Capítulo 2

Medidas descritivas para distribuições de freqüências

Neste capítulo, apresentamos medidas que descrevem aspectos importantes das distribuições de freqüências.

Uma classe de medidas descreve a posição ou localização das distribuições de freqüências, em particular medidas de centralidade, e aparecem na primeira seção.

Outro aspecto importante, visto na seção seguinte, é a largura das distribuições de freqüências.

Nas duas últimas seções abordamos a dependência e associação linear entre variáveis.

2.1 Medidas de posição

Apresentamos a seguir algumas medidas de posição de distribuições de freqüências. Estas medidas indicam onde na escala de valores se distribuem os valores da distribuição de freqüências, levando em conta os pesos (ou massas) das freqüências de cada valor de alguma forma.

Começaremos com a *média*, que é o centro de massa da distribuição. Em seguida, será a vez da *mediana*, também uma medida de centralidade. Depois, falaremos dos *quantis*, medidas de posição não centrais.

2.1.1 Média

Dada uma variável quantitativa discreta X, isto é, tal que o conjunto de valores possíveis de X é um conjunto discreto de números

$$\mathcal{V} = \{x_1, x_2, \ldots\}, \tag{2.1}$$

e com função de freqüência

$$P(X = x_i);\ i = 1, 2, \ldots, \tag{2.2}$$

a *média* de X é definida por

$$M(X) = \sum_{i=1,2,\ldots} x_i\, P(X = x_i). \tag{2.3}$$

Isto é, $M(X)$ é a média dos valores de X ponderada pelas respectivas freqüências.

Uma outra notação para $M(X)$ é $\bar{X}$.

Em termos da população Π, podemos obter $M(X)$ como a média aritmética dos valores de X sobre os indivíduos de Π, isto é,

$$M(X) = \frac{1}{N}\sum_{I\in\Pi} X(I). \tag{2.4}$$

De fato,

$$\begin{aligned}
\frac{1}{N}\sum_{I\in\Pi} X(I) &= \frac{1}{N}\sum_{i=1,2,\ldots}\ \sum_{\substack{I\in\Pi:\\ X(I)=x_i}} X(I) = \sum_{i=1,2,\ldots}\frac{1}{N}\sum_{\substack{I\in\Pi:\\ X(I)=x_i}} x_i \\
&= \sum_{i=1,2,\ldots} x_i \frac{1}{N}\sum_{\substack{I\in\Pi:\\ X(I)=x_i}} 1 = \sum_{i=1,2,\ldots} x_i \frac{N_i}{N} \\
&= \sum_{i=1,2,\ldots} x_i\, P(X = x_i) = M(X), \qquad (2.5)
\end{aligned}$$

onde $\sum_{I\in\Pi:X(I)=x_i} 1 = N_i$ é o número de indivíduos I na população Π com X igual a x_i.

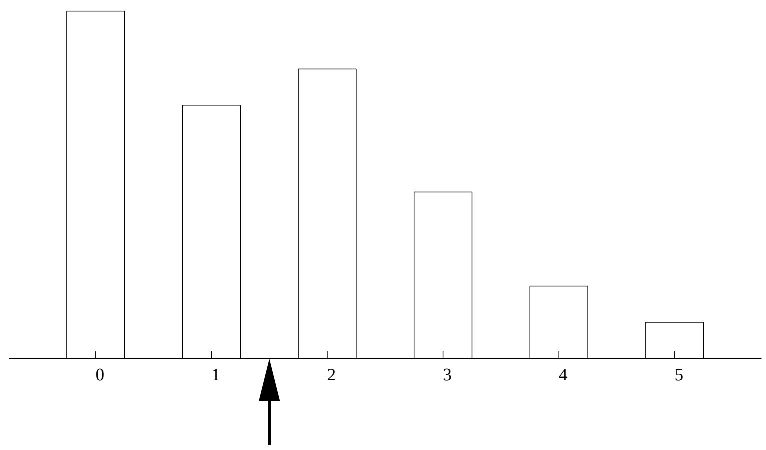

Figura 2.1 Representação gráfica de W no Exemplo 1.1. Barras podem ser vistas como pesos com massas proporcionais às respectivas alturas; a seta, fiel da balança equilibrada, indica a posição da média.

Exemplo 2.1 *Vamos calcular as médias de idade e de número de filhos no Exemplo 1.1. Da Tabela 1.3 e (2.3)*

$$\begin{aligned} M(W) &= 0 \times 0.30 + 1 \times 0.22 + 2 \times 0.25 + 3 \times 0.14 + 4 \times 0.06 + 5 \times 0.03 \\ &= 0 + 0.22 + 0.50 + 0.42 + 0.24 + 0.15 = 1.53. \end{aligned} \tag{2.6}$$

Para achar $M(V)$, vamos calcular a média aritmética sobre a população dada por (2.4) (veja a Figura (1.1).

$$M(V) = \frac{1}{36}(26 + 32 + \ldots + 42) = \frac{1245}{36} = 34.58 \tag{2.7}$$

Se pensarmos na distribuição de freqüências de dada variável quantitativa X como uma distribuição de pesos ou massas (as freqüências) na escala numérica, então a média de X é o centro de massa da distribuição, isto é, é o ponto de equilíbrio da distribuição de pesos; em outras palavras, é o ponto em que a distribuição de pesos se equilibra. Veja Figura 2.1.

Média para variáveis contínuas

No caso de X ser contínua, isto é, quando a distribuição de freqüências de X for dada por uma função de densidade de freqüência f_X, então a média é dada por

$$M(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx. \tag{2.8}$$

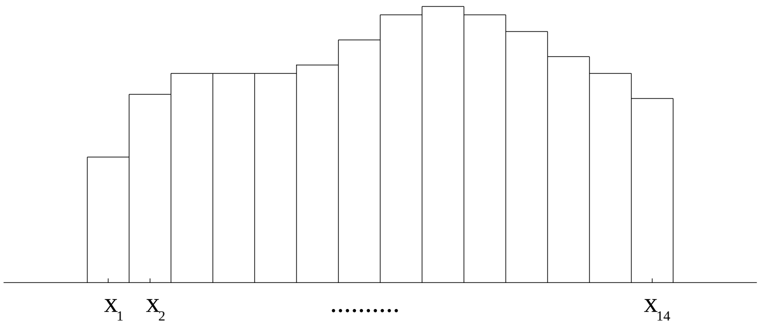

Figura 2.2

Para entender a fórmula acima, vamos considerar um histograma por classes aproximando o histograma pontual dado por f_X. Vamos ilustrar a situação com o gráfico da Figura 1.6, em que o histograma por classes aproxima o histograma pontual dado pela curva.

