Probabilidade e Estatística

L. N. Queiroz Xavier

5 de janeiro de 2024

1 Probabilidade

1.1 Espaço amostral e eventos

Estas notas são baseadas na referência [1]. Um *espaço amostral* Ω contém todos os resultados possíveis de um experimento. Por exemplo, se jogarmos uma moeda duas vezes, ao fim desse experimento podemos observar os seguintes resultados:

$$HH, HT, TH, TT,$$

onde H é cara e T é coroa. Logo, $\Omega = \{HH, HT, TH, TT\}$. Um resultado $\omega \in \Omega$ é chamado de um *resultado amostral*, uma *realização* ou um *elemento*. Um *evento* $A \subset \Omega$ é um sub-espaço de Ω. Por exemplo, jogando uma moeda duas vezes, o evento *obter cara na última jogada* é dado pelo sub-espaço $A = \{HH, TH\}$.

Dado o evento A, o evento $A^c = \{\omega \in \Omega | \omega \notin A\}$ pode ser visto como o evento *A não acontece*. Por exemplo, jogando uma moeda duas vezes, o evento *não obter cara na última jogada* é igual a $A^c = \{HT, TT\}$. Dados dois eventos A e B, o evento $A \cup B$ pode ser visto como o evento *A ou B*, enquanto que o evento $A \cap B$ pode ser visto como o evento *A e B*. Por exemplo, considere os eventos $A =$ *obter cara duas vezes* e $B =$ *obter coroa duas vezes*. O evento $A \cup B$ é *obter cara duas vezes ou obter coroa duas vezes*, e é dado por $\{HH, TT\}$. O evento $A \cap B$ é *obter cara duas vezes e obter coroa duas vezes*, que obviamente é igual a $\emptyset$.

1.2 Probabilidade

Seja Ω um espaço amostral. A cada evento $A \subset \Omega$ associamos um número real $P(A)$, a *probabilidade* de A. O mapa $P : \Omega \to [0, 1]$ é chamado de

distribuição de probabilidade ou *medida de probabilidade*, e deve ser tal que:

$$P(A) \geq 0, \tag{1.1}$$

$$P(\Omega) = 1, \tag{1.2}$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \tag{1.3}$$

$\forall A \in \Omega$ e para todo conjunto de eventos $\{A_i\}_{i=1}^{\infty}$ onde $A_i \cap A_j = \emptyset$.

Podemos interpretar o significado do número real $P(A)$ de basicamente duas maneiras. A primeira forma de entender o significado de $P(A)$ é como uma *frequência de ocorrências*. Se repetirmos o experimento cujo espaço amostral é Ω várias vezes, o número de vezes que o evento A ocorre em proporção ao número de repetições do experimento tende a $P(A)$. Por exemplo, quando dizemos que a probabilidade de tirar cara jogando uma moeda é $1/2$, o que queremos dizer é que, se jogarmos a moeda várias vezes, metade das vezes obteremos cara. No entanto, esta interpretação falha quando estamos lidando com probabilidades de eventos que não podem ser facilmente repetidos. Por exemplo, quando um médico diz que a probabilidade de um paciente estar doente é de 40%, é estranho atribuir uma noção de frequência a este número. Isto nos leva à segunda forma de interpretar $P(A)$, como um *grau de crença*. O número $P(A)$ significa então o quanto acreditamos que o evento A é verdadeiro, com $P(A) = 1$ no caso em que temos certeza que A acontece e $P(A) = 0$ quando temos certeza que A não acontece. Assim, no exemplo do médico, a probabilidade de 40% significa que o médico não acredita muito que o paciente esteja doente.

Dos axiomas da probabilidade, podemos deduzir várias propriedades importantes. Primeiramente, por definição $P(\emptyset) = 0$ ("sempre acontece algo"). Considere A e B dois eventos disjuntos, i.e., $A \cap B = \emptyset$. Considere o conjunto $\{A_i\}_{i=1}^{\infty}$ onde $A_1 = A$, $A_2 = B$ e $A_i = \emptyset$, para todo $i = 3, ...$ Do terceiro axioma, segue que, como $\bigcup_{i=1}^{\infty} A_i = A \cup B$,

$$P(A \cup B) = P(A) + P(B) \tag{1.4}$$

Ou seja, a probabilidade de A ou B é a soma das probabilidades de cada evento, se A e B forem disjuntos. Por exemplo, a probabilidade de tirar duas cara consecutivas ou duas coroas consecutivas jogando uma moeda duas vezes é $P(HH) + P(TT) = 1/2$.

Outra propriedade importante da probabilidade é a seguinte. Considere $A \in \Omega$. Obviamente, $A \cup A^c = \Omega$. Segue então que

$$1 = P(A) + P(A^c).$$

Logo, dada a probabilidade de A, a probabilidade de A não acontecer é

$$P(A^c) = 1 - P(A). \tag{1.5}$$

Uma última propriedade da probabilidade, desta vez menos óbvia. Considere A e B dois eventos quaisquer. Note que

$$A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B),$$

e que os eventos entre parênteses são disjuntos entre si, o que é fácil ver utilizando diagramas de Venn. Temos que

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$$

$$= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) + P(A \cap B) - P(A \cap B)$$

$$= P\left((A \cap B^c) \cup (A \cap B)\right) + P\left((A^c \cap B) \cup (A \cap B)\right) - P(A \cap B).$$

Note que $(A \cap B^c) \cup (A \cap B) = A$ e $(A^c \cap B) \cup (A \cap B) = B$, logo

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1.6}$$

Por exemplo, a probabilidade de obter cara na primeira jogada ou coroa na segunda jogada de uma moeda é $P(\{HH, HT\}) + P(\{HT, TT\}) - P(HT) = 3/4$.

1.3 Eventos independentes

Considere o evento $A =$ *tirar cara* jogando uma moeda. Qual é a probabilidade de tirar duas caras jogando uma moeda duas vezes? Note que o evento A é *independente* de si mesmo, isto é, tirar cara uma vez não afeta a chance de tirar cara outra vez. Interpretando probabilidades como frequências, temos que, se jogarmos uma moeda, o evento A acontece $P(A)$ vezes. O evento $A \cap A$ ocorre $P(A \cap A)$ vezes. Intuitivamente, $P(A \cap A) = P(A)P(A)$, pois, jogando a moeda uma vez, tiramos cara $P(A)$ vezes e, dessas $P(A)$ vezes, jogando a moeda novamente, tiramos cara $P(A)$ vezes.

