas e exaustivas”:

$$\sum_k P(x_k) = 1 \qquad \int P(x)dx = 1$$

regra de normalização das probabilidades

- marginalização: quando se tem um conjunto y_k de proposições “MEE”

$$P(x) = \sum_k P(x, y_k)$$

ou

$$P(x) = \int P(x, y)dy = \int P(x|y) \times P(y)dy$$

muito útil!

as regras fundamentais das probabilidades: algumas consequências

- aplicação da regra do produto: $P(x, y) = P(x|y) \times P(y)$
  - vamos supor que x e y são variáveis independentes:

    $$P(x|y) = P(x) \text{ e } P(y|x) = P(y)$$

  - portanto,

    $$P(x, y) = P(x|y)P(y) = P(x)P(y)$$

    a probabilidade de duas proposições independentes é o produto das probabilidades de cada uma

- aplicação da regra do produto:

$$P(x, y, z) = P(x|y, z)P(y, z)$$

$$P(y, z) = P(y|z)P(z)$$

logo,

$$P(x, y, z) = P(x|y, z)P(y|z)P(z)$$

o teorema de Bayes e a análise de dados

o teorema de Bayes oferece um procedimento lógico para condução da análise estatística: considere a análise de um conjunto de dados D com um modelo M que tem parâmetros w:

- aprendizado/inferência de parâmetros

$$P(w|D,M) = \frac{P(D|w,M)P(w|M)}{P(D|M)} = \frac{\text{verossimilhanca dos dados} \times \text{prior dos parametros}}{\text{evidencia}}$$

- predição (de um novo dado x)

$$P(x|D,M) = \int P(x|w,D,M)P(w|D,M)dw$$

- comparação de modelos

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

- veremos isso melhor nas próximas aulas

combinando distribuições

- muitas vezes conhecemos a distribuição de uma variável mas queremos saber a distribuição de uma quantidade derivada, $y = f(x)$
  ex: temos a distribuição de magnitudes e queremos a distribuição de fluxos
- $P(x)$: pdf de x
- $P(y)$: pdf de y
- a densidade de probabilidades é conservada,

$$P(x)dx = P(y)dy, \quad \text{logo} \quad P(y) = \left|\frac{dx}{dy}\right| P(x)$$

(o módulo é para assegurar $P \geq 0$)

- exemplo: suponha que $P(x) = \exp(-x)$ (com $x > 0$) e queremos $P(y)$, onde $y = \ln x$
  - como $x = \exp(y)$,

$$P(y) = P(x)/|dy/dx| = x \exp(-x)$$

ou

$$P(y) = \exp(y - \exp(y))$$

- esta técnica se torna difícil de aplicar para mais de uma variável
- cuidado se $f(x)$ não é monotônica!

exercícios sobre probabilidades

jogamos duas moedas: qual é a probabilidade de sair duas caras?

cada moeda pode dar:
cara (H, heads) ou coroa (T, tails)

- espaço amostral: o conjunto dos resultados possíveis
- o espaço amostral é $S = \{(H,T),(H,H),(T,H),(T,T)\}$
- seja E o evento "sair duas caras": $E = \{(H,H)\}$
- então, $P(E) = n(E)/n(S) = 1/4$

outro jeito de resolver, considerando que cada jogada de uma moeda é independente das demais:

- probabilidade de sair uma cara ao se jogar a moeda: 1/2
- probabilidade de sair uma cara ao se jogar novamente a moeda: 1/2
- probabilidade de sair duas caras: $1/2 \times 1/2 = 1/4$

este problema pode ser modelado com a **distribuição de probabilidades binomial**

exercícios sobre probabilidades

uma caixa contém 9 bolas:
4 azuis, 2 amarelas e 3 vermelhas

- tiramos uma bola ao acaso e a retornamos à caixa (amostragem COM substituição)
- repetimos isso 3 vezes
- qual é a probabilidade de termos tirado 2 bolas azuis e 1 vermelha?