Temos ali 14 intervalos ou classes em que a escala é dividida. Vamos denotá-las $\mathcal{C}_1, \mathcal{C}_2, \ldots$. Seja x_i o ponto médio, e Δx_i o comprimento de $\mathcal{C}_i$, $i = 1, 2, \ldots$. Veja a Figura 2.2. Então é razoável, seguindo a idéia do caso discreto, definir $M(X)$ de forma *aproximada* como

$$M(X) \simeq \sum_{i=1,2,\ldots} x_i \, P(X \in \mathcal{C}_i). \tag{2.9}$$

Mas $P(X \in \mathcal{C}_i)$, a freqüência com que X assume valores em $\mathcal{C}_i$, que é a área da i-ésima barra do histograma por classes, correspondente a $\mathcal{C}_i$, é aproximadamente igual a

$$P(X \in \mathcal{C}_i) = h_i \, b_i \simeq f_X(x_i) \, \Delta x_i, \tag{2.10}$$

onde h_i e b_i são respectivamente a altura e a base da i-ésima barra do histograma por classes. (2.9) fica então

$$M(X) \simeq \sum_{i=1,2,\ldots} x_i \, f_X(x_i) \, \Delta x_i. \tag{2.11}$$

Acontece agora que a expressão à direita em (2.11) é uma aproximação para a expressão à direita em (2.8) (no sentido de que a aproximação é tanto melhor quanto mais *fina* for a divisão em classes, e no limite em que o comprimento das classes vai a zero, a soma em (2.11) tende à integral em (2.8)[1]).

[1] Isto será precisado no curso de Cálculo

Exemplo 2.2 *Vamos calcular a média de X no Exemplo 1.3. De (2.8)*

$$M(X) = \int_0^1 x\,3x^2\,dx = 3\int_0^1 x^3\,dx = 3\frac{1^4 - 0^4}{4} = \frac{3}{4}. \qquad (2.12)$$

Exemplo 2.3 *Se $X \sim N(\mu, \sigma^2)$, então a partir de (2.8) e (1.20), podemos obter*[2]

$$M(X) = \mu, \qquad (2.13)$$

e com isto temos uma interpretação a mais para o parâmetro μ. Além de ser o ponto de simetria da distribuição, μ é também (ou por isto) a média, ou centro de massa da distribuição normal.

No caso de termos a distribuição de freqüências de uma variável X dada por classes, podemos usar o lado direito de (2.9) como definição $M(X)$.

Observação 2.4 *A média obtida desta forma* não *equivale àquela obtida como média aritmética dos valores na população, mas é uma aproximação daquela. De fato, tudo se passa como se substituíssemos os valores populacionais dentro de cada classe pelo ponto médio da classe.*

Entretanto, a média dada pelo lado direito de (2.9) coincide com a média obtida usando (2.8), com a função densidade de freqüência dada pelas densidades de freqüências das classes como em (1.17).

Exemplo 2.5 *No Exemplo 1.1, vamos calcular a média de Z a partir da tabela de freqüências de Z (veja Tabela 1.3 na página 4).*

$$\begin{aligned} M(Z) &\simeq 6 \times 0.28 + 10 \times 0.33 + 14 \times 0.22 + 18 \times 0.14 + 22 \times 0.03 \\ &= 11.24 \qquad (2.14) \end{aligned}$$

Média de funções de uma ou mais variáveis

Sejam $X_1, \ldots, X_n$ variáveis definidas em certa população Π, tomando valores nos conjuntos $\mathcal{V}_1, \ldots, \mathcal{V}_n$ respectivamente, e seja $h : \mathcal{V}_1 \times \ldots \times \mathcal{V}_n \to \mathbb{R}$ uma dada função numérica. Então

$$Y := h(X_1, \ldots, X_n) \qquad (2.15)$$

é uma variável numérica definida em Π. Estamos interessados na média de Y.

[2]Este cálculo *não* decorre das propriedades da integral vistas no Apêndice, e por isto não daremos importância aos detalhes.

Exemplo 2.6 *No Exemplo 1.1, suponha que a direção da companhia decida conceder um abono para cobrir gastos com saúde aos funcionários. O abono A é uma função do sexo e número de filhos calculado da seguinte forma.*

$$A = 1_f(X) + W, \tag{2.16}$$

onde 1_f *é a função indicadora de "feminino", isto é,* $1_f : \{f, m\} \to \{0, 1\}$ *é tal que*

$$1_f(f) = 1, \quad 1_f(m) = 0. \tag{2.17}$$

(Neste caso, $A = h(X, W)$*, onde* $h(x, w) = 1_f(x) + a(w)$*,* $x \in \{f, m\}$*,* $w = 0, 1, \ldots$*.) A direção da companhia quer saber qual será o abono médio por funcionário.*

Uma forma de calcular $M(Y)$ seria primeiro achar a função de freqüência de Y, e depois aplicar a definição em (2.3). Mas há uma outra forma, descrita a seguir, evitando este cálculo, usando apenas a função de freqüência conjunta de $X_1, \ldots, X_n$.

Proposição 2.7

$$M(Y) = M(h(X_1, \ldots, X_n)) = \sum_{x_1,\ldots,x_n} h(x_1, \ldots, x_n)\, P(X_1 = x_1, \ldots, X_n = x_n) \tag{2.18}$$

Demonstração Da definição (2.3),

$$M(Y) = \sum_{i=1,2,\ldots} y_i\, P(Y = y_i). \tag{2.19}$$

Para $i = 1, 2, \ldots$ seja

$$\Gamma_i = \{(x_1, \ldots, x_n) \in \mathcal{V}_1 \times \ldots \times \mathcal{V}_n :\, h(x_1, \ldots, x_n) = y_i\}.$$

Então

$$P(Y = y_i) = P((X_1, \ldots, X_n) \in \Gamma_i) = \sum_{(x_1,\ldots,x_n)\in\Gamma_i} P(X_1 = x_1, \ldots, X_n = x_n). \tag{2.20}$$

De (2.19,2.20) concluímos que

$$\begin{aligned}
M(Y) &= \sum_{i=1,2,\ldots} y_i \, P(Y = y_i) \\
&= \sum_{i=1,2,\ldots} \; \sum_{(x_1,\ldots,x_n)\in\Gamma_i} h(x_1,\ldots,x_n) P(X_1 = x_1,\ldots,X_n = x_n). \\
&= \sum_{(x_1,\ldots,x_n)\in\mathcal{V}_1\times\ldots\times\mathcal{V}_n} h(x_1,\ldots,x_n) P(X_1 = x_1,\ldots,X_n = x_n). \quad \square
\end{aligned}$$

