AGA 0505 - Análise de Dados em Astronomia

3. Probabilidades

Laerte Sodré Jr.

1o. semestre, 2024

aula de hoje:

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

A teoria das probabilidades é o senso comum reduzido ao cálculo

Pierre-Simon Laplace

dados e probabilidades

- descrevemos os dados (medidas, observações) como *distribuições de probabilidades*
- dados podem ser complexos, variáveis e, sobretudo, têm incertezas
  ex. de incerteza: os erros nas medidas
- a teoria das probabilidades provê um meio de modelar os dados e quantificar as incertezas

- duas visões sobre a natureza das probabilidades:
  - *frequentista (clássica)*: probabilidade como medida de frequência
  - *bayesiana*: probabilidade como medida de plausibilidade
- podemos considerar probabilidade como medida de frequência como um caso particular de medida de plausibilidade

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)

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

- 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

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

- 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

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

- exemplo: dado de 6 faces não-viezado

$$P(i) = \frac{1}{6} \qquad (\text{para i} = 1, 2..., 6)$$

$$\sum_{i=1}^{6} P(i) = 1$$

as regras fundamentais das probabilidades: algumas consequências

- marginalização: quando se tem um conjunto y_k de proposições "mutuamente exclusivas e exaustivas" (MEE)

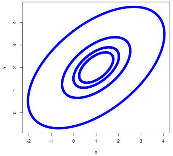
$$P(x) = \sum_k P(x, y_k)$$

ou

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

muito útil!

- exemplo: gaussiana bivariada



$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 \right.\right.$$

$$\left.\left. -2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

$$P(x) = \int P(x,y)dy \sim N(\mu_x, \sigma_x)$$

as regras fundamentais das probabilidades: algumas consequências

- (1) 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

- (2) 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,

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

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
- $P(x)$: distribuição de probabilidades de x
- $P(y)$: distribuição de probabilidades 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

- Um astrônomo pede a um colega para revisar um artigo que ele escreveu. A probabilidade do colega não aceitar é 1/3. A probabilidade do artigo ser aceito se o colega revisar é 1/2, e 1/4 no caso contrário.

- Qual é a probabilidade do artigo ser aceito, $P(A)$?
  - seja A:aceito, -A: rejeitado, R: revisou -R: não revisou
  - temos que
    $P(-R) = 1/3 \longrightarrow P(R) = 2/3$
    $P(A|R) = 1/2$
    $P(A|-R) = 1/4$
  - logo,
    $P(A) = P(A|R).P(R) + P(A|-R)P(-R) = 1/2 \times 2/3 + 1/4 \times 1/3 = 5/12$

exercícios sobre probabilidades

- Um astrônomo pede a um colega para revisar um artigo que ele escreveu. A probabilidade do colega não aceitar é 1/3. A probabilidade do artigo ser aceito se o colega revisar é 1/2, e 1/4 no caso contrário.

- Suponha que o artigo foi rejeitado. Qual é a probabilidade do colega não ter revisado o artigo, $P(-R|-A)$?
  - probabilidade conjunta de o colega não ter revisado e o artigo não ter sido aceito:
    $P(-R,-A) = P(-R|-A).P(-A) = P(-A|-R).P(-R)$
    $P(-R|-A) = P(-A|-R).P(-R)/P(-A)$
    $P(-A) = P(-A|R).P(R) + P(-A|-R).P(-R)$
  - note que
    $P(-A|R) = 1 - P(A|R)$ e
    $P(-A|-R) = 1 - P(A|-R)$
  - e, portanto, $P(A) = 3/4 \times 1/3/(3/4 \times 1/3 + 1/2 \times 2/3) = 3/7 \simeq 0.43$

exercícios sobre probabilidades

- Num levantamento de 200 galáxias, deseja-se verificar se a presença de formação estelar depende do ambiente em que a galáxia está: se em um aglomerado ou no "campo". O evento A indica se a galáxia está num aglomerado e o B se ela tem formação estelar. Metade das galáxias da amostra está em cada ambiente e a fração com formação estelar nos aglomerados é 40%, contra 80% no campo.
- Qual é a probabilidade de que uma galáxia tenha formação estelar e esteja num aglomerado, $P(B, A)$?
  - seja: A: galáxia em aglomerado, -A: galáxia no campo, B: galáxia com formação estelar, -B: galáxia quiescente (sem formação estelar)
  - temos que
    $P(A) = P(-A) = 1/2$
    $P(B|A) = 0.4$
    $P(B|-A) = 0.8$
  - logo,
    $P(B, A) = P(B|A)P(A) = 0.4 \times 0.5 = 0.2$
- Qual é a probabilidade de se ter uma galáxia com formação estelar nessa amostra?
  $P(B) = P(B|A)P(A) + P(B|-A)P(-A) = 0.4 \times 0.5 + 0.8 \times 0.5 = 0.6$

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
  $S = \{(H,T),(H,H),(T,H),(T,T)\}$
  e contém $n(s) = 4$ elementos
- E: evento "sair duas caras": (H,H)
- S mostra que E contém $n(E) = 1$ elemento
- 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?

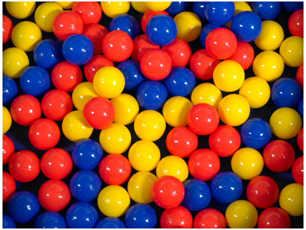


Table 1: Espaço amostral

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

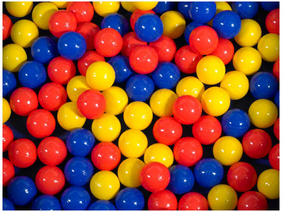
3 x (3/9 x 4/9 x 4/9) = 16/81 $\simeq$ 0,1975

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?

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

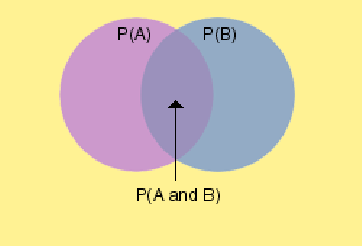


diagrama de Venn

- 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,
  $P(V \text{ ou } 3) = 26/52 + 4/52 - 2/52 =$
  $= 28/52 = 7/13 \simeq 0.538\ldots$

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

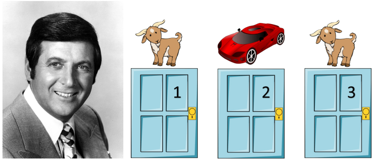
o problema de Monty Hall

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?

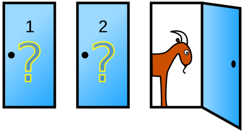


o problema de Monty Hall

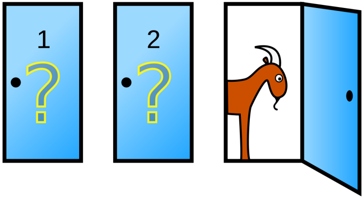
- digamos que voce escolheu a porta 1 e a Ferrari está na 2
- MH então abre a 3, mostrando um bode
- você, sem saber onde está a Ferrari, continua com a porta 1 ou muda para a 2?
- vamos ver com qual ação você tem maior probabilidade de ganhar a Ferrari!
- $P(ci)$: probabilidade a priori de que o carro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