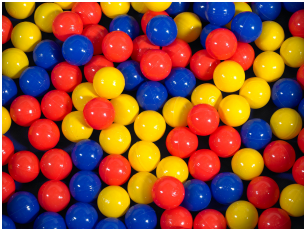


Table 1: Espaço amostral

possibilidade	cores	probabilidades	probabilidade cumulativa
01	AZ,AZ,AZ	4/9 x 4/9 x 4/9 = 0,0878	0,0878
02	AZ,AZ,AM	4/9 x 4/9 x 2/9 = 0,0439	0,1317
03	**AZ,AZ,V**	**4/9 x 4/9 x 3/9 = 0,0658**	**0,1975**
04	AZ,AM,AZ	4/9 x 2/9 x 4/9 = 0,0439	0,2414
05	AZ,AM,AM	4/9 x 2/9 x 2/9 = 0,0219	0,2634
06	AZ,AM,V	4/9 x 2/9 x 3/9 = 0,0329	0,2963
07	**AZ,V,AZ**	**4/9 x 3/9 x 4/9 = 0,0658**	**0,3621**
08	AZ,V,AM	4/9 x 3/9 x 2/9 = 0,0329	0,3951
09	AZ,V,V	4/9 x 3/9 x 3/9 = 0,0494	0,4444
10	AM,AZ,AZ	2/9 x 4/9 x 4/9 = 0,0439	0,4883
11	AM,AZ,AM	2/9 x 4/9 x 2/9 = 0,0219	0,5103
12	AM,AZ,V	2/9 x 4/9 x 3/9 = 0,0329	0,5432
13	AM,AM,AZ	2/9 x 2/9 x 4/9 = 0,0219	0,5652
14	AM,AM,V	2/9 x 2/9 x 3/9 = 0,0165	0,5816
15	AM,V,AZ	2/9 x 3/9 x 4/9 = 0,0329	0,6145
16	AM,V,AM	2/9 x 3/9 x 2/9 = 0,0165	0,6310
17	AM,V,V	2/9 x 3/9 x 3/9 = 0,0247	0,6557
18	**V,AZ,AZ**	**3/9 x 4/9 x 4/9 = 0,0658**	**0,7215**
19	V,AZ,AM	3/9 x 4/9 x 2/9 = 0,0329	0,7545
20	V,AZ,V	3/9 x 4/9 x 3/9 = 0,0494	0,8038
21	V,AM,AZ	3/9 x 2/9 x 4/9 = 0,0329	0,8368
22	V,AM,AM	3/9 x 2/9 x 2/9 = 0,0165	0,8532
23	V,AM,V	3/9 x 2/9 x 3/9 = 0,0247	0,8779
24	V,V,AZ	3/9 x 3/9 x 4/9 = 0,0494	0,9273
25	V,V,AM	3/9 x 3/9 x 2/9 = 0,0247	0,9520
26	V,V,V	3/9 x 3/9 x 3/9 = 0,0370	0,9890
27	AM,AM,AM	2/9 x 2/9 x 2/9 = 0,0110	1,0000

as linhas em negrito mostram os 3 casos em que tiramos 2 bolas azuis e 1 vermelha; portanto, a probabilidade total de tirarmos 2 bolas azuis e 1 vermelha é de

3 x (3/9 x 4/9 x 4/9) = 16/81 $\simeq$ 0,1975

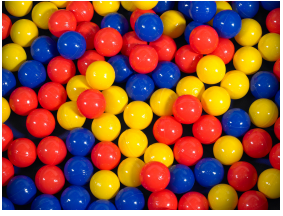
exercícios sobre probabilidades

a mesma caixa:
4 bolas azuis, 2 amarelas e 3 vermelhas

- tiramos uma bola ao acaso e NÃO a retornamos à caixa (amostragem SEM substituição)
- repetimos isso 3 vezes
- qual é a probabilidade de termos tirado 2 bolas azuis e 1 vermelha?