Voltando ao Exemplo 2.6, temos que

$$\begin{aligned}
M(A) &= (1_f(f)+0)\,P(X=f,W=0) + (1_f(f)+1)\,P(X=f,W=1) \\
&+ (1_f(f)+2)\,P(X=f,W=2) + (1_f(f)+3)\,P(X=f,W=3) \\
&+ (1_f(f)+4)\,P(X=f,W=4) + (1_f(f)+5)\,P(X=f,W=5) \\
&+ (1_f(m)+0)\,P(X=m,W=0) + (1_f(m)+1)\,P(X=m,W=1) \\
&+ (1_f(m)+2)\,P(X=m,W=2) + (1_f(m)+3)\,P(X=m,W=3) \\
&+ (1_f(m)+4)\,P(X=m,W=4) + (1_f(m)+5)\,P(X=m,W=5) \\
&= 1\times P(X=f,W=0) + 2\times P(X=f,W=1) \\
&+ 3\times P(X=f,W=2) + 4\times P(X=f,W=3) \\
&+ 5\times P(X=f,W=4) + 6\times P(X=f,W=5) \\
&+ 0\times P(X=m,W=0) + 1\times P(X=m,W=1) \\
&+ 2\times P(X=m,W=2) + 3\times P(X=m,W=3) \\
&+ 4\times P(X=m,W=4) + 5\times P(X=m,W=5) \\
&= 1\times 0.11 + 2\times 0.14 + 3\times 0.11 + 4\times 0.06 + 5\times 0.03 + 6\times 0 \\
&+ 0\times 0.19 + 1\times 0.08 + 2\times 0.14 + 3\times 0.08 + 4\times 0.03 + 5\times 0.03 \\
&= 1.98 \qquad (2.21)
\end{aligned}$$

Estivemos neste parágrafo implicitamente supondo que as variáveis X_1, ..., X_n são discretas, no sentido que os conjuntos $\mathcal{V}_1,\ldots,\mathcal{V}_n$ são finitos ou infinitos enumeráveis. Contudo, a Proposição 2.7 admite extensão para outros tipos de variáveis. A seguir enunciamos um caso específico simples.

Proposição 2.8 *Suponha que X seja uma variável contínua com função de densidade de freqüência f_X e $h:\mathbb{R}\to\mathbb{R}$. Então*

$$M(h(X)) = \int_{-\infty}^{\infty} h(x)\, f_X(x)\, dx. \qquad (2.22)$$

Exemplo 2.9 *No Exemplo 1.3, suponha que* $h(x) = x^2$. *Então, de (2.22),*

$$\begin{aligned} M(h(X)) &= M(X^2) = \int_{-\infty}^{\infty} h(x)\, f_X(x)\, dx = \int_0^1 x^2 3x^2\, dx \\ &= 3\int_0^1 x^4\, dx = \frac{3}{5}(1^5 - 0^5) = \frac{3}{5}. \end{aligned} \tag{2.23}$$

Corolário 2.10 *Sejam* $X_1, \ldots, X_n$ *variáveis numéricas definidas em certa população* Π, *e* $a_1, \ldots, a_n$ *constantes arbitrárias. Temos*

$$M(a_1X_1 + \ldots + a_nX_n) = a_1M(X_1) + \ldots + a_nM(X_n). \tag{2.24}$$

Demonstração De (2.14) com $h(x_1, \ldots, x_n) = a_1x_1 + \ldots + a_nx_n$ temos

$$\begin{aligned} & M(a_1X_1 + \ldots + a_nX_n) \\ = & \sum_{x_1,\ldots,x_n} (a_1x_1 + \ldots + a_nx_n)\, P(X_1 = x_1, \ldots, X_n = x_n) \\ = & \left(\sum_{x_1,\ldots,x_n} a_1x_1\, P(X_1 = x_1, \ldots, X_n = x_n)\right) \\ & + \ldots + \left(\sum_{x_1,\ldots,x_n} a_nx_n\, P(X_1 = x_1, \ldots, X_n = x_n)\right) \\ = & \; a_1\left(\sum_{x_1,\ldots,x_n} x_1\, P(X_1 = x_1, \ldots, X_n = x_n)\right) \\ & + \ldots + a_n\left(\sum_{x_1,\ldots,x_n} x_n\, P(X_1 = x_1, \ldots, X_n = x_n)\right) \\ = & \; a_1\left(\sum_{x_1} x_1 \sum_{x_2,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n)\right) \\ & + \ldots + a_n\left(\sum_{x_n} x_n \sum_{x_1,\ldots,x_{n-1}} P(X_1 = x_1, \ldots, X_n = x_n)\right). \end{aligned} \tag{2.25}$$

Da Proposição 1.20, temos que

$$\begin{aligned} \textstyle\sum_{x_2,\ldots,x_n} P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1), \\ \ldots, \\ \textstyle\sum_{x_1,\ldots,x_{n-1}} P(X_1 = x_1, \ldots, X_n = x_n) = P(X_n = x_n). \end{aligned} \tag{2.26}$$

Substituindo (2.26) em (2.25), temos finalmente

$$\begin{aligned} & M(a_1X_1 + \ldots + a_nX_n) \\ = & \ a_1 \sum_{x_1} x_1 \, P(X_1 = x_1) + \ldots + a_n \sum_{x_n} x_n \, P(X_n = x_n) \qquad (2.27) \\ = & \ a_1M(X_1) + \ldots + a_nM(X_n). \qquad (2.28) \end{aligned}$$

Exemplo 2.11 *Voltando ao Exemplo 2.6, suponha que a direção da companhia considere uma outra forma de calcular o abono, em função agora da idade e número de filhos dos funcionários. O abono A' é calculado da seguinte forma.*

$$A' = 0.05V + W. \tag{2.29}$$

Neste caso, de (2.24)

$$M(A') = 0.05M(V) + M(W) \overset{(2.6,2.7)}{=} 0.05 \times 34.58 + 1.53 = 3.26. \tag{2.30}$$

Podemos também usar (2.24) para simplificar o cálculo de $M(A)$ (veja (2.16,2.21)).

$$M(A) \overset{(2.24)}{=} M(1_f(X)) + M(W) \tag{2.31}$$

Aplicando a Proposição 2.7 com $n = 1$, $X_1 = X$ e $h = 1_f$, temos

$$\begin{aligned} M(1_f(X)) & = 1_f(f) \, P(X = f) + 1_f(m) \, P(X = m) \\ & \overset{(2.17,1.6)}{=} 1 \times 0.44 + 0 \times 0.56 = 0.44. \qquad (2.32) \end{aligned}$$

Voltando a (2.31), concluimos que

$$M(A) = 0.44 + 1.53 = 1.97.^3 \tag{2.33}$$

O seguinte fato é óbvio.

Proposição 2.12 *Se uma variável X for constante, isto é, se $X(I) = c$ para todo $I \in \Pi$, então*

$$M(X) = c. \tag{2.34}$$

[3] A diferença na segunda casa decimal em relação a (2.21) se deve a erros de aproximação.

