5 Caracterização Estatística de Variáveis

ESQUEMA DO CAPÍTULO

5.1 Introdução

- Denomina-se ***variável aleatória*** aquela em que é impossível determinar o seu valor a priori;
- Para estudar uma variável aleatória ***identificam-se:***
  - *valores* que ela pode assumir;
  - a distribuição de *probabilidade*;
- Será apresentada uma forma de caracterização de variáveis aleatórias do mesmo tipo em classes denominadas ***modelos***;
- Serão estudados os seguinte ***modelos***:
  - Modelo de Poisson, para variáveis *discretas*;
  - Modelo de Gauss, para variáveis *contínuas*.

5.2 Conceitos fundamentais

Exemplo 5.1: Número de meninos

Seja X o número de ***meninos*** em uma família com duas crianças. Os valores possíveis dessa variável aleatória são 0, 1 e 2. A probabilidade das crianças serem do mesmo sexo é ¼ e indicamos Pr[X=0] = ¼, para duas meninas e Pr[X=2] = ¼, para dois meninos. A probabilidade de apenas um menino é 2/4 e indicaremos pro Pr[x=1]=2/4=1/2. A distribuição de probabilidade é apresentada na Tabela 5.1.

Tabela 5.1: Distribuição de probabilidade do número de meninos em uma família com duas crinças.

Número de meninos (x)	$Pr(X = x)$
0	1/4
1	2/4
2	1/4
Total	1

5.2 Conceitos fundamentais (cont.)

Exemplo 5.2: Lançamento de um dado

Seja Y o ***número da face*** voltada para cima, em um lançamento de um dado equilibrado. Os valores possíveis para Y são 1, 2, 3, 4, 5 e 6. A distribuição de probabilidade é apresentada na Tabela 5.2.

Tabela 5.2: Distribuição de probabilidade dos resultados do lançamento de um dado.

Resultado (y)	$Pr(Y = y)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

5.2 Conceitos fundamentais (cont.)

Exemplo 5.4: Pressão arterial

A Figura 5.1 mostra a distribuição da ***pressão diastólica*** de 158.906 pessoas entre 30 e 69 anos de 14 comunidades nos Estados Unidos (*Circulation Research*, 1977). Para pessoas, mesmo pertencendo a um ***grupo homogêneo***, são observados valores diferentes de pressão arterial, daí ser considerada uma variável aleatória.

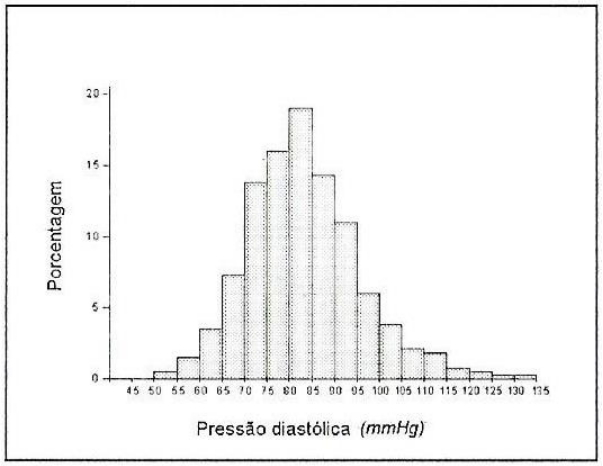

Figura 5.1: Distribuição da pressão diastólica (*mmHg*)

5.2 Conceitos fundamentais (cont.)

População e amostra

- Uma variável aleatória e sua distribuição é denominada ***população*** (no sentido estatístico, não no sentido de um conjunto de pessoas);
- Um conjunto de observações extraídas da população é a ***amostra***, obtida através do processo de amostragem;
- Na prática, dispomos para estudo apenas de uma amostra, a partir da qual fazemos ***inferência*** para a população;
- Baseado em amostras, estimamos ***parâmetros*** de interesse, *p.e.*:
  - Média;
  - Variância;
  - Proporção.

5.3 Modelo de Poisson

- O modelo de Poisson é usado para descrever variáveis aleatórias provenientes de ***contagens***;
- Suponha que X represente o ***número de ocorrências*** de um evento em um período de tempo;
- Tais variáveis seguem o modelo cuja ***distribuição de probabilidade*** é dada por:

$$Pr(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,...$$

em que $e \approx 2{,}718$, $x! = 1 \times 2 \times ... \times x$ e $\lambda > 0$ é a taxa.

5.3 Modelo de Poisson (cont.)

- A Figura 5.4 ilustra a ***distribuição de Poisson*** para alguns valores da taxa λ.