Esse exemplo serve como intuição para a definição de eventos independentes. Dois eventos A e B são *independentes* se

$$P(A \cap B) = P(A)P(B). \tag{1.7}$$

Note que eventos disjuntos A e B com probabilidades não-nulas $P(A)$, $P(B)$ não podem ser independentes. De fato, $P(A \cap B) = P(\emptyset) = 0$, mas $P(A)P(B) >$

0. Intuitivamente, eventos como eventos disjuntos são mutuamente exclusivos (ou A acontece ou B acontece), se um acontece, sabemos imediatamente que o outro não ocorre, e temos assim uma dependência entre os eventos.

Exemplo. Se jogarmos uma moeda 10 vezes, qual a probabilidade de obtermos pelo menos uma cara? Seja $A =$ *pelo menos uma cara.* Obviamente, $A^c =$ *só coroas*, e segue que

$$P(A^c) = P(T_1 \cap T_2 \cap ... \cap T_{10}),$$

onde $T_i =$ *coroa na jogada i.* Como cada evento T_i é independente,

$$P(A^c) = \prod_{i=1}^{10} P(T_i) = \frac{1}{2^{10}},$$

e segue que

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{2^{10}} \approx 0.999,$$

i.e., a probabilidade de obter pelo menos uma cara é muito alta.

Exemplo. Considere duas pessoas tentando acertar uma cesta de basquete. A pessoa 1 acerta uma cesta com probabilidade 1/3, enquanto a pessoa 2 acerta uma cesta com probabilidade 1/4. Qual a probabilidade da pessoa 1 acertar a cesta antes da pessoa 2? Seja $A =$ *pessoa 1 acerta a cesta antes da pessoa 2.* Seja $A_i =$ *a primeira cesta é feita pela pessoa 1 na tentativa i.* Note que os eventos A_i são disjuntos, pois por exemplo se A_1 acontece, A_2 não pode ser verdade pois a pessoa 1 já acertou a cesta na tentativa 1. Note ainda que $A = \cup_{i=1}^{\infty} A_i =$ *pessoa 1 acerta primeiro na tentativa 1 ou pessoa 1 acerta primeiro na tentativa 2 ou ...* Logo,

$$P(A) = \sum_{i=1}^{\infty} P(A_i).$$

Note que $P(A_1) = 1/3$. O evento A_2 acontece se a pessoa 1 e a pessoa 2 errarem na primeira tentativa e a pessoa 1 acertar na segunda tentativa. Cada um desses eventos são independentes, o que significa que a probabilidade de A_2 acontecer é

$$\begin{aligned} P(A_2) &= P(\text{pessoa 1 erra})P(\text{pessoa 2 erra})P(\text{pessoa 1 acerta}) \\ &= (1 - 1/3)(1 - 1/4)(1/3) \\ &= (2/3)(3/4)(1/3) = (1/2)(1/3) \end{aligned}$$

e cada evento A_i deve seguir essa lógica. Segue que $P(A_i) = (1/2)^{i-1}(1/3)$ e

$$P(A) = \frac{1}{3} \sum_{i=1}^{\infty} \frac{1}{2^{i-1}} = \frac{2}{3}.$$

1.4 Probabilidade condicional

Considere dois eventos $A, B \subset \Omega$. Qual é a probabilidade de A, dado que sabemos que B acontece? Intuitivamente, se os eventos forem independentes, devemos ter que a probabilidade de A não deve ser afetada pelo acontecimento de B. Logo, se $P(A|B)$ é a probabilidade de A dado que B acontece, se A e B forem independentes, $P(A|B) = P(A)$.

E se os eventos não forem independentes? Para construirmos nossa intuição, vamos considerar a interpretação onde probabilidades representam frequências, e que todos os eventos em Ω são igualmente prováveis. Assim,

$$P(A|B) = \frac{\#\text{ de vezes que A acontece dado que B acontece}}{\#\text{ de eventos possíveis dado que B acontece}}.$$

Se sabemos que B acontece, e queremos a probabilidade $P(A|B)$, temos que contar quantas vezes A e B acontecem, i.e.,

$$\#\text{ de vezes que A acontece dado que B acontece} = P(A \cap B)N,$$

onde N é o número de eventos possíveis $N = |\Omega|$. Mais ainda, o número de eventos possíveis em que B acontece é igual a $P(B)N$. Logo,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

O caso onde os eventos possuem probabilidade uniforme nos motiva a definir a *probabilidade condicional* de um evento dado outro. Se $P(B) > 0$, a probabilidade de A dado que B ocorre é

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.8}$$

Em geral, $P(A|B) \neq P(B|A)$. De fato, isto acontece apenas quando $P(A) = P(B)$. Por exemplo, a probabilidade de eu ter espirros dado que estou gripado é próxima de 1, mas a probabilidade de eu estar gripado dado que tenho espirros não é próxima de 1, pois os espirros podem ser causados por outras coisas.

Exemplo. Um teste médico para uma doença D tem resultados positivos e negativos com as seguintes probabilidades:

	D	D^c
+	0.009	0.099
-	0.001	0.891

A probabilidade de testar positivo dado que se tem a doença é

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9.$$

Já a probabilidade de testar negativo dado que não se tem a doença é

$$P(-|D^c) = \frac{P(- \cap D^c)}{P(D^c)} = \frac{0.891}{0.099 + 0.891} = 0.9.$$

Logo, o teste parece razoável. Para confirmar, vamos calcular a probabilidade de se ter a doença dado que o teste dá positivo. Temos que

$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{0.009}{0.009 + 0.099} = 0.083,$$

o que é incrivelmente baixo! Apenas 8 em cada 100 pessoas que testam positivo estão realmente doentes. A lição é: cuidado com probabilidades condicionais!