Combinando (2.24) e (2.34), temos a propriedade de *linearidade* da média a seguir.

$$M(a_0 + a_1X_1 + \ldots + a_nX_n) = a_0 + a_1M(X_1) + \ldots + a_nM(X_n). \tag{2.35}$$

Observação 2.13 *A propriedade (2.35) vale para todo tipo de família de variáveis, e não apenas para as discretas, para as quais ela foi demonstrada.*

Observação 2.14 *De (2.35) para $n = 1$, temos que dada uma variável X e uma função linear $h(x) = a + bx$, então*

$$M(h(X)) = h(M(X)), \tag{2.36}$$

isto é, M e h comutam. *Esta é uma propriedade da média válida para funções lineares* apenas. *Em geral, ela* não *vale para outros tipos de função. Discutimos o caso quadrático adiante; veja a Observação 2.39.*

Observação 2.15 *Como em (2.4), podemos calcular $M(h(X_1, \ldots, X_n))$ como uma média aritmética sobre os valores de $h(X_1, \ldots, X_n)$ na população.*

$$M(h(X_1, \ldots, X_n)) = \frac{1}{N}\sum_{I\in\Pi} h(X_1(I), \ldots, X_n(I)). \tag{2.37}$$

Média condicional

Dada uma dupla de variáveis (X, Y), sendo a primeira quantitativa, com função de distribuição conjunta dada por $P(X = \cdot, Y = \cdot)$ (veja a Definição 1.2), seja $P(X = \cdot|Y = \cdot)$ a função de distribuição condicional de X dado Y (veja(1.70)). Então dado um valor possível y de Y, a *média condicional de X dado $Y = y$* é dada pela média da distribuição condicional de X dado $Y = y$, isto é,

$$M(X|Y = y) = \sum_x x\, P(X = x|Y = y). \tag{2.38}$$

Consideremos a função $h : \mathcal{V}_Y \to \mathbb{R}$: $h(y) = M(X|Y = y)$. Então $h(Y)$ será denotada por $M(X|Y)$ e denominada *média condicional de X dado Y*. (Note que se trata de uma função de Y.) $M(X|Y)$ tem a seguinte propriedade.

Proposição 2.16

$$M[M(X|Y)] = M(X) \tag{2.39}$$

Demonstração Pela definição da média condicional,

$$\begin{aligned}
M[M(X|Y)] &= \sum_y M(X|Y=y)\,P(Y=y) \\
&= \sum_y \left\{ \sum_x x\,P(X=x|Y=y) \right\} P(Y=y) \\
&= \sum_y \sum_x x\,P(X=x|Y=y)\,P(Y=y) \\
&= \sum_x x \left\{ \sum_y P(X=x|Y=y)\,P(Y=y) \right\} \\
&\overset{(1.74)}{=} \sum_x x\,P(X=x) = M(X). \quad \square
\end{aligned}$$

Exemplo 2.17 *Suponha que X, Y tenham distribuição conjunta dada pela Tabela 1.7. Como discutimos no Exemplo 1.26, a distribuição condicional de X dado Y é dada pela Tabela 1.8. Vamos calcular as médias condicionais de X dado $Y = y$, $y = a, b, c$. Da Tabela 1.8 e da definição (2.38), temos*

$$\begin{aligned}
M(X|Y=a) &= 1 \times 0.20 + 2 \times 0.40 + 3 \times 0.30 + 4 \times 0.10 = 2.30, \\
M(X|Y=b) &= 1 \times 0.20 + 2 \times 0.44 + 3 \times 0.18 + 4 \times 0.18 = 2.34, \\
M(X|Y=c) &= 1 \times 0.20 + 2 \times 0.30 + 3 \times 0.35 + 4 \times 0.15 = 2.45.
\end{aligned}$$

De (2.39), temos então

$$\begin{aligned}
M(X) = M[M(X|Y)] &= \sum_{y=a,b,c} M(X|Y=y)\,P(Y=y) \\
&= 2.30 \times 0.30 + 2.34 \times 0.50 + 2.45 \times 0.20 \\
&= 2.35. \qquad (2.40)
\end{aligned}$$

A distribuição marginal de Y utilizada em (2.40) foi obtida da marginal apropriada da Tabela 1.7.

Outra propriedade importante da média condicional é relevante na seguinte situação. Um dos motivos de nos interessarmos na distribuição conjunta de variáveis é como uma forma de entender a associação entre elas. Suponha que (X, Y) são as variáveis (X numérica).

Neste caso, é natural procurar uma função numérica de Y como *estimador* de X. Isto é, queremos $H : \mathcal{V}_Y \to \mathbb{R}$ tal que $H(Y)$ esteja *próxima* de X de alguma forma. Não podemos esperar em geral que $H(Y)$ seja perfeitamente igual a X sempre (isto equivaleria a pedir que $H(y) = x$ para todo par (x, y) de valores de (X, Y)).

Vamos então, como em outras situações semelhantes que veremos adiante, estabelecer uma distância entre $H(Y)$ e X, e em seguida procurar H que minimize esta distância entre todas as possibilidades.

A distância que consideraremos é o *desvio quadrático médio* de $H(Y)$ em relação a X:

$$DQ[H(Y); X] = M[(H(Y) - X)^2] = \sum_{x,y} (H(y) - x)^2 P(X = x, Y = y). \tag{2.41}$$

Proposição 2.18 *A média condicional de X dado Y é a função de Y que minimiza o desvio quadrático médio em relação a X, isto é*

$$DQ[M(X|Y); X] = \min_H DQ[H(Y); X], \tag{2.42}$$

e se H for tal que $H(y) \neq M(X|Y = y)$ para algum valor y de Y, então

$$DQ[H(Y); X] > DQ[M(X|Y); X].$$

Esta é a segunda propriedade da média condicional a que aludimos.

2.1.2 Mediana

A *mediana* de uma dada variável é a grosso modo o/um valor que divide a distribuição daquela variável em *duas metades*. Apesar de fazer sentido para qualquer variável ordinal (para as quais há uma ordem natural no conjunto de valores tomados), vamos restringir atenção ao caso quantitativo.

Seja X uma variável numérica. Segundo o parágrafo acima, a *mediana de X* deveria ser um número $m = m(X)$ com a propriedade de que metade

W'	freqüência
0	0.30
1	0.20
2	0.25
3	0.16
4	0.06
5	0.03

Tabela 2.1

da distribuição de X se localiza acima de m e metade da distribuição de X se localiza abaixo de m. Em suma,

$$P(X < m) = P(X > m) = 0.50. \tag{2.43}$$

Mas há um problema com (2.43). Nem sempre existe m que a satisfaça. Considere por exemplo a variável W no Exemplo 1.1. Se $m \leq 1$, então

$$P(W < m) \leq P(X < 1) = 0.30 < 0.50;$$

e se $m > 1$, então

$$P(W < m) \geq P(W \leq 2) = 0.52 > 0.50.$$

Em face deste e outros problemas em dotar a idéia da mediana de uma definição rigorosa, chegamos ao seguinte.

Definição 2.19 *Dada uma variável numérica X, a* mediana *de X, denotada $m = m(X)$, é qualquer número m satisfazendo*

$$P(X \leq m) \geq 0.50 \text{ e } P(X \geq m) \geq 0.50. \tag{2.44}$$

Com esta definição, podemos verificar que $m(W) = 1$ no exemplo que acabamos de discutir. Note que 1 é o único valor a satisfazer a condição de ser mediana neste caso. Mas há exemplos em que há mais de um valor satisfazendo (2.44). Considere a variável W' com distribuição dada pela Tabela 2.1. Neste caso qualquer número do intervalo $[1, 2]$ satisfaz (2.44), e logo é uma mediana.