- possibilidades:
  P(A,A,V)=4/9×3/8×3/7=36/504
  P(A,V,A)=4/9×3/8×3/7=36/504
  P(V,A,A)=3/9×4/8×3/7=36/504

- probabilidade de tirarmos 2 bolas azuis e 1 vermelha:
  $3 \times 36/504 = 1/14 \simeq 0.2142$



exercícios sobre probabilidades

- probabilidade de A ou B:
  dados 2 eventos, A e B, qual é a probabilidade de se ter A OU B?

$$P(A \text{ ou } B) = P(A) + P(B) - P(A \text{ e } B)$$

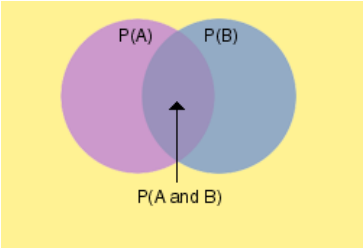


diagrama de Venn

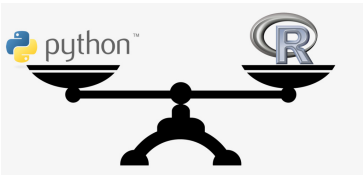
- exemplo: qual é a probabilidade de se tirar uma carta vermelha ou um 3 em um baralho de 52 cartas (4 naipes de 13 cartas)?
  $P(V) = 26/52$; $P(3) = 4/52$;
  $P(V, 3) = 2/52$
  e, portanto,
  $P(V \text{ ou } 3) = 26/52 + 4/52 - 2/52 =$
  $= 28/52 = 7/13 \simeq 0.538...$

- eventos mutuamente exclusivos: se um ocorre o outro não ou os dois não podem ocorrer ao mesmo tempo
  $P(A \text{ e } B) = 0$

exercícios sobre probabilidades

- 40% dos estudantes da classe disseram conhecer R e Python.
- 60% dos estudantes disseram conhecer Python.

- Qual é a probabilidade de um estudante que conhece Python conhecer também R?



- seja A conhecer Python e B conhecer R

- dados: $P(A, B) = 0.4 \quad P(A) = 0.6$

- o que se quer: $P(B|A)$

- probabilidades condicionais: $P(A, B) = P(B|A)P(A)$

- logo, $P(B|A) = P(A, B)/P(A) = 0.4/0.6 = 2/3 \simeq 0.67$

tabela de contingência ou matriz de confusão

- vamos considerar duas variáveis discretas que podem assumir dois valores cada uma (0 ou 1), produzindo 4 resultados possíveis
- ex.: teste médico (Ivezić+, MLA)
  - T: resultado do teste
    0- negativo, 1-positivo
  - D: estado de saúde do paciente

    0- não tem a doença, 1: tem a doença
- espaço amostral:
  (T, D) ={ (0,0),(0,1),(1,0),(1,1)}
- vamos supor que conhecemos as probabilidades de cada evento:
  - $P(T = 0|D = 0)$: *True Negative*
  - $P(T = 1|D = 0) = p_{FP}$: *False Positive*
  - $P(T = 0|D = 1) = p_{FN}$: *False Negative*
  - $P(T = 1|D = 1)$: *True Positive*
  - normalizações:
    $P(T = 0|D = 0) + P(T = 1|D = 0) = 1$

    $P(T = 0|D = 1) + P(T = 1|D = 1) = 1$
- espera-se que p_{FP} e p_{FN} sejam pequenos

matriz de confusão

D \ T	0	1
0	TN $1\text{-}p_{FP}$	FP p_{FP}
1	FN p_{FN}	TP $1\text{-}p_{FN}$

tabela de contingência ou matriz de confusão

suponha que um teste deu positivo $(T = 1)$
qual é a probabilidade de que o paciente tenha contraido a doença?

- queremos $P(D = 1|T = 1)$
- vamos supor que a probabilidade a priori dessa doença é $P(D = 1) = p_D$
  a probabilidade de não ter a doença é então $P(D = 0) = 1 - p_D$
- teorema de Bayes:

$$P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1)}$$

$$P(T = 1|D = 1)P(D = 1) = (1 - p_{FN})p_D$$

$$P(T = 1) = P(T = 1|D = 0)P(D = 0) + P(T = 1|D = 1)P(D = 1) =$$

$$= p_{FP}(1 - p_D) + (1 - p_{FN})p_D$$

$$P(D = 1|T = 1) = \frac{p_D - p_{FN}p_D}{p_D + p_{FP} - p_D(p_{FP} + p_{FN})} \simeq \frac{p_D}{p_D + p_{FP}}$$

