Introdução às Probabilidades e Estatística

Resumo

João Matos

Capítulo 1

Probabilidade

1.1 Massa de probabilidade (variáveis discretas)

1.1.1 Modelo Binomial

É a distribuição do número de sucessos obtidos depois de n ensaios de Bernoulli independentes de parâmetros $p \in [0, 1]$, ou seja, é a distribuição da soma de n variáveis aleatórias independentes da distribuição de Bernoulli de mesmo parâmetro. O modelo binomial é utilizado quando se verificam as seguintes propriedades:

- Cada tentativa tem exclusivamente como resultado duas possibilidades, sucesso ou fracasso;
- Cada tentativa é independente das demais;
- A probabilidade de sucesso p a cada tentativa permanece constante independente das demais;
- A variável de interesse, ou pretendida, é o número de sucessos k nas n tentativas.

As suas variáveis aleatórias são dadas por:

$$X \cap B(n, p)$$

E a sua f.m.p é dada por:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, 1, 2, ..., n$$

O seu valor médio é $E(X) = np$ e a variância é $var(X) = np(1 - p)$

A soma de variáveis aleatórias independentes com distribuição binomial é calculada como:

$$\sum_{i=1}^{n} X_i \cap B\left(\sum_{i=1}^{n} n_i, p\right)$$

1.1.2 Modelo Geométrico

É a distribuição que modela o tempo de espera do primeiro sucesso de uma de ensaios de Bernoulli independentes com probabilidade de sucesso $p \in [0, 1]$. É a única distribuição discreta que possui a propriedade de falta de memória.
As suas variáveis aleatórias são dadas por:

$$X \cap G(p)$$

E a sua f.m.p é dada por:

$$P(X = k) = (1 - p)^{k-1}p, k = 1, 2, ...$$

O seu valor médio é $E(X) = \frac{1}{p}$ e a variância é $var(X) = \frac{1-p}{p^2}$

1.1.3 Modelo Binomial Negativo

Esta distribuição indica o número de tentativas necessárias para obter r sucessos de igual probabilidade p ao fim de n experimentos de Bernoulli, sendo a última tentativa um sucesso. As suas variáveis aleatórias são dadas por:

$$X \cap BN(r, p)$$

E a sua f.m.p é dada por:

$$P(X = k) = \binom{k-1}{r-1} p^r (1 - p)^{k-r}, k = r, r + 1, ...$$

O seu valor médio é $E(X) = \frac{r}{p}$ e a variância é $var(X) = \frac{r(1-p)}{p^2}$

1.1.4 Modelo Hipergeométrico

A distribuição hipergeométrica modela uma retirada simultânea de n bolas uma urna contendo uma proporção pN de bolas vencedoras e uma proporção $(1-p)N$ de bolas perdedoras para um número total N de bolas. A distribuição descreve o número de bolas vencedoras extraídas.

As suas variáveis aleatórias são dadas por:

$$X \cap H(N, n, p)$$

E a sua f.m.p é dada por:

$$P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{M}}, k=max(0,n-(N-M)),...,min(M,n)$$

O seu valor médio é $E(X)=np$ e a variância é $var(X)=np(1-p)\frac{N-n}{N-1}$

1.1.5 Modelo de Poisson

A distribuição de Poisson descreve o comportamento do número de eventos que ocorrem em um espaço determinado de tempo.

As suas variáveis aleatórias são dadas por:

$$X \cap P(\lambda)$$

E a sua f.m.p é dada por:

$$P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}, k=0,1,2,... \quad (\lambda>0)$$

O seu valor médio é $E(X)=\lambda$ e a variância é $var(X)=\lambda$

A soma de variáveis aleatórias independentes com distribuição de Poisson é calculada como:

$$\sum_{i=1}^{n} X_i \cap P\left(\sum_{i=1}^{n}\lambda_i\right)$$

1.2 Distribuição (variáveis contínuas)

1.2.1 Modelo Uniforme

A distribuição uniforme corresponde á a probabilidade de se gerar qualquer ponto num intervalo contido no espaço amostral que é proporcional ao tamanho do intervalo, visto que na distribuição uniforme a $f(x)$ é igual para qualquer valor de x no intervalo considerado. Por exemplo, se considerarmos um intervalo em x de zero à dez, ($x \in [0,10]$), e assumirmos que temos uma distribuição uniforme nesse intervalo, a probabilidade de f(x) no intervalo 2<x<5 é igual a probabilidade de f(x) no intervalo 5 < x < 8 pois sabemos que a distribuição é uniforme nesses intervalos e possuímos os intervalos com o mesmo tamanho.

