Variável aleatória discreta

Gilberto Pereira Sassi

Universidade Federal da Bahia
Instituto de Matemática e Estatística
Departamento de Estatística

Da amostra para a população: variável quantitativa discreta

Objetivo

Atribuir probabilidades para valores de uma variável quantitativa discreta usando a teoria de probabilidades.

Seja X uma variável quantitiva discreta em amostra de tamanho n com tabela de distribuição dada por

Tabela 1: Tabela de distribuição de frequência para uma variável quantitativa discreta

X	frequência	frequência relativa	porcentagem
x_1	n_1	$f_1 = {}^{n_1}/_{n}$	$100 \cdot f_1\%$
x_2	n_2	$f_2 = {}^{n_2}/_{n}$	$100 \cdot f_2\%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = {}^{n_k}/_{n}$	$100 \cdot f_k\%$
Total	n	1	100

As afirmações usando $f_1, \ldots, f_k$ são válidas apenas para a amostra e, no máximo, são aproximações para a população. Então, a ideia é substituir a frequência relativa f_i pela probabilidade $f(x_i)$ de X ser igual a x_i na população.

Variável aleatória discreta e função de probabilidade

Definição

- Considere um fenômeno aleatório com espaço amostral Ω e probabilidade $P(\cdot)$;
- $X : \Omega \to \mathbb{Z}$ é chamada de variável aleatória discreta;
- Suponha que os valores possível dessa variável é $\{x_1, x_2, x_3, x_4, x_5, \cdots\}$. A função dada por

$$\begin{aligned} f(x_i) &= P(X = x_i) \\ &= P\{\omega \in \Omega \mid X(\omega) = x_i\}, \end{aligned}$$

  é chamada de função de probabilidade.

Observações

- O conjunto de todos os valores possíveis de uma variável aleatória discreta X é chamada de suporte e usamos a notação $\chi = \{x_1, \ldots, x_k, \ldots\}$.
- Em situações práticas, não nos preocupamos com o espaço amostral Ω, e focamos nossa atenção em estabelecer o suporte e a função de probabilidade da variável aleatória discreta.

Propriedades da função de probabilidade

Note que

- $0 \leq f(x_i) \leq 1$;
- $f(x_1) + f(x_2) + f(x_3) + f(x_4) + \cdots = 1$.

Para caracterizar uma variável aleatória discreta, precisamos estabelecer:

- valores possíveis da variável aleatória discreta $x_1, x_2, x_3, \ldots$;
- A função de probabilidade para cada valor possível da variável aleatória discreta.

Seja $B \subset \{x_1, x_2, x_3, \ldots\}$, então

$$P(X \in B) = \sum_{x \in B} f(x)$$

Função de distribuição acumulada

Uma outra abordagem para calcular probabilidades de uma variável aleatória é usar somas acumuladas.

Função de Distribuição Acumulada

$$
\begin{aligned}
F(x) &= P(X \leq x) \\
&= P\{\omega \in \Omega \mid X(\omega) \leq x\} \\
&= f(x_1) + f(x_2) + \cdots + f(\lfloor x \rfloor)
\end{aligned}
$$

em que $\lfloor x \rfloor$ é a função “arrendonda x para baixo”.

Para especificar completamente uma variável aleatória discreta precisamos estabelecer

1. o suporte da variável aleatória discreta;
2. a função de probabilidade ou a função de distribuição acumulada.

Note que podemos derivar a função de probabilidade usando a função de distribuição acumulada, e vice-versa.

Exemplo

Considere o fenômeno aleatório que consiste no lançamento de duas moedas “justas” ou “normais”. Qual o espaço amostral? Usando o princípio da equiprobabilidade, qual seria a probabilidade de sair ao menos uma cara? Considere a variável aleatória discreta $X : \Omega \to \mathbb{R}$ em que

$$X(\omega) = \text{Número de caras em } \omega.$$

Encontre a função de probabilidade e a função de distribuição acumulada de X.

Solução: Note que o espaço amostral desse fenômeno aleatório é $\Omega = \{cc, kc, ck, kk\}$, em que c representa cara e k representa coroa. Então, usando o princípio da equiprobabilidade, temos que

ω	$P(\{\omega\})$	$X(\omega)$
cc	$1/4 = 0,25$	2
kc	$1/4 = 0,25$	1
ck	$1/4 = 0,25$	1
kk	$1/4 = 0,25$	0

Ou seja,

$$\begin{aligned} f(0) &= P(X = 0) = P(\{kk\}) = 1/4 = 0,25, \\ f(1) &= P(X = 1) = P(\{ck, kc\}) = 2/4 = 0,5, \\ f(2) &= P(X = 2) = P(\{cc\}) = 1/4 = 0,25. \end{aligned}$$

Note que o suporte da variável aleatória X é $\chi = \{0, 1, 2\}$.