Definição 2.20 (Complementar) *Com o objetivo de ter um valor único para a mediana no caso em houver um intervalo de valores para m segundo (2.44), seguindo um princípio de simetria, tomaremos como mediana o valor médio do intervalo em questão, isto é, sendo*

$$\mathcal{M} = \{m : m \text{ satisfaz (2.44)}\} \tag{2.45}$$

e $\underline{m} = \min \mathcal{M}$ *e* $\bar{m} = \max \mathcal{M}$*, então*

$$m = \frac{\underline{m} + \bar{m}}{2}. \tag{2.46}$$

No último exemplo, aplicando a Definição 2.20, teremos $\mathcal{M}(W') = [1, 2]$, $\underline{m}(W') = 1$, $\bar{m}(W') = 2$, e logo $m(W') = 1.5$.

Observação 2.21 *Podemos expressar a condição (2.44) em termos da função de distribuição acumulada. Note que a segunda expressão em (2.44) é equivalente a* $P(X < m) \leq 0.50$. *Seja* F_X *a função de distribuição de* X*; então,* m *é uma mediana de* X *se e só se*

$$F_X(m) \geq 0.50 \text{ e } F_X(m-) \leq 0.50, \tag{2.47}$$

onde $F_X(x-) = P(X < x)$ *para dado* $x \in \mathbb{R}$ *(conforme a Observação 1.19).*

Desta última observação, podemos obter a mediana a partir do gráfico de F_X da seguinte forma. Faça o gráfico de F_X, colocando segmentos de reta nos saltos, se houver. Veja a Figura 2.3. Projete o ponto 0.50 no eixo das ordenadas horizontalmente no gráfico. Pode haver apenas um ponto projetado no gráfico, como na Figura 2.3, ou pode haver um intervalo, como na Figura 2.4. No primeiro caso, a mediana é a projeção vertical no eixo das abscissas do ponto no gráfico. No segundo caso, é o intervalo projetado da mesma forma, ou seguindo a definição complementar, o seu ponto médio.

Observação 2.22 *No caso em que a variável* X *for contínua, tomando valores num intervalo* (a, b)*, com densidade* f_X *positiva em* (a, b)*, temos que existe apenas uma mediana, que pode ser obtida da função inversa de* F_X*, bem definida em* $(0, 1)$*, calculada em* 0.50*.*

De fato, note que no caso acima, $F_X(x) = \int_a^x f_X(x)\, dx$ *é estritamente crescente em* (a, b)*, pois* f_X *é positiva neste intervalo por hipótese. Logo existe a inversa* $F_X^{-1} : (0, 1) \to (a, b)$*. Como* F_X *é estritamente crescente em*

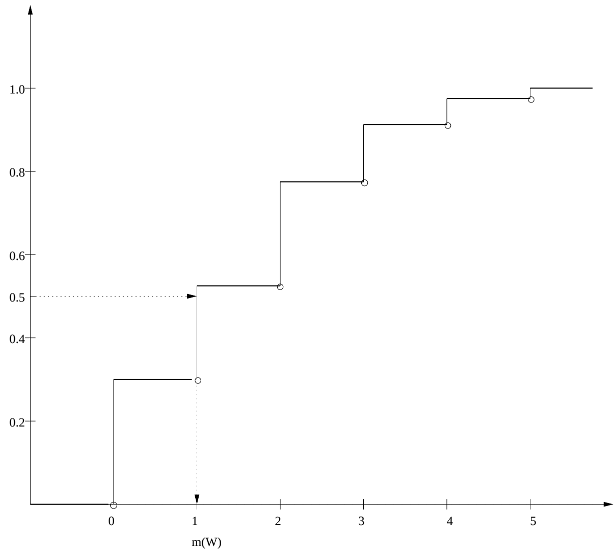

Figura 2.3 Gráfico de F_W com indicação da mediana.

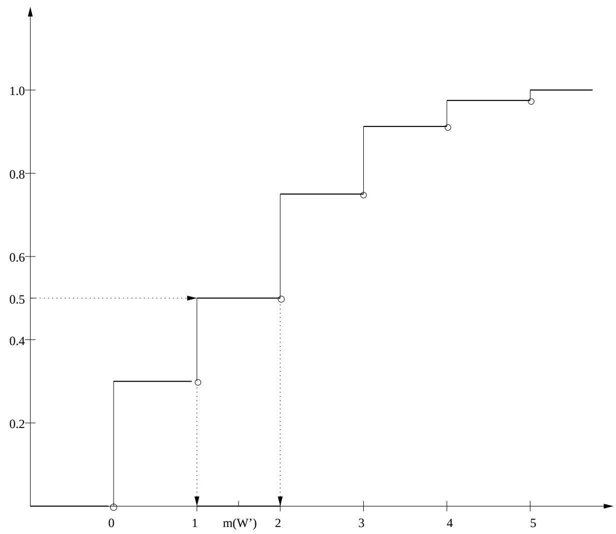

Figura 2.4 Gráfico de $F_{W'}$ com indicação da mediana.

(a,b), e $F_X(a)=0$ e $F_X(b)=1$ (já que a e b são os extremos do intervalo em que X toma valores), então existe apenas um número m tal que

$$F_X(m) = 0.50, \tag{2.48}$$

e aplicando a inversa nos dois lados de (2.48), temos

$$m(X) = F_X^{-1}(0.50). \tag{2.49}$$

Para X do Exemplo 1.3 (veja (1.59)), temos $F_X(x) = x^3$ em $(0,1)$, logo $F_X^{-1}(x) = \sqrt[3]{x}$ em $(0,1)$. Portanto, de (2.49),

$$m(X) = \sqrt[3]{0.50} = 0.79. \tag{2.50}$$

Mediana de dados populacionais

De posse dos dados populacionais da variável X, isto é, da lista de indivíduos $\Pi = \{I_1, \ldots, I_N\}$ e dos respectivos valores de X,

$$\mathcal{D}_X = \{X(I_1), \ldots, X(I_N)\},$$

a mediana de X pode ser calculada da seguinte forma.

Em primeiro lugar devemos ordenar de forma crescente os valores de $\mathcal{D}_X$, respeitando a multiplicidade, obtendo o conjunto ordenado

$$\mathcal{O}_X = \{y_1, \ldots, y_N\}.$$

Por exemplo, se $N = 5$ e $\mathcal{D}_X = \{1,3,5,4,1\}$, então, $\mathcal{O}_X = \{1,1,3,4,5\}$.

Há dois casos. Se N for ímpar, então

$$m(X) = y_{(N+1)/2}; \tag{2.51}$$

e se N for par, então

$$m(X) = \frac{y_{N/2} + y_{(1+N/2)}}{2}. \tag{2.52}$$

No exemplo que acamos de ver, $N = 5$, ímpar, logo de (2.51)

$$m(X) = y_3 = 3. \tag{2.53}$$

Se tivéssemos $\mathcal{D}_X = \{1,3,5,2,4,1\}$, então, $N = 6$, par, e

$$\mathcal{O}_X = \{1,1,2,3,4,5\},$$

logo de (2.52),

$$m(X) = \frac{y_3 + y_4}{2} = \frac{2+3}{2} = 2.5. \tag{2.54}$$

Comparação entre média e mediana

Ambas, média e mediana, são medidas de centralidade da distribuição de dada variável X, isto é, indicam onde se localiza o centro da distribuição de X.

