1 Estatística e Probabilidades

Inferência Estatística consiste na generalização das informações a respeito de uma amostra, para a sua população.

A Probabilidade considera modelos para estimar informações sobre instâncias. É um processo de dedução lógica.

A Estatística considera informações sobre instâncias pra gerar um modelo para toda a população. É um processo de raciocínio indutivo.

1.1 Exemplo da diferença da média da população para a média amostral.

Considere um dado de seis lados. Qual a média esperada para jogadas desse dado?

$$\frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3,5$$

Suponha que joguei o dado 5 vezes e obtive: 2, 3, 3, 6, 1. O que é plenamente possível.

A média amostral é dada por

$$\frac{2+3+3+6+1}{5} = 3,0$$

Assim, $\mu = 3,5$ e $\bar{x} = 3,0$

2 Probabilidade

Probabilidade caracteriza um fenômeno aleatório e é um modelo para a freqüência que ocorre um evento quando se tende a um número infinito de experimentos, jogadas, amostras.

Seja A um evento, então:

1. $P(A) \geq 0$
2. Se $A \cap B = \emptyset$, então $P(A \cup B) = P(A) + P(B)$.
3. Seja S o espaço amostral, então $P(S) = 1$.

Outras propriedades:

- $P(\emptyset) = 0$.
- $\bar{A}$ é complemento de A, então $P(A) = 1 - P(\bar{A})$.
- E é claro que $P(A \cup\ B) = P(A) + P(B) - P(A \cap B)$.

Quando $A \cap B = \emptyset$ dizemos que os eventos A e B são mutuamente exclusivos.

2.1 Probabilidade e variável aleatória

Definimos a distribuição de probabilidades ou função de densidade de probabilidade (pdf - probability density function) sobre pontos da reta de Borel.

No caso de variáveis discretas, o valor da função de densidade de probabilidade corresponde à freqüência relativa de que o resultado de um experimento seja igual ao argumento da função.

$$P(X = 5) = f(5)$$

No caso de variáveis contínuas, o valor da densidade de probabilidade é tal que a integral da função sobre um intervalo corresponda à freqüência relativa do resultado de um experimento caia dentro do intervalo.

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

2.2 Variável Aleatória de Bernoulli

Variável aleatória de Bernoulli apresenta como possíveis valores 0 ou 1. Isto é, o espaço amostral é binário $= \{0, 1\}$.

Distribuição de Bernoulli

$$Bern(x; \alpha) = \begin{cases} 1 - \alpha & \text{se } x = 0 \\ \alpha & \text{se } x = 1 \\ 0 & \text{caso contrário} \end{cases}$$

Em geral utilizamos $p = \alpha$ e $q = 1 - \alpha$.

Exemplo: Quantos compradores levam monitores de CRT?

$$P(1) = 0,2$$

$$P(0) = 0,8$$

(soma deve ser 1)

$$P(x) = Bern(x; 0,2)$$

α é uma parâmetro, isto é, uma quantidade que define a distribuição dentre uma família de distribuições.

Exemplo: (Devore) Quantos nascimentos até nascer um menino?

$$\begin{array}{c} P(B) = p \\ P(G) = 1 - p \\ \hline p(1) = P(B) = p \\ p(2) = P(G) \cdot P(B) = (1 - p)p \\ p(3) = P(G)P(G)P(B) = (1 - p)^2 p \\ p(x) = \begin{cases} (1 - p)^{x-1} p & x = 1, 2, 3, \ldots \\ 0 & \text{caso contrário} \end{cases} \end{array}$$

3 Função de densidade acumulada

A função de densidade acumulada (cdf - cumulative density function) é definida para variáveis discretas como

$$F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$$

No caso contínuo, a definição é a seguinte:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(y)dy$$

Assim, a probabilidade de um intervalo pode ser obtida por:

$$P(a < X < b) = F(b) - F(a)$$

Os casos contínuos e discretos podem ser unificados utilizando ou funções impulso de Dirac ou definição da integral de Lebesgue sobre espaços mensuráveis (incluindo σ-álgebras).