Exemplo – continuação

- Para $x < 0$, temos que

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega \mid X(\omega) \leq x < 0\}) = P(\emptyset) = 0;$$

- Para $0 \leq x < 1$, temos que

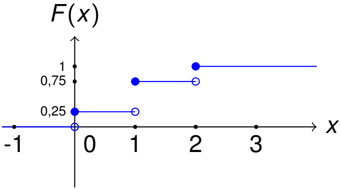
$$F(x) = P(X \leq x) = f(0) = 0,25;$$

- Para $1 \leq x < 2$, temos que

$$F(x) = P(X \leq x) = f(0) + f(1) = 0,25 + 0,5 = 0,75;$$

- Para $x \geq 2$, temos que

$$F(x) = P(X \leq x) = f(0) + f(1) + f(2) = 1.$$



Distribuição uniforme discreta

Motivação

Algumas variáveis e quantidades aparecem com frequência e a literatura estatística já estabeleceu funções de probabilidade e funções de distribuição acumulada.

Distribuição uniforme discreta

Seja X uma variável aleatória discreta assumindo valores $j, \ldots, k$. Dizemos que X segue o modelo uniforme discreto se cada um dos valores $j, \ldots, k$ tem função de probabilidade $\dfrac{1}{k-j+1}$. Ou seja, a função de probabilidade de X é dada por

$$f(i) = \frac{1}{k-j+1}, \quad i = j, \ldots, k.$$

Notação: $X \sim U_D[j, k]$.

Exemplo

Uma rifa tem 100 bilhetes numerados de 1 a 100. Tenho 5 bilhetes consecutivos numerados de 21 a 25 e meu colega tem outros 5 bilhetes com os números 1, 11, 29, 68 e 93. Quem tem mais chance de ganhar?

Solução: Seja X a variável aleatória discreta que é um número sorteado. Então, $X \sim U_D[1,100]$, e temos as seguintes probabilidades

- Probabilidade de ter comprado um bilhete premiado:

$$\begin{aligned} P(X \in \{21,22,23,24,25\}) &= f(21)+f(22)+f(23)+f(24)+f(25) \\ &= \frac{1}{100}+\frac{1}{100}+\frac{1}{100}+\frac{1}{100}+\frac{1}{100} = \frac{5}{100} \\ &= \frac{1}{20} = 0,05. \end{aligned}$$

- Probabilidade do meu amigo ter comprado um bilhete premiado:

$$\begin{aligned} P(X \in \{1,11,29,69,93\}) &= f(1)+f(11)+f(29)+f(69)+f(93) \\ &= \frac{1}{100}+\frac{1}{100}+\frac{1}{100}+\frac{1}{100}+\frac{1}{100} = \frac{5}{100} \\ &= \frac{1}{20} = 0,05. \end{aligned}$$

Distribuição Bernoulli

Ensaios de Bernoulli são fenômenos aleatórios com 2 resultados possíveis, chamados de sucesso e fracasso. A variável X que atribui 1 ao sucesso e zero ao fracasso é chamado de distribuição Bernoulli. Mais precisamente, seja p a probabilidade de sucesso, então a função de probabilidade de X é dada por

$$f(1) = p;$$
$$f(0) = 1 - p.$$

Notação: $X \sim Bernoulli(p)$.

Exemplo

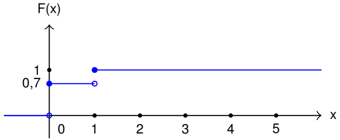
Assuma que a prevalência de infecção pelo vírus HIV em país da África Subsariana seja 30%. Considere o fenômeno aleatório que consiste de prever se um novo paciente está infectado. Qual o modelo probabilístico adequado neste contexto? Qual a função de probabilidade? Qual a função de distribuição acumulada?

Solução: Considere sucesso o paciente estar infectado com o vírus HIV. Então, temos um ensaio de Bernoulli com probabilidade de sucesso $0,3$, e a variável aleatória discreta associada é $X \sim Bernoulli(0,3)$.