1.5 Teorema de Bayes

Teorema 1.5.1. *(A Lei da Probabilidade Total). Seja Ω um espaço amostral e seja $A_1, ..., A_k$ uma partição de Ω, i.e., $A_i \cap A_j = \emptyset$ $\forall i, j$ e $\cup_{i=1}^k A_i = \Omega$. Para qualquer evento $B \subset \Omega$,*

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i). \tag{1.9}$$

Demonstração. Sejam $C_i = B \cap A_i$, $\forall i = 1, ..., k$. Note que

$$C_i \cap C_j = B \cap (A_i \cap A_j) \cap B = \emptyset,$$

$\forall i, j = 1, ..., k$. Mais ainda,

$$\cup_{i=1}^k C_i = B \cap \left(\cup_{i=1}^k A_i\right) = B \cap \Omega = B.$$

Assim,

$$P(B) = P(\cup_{i=1}^k C_i) = \sum_{i=1}^{k} P(C_i) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(B|A_i)P(A_i),$$

onde utilizamos a (1.8). □

Teorema 1.5.2. *(Teorema de Bayes). Seja $A_1, ..., A_k$ uma partição de Ω tal que $P(A_i) \neq 0$ $\forall i$. Se $P(B) \neq 0$, então para cada $i = 1, ..., k$,*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}. \tag{1.10}$$

Demonstração. Utilizando a (1.8) e a (1.9), temos que

$$P(A_i|B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}.$$

□

Chamamos $P(A_i)$ de *prior* e $P(A_i|B)$ de *posterior*, i.e., $P(A_i)$ é a probabilidade de A_i antes de sabermos que B acontece, enquanto que $P(A_i|B)$ é a probabilidade de A_i depois que sabemos que B acontece.

Exemplo. Em uma cidade, existem apenas bibliotecários e fazendeiros, em uma proporção de 1 : 20. Dos bibliotecários, 40% são pessoas calmas, silenciosas e organizadas, enquanto 20% dos fazendeiros possuem essas características. Roberto é uma pessoa calma e silenciosa, que gosta de organização e não é muito fã de agitação. É mais provável Roberto ser um bibliotecário ou um fazendeiro?

Baseado na descrição, muitas pessoas podem se inclinar a dizer que Roberto provavelmente é um bibliotecário. No entanto, devemos levar em conta o número de fazendeiros que existem na região em relação ao número de bibliotecários. Seja A o evento *Roberto é um bibliotecário* e B o evento *Roberto é calmo, silencioso e organizado*. Queremos encontrar $P(A|B)$. Das informações do problema, a probabilidade de uma pessoa ser calma, silenciosa e organizada dado que ela é bibliotecária é

$$P(B|A) = \frac{\#\text{ de pessoas calmas, silenciosas, organizadas e bibliotecárias}}{\#\text{ total de pessoas}}$$

$$= \frac{0.4}{\#\text{ total de pessoas}}.$$

A probabilidade de uma pessoa ser bibliotecária é

$$P(A) = \frac{\#\text{ de bibliotecários}}{\#\text{ total de pessoas}} = \frac{1}{21}.$$

A probabilidade de uma pessoa ser calma, silenciosa e organizada dado que ela é fazendeira é

$$P(B|A^c) = \frac{0.2}{\#\text{ total de pessoas}},$$

e a probabilidade de uma pessoa ser fazendeira é

$$P(A^c) = \frac{20}{21}.$$

Do teorema de Bayes, sendo $N = \#$ total de pessoas, segue que

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{(0.4/N)(1/21)}{(0.4/N)(1/21) + (0.2/N)(20/21)} \\ &= \frac{0.4/21}{0.4/21 + 0.2(20/21)} \\ &\approx \frac{0.01905}{0.20952} \\ &\approx 0.09092. \end{aligned}$$

Ou seja, a probabilidade de que Roberto seja um bibliotecário é de apenas 9%, contra 91% de chance de que Roberto seja um fazendeiro.

2 Variáveis aleatórias

Dado um espaço amostral Ω, uma *variável aleatória* X é um mapa

$$X : \Omega \to \mathbb{R}.$$

Isto é, uma variável aleatória associa a cada resultado $\omega \in \Omega$ um número real $X(\omega) \in \mathbb{R}$. Por exemplo, considere que jogamos uma moeda duas vezes. Temos que $\Omega = \{HH, TT, TH, HT\}$. Seja $X(\omega)$ o número de caras em ω. Logo, $X(TT) = 0$, $X(HH) = 2$ e $X(TH) = X(HT) = 1$.

Seja X uma variável aleatória e $I \subset \mathbb{R}$. Temos que a imagem inversa $X^{-1}(I) = \{\omega \in \Omega | X(\omega) \in I\}$. Definimos então a probabilidade $P(X \in I)$ como

$$P(X \in I) = P(X^{-1}(I)),$$

e a probabilidade $P(X = x)$ como

$$P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega | X(\omega) = x\}).$$

Por exemplo, no caso em que jogamos uma moeda duas vezes, sendo X o número de caras, temos que $P(X = 0) = P(TT) = 1/4$, $P(X = 1) = P(\{TH, HT\}) = 1/2$ e $P(X = 2) = P(HH) = 1/4$.

2.1 Funções de distribuição e funções de probabilidade

Seja X uma variável aleatória. A *função de distribuição acumulada* (fda), ou simplesmente função de distribuição, é o mapa $F_X : \mathbb{R} \to [0, 1]$ tal que

$$F_X(x) = P(X \leq x). \tag{2.11}$$

Isto é, $F_X(x)$ é a probabilidade de $X \leq x$.

Exemplo. Considere que jogamos uma moeda duas vezes e seja X o número de caras. Temos que $X \in \{0, 1, 2\}$. Segue que $P(X = 0) = P(X = 2) = 1/4$ e $P(X = 1) = 1/2$. Vamos calcular a função de distribuição acumulada dessa variável. Temos que, para $F_X(x) = 0$ para $x < 0$, pois $P(X < 0) = 0$. Para $x \geq 2$, temos que $F_X(x) = P(X \leq 2) = 1$, pois X assume apenas valores menores ou iguais a 2. Para $0 \leq x < 1$, temos que $F_X(x) = 1/4$. Finalmente, para $1 \leq x \leq 2$, temos que $F_X(x) = 3/4$. Segue que a fda é dada por

$$F_X(x) = \begin{cases} 0, & \text{se} \quad x < 0, \\ 1/4, & \text{se} \quad 0 \leq x < 1, \\ 3/4, & \text{se} \quad 1 \leq x < 2, \\ 1, & \text{se} \quad x \geq 2. \end{cases} \tag{2.12}$$

Uma variável aleatória X é dita *discreta* se assume valores em um conjunto contável $\{x_1, x_2, ...\}$. A *função de probabilidade* ou *função massa de probabilidade* f_X associada a X é dada por $f_X(x) = P(X = x)$, i.e., para cada x, a função de probabilidade $f_X(x)$ é a probabilidade de $X = x$. Segue que $f_X(x) \geq 0$ para todo x e

$$\sum_i f_X(x_i) = 1.$$

A fda associada a X é dada por

$$F_X(x) = P(X \leq x) = \sum_{y \leq x} f_X(y).$$

Exemplo. A função de probabilidade associada a jogar uma moeda duas vezes é

$$f_X(x) = \begin{cases} 1/4, & \text{se} \quad x = 0, \\ 1/2, & \text{se} \quad x = 1, \\ 1/4, & \text{se} \quad x = 2, \\ 0, & \text{caso contrário.} \end{cases} \tag{2.13}$$