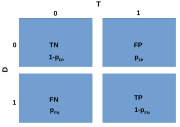
logo,

$$P(D = 1|T = 1) \simeq \frac{p_D}{p_D + p_{FP}}$$

só podemos diagnosticar uma doença de forma confiável se $p_{FP} << p_D$

- se $p_{FP} >> p_D$, $P(D = 1|T = 1) << 1$ e o teste não produz evidência conclusiva
- p_{FN} não é tão importante desde que não seja muito maior que os outros parâmetros
- recomendação do Zeljko+ (MLA):
  se for fazer um teste fique de olho em p_{FP}!
- doenças raras: p_D é muito pequeno, e para um teste positivo ser confiável tem que ter p_{FP} muito pequeno!

matriz de confusão



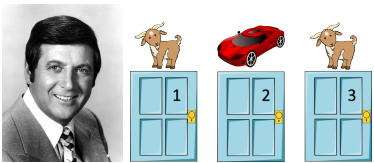
o problema de Monty Hall

um programa de auditório:

- você tem 3 portas na sua frente: uma contém uma Ferrari e as duas outras um bode em cada uma
- MH pede para você escolher uma delas: você escolhe, por exemplo, a 1
- antes de abrir a porta que você escolheu, MH (que sabe onde o carro está), escolhe uma das portas com um bode e a abre; sobra assim uma outra porta fechada

- MH então te pergunta: quer trocar de porta ou não quer? Você quer trocar a porta que escolheu por esta que sobrou ou não?

- o que é mais vantajoso fazer?



o problema de Monty Hall

- digamos que voce escolheu a porta 1 e a Ferrari está na 2
- MH então abre a 3, mostrando um bode
- você, sem saber onde está a Ferrari, continua com a porta 1 ou muda para a 2?
- vamos ver com qual ação você tem maior probabilidade de ganhar a Ferrari!
- $P(ci)$: probabilidade a priori de que o carro esteja atrás da porta i:

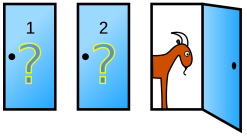
  $P(c1) = P(c2) = P(c3) = 1/3$
- $P(i)$: probabilidade de que MH abra a porta i
- como MH abriu a porta 3, precisamos comparar $P(c1|3)$ com $P(c2|3)$
- probabilidade de o carro estar na porta 1 dado que MH abriu a porta 3:

  $P(c1|3) = P(3|c1)P(c1)/P(3)$
- probabilidade de o carro estar na porta 2 dado que MH abriu a porta 3

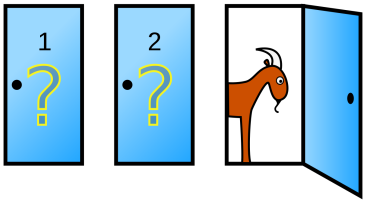
  $P(c2|3) = P(3|c2)P(c2)/P(3)$

  onde

  $P(3) = P(3|c1)P(c1) + P(3|c2)P(c2) + P(3|c3)P(c3)$
- qual é maior? $P(c1|3)$ ou $P(c2|3)$?



o problema de Monty Hall



- qual é maior? $P(c1|3)$ ou $P(c2|3)$?

- probabilidades relevantes:

  $P(c1) = P(c2) = P(c3) = 1/3$

  $P(c1|3) = P(3|c1)P(c1)/P(3)$

  $P(c2|3) = P(3|c2)P(c2)/P(3)$

  $P(3) = P(3|c1)P(c1) + P(3|c2)P(c2) + P(3|c3)P(c3)$

- mas
  $P(3|c1) = 1/2$: ele tinha duas portas para abrir
  $P(3|c2) = 1$: como o carro estava em 2, ele só podia abrir a 3
  $P(3|c3) = 0$: 3 tem um bode, por isso MH a abre

- logo,
  $P(c1|3) = (1/2 \times 1/3)/(1/2 \times 1/3 + 1 \times 1/3 + 0 \times 1/3) = 1/3$

- se trocar de porta:
  $P(c2|3) = P(3|c2)P(c2)/P(3) = 2/3$

  é vantajoso trocar!!

o problema de Monty Hall

- quando você escolheu uma das portas, a chance de ter um bode atrás dela era 2/3
- se você muda de porta, a chance de ter uma Ferrari atrás dela é 2/3
- logo, convém mudar de porta!