Propriedades

- $F(-\infty) = 0$
- $F(+\infty) = 1$
- A cdf é sempre crescente.
- A cdf é diferenciável à direita

3.1 Amostras aleatórias sintéticas

Para fins de simulação, se possuímos um gerador de números pseudo-aleatórios entre 0 e 1 (exclusive) e com distribuição uniforme, podemos utilizar a cdf para obter números aleatórios sorteados de acordo com uma determinada distribuição de probabilidades.

Se F é a cdf da distribuição de que queremos obter amostras, então a probabilidade de obtermos um valor no intervalo (a, b) é $F(b) - F(a)$. Como F varia de 0 a 1, assim como o nosso gerador de números aleatórios, e é crescente, então se obtivermos um valor sorteado uniformemente entre $F(a)$ e $F(b)$ podemos considerar como um valor no intervalo (a, b) sorteado de acordo com a distribuição almejada.

Assim, basta sortear uniformemente um valor x entre 0 e 1 e aplicar a inversa da cdf $y = F^{-1}(x)$.

Mostrar que a variável aleatória y tem cdf F se x for variável aleatória uniforme.

3.2 Mediana e quantis dada a densidade acumulada

A mediana de uma distribuição corresponde ao valor que separa 50% da probabilidade, assim:

$$\tilde{x} = F^{-1}(50\%)$$

Da mesma forma qualquer quantil (quartis ou percentis) podem ser obtidos.

3.3 Obtendo a pdf a partir da cdf

Lembrar que

$$f(x) = F'(x)$$

4 Esperança

Esperança é o valor médio esperado de uma variável aleatória.

$$E(x) = \sum_{x \in D} x \cdot p(x)$$

No caso contínuo,

$$E(x) = \int_{-\infty}^{+\infty} x \cdot p(x) dx$$

Exemplo (Devore)
Crianças são distribuídas na escala Apgar de 0 a 10.

Apgar	0	1	2	3	4	5	6	7	8	9	10
%	0,002	0,001	0,002	0,005	0,02	0,04	0,18	0,37	0,25	0,12	0,01

$$E(x) = \mu = 0 \cdot 0,002 + 1 \cdot 0,001 + \ldots + 10 \cdot 0,01 = 7,15$$

Exemplo (Devore)
X é o número de entrevistas pelas quais um estudante passa antes de conseguir um emprego.

$$p(x) = \begin{cases} \frac{k}{x^2} & x = 1, 2, 3, \ldots \\ 0 & \text{caso contrário} \end{cases}$$

k é tal que $\sum_{x=1}^{\infty} \frac{k}{x^2} = 1$ e não precisa ser calculado (basta ver que é finito).

$$E(x) = \frac{1 \cdot k}{1} + \ldots + \frac{x \cdot k}{x^2} + \ldots = \sum_{x=1}^{\infty} \frac{k}{x} \to \infty$$

Trata-se do somatório da série harmônica que não converge. Dessa forma, a média não é uma boa medida para caracterizar esse tipo de distribuição.

4.1 Esperança de uma função

$$E[f(x)] = \int_{x \in D} f(x) p(x) dx$$

Propriedade de operador linear:

$$E[aX + b] = aE[X] + b$$

4.2 Variância da distribuição

Seja μ o valor médio esperado dado por

$$\mu = E(x)$$

A variância é o valor esperado para o quadrado dos desvios

$$Var(x) = E[(x - \mu)^2]$$

Outras fórmulas que podem ser utilizadas para obter a variância:

$$Var(x) = \int_{x \in D} (x - \mu)^2 p(x) dx = E[x^2] - E[x]^2$$

4.3 Momentos estatísticos

Além da média e variância, é possível definir descritores de ordem mais alta da distribuição. O momento de ordem n é definido como a esperança de x^n.

$m_0 = E(x^0) = E(1) = \int p(x)dx = 1$
$m_1 = E(x) = \mu$
$m_2 = E(x^2)$
$m_3 = E(x^3)$

A partir dos momentos de ordem 2, podem-se utilizar momentos baseados nos desvios em relação à média. Esses são momentos centrais.