Uma diferença importante entre estas medidas é que a média leva em conta todos os valores tomados por X, enquanto a mediana depende apenas dos valores centrais da distribuição de X.

Por exemplo, se no último conjunto de dados acima tivermos 50 ao invés de 5, a alteração na média é substancial, enquanto a mediana não sofre modificação.

Isto pode ser considerado uma virtude da mediana. Ela não é afetada por valores extremos, que muitas vezes não são representativos, e podem até ser devidos a erros de anotação.

A virtude da média é que ela é mais simples de calcular e de manipular, e tem boas propriedades, como a linearidade (veja (2.35)), que a mediana não tem.

2.1.3 Quantis

Os quantis são medidas de posição como a mediana, mas não de centralidade. Informalmente, o *quantil* p de uma variável numérica X é um valor q_p tal que uma porcentagem p da distribuição de X encontra-se abaixo de q_p, e uma porcentagem $1-p$ da distribuição de X encontra-se acima de q_p.

Para tornar esta idéia uma definição formal, encontramos as mesmas dificuldades do que no caso da mediana (veja a discussão acima da Definição 2.19). A solução é semelhante à achada para aquele caso.

Definição 2.23 *Dada uma variável numérica X, o* quantil p *de X, denotado $q_p = q_p(X)$, é qualquer número q satisfazendo*

$$P(X \leq q) \geq p \text{ e } P(X \geq q) \geq 1-p. \tag{2.55}$$

Observação 2.24 *Podemos expressar as condição (2.55) em termos da função de distribuição acumulada. q é um quantil p se e só se*

$$F_X(q) \geq p \text{ e } F_X(q-) \leq p. \tag{2.56}$$

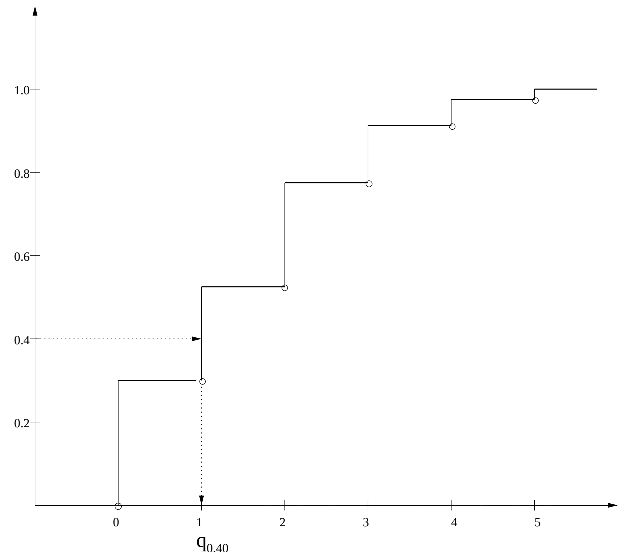

Figura 2.5 Gráfico de F_W com indicação do quantil 0.40.

Como no caso da mediana, podemos ter mais de um quantil p segundo a definição acima. Seguinto um princípio de simetria, fazemos a seguir uma escolha de um único quantil p quando houver multiplicidade.

Definição 2.25 (Complementar) *Se houver mais de um quantil p, então haverá um intervalo de possibiliadades. Seja*

$$\mathcal{Q} = \{m : m \text{ satisfaz (2.55)}\} \tag{2.57}$$

e $\underline{q} = \min \mathcal{M}$ *e* $\bar{q} = \max \mathcal{M}$, *então*

$$q_p = (1-p)\underline{q} + p\bar{q}. \tag{2.58}$$

Podemos ainda obter o quantil p graficamente de forma similar ao que fizemos com a mediana (veja a discussão abaixo da Observação 2.21 e as Figuras 2.3 e 2.4). No presente caso, devemos localizar p no eixo das ordenadas e fazer as projeções como no caso da mediana. Veja as Figuras 2.5 e 2.6 em que consideramos o quantil 0.40 da variável W do Exemplo 1.1, e o quantil 0.75 da variável W' com distribuição dada pela Tabela 2.1.

Observação 2.26 *No caso em que a variável X for contínua, tomando valores num intervalo (a, b), com densidade f_X positiva em (a, b), temos que*

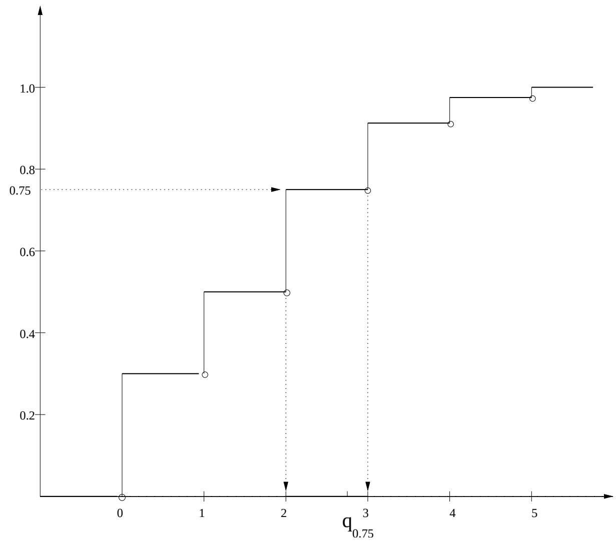

Figura 2.6 Gráfico de $F_{W'}$ com indicação do quantil 0.75.

existe apenas um quantil p *para todo* $p \in (0,1)$, *que pode ser obtida de* F_X^{-1}, *como no caso da mediana (veja a Observação 2.22). Neste caso, o quantil* p *satisfaz*

$$q_p(X) = F_X^{-1}(p). \tag{2.59}$$

Para X *do Exemplo 1.3 (veja (1.59)), temos* $F_X(x) = x^3$ *em* $(0,1)$, *logo* $F_X^{-1}(x) = \sqrt[3]{x}$ *em* $(0,1)$. *Portanto, de (2.59),*

$$q_p(X) = \sqrt[3]{p}. \tag{2.60}$$

Por exemplo, o quantil 0.67 *vale*

$$q_{0.67}(X) = \sqrt[3]{0.67} = 0.88. \tag{2.61}$$

2.1.4 Quartis e diagrama de caixa

Os quantis 0.25 e 0.75 de dada variável X são chamados de *primeiro* e *terceiro quartis*, e denotados Q_1 e Q_3, respectivamente.

Junto com a mediana de X, eles entram numa representação gráfica da distribuição de X chamada de *diagrama de caixa*, que passamos a descrever[4].

[4]Nossa definição é uma adaptação daquela para conjuntos de dados comumente encontrada na literatura.