Uma variável aleatória X é dita *contínua* se existe uma função f_X tal que $f_X(x) \geq 0$ para todo x,

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

e, para quaisquer $a < b$,

$$P(a < X < b) = \int_a^b f_X(x)dx.$$

A função f_X é chamada *função de densidade de probabilidade* (pdf). A função de distribuição acumulada relacionada a X é tal que

$$F_X(x) = \int_{-\infty}^{x} f_X(x')dx'.$$

Exemplo. Considere a fdp

$$f_X(x) = \begin{cases} 1, & \text{se} \quad 0 \leq x \leq 1, \\ 0, & \text{caso contrário.} \end{cases} \tag{2.14}$$

Uma variável aleatória X com tal distribuição é dita *uniforme*. A fda associada é fácil de ser obtida:

$$F_X(x) = \begin{cases} 0, & \text{se} \quad x < 0, \\ x, & \text{se} \quad 0 \leq x \leq 1, \\ 1, & \text{se} \quad x > 1. \end{cases} \tag{2.15}$$

Dada uma variável aleatória X de fda F, sua fda inversa ou *função de quantil* é definida como

$$F^{-1}(q) = \inf\{x | F(x) > q\}. \tag{2.16}$$

A interpretação é a seguinte: $F^{-1}(q)$ nos dá o menor valor de x tal que a probabilidade de $X \leq x$ é maior que q. $F^{-1}(1/4)$ é o *primeiro quartil*, $F^{-1}(1/2)$ é a *mediana* ou *segundo quartil* e $F^{-1}(3/4)$ é o *terceiro quartil*.

2.2 Algumas distribuições discretas importantes

Distribuição pontual. X possui distribuição pontual no ponto a, isto é $X \sim \delta_a$, se $P(X = a) = 1$. Neste caso, $f(x) = \delta(x - a)$ e

$$F(x) = \begin{cases} 0, & \text{se} \quad x < a, \\ 1, & \text{se} \quad x \geq a. \end{cases} \tag{2.17}$$

Distribuição uniforme discreta. Seja $k > 1$ um inteiro. X possui distribuição uniforme no conjunto $\{1, ..., k\}$ se sua função de probabilidade for

$$f(x) = \begin{cases} \frac{1}{k} & \text{se} \quad x = 1, ..., k, \\ 0 & \text{caso contrário.} \end{cases} \tag{2.18}$$

Distribuição de Bernoulli. Considere que X representa um evento com resultado binário, como jogar uma moeda. Então, $P(X = 0) = p$ e $P(X = 1) = 1 - p$, para alguma probabilidade $p \in [0, 1]$. Dizemos que X possui distribuição de Bernoulli, isto é $X \sim Bernoulli(p)$. A função de probabilidade é

$$f(x) = p^x(1 - p)^{1-x}, \tag{2.19}$$

para $x \in \{0, 1\}$.

Distribuição binomial. Considere que jogamos uma moeda n vezes. A probabilidade do resultado ser cara é $p \in [0, 1]$. Seja X o número de caras. Podemos derivar a distribuição de X a partir da distribuição de Bernoulli. Suponha que obtivemos x caras em n tentativas. A probabilidade desse evento é $p^x(1 - p)^{n-x}$, de acordo com a distribuição de Bernoulli. Note que podemos obter x em n de $\binom{n}{x}$ maneiras. Logo, a função de probabilidade de X é

$$f(x) = \begin{cases} \binom{n}{x} p^x(1 - p)^{n-x} & \text{para} \quad x = 0, ..., n, \\ 0 & \text{caso contrário.} \end{cases} \tag{2.20}$$

Distribuição geométrica. Considere que X é o número de jogadas de moeda necessárias para se obter uma cara. Segue que X é distribuída geometricamente, com função de probabilidade

$$f(x) = p(1 - p)^{x-1}, \ x \geq 1. \tag{2.21}$$

Distribuição de Poisson. Suponha que temos um evento binário que acontece raramente em várias tentativas, como o decaimento radioativo (o elemento decai no tempo t ou não). Tal evento é modelado pela distribuição de Poisson, que pode ser obtida através da distribuição binomial.

Considere que não conhecemos nem a probabilidade p nem o número de tentativas, além de sabermos que p é pequena e n é grande. No entanto, a frequência $\lambda = np$ de ocorrências é conhecida. Substituindo $p = \lambda/n$ na distribuição binomial, temos

$$f(x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}. \tag{2.22}$$

Tomando o limite $n \to \infty$, temos que

$$\begin{aligned}\lim_{n\to\infty} f(x) &= \lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n!}{(n-x)!} \frac{1}{n^x} \left(1-\frac{\lambda}{n}\right)^{n-x}.\end{aligned}$$

O primeiro termo é

$$\begin{aligned}\lim_{n\to\infty} \frac{n!}{n^x(n-x)!} &= \lim_{n\to\infty} \frac{n(n-1)...(n-x)(n-x-1)...1}{n^x(n-x)(n-x-1)...1} \\ &= \lim_{n\to\infty} \frac{n(n-1)...(n-x+1)}{n^x} \\ &= \lim_{n\to\infty} \frac{n}{n}\frac{n-1}{n}...\frac{n-x+1}{n} \\ &= 1.\end{aligned}$$

O segundo termo é

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n-x}.$$

Definindo $y = -n/\lambda$, temos que

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n-x} = \lim_{y\to-\infty} \left(1+\frac{1}{y}\right)^{-\lambda y} = e^{-\lambda}.$$

Ou seja, a distribuição de Poisson é dada por

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}. \tag{2.23}$$

2.3 Algumas distribuições contínuas importantes

Distribuição uniforme. X é uniformemente distribuída no intervalo $[a, b]$ se segue a seguinte distribuição:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{se} \quad x \in [a,b], \\ 0, & \text{caso contrário.} \end{cases} \tag{2.24}$$

Distribuição normal. A distribuição normal de média μ e desvio padrão σ é dada pela fdp

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma}(x-\mu)^2}, \tag{2.25}$$

onde $\sigma > 0$. Quando X é distribuída de acordo com a distribuição normal, dizemos que $X \sim N(\mu, \sigma^2)$.