$\mu_2 = E[(x - \mu)^2] = \sigma^2$
$\mu_3 = E[(x - \mu)^3]$

Duas medidas importantes para caracterizar uma distribuição não-normal são os coeficientes de skewness e de kurtosis. No caso do skewness, coeficiente próximo de zero significa simetria, caso contrário, uma tendência à esquerda para números negativos e, à direita para números positivos.

$$\text{skewness} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$$

A kurtosis mede a concentração próxima a média (ou pico). No caso da normalidade, o valor é 3. Menos que 3, a distribuição é mais achatada chamada platykurtic. Maior que 3, o pico é mais acentuado e a distribuição é chamada leptokurtic.

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2}$$

ou

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2} - 3$$

4.4 Desigualdades interessantes sobre momentos

Desigualdade de Chebyshev se aplica a qualquer distribuição.

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

Uma interpretação pode ser obtida para $a = k\sigma$

$$P(|X - \mu| \geq \sigma) \leq \frac{1}{k^2}$$

A probabilidade do valor de X cair numa distância maior ou igual a k desvios-padrão da média é de no máximo $\frac{1}{k^2}$. Isso para qualquer tipo de distribuição. Para 3 sigmas, a probabilidade é menor ou igual a 1/9. Para 6 sigmas, a probabilidade é no máximo 1/36 ou 2,7%.

Desigualdade de Markov se aplica a variáveis não-negativas.

$$P(X \geq a) \leq \frac{\mu}{a}$$

Em ambos os casos, $a > 0$.

4.5 Entropia

A entropia é uma medida da aleatoriedade de uma distribuição, definida como

$$H(X) = E\left[\ln \frac{1}{P(X)}\right] = -\int_{x \in D} p(x) \ln p(x) dx$$

Se o logaritmo for na base 2, a unidade de medida é o bit. (Para ln, diz-se que é o nit).

Verificar que $\lim_{x \to 0} xlnx = 0$.

Considere uma variável aleatória de Bernoulli com probabilidade p de sucesso. Pela definição de entropia (vamos utilizar log na base 2),

$$H(p) = -p \lg p - (1 - p) \lg(1 - p)$$

Pelos limites, temos que

$$H(0) = H(1) = 0$$

Interpretação: total determinismo se 100% de chance de ser 1 ou de ser 0.

Exemplo (Mitzenmacher e Upfal): Entropia de duas moedas viciadas, uma com 3/4 de probabilidade de ser coroa e outra com 7/8.

$$H(\frac{3}{4}) = -\frac{3}{4} \lg \frac{3}{4} - \frac{1}{4} \lg \frac{1}{4} \approx 0,8113$$

$$H(\frac{7}{8}) = -\frac{7}{8} \lg \frac{7}{8} - \frac{1}{8} \lg \frac{1}{8} \approx 0,5436$$

A primeira moeda é aquela que apresenta distribuição com maior entropia. Logo, menos se pode dizer sobre o resultado obtido antes de observá-lo.

Agora, queremos determinar p para que a entropia seja máxima.

$$\frac{\partial H(p)}{\partial p} = -\lg p + \lg(1-p) = \lg \frac{1-p}{p}$$

Assim, $\lg p = \lg(1-p)$ que acontece quando p vale 1/2 e $H(1/2) = 1$ bit.
O lançamento de uma moeda não-viciada tem a aleatoriedade de um bit.
Suponha uma roleta de 8 posições de probabilidade uniforme. Calcular a entropia:

$$H = 8 \times \left(-\frac{1}{8} \lg \frac{1}{8}\right) = 3$$

São necessários 3 bits para codificar o resultado da roleta.