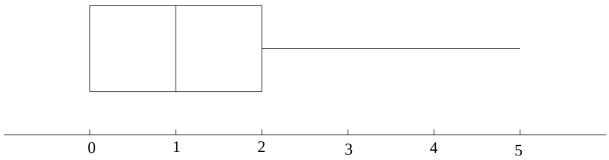

Figura 2.7 Diagrama de caixa de W

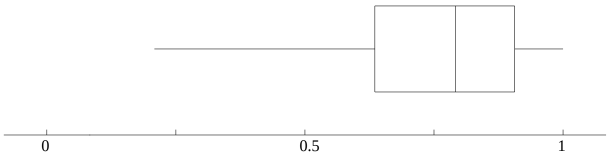

Figura 2.8 Diagrama de caixa de X

A idéia é indicar o centro da distribuição através da mediama m, e a região central da distribuição através de Q_1 e Q_3. Note da definição dos quartis que entre eles concentra-se grosso modo 50% da distribuição.

A porção central da distribuição é então representada por uma caixa com extremidades nos quartis e indicação central na mediana.

As caudas da distribuição são representadas por segmentos a partir dos quartis de comprimento igual ao mínimo entre $1.5(Q_3 - Q_1)$ e

1. à direita, a distância entre Q_1 e o mínimo da distribuição;

2. à esquerda, a distância entre Q_3 e o máximo da distribuição.

Nas Figuras 2.7, 2.8 e 2.9, exibimos os diagramas de caixa das distribuições de W do Exemplo 1.1, X do Exemplo 1.3 e da distribuição normal padrão. Indicamos os cálculos envolvidos nestes diagramas a seguir.

Já tínhamos achado $m(W) = 1$ (veja a Figura 2.3). Da mesma forma, podemos encontrar $Q_1(W) = q_{0.25}(W) = 0$ e $Q_3(W) = q_{0.75}(W) = 2$. A caixa do diagrama de caixa de W tem pois extremidades em 0 e 2 e uma

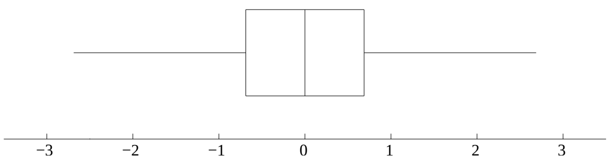

Figura 2.9 Diagrama de caixa da distribuição normal padrão

indicação central em 1. Há um segmento a partir de 2 de comprimento $1.5(Q_3 - Q_1) = 1.5 \times 2 = 3$, que coincide com a distância ao valor máximo da distribuição. Não há valores da distribuição abaixo de 0, por isto não colocamos segmento aí. Veja a Figura 2.7.

Já tínhamos achado $m(X) = 0.79$ (veja (2.50)). De forma similar, achamos $Q_1(X) = q_{0.25}(X) = \sqrt[3]{0.25} = 0.63$ e $Q_3(X) = q_{0.75}(X) = \sqrt[3]{0.75} = 0.91$. A caixa do diagrama de caixa de X tem pois extremidades em 0.63 e 0.91 e uma indicação central em 0.79. Há um segmento a partir de 0.91 até 1, o máximo da distribuição de X. Há outro segmento de comprimento $1.5(Q_3 - Q_1) = 1.5 \times 0.28 = 0.42$ a partir de 0.63. Veja a Figura 2.8.

Por último, o caso da normal padrão, digamos Z. Note em primeiro lugar que a função de distribuição de Z, F_Z, coincide com a função A (veja (1.31)), e logo F_Z^{-1} coincide com B (veja (1.34)). Logo podemos obter $m(Z) = B(0.50) = 0$ e $Q_3(Z) = q_{0.75}(Z) = B(0.75) = 0.67$ consultando a tabela de A ao contrário, como fizemos no parágrafo sobre freqüências inversas da Subseção 1.2.1. Usando a simetria da distribuição normal padrão (como já o fizemos no parágrafo que acabamos de mencionar), temos que para todo $\alpha \in (0, 1)$,

$$B(\alpha) = -B(1 - \alpha), \tag{2.62}$$

do que concluímos que $Q_1(Z) = q_{0.25}(Z) = B(0.25) = -B(0.75) = -0.67$. Logo, a caixa do diagrama de caixa de Z tem extremidades em -0.67 e 0.67 e uma indicação central em 0. Há segmentos a partir de -0.67 e 0.67 de comprimento $1.5(Q_3 - Q_1) = 1.5 \times 1.34 = 2.01$. Veja a Figura 2.9.

2.2 Medidas de dispersão

Nesta seção apresentamos algumas medidas de largura para distribuições de freqüências. Elas descrevem a variabilidade nos valores da distribuição. Todas envolvem tomar distâncias entre os valores da variável em questão. Portanto, elas se aplicam somente ao caso de variáveis numéricas, em que esta noção de distância é natural.

Amplitude

Uma tal medida é a *amplitude* da distribuição, definida como a distância, ou *desvio*, entre o máximo valor possível e o mínimo valor possível. Isto é, sendo X uma variável numérica cujo conjunto de valores possíveis é $\mathcal{V}$. Então

$$A(X) = \max \mathcal{V} - \min \mathcal{V} \tag{2.63}$$

é a amplitude de X.

Esta medida é em geral muito grosseira, levando as freqüências da distribuição pouco em conta, e por é isto pouco utilizada.

Exemplo 2.27 *É fácil de ver que para as variáveis do Exemplo 1.1, temos*

$$A(V) = 48-20 = 28; \quad A(W) = 5-0 = 5; \quad A(Z) = 23.3-4 = 19.3. \tag{2.64}$$

Desvio interquartis

Como indica o nome, trata-se do desvio entre o terceiro e primeiro quartis. Isto é, sendo X uma variável numérica cujos primeiro e terceiro quartis são dados por Q_1 e Q_3 respectivamente, então

$$DI(X) = Q_3 - Q_1 \tag{2.65}$$

é o desvio interquartis de X.

Exemplo 2.28 *Para W no Exemplo 1.1, temos que $Q_1 = 0$ e $Q_3 = 2$ (veja o final da seção anterior). Então*

$$DI(W) = 2 - 0 = 2. \tag{2.66}$$

Para X no Exemplo 1.3, temos que $Q_1 = 0.63$ e $Q_3 = 0.91$. Então

$$DI(X) = 0.91 - 0.63 = 0.28. \tag{2.67}$$

Desvio absoluto médio

A idéia para as duas medidas a seguir é tomar médias sobre os desvios dos possíveis valores de uma distribuição de freqüência entre si, ou, em ambos os casos, em relação a um valor central, a média, neste casos.

Seja X uma variável numérica com média $M(X)$.