As suas variáveis aleatórias são dadas por:

$$X \cap U(a, b)$$

E a sua f.d.p. é dada por:

$$F(x) = \frac{x-a}{b-a}, a \leq x < b$$

O seu valor médio é $E(X) = \frac{a+b}{2}$ e a variância é $var(X) = \frac{(b-a)^2}{12}$

1.2.2 Modelo Exponencial

A distribuição Exponencial utiliza-se, na prática, para modelar o tempo de espera até à ocorrência de certo acontecimento, o tempo de vida de determinados objetos ou a duração de certos serviços, como, por exemplo, o tempo de vida de uma componente eletrónica, o tempo que um cliente demora a ser atendido, a duração de uma chamada telefónica, etc. A distribuição Exponencial modela o tempo que decorre entre ocorrências sucessivas de um acontecimento, quando o número de ocorrências de um acontecimento num certo intervalo de tempo tem distribuição de Poisson. A distribuição Exponencial é caracterizada pela "falta de memória", i.e., é a única distribuição contínua que verifica a seguinte propriedade:

$$P(X > t + s | X > t) = P(X > x), \forall t > 0, \forall x > 0$$

As suas variáveis aleatórias são dadas por:

$$X \cap Exp(\lambda), \lambda > 0$$

E a sua f.d.p. é dada por:

$$F(x) = 1 - e^{-\lambda x}, x > 0$$

O seu valor médio é $E(X) = \frac{1}{\lambda}$ e a variância é $var(X) = \frac{1}{\lambda^2}$

1.2.3 Modelo Normal

As suas variáveis aleatórias são dadas por:

$$X \cap N(\mu, \sigma)$$

E a sua f.d.p. é dada por:

$$\Phi\left(\frac{x-\mu}{\sigma}\right), x \in \mathbb{R}$$

O seu valor médio é $E(X) = \mu$ e a variância é $var(X) = \sigma^2$

A soma de variáveis aleatórias independentes com distribuição de Normal é calculada como:

$$S_n = \sum_{i=1}^{n} c_i X_i \cap N\left(\sum_{i=1}^{n} c_i \mu_i, \sqrt{\sum_{i=1}^{n} c_i^2 \sigma_i^2}\right), \forall c_1, c_2, ..., c_n \in \mathbb{R}$$

1.3 Pares aleatórios discretos

1.3.1 Função massa de probabilidade condicional

Seja (X, Y) um par aleatório discreto que assume os valores (x_i, y_j). A f.m.p. condicional de X dado $Y = y_j$ é definida por:

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{.j}}, i = 1, ..., k$$

desde que $P(X = x_i) > 0$

Capítulo 2

Estatística

2.1 Estatísticas de dispersão (População)

2.1.1 Valor médio

Para variáveis discretas, o valor médio calcula-se na forma:

$$E(X) = \sum_i x_i p_i$$

Por exemplo, o número de faces obtidas em dois lançamentos de uma moeda:

Nº de sucessos	0	1	2
Probabilidade	0.25	0.5	0.25

O seu valor médio é dado por:

$$E(X) = 0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1$$

Para variáveis continuas obtém-se da forma:

$$\int_{-\infty}^{+\infty} x f(x)\, dx$$

2.1.2 Variância

A variância de uma população obtêm-se calculando:

$$var(x) = E(X^2) - E^2(X)$$

2.1.3 Desvio padrão

O desvio padrão de uma população obtêm-se calculando:

$$\sigma_X = \sqrt{var(x)}$$

2.2 Estatísticas de dispersão (Amostra)

2.2.1 Variância

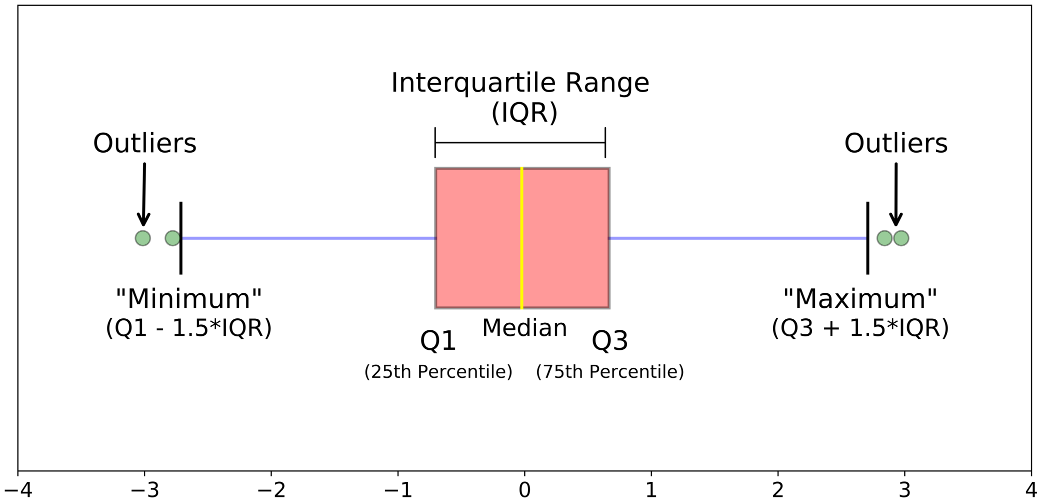
A variância de uma amostra é calculada da forma:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2\right)$$

Se tivermos uma amostra de uma população X com valor médio μ e variãncia σ^2,

$$E(\bar{X}) = \mu \qquad var(\bar{X}) = \frac{\sigma^2}{n}$$

2.3 Box plot



Em que Minimum representa a barreira inferior, $BI = Q_1 - 1.5 \times (Q_3 - Q_1)$ e Maximum representa a barreira superior, $BS = Q_3 + 1.5 \times (Q_3 - Q_1)$

2.4 Teorema do limite central (T.L.C.)

Este teorema afirma que quando o tamanho da amostra aumenta, a distribuição amostral da sua média aproxima-se cada vez mais de uma distribuição normal. O TLC permite afirmar que, para n suficientemente grande, se verifica:

$$P\left(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq z\right) \approx \Phi(z)$$

Isto é, a distribuição (desconhecida) de $\bar{X}$ pode ser aproximada pela distribuição $N(\mu, \sigma/\sqrt{n})$. Analogamente:

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) \approx \Phi(z)$$

A aproximação é considerada razoável para n $>$ 30.

2.4.1 TLC em distribuição Binomial

Se $X \cap B(n,p)$, a v.a. X pode ser considerada como sendo a soma de n variáveis aleatórias i.i.d. com distribuição $B(1,p)$. Para n suficientemente grande tem-se:

$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \leq z\right) \approx \Phi(z)$$

i.e., a distribuição binomial $B(n,p)$ da v.a. X pode ser aproximada pela distribuição normal $N(np, \sqrt{np(1-p)})$. A aproximação é considerada satisfatória quando $np > 10$ e $n(1-p) > 10$.

2.4.2 TLC em distribuição de Poisson

Se $X \cap P(\lambda)$, a v.a. X pode ser considerada como sendo a soma de n variáveis aleatórias i.i.d. com distribuição $P(\lambda/n)$. Para λ suficientemente grande tem-se:

$$P\left(\frac{X - \lambda}{\sqrt{\lambda}} \leq z\right) \approx \Phi(z)$$

i.e., a distribuição de Poisson $P(\lambda)$ da v.a. X pode ser aproximada pela distribuição normal $N(\lambda, \sqrt{\lambda})$. A aproximação considera-se satisfatória para $\lambda > 20$.

2.5 Testes de Hipóteses e Intervalos de confiança

z A forma z_x representa a tabela da normal (parte final) com o valor x

t A forma $t_{x;y}$ representa a tabela da t de student, linha x coluna y

p̂ Recolhida uma amostra de dimensão n da população, seja X o número de elementos da amostra que pertencem a uma dada categoria. p̂ representa a proporção de elementos de uma amostra que pertencem a essa categoria.

$$\hat{p} = \frac{X}{n}$$

m/n/M/N Número de elementos da amostra / população

S/σ Desvio padrão da amostra / população

S^2/σ^2 Variância da amostra / população

$\bar{X}/\mu$ Média da amostra / população