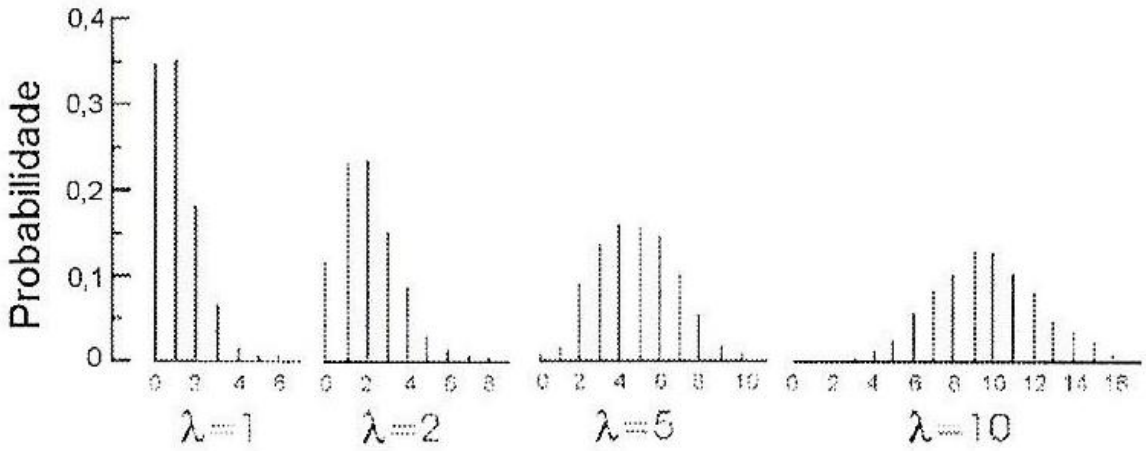

Figura 5.1: Distribuição de Poisson para valores selecionados de λ

5.3 Modelo de Poisson (cont.)

Exemplo 5.6: Chegadas de pacientes em um pronto socorro
Suponha que o número de pacientes que chegam a um ponto socorro de uma pequena cidade durante a madrugada tenha distribuição de Poisson com média 3 por madrugada (λ=3). Para obtenção da distribuição de probabilidade apresentada na Tabela 5.3, a probabilidade de chegadas de *x* pacientes é calculada como:

$$Pr(X = x) = \frac{e^{-3}3^{x}}{x!}, x = 0,1,...$$

Tabela 5.3: Distribuição de Poisson com parâmetro λ = 3

x	$Pr(X = x)$	*x*	$Pr(X = x)$
0	0,0498	7	0,0216
1	0,1494	8	0,0081
2	0,2240	9	0,0027
3	0,2240	10	0,0008
4	0,1680	11	0,0002
5	0,1008	12	0,0001
6	0,0504	≥13	≈ 0

5.3 Modelo de Poisson (cont.)

Exemplo 5.6: Chegadas de pacientes em um pronto socorro (cont.)

Pelo modelo de Poisson pode-se concluir que:

1. É pouco provável que nenhum paciente chegue em determinada madrugada, pois $Pr[X = 0] = 0,0498$;
2. É muito provável que pelo menos um paciente chegue em determinada madrugada, pois $Pr[X \geq 1] = 1 - 0,0498 = 0,9502$;
3. É improvável que cheguem 13 ou mais pacientes durante a madrugada, uma vez que $Pr[X \geq 13] \approx 0$;
4. A maior concentração de chegadas está em torno de $x = 3$, que é o número médio de chegadas;
5. Em um mês, esperam-se $\lambda = 30 \times 3 = 90$ pacientes.

5.3 Modelo de Poisson (cont.)

Exemplo 5.7: Número de consultas médicas

Em um plano de saúde com 5.694 filiados, ao fim de um ano fizeram-se 13.098 consultas, segundo a distribuição mostrada na Tabela 5.4.

Tabela 5.4: Distribuição de consultas anuais de um plano de saúde

Nº de consultas	Frequência	Nº de consultas	Frequência
0	589	5	304
1	1274	6	126
2	1542	7	39
3	1144	8	10
4	663	9	3

A estimativa do parâmetro λ é:

$$\lambda = \frac{\text{nº de consultas}}{\text{nº total de associados}} = \frac{0\times 589 + 1\times 1274 + 2\times 1542 + \cdots + 9\times 3}{598 + 1274 + 1542 + \cdots + 3} = \frac{13.098}{5.694} \cong 2,3$$

5.4 Modelo de Gauss

Curva de Gauss

- Caracterizada por dois parâmetros, μ (média) e σ (desvio-padrão);
- Se a variável aleatória X tem distribuição gaussiana, é representada por $X \sim N(\mu,\sigma)$;
- Definida matematicamente pela equação:

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

- Representada graficamente na Figura 5.5.