Distribuição t. X possui distribuição t com ν graus de liberdade, i.e., $X \sim t_\nu$, se

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}. \tag{2.26}$$

A distribuição t é similar à distribuição normal, mas possui os extremos mais largos. De fato, quando $\nu \to \infty$, recuperamos a distribuição normal. Com $\nu = 1$, obtemos a *distribuição de Cauchy*

$$f(x) = \frac{1}{\pi(1 + x^2)}. \tag{2.27}$$

Distribuição exponencial. X possui distribuição exponencial com parâmetro $\beta > 0$ se

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \tag{2.28}$$

onde $x > 0$.

Distribuição χ^2. X possui distribuição χ^2 com p graus de liberdade, i.e., $X \sim \chi^2_p$, se

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} e^{-x/2}, \tag{2.29}$$

para $x > 0$. Para $p = 1$, a distribuição χ^2 é a distribuição da variável $Z = X^2$, onde $X \sim N(0, 1)$. Para qualquer p, χ^2_p é a distribuição da variável $Z = \sum_{n=1}^{p} X_n^2$, onde cada X_n é normalmente distribuída.

2.4 Distribuições bivariadas, distribuições marginais, independência e distribuições condicionais

Dado um par de variáveis aleatórias discretas X e Y, definimos a função de massa conjunta $f(x, y) = P(X = x \text{ e } Y = y) = P(X = x, Y = y)$. O caso contínuo é semelhante: dadas duas variáveis aleatórias contínuas X e Y, a função densidade de probabilidade do par (X, Y), $f(x, y)$, é tal que é positiva para todo (x, y), é normalizada e

$$P((X, Y) \in A) = \int_A f(x, y) dx dy.$$

Dada uma distribuição conjunta $f(x, y)$ das variáveis discretas (X, Y), a distribuição marginal de X é dada por

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f(x, y).$$

Já a distribuição marginal de Y é dada por

$$f_Y(y) = P(Y = y) = \sum_x P(X = x, Y = y) = \sum_x f(x, y).$$

No caso contínuo, fazemos um processo semelhante: as densidades marginais são obtidas por integração, i.e.,

$$f_X(x) = \int f(x, y)dy, \ f_Y(y) = \int f(x, y)dx.$$

Independência. Duas variáveis aleatórias X e Y são independentes se

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B). \tag{2.30}$$

Para variáveis contínuas, temos que se X e Y têm distribuição conjunta $f(x, y)$, elas são independentes se, e somente se,

$$f(x, y) = f(x)f(y), \tag{2.31}$$

para todo x, y.

Distribuição condicional. Sejam X e Y duas variáveis discretas. A função massa de probabilidade condicional de X dado que observamos $Y = y$ é dada por

$$f(x|y) = \frac{P(X = x, Y = y)}{P(y = y)} = \frac{f(x, y)}{f_Y(y)}. \tag{2.32}$$

No caso contínuo, a fdp é

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}, \tag{2.33}$$

e a probabilidade de observar $X \in A$ dado que $Y = y$ é

$$P(X \in A|Y = y) = \int_A f(x|y)dx. \tag{2.34}$$

Distribuições multivariadas. Se temos muitas variáveis aleatórias $X_1, ..., X_n$, podemos agrupá-las em um vetor aleatório $X = (X_1, ..., X_n)$. Dada a fdp

$f(x_1, ..., x_n)$, podemos extender a mesmas definições anteriores para o caso multivariado. As variáveis $X_1, ..., X_n$ são independentes se, para todos $A_1, ..., A_n$,

$$P(X_1 \in A_1, ..., X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

Para atestar independência, é suficiente verificar se

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i).$$

As variáveis $X_1, ..., X_n$ são *IID* (abreviação para *independentes e identicamente distribuídas*) se são idependentes e possuem a mesma distribuição $f(x)$. O vetor $X = (X_1, ..., X_n)$ é uma *amostra aleatória* da distribuição f.

Como exemplo de distribuição multivariada discreta, temos a *distribuição multinomial*. Considere pegar uma bola de uma caixa cheia de bolas, onde cada bola pode possuir uma de k cores. Seja $p = (p_1, ...p_k)$ o vetor cuja componente i é a probabilidade de se pegar uma bola da cor i. Pegue n bolas independentemente, com reposição, e seja $X = (X_1, ..., X_n)$, onde X_i é o número de bolas da cor i pegas. X possui distribuição multinomial (n, p), com função de probabilidade

$$f(x) = \binom{n}{x_1...x_k} p_1^{x_1}...p_k^{x_k}, \tag{2.35}$$

onde

$$\binom{n}{x_1...x_k} = \frac{n!}{x_1!...x_k!}.$$

Como exemplo de distribuição multivariada contínua, temos a *distribuição multivariada Normal*. Um vetor X possui distribuição multivariada Normal se possui densidade

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{2.36}$$

onde $x = (x_1, ..., x_n)$, $\mu = (\mu_1, ..., \mu_n)$ e Σ é uma matriz $k \times k$ simétrica e positivo-definida.

3 Inferência estatística

3.1 Estimativa pontual

Dada uma série de observações $X_1, X_2, ..., X_n$, utilizamos inferência estatística para determinar qual distribuição F gerou tais dados. Normalmente, estamos interessados apenas em algumas propriedades de F, como seu valor esperado.

Por exemplo, se supormos que $X_1, X_2, ..., X_n \sim N(\mu, \sigma)$, para determinar qual distribuição normal gera tais dados, precisamos inferir a média e o desvio padrão a partir dessas amostras. Nesse caso, estamos tentando determinar um *modelo paramétrico* para os dados em questão.

Sejam $X_1, X_2, ..., X_n$ observações IID tiradas de uma distribuição F. Uma *estimativa pontual* $\hat{\theta}_n$ de algum parâmetro θ é uma função $\hat{\theta}_n = g(X_1, ..., X_n)$. O *viés* (bias) de um estimador é dado por

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta, \tag{3.37}$$

e dizemos que um estimador não é enviesado (unbiased) se $\mathbb{E}_\theta(\hat{\theta}_n) = \theta$.

Seja $\hat{\theta}_n$ uma estimativa pontual de θ. $\hat{\theta}_n$ é *consistente* se $\hat{\theta}_n \to \theta$ quando $n \to \infty$.