O suporte para X é $\chi = \{0,1\}$, e a função de probabilidade é $f(0) = 1 - 0,3 = 0,7$ e $f(1) = 0,3$.

A função de distribuição acumulada para

- $x < 0$ é $F(x) = P(X \leq x < 0) = 0$
- $0 \leq x < 1$ é $F(x) = P(X \leq x) = f(0) = 0,7$;
- $x \geq 1$ é $F(x) = 1$.



Distribuição binomial

Considere o fenômeno aleatório que consiste da repetição de n ensaios de Bernoulli independentes e todos com a mesma probabilidade de sucesso p. A variável aleatória que conta o número total de sucessos é denominada de distribuição binomial com parâmetros n e p e sua função de probabilidade é dada por

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n,$$

em que $\binom{n}{k}$ é chamado de coeficiente binomial e é dado por

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

em que $n! = n \cdot (n-1) \cdot (n-2) \ldots 1$.
Notação: $X \sim b(n, p)$.

Exemplo

Sabe-se que a eficiência de uma vacina é de 80%. Um grupo de três indivíduos é sorteado dentre a população vacinada e submetido a testes para averiguar se a imunização foi efetivada. Qual o distribuição de probabilidade adequada para este caso? Encontre a função de probabilidade e a função de distribuição acumulada.

Solução: Considere sucesso a imunização do indivíduo, então temos três repetições de um ensaio de Bernoulli com probabilidade de sucesso $0,8$. Ou seja, a variável X, número de indivíduos imunizados, tem distribuição binomial com parâmetros $n = 3$ e $p = 0,8$.

O suporte de X é $\chi = \{0, 1, 2, 3\}$ e a função de probabilidade é dada por

$$f(0) = \binom{3}{0} 0,8^0 (1 - 0,8)^3 = 0,08$$

$$f(1) = \binom{3}{1} 0,8^1 (1 - 0,8)^2 = 0,096$$

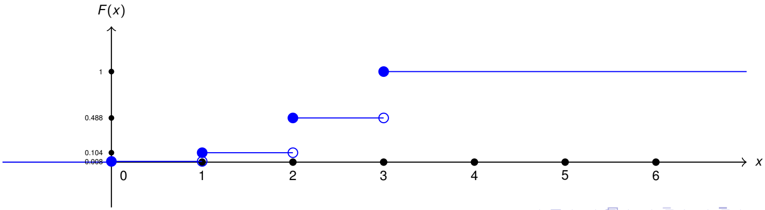
$$f(2) = \binom{3}{2} 0,8^2 (0 - 0,8)^1 = 0,384$$

$$f(3) = \binom{3}{3} 0,8^3 (0 - 0,8)^0 = 0,512$$

Exemplo – continuação

Para encontrar a função de distribuição acumulada, precisamos dividir em casos:

- Para $x < 0$, temos que
$$F(x) = P(X \leq x) = P\left(\{\omega \in \Omega \mid X(\omega) \leq x < 0\}\right) = P(\emptyset) = 0;$$
- Para $0 \leq x < 1$, temos que
$$F(x) = P(X \leq x) = f(0) = 0,08;$$
- Para $1 \leq x < 2$, temos que
$$F(x) = P(X \leq x) = f(0) + f(1) = 0,08 + 0,096 = 0,104;$$
- Para $2 \leq x < 3$, temos que
$$F(x) = P(X \leq x) = f(0) + f(1) + f(2) = 0,08 + 0,096 + 0,384 = 0,488;$$
- Para $x \geq 3$, temos que
$$F(x) = P(X \leq x) = f(0) + f(1) + f(2) + f(3) = 0,08 + 0,096 + 0,384 + 0,512 = 1;$$



Distribuição Poisson

Distribuição de probabilidade utilizada em fenômenos aleatórios que consistem contar o número de ocorrências de um evento em um intervalo de tempo. Neste modelo, λ é a frequência média ou esperada de ocorrências do evento no intervalo de tempo. A variável aleatória discreta X, número de ocorrências no intervalo de tempo, tem distribuição de Poison com parâmetro $\lambda > 0$, e sua função de probabilidade é dada por

$$f(k) = \frac{e^{-\lambda}\lambda^{k}}{k!}, \quad k = 0, 1, 2, 3, \ldots$$

Notação: $X \sim Poison(\lambda)$.

Exemplo

Suponha que uma unidade básica de saúde de um bairro de classe média realiza em média 10 atendimentos em dias de segunda-feira. Qual a probabilidade desta UBS atender, na próxima segunda-feira, no máximo 5 cidadãos?