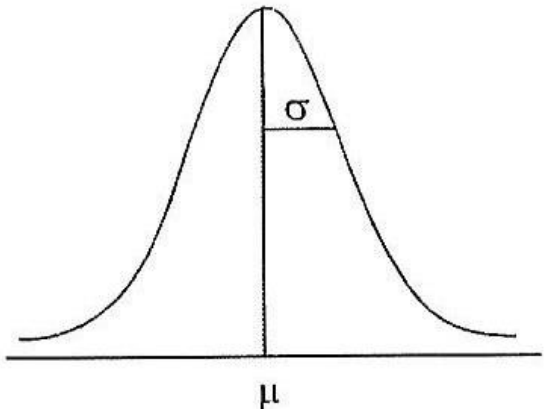

Figura 5.5: Curva de Gauss

5.4 Modelo de Gauss (cont.)

Curva de Gauss padrão

- Distribuição gaussiana com média $\mu = 0$ e desvio-padrão $\sigma = 1$;
- Se a variável aleatória Z tem distribuição gaussiana padrão, é representada por $Z \sim N(0,1)$;
- Seu gráfico está na Figura 5.6.

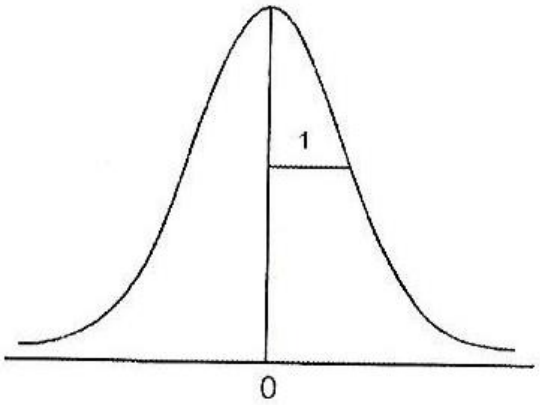

Figura 5.5: Curva de Gauss padrão

5.4 Modelo de Gauss (cont.)

Curva de Gauss padrão (cont.)

- Fundamental conhecer e saber usar a tabela da curva gaussiana padrão, reproduzida na Tabela A3 do Apêndice.

Tabela A3: Distribuição gaussiana para valores positivos: $Pr(X \leq x)$ (continuação)

x	Segunda casa decimal de x									
	0	1	2	3	4	5	6	7	8	9
0,0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0,1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0,2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0,3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0,4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879

- Esta tabela dá, para cada numero x entre -3 e +3 a área abaixo da curva gaussiana padrão no intervalo $(-\infty, z]$, ou seja fornece o valor da probabilidade $\Pr[Z < z]$.

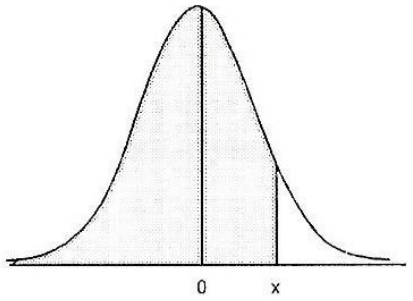

Figura 5.7: Área fornecida pela tabela da curva de Gauss padrão

5.4 Modelo de Gauss (cont.)

Exemplo 5.8: Cálculo de probabilidades na gaussiana padrão

1. $Pr[Z \leq 1]$ e $Pr[Z \leq -1]$
   Os valores são obtidos diretamente da tabela, ou seja, $Pr[Z \leq 1] = 0,8413$ e $Pr[Z \leq -1] = 0,1586$;
2. $Pr[-1 \leq Z \leq 1]$
   Pela expressão $Pr[-1 \leq Z \leq 1] = Pr[Z \leq 1] - Pr[Z \leq -1] = 0,8413 - 0,1586 = 0,6827$; em outras palavras, o intervalo centrado na média, com amplitude de 2 desvios-padrão, engloba 68,27% da distribuição;
3. $Pr[1 \leq Z \leq 2]$
   Similarmente, pela expressão $Pr[1 \leq Z \leq 2] = Pr[Z \leq 2] - Pr[Z \leq 1] = 0,9773 - 0,8413 = 0,1360$;
4. $Pr[Z \geq 2,33]$
   Para este cálculo, usa-se o fato de que a curva gaussiana é simétrica, portanto $Pr[Z \geq 2,33] = Pr[Z \leq -2,33] = 0,0099$;