A distribuição de $\hat{\theta}_n$ é chamada *distribuição amostral*. O desvio padrão de $\hat{\theta}_n$ é chamado de *erro padrão*:

$$\text{se} = \text{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}.$$

Exemplo. Considere $X_1, ..., X_n \sim Bernoulli(p)$. Seja

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

uma estimativa do parâmetro p da distribuição de Bernoulli, que é a probabilidade do evento binário. Temos que

$$\mathbb{E}(\hat{p}_n) = \frac{1}{n}\sum_i \mathbb{E}(X_i).$$

Como

$$\mathbb{E}(X) = \sum_{x\in\{0,1\}} xf(x) = \sum_{x\in\{0,1\}} xp^x(1-p)^{1-x} = p,$$

para $X \sim Bernoulli(p)$,

$$\frac{1}{n}\sum_i \mathbb{E}(X_i) = p,$$

e assim $\hat{p}$ é não-enviesado e consistente.

É comum utilizar o *erro quadrado médio* MSE para avaliar a qualidade de uma estimativa pontual. Definimos

$$\text{MSE} = \mathbb{E}_\theta\left((\hat{\theta}_n - \theta)^2\right), \tag{3.38}$$

onde o valor esperado é tomado sobre a distribuição que gerou os dados. O erro quadrado médio pode ser escrito como a soma de dois termos: o quadrado do viés do estimador e a variância do mesmo, isto é

$$\text{MSE} = \text{bias}(\hat{\theta}_n)^2 + \mathbb{V}_\theta(\hat{\theta}_n). \tag{3.39}$$

De fato, seja $\bar{\theta} = \mathbb{E}_\theta(\hat{\theta}_n)$. Temos que

$$\mathbb{E}_\theta\left((\hat{\theta}_n - \theta)^2\right) = \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta} + \bar{\theta} - \theta)^2\right)$$
$$= \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta})^2 + 2(\hat{\theta}_n - \bar{\theta})(\bar{\theta} - \theta) + (\bar{\theta} - \theta)^2\right).$$

Como $\bar{\theta}$ e θ são constantes (não são variáveis aleatórias),

$$\text{MSE} = \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta})^2\right) + 2\mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta})(\bar{\theta} - \theta) + (\bar{\theta} - \theta)^2.$$

Temos que

$$\mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}) = \mathbb{E}_\theta(\hat{\theta}_n) - \bar{\theta} = 0,$$

e

$$\mathbb{V}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta\left((\hat{\theta}_n - \bar{\theta})^2\right),$$

logo,

$$\text{MSE} = \text{bias}(\hat{\theta}_n)^2 + \mathbb{V}_\theta(\hat{\theta}_n).$$

Um intervalo de confiança $1-\alpha$ para o parâmetro θ é um intervalo $C_n = (a, b)$, onde $a = a(X_1, ..., X_n)$ e $b = b(X_1, ..., X_n)$ são funções dos dados, tal que $P(\theta \in C_n) \geq 1 - \alpha$. Ou seja, a probabilidade de que $\theta \in C_n$ é $1 - \alpha$. Intuitivamente, em uma visão frequentista, se repertirmos o experimento diversas vezes e medirmos θ, C_n conterá θ $(1 - \alpha)\%$ das vezes.

3.2 Estimação paramétrica

Suponha que coletamos amostras $X_1, ..., X_n$, que assumimos que obedecem a uma distribuição $f(x; \theta)$, onde θ é um parâmetro conhecido. Por exemplo, se $X_1, ..., X_n \sim N(\mu, \sigma)$, para determinarmos qual distribuição $N(\mu, \sigma) \in \{N(\mu, \sigma) : \mu, \sigma \in \mathbb{R}, \sigma > 0\}$ gerou os dados, precisamos de uma maneira de *estimar* μ e σ a partir dos dados gerados. Para isso, usamos técnicas de *estimação paramétrica*, que vamos detalhar a seguir.

Parâmetro de interesse. Na maior parte das vezes, estamos interessados em estimar apenas um determinado parâmetro de uma distribuição, enquanto os outros não nos interessam tanto. Por exemplo, se $X_1, ..., X_n \sim N(\mu, \sigma)$, o

parâmetro desconhecido da distribuição é $\theta = (\mu, \sigma)$. Se estamos interessados apenas em μ, dizemos que$\mu = T(\theta)$ é o nosso *parâmetro de interesse*.

Exemplo. Seja $X_1, ..., X_n \sim N(\mu, \sigma)$. Então, $\theta = (\mu, \sigma)$ é o parâmetro que devemos estimar. Suponha que os dados correspondem ao score de pacientes em testes de sangue, e que estamos interessados na fração de pacientes cujo score é maior que 1. Isto é, queremos

$$\tau = P(X > 1) = 1 - P(X < 1)$$

$$= 1 - P\left(\frac{X-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right) = 1 - P\left(Z < \frac{1-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1-\mu}{\sigma}\right),$$

e o parâmetro de interesse é

$$\tau = T(\mu, \sigma) = 1 - \Phi\left(\frac{1-\mu}{\sigma}\right).$$

Exemplo. Seja $X_1, ..., X_n \sim \text{Gamma}(\alpha, \beta)$. A distribuição Gamma é normalmente utilizada para modelar tempos de vida de pessoas, animais e produtos eletrônicos. Suponha que estamos interessados no tempo de vida médio. Queremos estimar

$$T(\alpha, \beta) = \mathbb{E}_\theta(X) = \int x f(x; \alpha, \beta) dx,$$

onde

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, x > 0.$$

Calculando a integral, temos que $T(\alpha, \beta) = \alpha\beta$.

Método dos momentos. Seja $X_1, ..., X_n \sim f(x; \theta)$, onde $\theta = (\theta_1, ..., \theta_k)$. Uma estimativa de θ é dada pelo *método dos momentos*, que funciona da seguinte forma: para $1 \leq j \leq k$, defina o *j-ésimo momento*

$$\alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j dF_\theta(x),$$

e o *j-ésimo momento amostral*

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Uma estimativa $\hat{\theta}_n$ para θ é encontrada pela solução do seguinte sistema de equações:

$$\begin{aligned}\alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ &\vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k\end{aligned}$$

Exemplo. Seja $X_1, ..., X_n \sim N(\mu, \sigma)$. Queremos estimar $\theta = (\mu, \sigma)$. Temos que

$$\begin{aligned}\alpha_1 &= \mathbb{E}(X) = \mu, \\ \alpha_2 &= \mathbb{E}(X^2) = \mathbb{V}(X) + \mu^2 = \sigma^2 + \mu^2.\end{aligned}$$

Segue que

$$\hat{\mu} = \frac{1}{n}\sum_i X_i \tag{3.40}$$

$$\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_i X_i^2, \tag{3.41}$$

o que nos dá as estimativas da média e do desvio padrão da distribuição normal para esse caso.