4.6 Distribuição de máxima entropia

Encontrar a distribuição de máxima entropia consiste em determinar a pdf $p(x)$ que maximiza H sob as restrições que regram as pdfs. Assim, procura-se maximizar

$$H = -\int_D p(x) \ln p(x) dx$$

sujeito à

$$\int_D p(x) dx = 1$$

Vamos procurar a pdf de máxima entropia, dado que conhecemos a média μ e a variância σ^2. As restrições são:

$$\mu = \int_{-\infty}^{+\infty} x p(x) dx, \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx$$

Formulando com multiplicadores de Lagrange, o novo funcional a minimizar é

$$F = \begin{array}{l} -\int_{-\infty}^{+\infty} p(x) \ln p(x) dx + \\ \lambda_1 \left(\int_{-\infty}^{+\infty} p(x) dx - 1\right) + \\ \lambda_2 \left(\int_{-\infty}^{+\infty} x p(x) dx - \mu\right) + \\ \lambda_3 \left(\int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx - \sigma^2\right) \end{array}$$

Derivando em função de p e igualando a zero, obtemos que

$$p(x) = e^{-1+\lambda_1+\lambda_2 x+\lambda_3 (x-\mu)^2}$$

Substituindo $p(x)$ nas restrições, determinamos os multiplicadores λ.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

5 A Distribuição Gaussiana (ou Normal)

Para média μ e variância σ^2, a distribuição normal é definida pela expressão:

$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

5.1 A Distribuição normal padrão

Para média zero e variância unitária (e desvio-padrão), definimos a distribuição normal padrão:

$$p(z) = N(z; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp -\frac{z^2}{2}$$

A função cumulativa de densidade da normal padrão é baseada na função de erro:

$$\Phi(z) = \int_{-\infty}^{z} N(y; 0, 1) dy$$

Qualquer distribuição normal pode ser padronizada utilizando a transformação linear:

$$Z = \frac{X - \mu}{\sigma}$$

5.2 Propriedade dos desvio-padrão da distribuição normal

A probabilidade de uma amostra ser obtida dentro de 1 desvio-padrão da média é dada por

$$\Phi(1) - \Phi(-1)$$

Vamos tabelar para alguns desvios-padrão de distância

k	dentro de $k\sigma$	fora	Chebyshev $1/k^2$
1	0,6826	0,3173	1
2	0,9545	0,0455	0,25
3	0,9973	0,0027	0,1111
6	0,9999	0,2e-8	0,0278

6 A Distribuição Binomial

6.1 Bernoulli trials

Experimentos de Bernoulli (jogadas, rodadas, tentativas)

- n experimentos chamados tentativas;
- resultado de cada experimento é sucesso S ou falha F;
- tentativas são independentes;
- probabilidade de sucesso (p) é constante de uma tentativa para outra.

Repetimos um experimento binomial de Bernoulli n vezes.
Quantas vezes foi obtido "sucesso", isto é, resposta 1?
Resultados possíveis e igualmente prováveis de 3 tentativas:
SSS SSF SFS SFF FSS FSF FFS FFF
Agrupar por número de sucessos

3	SSS
2	SSF SFS FSS
1	SFF FSF FFS
0	FFF

A distribuição binomial é definida por

$$P(x) = Bin(x; n, p) = \begin{cases} \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, & x = 0, 1, 2, \ldots, n \\ 0, & \text{caso contrário} \end{cases}$$

Lembrando números binomiais:

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{(n-x)!x!}$$

A esperança e a variância de uma distribuição binomial são dadas por:

$$E(x) = n \cdot p$$

$$Var(x) = np(1-p)$$

A distribuição binomial pode aproximar uma normal com média np e variância $np(1-p)$.

7 Distribuição uniforme

Na distribuição uniforme discreta, cada elemento do espaço amostral é igualmente provável.