5.4 Modelo de Gauss (cont.)

Exemplo 5.8: Cálculo de probabilidades na gaussiana padrão (cont.)
Para calcular áreas sob a curva de Gauss com parâmetros μ e σ, basta ***reduzir*** o problema ao uso da curva padrão, através do resultado:

$$X \sim N(\mu,\sigma) \Rightarrow Z = \frac{\text{X-}\mu}{\sigma} \sim N(0,1)$$

Por exemplo:

$$\Pr[X \le a] = \Pr\left[\frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma}\right] = \Pr\left[Z \le \frac{a-\mu}{\sigma}\right]$$

5.4 Modelo de Gauss (cont.)

Exemplo 5.8: Cálculo de probabilidades na gaussiana padrão (cont.)
Frequentemente, o interesse é nos intervalos cujos limites são média ± 1 desvio-padrão, média ± 2 desvios-padrão e média ± 3 desvios-padrão:

$$\Pr[\mu - \sigma \leq X \leq \mu + \sigma] = \Pr[-1 \leq Z \leq +1] = 0{,}6826$$
$$\Pr[\mu - 2\sigma \leq X \leq \mu + 2\sigma] = \Pr[-2 \leq Z \leq +2] = 0{,}9544$$
$$\Pr[\mu - 3\sigma \leq X \leq \mu + 3\sigma] = \Pr[-3 \leq Z \leq +3] = 0{,}9973$$

Em outras palavras, em uma distribuição gaussiana, cerca de 68% da população está entre a média e o desvio-padrão; aproximadamente 95%, entre a média e dois desvio-padrão; e praticamente toda a população (99,7%), entre a média e 3 desvios-padrão.

5.4 Modelo de Gauss (cont.)

Exemplo 5.9: Pressão sistólica em jovens saudáveis

Suponha que a pressão arterial sistólica em pessoas jovens gozando de boa saúde tenha distribuição gaussiana $X \sim N(\mu = 120, \sigma = 10)$;

1. Qual é a probabilidade de se encontrar uma pessoa com pressão sistólica acima de 140 *mmHg*?

$$\Pr[X \geq 140] = \Pr\left[\frac{X - 120}{10} \geq \frac{140 - 120}{10}\right] = \Pr[Z \geq 2]$$

$$= 1 - \Pr[X \leq 2] = 1 - 0,9772 = 0,0228$$

Ou seja, apenas 2,28% das pessoas jovens e sadias têm pressão sistólica acima de 140 *mmHg*.

5.4 Modelo de Gauss (cont.)

Exemplo 5.9: Pressão sistólica em jovens saudáveis (cont.)

2. Quais são os limites de um ***intervalo simétrico*** em relação à média que engloba 95% dos valores das pressões sistólicas de pessoas jovens e sadias?

$$\Pr\left[a \le X \le b\right] = \Pr\left[\frac{a-120}{10} \le \frac{X-120}{10} \le \frac{b-120}{10}\right] = 0,95$$

Ou seja, Pr[a′ ≤ Z ≤ b′] = 0,95. Escolhendo-se uma solução simétrica, −a′ = b′, tem-se que Pr[Z ≤ b′] = 0,950+0,025 = 0,975. Da tabela da gaussiana, temos que b′ ≈ 1,96. Portanto, $a = 120 - 1,96 \times 10 = 100,4$ e $b = 120 + 1,96 \times 10 = 139,6$.

5.5 Verificação da adequação do modelo

- Após a escolha de um modelo para descrever uma situação, é necessário verificar a sua ***adequação*** aos dados;
- Um critério preciso para verificar se há uma boa aderência de um conjunto de dados ao modelo gaussiano é através do ***gráfico*** Q-Q *plot*, disponível em *softwares* estatísticos.

5.5 Verificação da adequação do modelo (cont.)

Exemplo 5.12: Pressão sistólica de estudantes

Os alunos de uma turma de 60 estudantes do sexo masculino mediram a pressão sistólica (em *mmHg*), uns dos outros e obtiveram os resultados da Tabela 5.7. Confirma-se a ***adequação*** do modelo gaussiano, pelas Figuras 5.10 e 5.11.