Maximum likelihood estimator. Sejam $X_1, ..., X_n$ amostras IID distribuídas de acordo com a fdp $f(x; \theta)$, onde θ é algum parâmetro desconhecido. Definimos a função de *likelihood* $L_n = L_n(\theta)$ como a fdp conjunta dos dados vista como uma função de θ:

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta). \tag{3.42}$$

Isto é, intuitivamente, a função de likelihood nos dá a probabilidade de observarmos os dados que observamos. Muitas vezes, utilizamos a *log-likelihood*

$$l_n(\theta) = \log(L_n(\theta)). \tag{3.43}$$

O *estimador de likelihood máxima* (maximum likelihood estimator, ou MLE), é as estimativa $\hat{\theta}_n$ do parâmetro θ que maximiza a função likelihood, i.e.,

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} L_n(\theta), \tag{3.44}$$

onde Θ é um conjunto de parâmetros possíveis. Maximizar L_n equivale a maximizar l_n e vice-versa, então é comum trocarmos likelihood por log-likelihood quando conveniente. Intuitivamente, o MLE maximiza a probabilidade de observarmos os dados que observamos, então essa definição é bastante razoável.

Exemplo. Suponha que $X_1, ..., X_n \sim Bernoulli(p)$. Segue que a função likelihood é dada por

$$\begin{aligned} L_n(p) &= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} \\ &= p^{X_1}(1-p)^{1-X_1}p^{X_2}(1-p)^{1-X_2}p^{X_3}(1-p)^{1-X_3}... \\ &= p^S(1-p)^{n-S}, \end{aligned}$$

onde $S = \sum_i X_i$. Logo,

$$L_n(p) = p^S(1-p)^{n-S}, \tag{3.45}$$

$$l_n(p) = S\log(p) + (n-S)\log(1-p). \tag{3.46}$$

O estimador MLE é dado encontrando o máximo da likelihood:

$$\begin{aligned} \frac{dL_n(p)}{dp} &= Sp^{S-1}(1-p)^{n-S} - p^S(n-S)(1-p)^{n-S-1} \\ &= p^{S-1}(1-p)^{n-S-1}\left[S(1-p) - p(n-S)\right] = 0, \end{aligned}$$

ou seja,

$$S - Sp - pn + Sp = 0,$$

e o MLE é dado por

$$\hat{p} = \frac{1}{n}S = \frac{1}{n}\sum_i X_i. \tag{3.47}$$

Exemplo. Seja $X_1, ..., X_n \sim Unif(0, \theta)$, onde devemos estimar o limite superior do intervalo da distribuição uniforme. Temos que

$$f(x;\theta) = \begin{cases} 1/\theta, \ 0 \leq x \leq \theta, \\ 0, \ x > \theta. \end{cases} \tag{3.48}$$

Considere que $\theta < \max(\{X_i\}_{i=1}^n)$. Logo, $f(X_i;\theta) = 0$ e $L(\theta) = 0$. Agora, se $\theta \geq \max(\{X_i\}_{i=1}^n)$, $L(\theta) = 1/\theta^n$. Logo,

$$L(\theta) = \begin{cases} 1/\theta^n, \ \theta \geq \max(\{X_i\}_{i=1}^n), \\ 0, \ \theta < \max(\{X_i\}_{i=1}^n). \end{cases} \tag{3.49}$$

É fácil ver que a função $L(\theta)$ atinge seu máximo em $\hat{\theta} = \max(\{X_i\}_{i=1}^n)$, que é o MLE. Logo, o intervalo da distribuição uniforme é estimado como sendo $[0, \max(\{X_i\}_{i=1}^n)]$.

3.3 Testes de hipótese

Suponha que queremos comprovar que fumar pode causar câncer. Realizamos um experimento onde coletamos dados de pessoas fumantes e não-fumantes. Se a incidência de câncer no grupo de fumantes for maior que a incidência no grupo de não-fumantes, ficamos dispostos a dizer que fumar de fato é um fator de risco. Neste exemplo de um *teste de hipótese*, temos duas hipóteses:

- H_0: a incidência de câncer no grupo de não fumantes é maior ou igual a incidência de câncer no grupo de fumantes;
- H_1: a incidência de câncer é maior no grupo de fumantes que no grupo de não-fumantes.

Seja θ = taxa de incidência de câncer em alguma população. Segue que

$$H_0 : \theta_{\text{não-fumantes}} \geq \theta_{\text{fumantes}}, \tag{3.50}$$

$$H_1 : \theta_{\text{não-fumantes}} < \theta_{\text{fumantes}}. \tag{3.51}$$

Os dados que coletamos devem ser utilizados para comprovar ou rejeitar a *hipótese nula* H_0. Suponha que X é a variável aleatória que representa os dados coletados, e que S seja o conjunto onde essa variável assume valores. Se os valores coletados de X estão dentro de um conjunto $R \subset S$, chamado *região de rejeição*, descartamos a hipótese nula, i.e.,

$$X \in R \Rightarrow \text{Rejeitamos } H_0, \tag{3.52}$$

$$X \notin R \Rightarrow \text{Mantemos } H_0. \tag{3.53}$$

Normalmente, a região de rejeição é dada por

$$R = \{x : T(x) > c\}, \tag{3.54}$$

onde T é a *estatística do teste* e c é um valor crítico. Por exemplo, podemos pensar na região crítica para o caso de incidência de câncer da seguinte forma: seja X o fato de certa pessoa estar doente ou não em dada população. Se a média de pessoas doentes foram altas o suficiente nessa população, podemos atestar que essa população está inclinada a ter câncer. A região de rejeição é o conjunto de respostas que dão uma média maior que um determinado valor crítico c.

Definimos o *poder* de um teste com região de rejeição R como sendo

$$\beta(\theta) = P_\theta(X \in R). \tag{3.55}$$

O *tamanho* de um teste é dado por

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta). \tag{3.56}$$

O poder do teste é a probabilidade, dado que o parâmetro θ é verdadeiro, de que X esteja dentro da região de rejeição. Isto é, é a probabilidade de rejeitar H_0 dado θ. Já o tamanho de um teste pode ser interpretado como a probabilidade máxima de que $X \in R$ (rejeitamos H_0) dado que H_0 é verdadeiro (θ pertence ao conjunto de parâmetros Θ_0).