No caso contínuo, a probabilidade é proporcional ao tamanho do intervalo (desde que dentro do intervalo em que a distribuição é definida). Para um intervalo $[A, B]$ utilizamos a definição:

$$f(x; A, B) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{caso contrário} \end{cases}$$

8 Caso multivariado

Definimos a função de densidade de probabilidade conjunta no caso de mais de uma variável:

$$p(x, y) = P(X = x \wedge Y = y)$$

Caso discreto:

$$P[(X,Y) \in A] = \sum_A p(x,y)$$

Caso contínuo:

$$P[(X,Y) \in A] = \int\int_A p(x,y)dx\,dy$$

Probabilidade de ocorrer uma instância dentro de um (hiper-)retângulo

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_l \leq X_l \leq b_l) = \int_{a_1}^{b_1} \ldots \int_{a_l}^{b_l} p(x_1, \ldots, x_l)dx_1 \ldots dx_l$$

8.1 Probabilidade marginal

Corresponde à soma de todas probabilidades conjuntas para um dado eixo

$$p_x(x) = \sum_y p(x,y)$$

$$p_y(y) = \sum_x p(x,y)$$

As funções p_x e p_y são funções de densidade de probabilidade marginal.

9 Independência estatística

Duas variáveis aleatórias são independentes se e só se

$$p(x,y) = p_x(x) \cdot p_y(y), \forall (x,y)$$

10 Distribuição multinomial

No caso binomial, definimos uma distribuição para o número de elementos obtidos para uma classe dentre duas. No caso multinomial, para M classes, temos a quantidade x_i correspondente ao número de elementos obtidos na classe i de um total de n elementos.

Primeiro notar que

$$\sum_{i=1}^{M} x_i = n$$

E que cada $x_i > 0$.

Após analisar n objetos (com reposição), a probabilidade de se obter x_1 amostras na classe C_1, x_2 na classe C_2 ... e x_M na classe C_M é dada por

$$p(x_1, x_2, \ldots, x_M) = \begin{cases} \frac{n!}{x_1!x_2!\ldots x_M!} {p_1}^{x_1} {p_2}^{x_2} \ldots {p_M}^{x_M}, & x_i = 0 \ldots n \\ 0, & \text{caso contrário} \end{cases}$$

onde p_i é a probabilidade do resultado de uma amostra estar na classe C_i.

11 Covariância e Correlação

Valor esperado no caso conjunto:

$$E[h(x,y)] = \int \int_{\Omega} h(x,y)p(x,y)dxdy$$

Covariância

$$Cov(x,y) = E[(x-\mu_x)\cdot(y-\mu_y)]$$

$$Cov(x,y) = \int \int_{\Omega} (x-\mu_x)(y-\mu_y)p(x,y)dxdy$$

$$Cov(x,y) = E[x\cdot y] - \mu_x \cdot \mu_y$$

Coeficiente de correlação

$$Corr(x,y) = \rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

Propriedade

$$Corr(aX+b, cY+d) = Corr(X,Y)$$

Se X e Y são independentes, $\rho = 0$.

12 Matriz de covariância

Seja $\sigma_x y = Cov(x,y)$.

Notar que $\sigma_x y = \sigma_y x$ e que $\sigma_x x = Cov(x,x) = E[(x-\mu_x)^2] = \sigma_x^2$

A matriz de covariância Σ é definida como

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_n^2 \end{pmatrix}$$

Para um vetor coluna de variáveis aleatórias $\mathbf{x}$ com vetor média μ

$$\Sigma = E_{\mathbf{x}}[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]$$

13 Gaussianas multivariadas

$$p(x) = \frac{1}{(2\pi)^{\frac{2}{2}}\|\Sigma\|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

14 Amostragem

$X_1 \dots X_n$ formam uma amostra aleatória de tamanho n se

1. X_i são todos independentes entre si;
2. Todo X_i possui a mesma distribuição de probabilidades.

A amostra é chamada i.i.d. ou independentes e identicamente distribuídos.

14.1 Distribuição da média amostral

$X_1 \dots X_n$ elementos amostrados de uma distribuição qualquer com média μ e desvio-padrão σ. Para a média $\bar{X}$ da amostra, temos que

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

14.2 Teorema do Limite Central

$X_1 \dots X_n$ elementos amostrados de uma distribuição qualquer com média μ e desvio-padrão σ.

Se n é suficientemente grande, $\bar{X}$ tem aproximadamente uma distribuição normal com $\mu_{\bar{X}} = \mu$ e $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.

Quanto maior n, melhor a aproximação.

Conseqüência: a Gaussiana é boa (em média) para aproximar ruído.