Tabela 5.7: Pressão sistólica (*mmHg*) de 60 estudantes

142	142	134	110	98	130
136	120	118	130	116	140
118	122	128	128	114	138
104	116	110	100	128	128
124	140	108	146	116	114
152	118	140	128	116	110
138	132	118	120	122	120
108	112	94	130	130	118
120	128	108	120	124	110
124	132	132	130	102	118

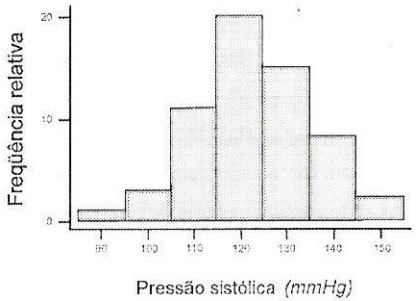

Figura 5.10: Histograma para a pressão sistólica (*mmHg*)

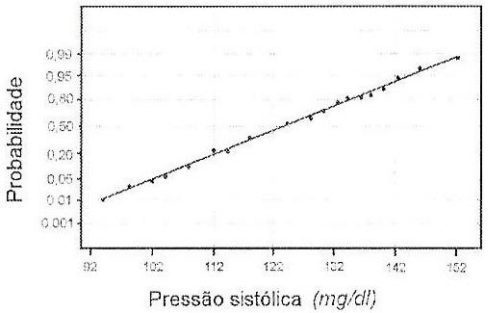

Figura 5.11: Q-Q *plot* para a pressão sistólica (*mmHg*)

5.6 Faixa de referência

A hemoglobina é um parâmetro hematológico de uso rotineiro na área da saúde. Tomando o histograma da Figura 5.13 como aproximação para a sua distribuição entre mulheres, vemos que valores menores que 12 *g*/*dl* são pouco comuns. É, pois, razoável considerar um diagnóstico de anemia ao se observar um paciente com valor abaixo de 11 *g*/*dl*.

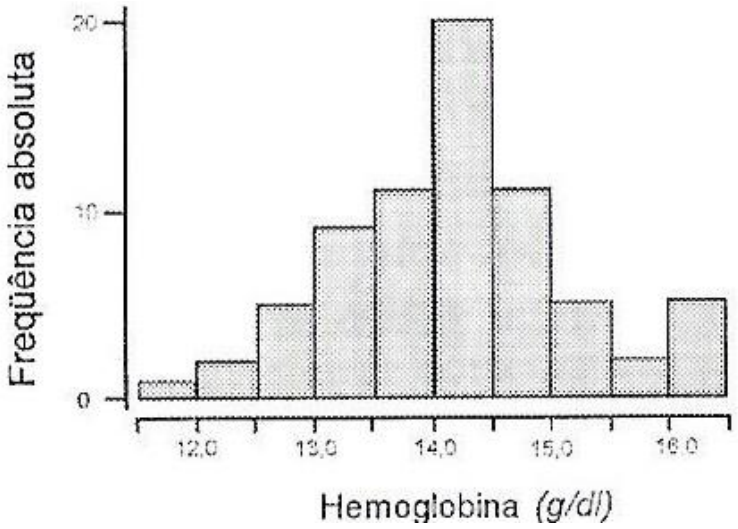

Figura 5.13: Histograma da hemoglobina (*mg*/*dl*)

5.6 Faixa de referência (cont.)

Obtenção da faixa de referência – Método da curva de Gauss

Na Seção 5.4 foi mostrado como se obter valores simétricos em relação à média μ, tais que a área sob a curva de Gauss, delimitada por tal intervalo, seja um valor pré-fixado. Em particular, sabemos que o intervalo (μ-2σ, μ+2σ) engloba 95,6% da área sob a curva gaussiana, e assim por diante.

As faixas de referência para as coberturas mais comuns são:

Cobertura	Faixa de referência
90%	$(\bar{x} - 1{,}64s;\ \bar{x} + 1{,}64s)$
95%	$(\bar{x} - 1{,}96s;\ \bar{x} + 1{,}96s)$
99%	$(\bar{x} - 2{,}58s;\ \bar{x} + 2{,}58s)$

5.6 Faixa de referência (cont.)

Exemplo 5.14: Teor de gordura fecal

Utilizando-se os dados da Tabela 3.1 do Cap. 3, referentes ao teor de gordura fecal de 43 crianças que não eram amamentadas pela mãe, calcula-se a média e o desvio-padrão ($\bar{x} = 2{,}303$, $s = 0{,}872$) e obtém-se os seguintes valores de referência, para uma cobertura aproximada de 95%:

Limite inferior de referência: $2{,}303 - 2(0{,}872) = 0{,}559$

Limite superior de referência: $2{,}303 + 2(0{,}872) = 4{,}047$