Exemplo. Suponha que $X_1, ..., X_n \sim N(\mu, \sigma)$, onde σ é conhecido, e queremos testar a hipótese $\mu > 0$. Temos então que $H_0 : \mu \leq 0$ e $H_1 : \mu > 0$. Os espaços que nos interessam são então $\Theta_1 = (0, \infty)$ e $\Theta_0 = (-\infty, 0]$. A estimativa de μ que consideraremos é a MLE $\hat{\mu} = \bar{X}$, a média amostral. Segue que rejeitaremos H_0 se $\bar{X} > c$, onde c é uma constante a ser definida. A região de rejeição é

$$R = \{(x_1, ..., x_n) : \frac{1}{n} \sum_{i=1}^{n} x_i > c\}.$$

O poder do teste é dado por

$$\begin{aligned} \beta(\mu) = P_\mu(X \in R) = P_\mu(\bar{X} > c) &= P(Z > \sqrt{n}(c - \mu)/\sigma) \\ &= 1 - \Phi(\sqrt{n}(c - \mu)/\sigma). \end{aligned} \tag{3.57}$$

O tamanho do teste é $\alpha = \sup_{\mu \in (-\infty, 0]} \beta(\mu)$. A cdf Φ é crescente em seu argumento, então seu máximo (menor limite superior) no intervalo $(-\infty, 0]$ acontece no maior valor do intervalo, i.e., $\alpha = \beta(0) = 1 - \Phi(\sqrt{n}c/\sigma)$. Dado que quermos um teste cuja probabilidade de rejeitarmos H_0 sendo que H_0 é verdade seja α, podemos escolher o limite c de modo que

$$\sqrt{n}c\sigma = \Phi^{-1}(1 - \alpha),$$

i.e.,

$$c = \sigma\Phi^{-1}(1 - \alpha)/\sqrt{n}.$$

Assim, rejeitamos H_0 quando

$$\bar{X} > \sigma\Phi^{-1}(1 - \alpha)/\sqrt{n}.$$

4 Teoria da Informação

Para quantificar a informação contida em uma distribuição, vamos utilizar nossa intuição. O acontecimento de um evento raro carrega mais informação

do que o acontecimento de um evento frequente. Por exemplo, o evento "o Sol nasceu hoje"não nos traz nenhuma informação nova, enquanto que o evento "aconteceu um eclipse solar hoje"é bastante informativo. Considerando isso, uma medida da informação que um evento carrega deve apresentar as seguintes características:

- A informação de um evento com probabilidade alta de ocorrência deve ser baixa;
- A informação de um evento com probabilidade baixa de ocorrência deve ser alta;
- A informação de eventos independentes deve ser aditiva. Por exemplo, saber que tiramos duas coroas em duas jogadas de uma moeda deve trazer mais informação que saber que tiramos uma cora em uma jogada de uma moeda.

Baseado nessas condições, podemos definir a *auto-informação* $I(x)$ de um evento $X = x$ como sendo

$$I(x) = -\log P(x), \tag{4.58}$$

onde $P(x)$ é a probabilidade de $X = x$.

Para quantificar a informação que uma distribuição de probabilidades carrega, utilizamos a *entropia de Shannon*

$$H(X) = \langle I(x) \rangle_{X \sim P} = - \langle \log P(x) \rangle_{x \sim P}, \tag{4.59}$$

onde $\langle . \rangle_{X \sim P}$ é o valor esperado com respeito à distribuição P. A entropia de Shannon é a quantidade de informação esperada de um evento distribuído de acordo com P. Distribuições que são quase determinísticas, onde o resultado é quase certo, possuem baixa entropia, enquanto que distribuições que são parecidas com a distribuição uniforme possuem alta entropia.

Exemplo. Considere a distribuição de Bernoulli com probabilidade p. Segue que a entropia de Shannon dessa distribuição é dada por

$$H(P) = -P(X = 0)\log P(X = 0) - P(X = 1)\log P(X = 1)$$

$$= -p\log p - (1 - p)\log(1 - p).$$

O gráfico de tal função é representado na figura 1. Note que para $p \approx 0$ ou $p \approx 1$, a entropia é mínima, pois os eventos $X = 0$ ou $X = 1$ são quase certos. Para $p \approx 1/2$, a entropia é máxima.

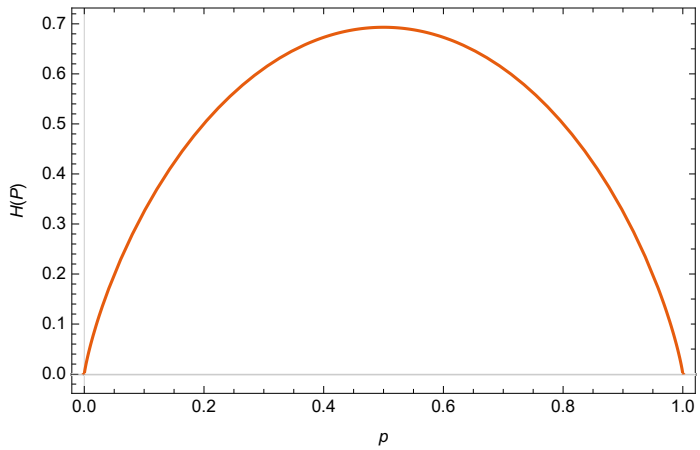

Figura 1: Entropia de Shannon da distribuição de Bernoulli.

Dadas duas distribuições P e Q sobre uma mesma variável aleatória, podemos medir o quão diferentes elas são utilizando a *divergência de Kullback-Leibler*

$$D_{KL}(P||Q) = \langle \log P(x) - \log Q(x) \rangle_{X \sim P} \,. \tag{4.60}$$

Como $D_{KL}(P||Q) \geq 0$ e como $D_{KL}(P||Q) = 0 \Leftrightarrow P = Q$, é comum utilizar $D_{KL}(P||Q)$ como uma medida de distância entre distribuições. No entanto, note que $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, o que significa que tal quantidade não define uma métrica no espaço de distribuições. É possível interpretar $D_{KL}(P||Q)$ como a quantidade de informação perdida quando utilizamos Q para aproximar P.

Referências

[1] Larry Wasserman. *All of Statistics: a Concise Course in Statistical Inference*. Springer Science & Business Media, 2013.