Solução: Note que estamos contando o número de atendimentos (ocorrência=atendimento) em um dia de semana (intervalo de tempo = segunda-feira). Então, a variável aleatória discreta X, número de atendimentos em segunda-feira, tem distribuição Poison com média $\lambda = 10$. Então,

$$\begin{aligned} P(X \leq 5) &= f(0) + f(1) + f(2) + f(3) + f(4) + f(5) \\ &= \frac{e^{-10}10^0}{0!} + \frac{e^{-10}10^1}{1!} + \frac{e^{-10}10^2}{2!} + \frac{e^{-10}10^3}{3!} + \frac{e^{-10}10^4}{4!} + \frac{e^{-10}10^5}{5!} \\ &= 0,07 \end{aligned}$$

Definição

Seja X uma variável aleatória discreta com suporte $\chi = \{x_1, x_2, x_3, \ldots\}$ e com função de probabilidade $f(x_1), f(x_2), f(x_3), \ldots$ Então,

- A média ou valor esperado ou esperança matemática de X é definida por

$$\begin{aligned} \mathrm{E}(X) &= x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + x_3 \cdot f(x_3) + \cdots \\ &= \mu; \end{aligned}$$

- A variância de X é definida por

$$\begin{aligned} \mathrm{Var}(x) &= (x_1 - \mu)^2 \cdot f(x_1) + (x_2 - \mu)^2 \cdot f(x_2) + (x_3 - \mu)^3 \cdot f(x_3) + \cdots \\ &= \sigma^2; \end{aligned}$$

- Para manter a mesma unidade da variável aleatória discreta, usamos o desvio padrão

$$\sigma = \sqrt{\mathrm{Var}(x)} = \sqrt{\sigma^2}$$

- A mediana de X é um valor Md tal que

$$P(X \geq Md) \geq 0,5 \text{ e } P(X \leq Md) \geq 0,5;$$

- A moda de X é valor x_i com maior valor de $f(x_i)$.

Exemplo

Uma pequena cirurgia dentária pode ser realizada por dois métodos diferentes cujos tempos de recuperação (em dias) são modeladas pelas variáveis aleatórias discretas X_1 e X_2. Admita que as funções de probabilidade são dadas por

Funções de probabilidade.

x	0	4	5	6	10
$f(x)$	0,2	0,2	0,2	0,2	0,2

Tabela 2: Método 1.

x	1	2	3	4	5
$f(x)$	0,4	0,15	0,15	0,15	0,15

Tabela 3: Método 2.

Calcule a média, a variância, a mediana e a moda para cada uma das variáveis. Qual método você recomendaria para um paciente que precis fazer esta cirurgia dentária?

Exemplo – solução

Método 1

- **Média** $\mu = 0 \cdot 0,2 + 4 \cdot 0,2 + 5 \cdot 0,2 + 6 \cdot 0,2 + 10 \cdot 0,2 = 5$
- **Mediana** Note que $P(md \leq 5) = 0,6 \geq 0,5$ e $P(md \geq 5) = 0,6 \geq 0,5$
- **Moda** $Mo = \{0, 4, 5, 6, 10\}$
- **Variância** $\sigma^2 = (0-5)^2 f(0) + (4-5)^2 f(4) + (5-5)^2 f(5) + (6-5)^2 f(6) + (10-5)^2 f(10) = 10,4$
- **Desvio Padrão** $\sigma = \sqrt{10,4} = 3,22$

Método 2

- **Média** $\mu = 1 \cdot 0,4 + 2 \cdot 0,15 + 3 \cdot 0,15 + 4 \cdot 0,15 + 5 \cdot 0,15 = 2,5$
- **Mediana** Note que $P(md \leq 2) = 0,55 \geq 0,5$ e $P(md \geq 2) = 0,6 \geq 0,5$
- **Moda** $Mo = 0$
- **Variância** $\sigma^2 = (1-2,5)^2 f(1) + (2-2,5)^2 f(2) + (3-2,5)^2 f(3) + (4-2,5)^2 f(4) + (5-2,5)^2 f(5) = 2,25$
- **Desvio Padrão** $\sigma = \sqrt{2,25} = 1,5$

Note que a média, a moda ou a mediana é menor para o método 2. Além disso, a variância e o desvio padrão para o segundo método também é menor, ou seja, a incerteza de quantos dias o paciente estará recuperado é menor para o método 2. Logo, deveríamos indicar o segundo método para o paciente.