Então $|X - M(X)|$ representa a função de X que para cada valor possível desta variável associa o *desvio absoluto* deste valor em relação à média $M(X)$. Isto é, $|X - M(X)| = h(X)$, onde $h : \mathbb{R} \to \mathbb{R}$, $h(x) = |x - M(X)|$. O *desvio absoluto médio* (da distribuição) de X é então definido como a média de $h(X)$.

$$DA(X) = M(|X - M(X)|). \tag{2.68}$$

No caso de X ser discreta, com função de freqüência $p_X = \{P(X = x_i);\, i = 1, 2, \ldots\}$, teremos

$$DA(X) = \sum_{i=1,2,\ldots} |x_i - M(X)|\, P(X = x_i). \qquad (2.69)$$

E no caso de X ser contínua, com função de densidade de freqüência f_X, teremos

$$DA(X) = \int_{-\infty}^{\infty} |x - M(X)|\, f_X(x)\, dx. \qquad (2.70)$$

Exemplo 2.29 *Para W no Exemplo 1.1, temos*

$$\begin{aligned} DA(W) &= |0 - 1.53| \times 0.30 + |1 - 1.53| \times 0.22 + |2 - 1.53| \times 0.25 \\ &+ |3 - 1.53| \times 0.14 + |4 - 1.53| \times 0.06 + |5 - 1.53| \times 0.03 \\ &= 1.15 \end{aligned}$$

Exemplo 2.30 *Para X no Exemplo 1.3,*

$$\begin{aligned} DA(X) &= \int_0^1 \left|x - \frac{3}{4}\right| 3x^2\, dx = 3\int_0^1 \left|x - \frac{3}{4}\right| x^2\, dx \\ &= 3\int_0^{\frac{3}{4}} \left(\frac{3}{4} - x\right) x^2\, dx + 3\int_{\frac{3}{4}}^1 \left(x - \frac{3}{4}\right) x^2\, dx \\ &= 3\int_0^{\frac{3}{4}} \frac{3}{4} x^2\, dx - 3\int_0^{\frac{3}{4}} x^3\, dx + 3\int_{\frac{3}{4}}^1 x^3\, dx - 3\int_{\frac{3}{4}}^1 \frac{3}{4} x^2\, dx \\ &= \frac{9}{4}\frac{\left(\frac{3}{4}\right)^3 - 0^3}{3} - 3\frac{\left(\frac{3}{4}\right)^4 - 0^4}{4} + 3\frac{1^4 - \left(\frac{3}{4}\right)^4}{4} - \frac{9}{4}\frac{1^3 - \left(\frac{3}{4}\right)^3}{3} \\ &= \frac{3}{4}\left(\left(\frac{3}{4}\right)^3 - \left(\frac{3}{4}\right)^4 + 1 - \left(\frac{3}{4}\right)^3 - \left(1 - \left(\frac{3}{4}\right)^4\right)\right) \\ &= \frac{3}{4}\left(2\left(\frac{3}{4}\right)^3 - 2\left(\frac{3}{4}\right)^4\right) = \frac{3}{2}\left(\frac{3}{4}\right)^3 \frac{1}{4} = \frac{81}{512} = 0.158. \end{aligned}$$

2.2.1 Variância e Desvio-padrão

Se acima no lugar de desvio absoluto, considerássemos desvio quadrático, obteríamos o *desvio quadrático médio* ou *variância* de X.

$$V(X) = M[(X - M(X))^2]. \qquad (2.71)$$

No caso discreto,

$$V(X) = \sum_{i=1,2,\ldots} (x_i - M(X))^2 \, P(X = x_i), \tag{2.72}$$

e no caso contínuo,

$$V(X) = \int_{-\infty}^{\infty} (x - M(X))^2 \, f_X(x) \, dx. \tag{2.73}$$

Exemplo 2.31 *Para W no Exemplo 1.1, temos*

$$\begin{aligned} V(W) &= (0-1.53)^2 \times 0.30 + (1-1.53)^2 \times 0.22 + (2-1.53)^2 \times 0.25 \\ &+ (3-1.53)^2 \times 0.14 + (4-1.53)^2 \times 0.06 + (5-1.53)^2 \times 0.03 \\ &= 1.85 \end{aligned}$$

Exemplo 2.32 *Para X no Exemplo 1.3,*

$$\begin{aligned} V(X) &= \int_0^1 \left(x - \frac{3}{4}\right)^2 3x^2 \, dx = 3 \int_0^1 \left(x^2 - \frac{3}{2}x + \frac{9}{16}\right) x^2 \, dx \\ &= 3 \int_0^1 x^4 \, dx - \frac{9}{2} \int_0^1 x^3 \, dx + \frac{27}{16} \int_0^1 x^2 \, dx \\ &= 3\frac{1}{5} - \frac{9}{2}\frac{1}{4} + \frac{27}{16}\frac{1}{3} = \frac{3}{5} - \frac{9}{8} + \frac{27}{48} = \frac{3}{80} = 0.0375. \end{aligned}$$

Neste caso, para obtermos uma medida de dispersão linear, isto é, nas mesmas unidades de medida da variável, e não em tais unidades ao quadrado, como no caso da variância, tomamos a raiz quadrada desta para obter o *desvio padrão*.

$$DP(X) = \sqrt{V(X)}. \tag{2.74}$$

Exemplo 2.33 *No Exemplo 1.1*

$$DP(W) = \sqrt{1.85} = 1.36. \tag{2.75}$$

Exemplo 2.34 *No Exemplo 1.3*

$$DP(X) = \sqrt{0.0375} = 0.19. \tag{2.76}$$

Exemplo 2.35 *Se* $X \sim N(\mu, \sigma^2)$*, então a partir de (2.73) e (1.20), podemos obter*[5]

$$V(X) = \sigma^2, \tag{2.77}$$

e logo $DP(X) = \sigma$*. Com isto podemos interpretar o parâmetro* σ *como o desvio padrão de* X*.*

Observação 2.36 *Da definição da variância, vemos que quanto maior a variabilidade da distribuição de* X*, maior sua variância. De fato, uma maior variabilidade significa maiores desvios em relação a média, e logo um maior valor médio destes desvios ao quadrado.*

Note ainda que a variância é sempre não negativa, pois os desvios ao quadrado são sempre não negativos, e que ela só se anula quando X *for uma constante. De fato, analisando por exemplo (2.72), como se trata de uma soma de termos não negativos, a variância só pode se anular se todos os termos da soma se anularem, e isto ocorrerá somente se todos os desvios em relação à média se anularem, o que significa que* X *é constante.*

As observações acima são obviamente válidas também para o desvio padrão (e também para o desvio absoluto e amplitude, mas não *para o desvio interquartis).*

Fórmula alternativa para a variância

Vamos expandir a expressão quadrática que entra na variância.

$$(X - M(X))^2 = X^2 - 2M(X)X + [M(X)]^2. \tag{2.78}$$

Tomando a média e usando sua linearidade (veja (2.35)), teremos

$$\begin{aligned} M[(X - M(X))^2] &= M\{X^2 - 2M(X)X + [M(X)]^2\} \\ &= M(X^2) - 2M(X)M(X) + [M(X)]^2 \\ &= M(X^2) - [M(X)]^2. \end{aligned} \tag{2.79}$$

(Note que na segunda igualdade usamos o fato que $M(X)$ é uma constante.) De (2.78) e (2.79), temos

$$V(X) = M(X^2) - [M(X)]^2. \tag{2.80}$$

[5]Este cálculo *não* decorre das propriedades da integral vistas no Apêndice, e por isto não daremos importância aos detalhes.

Exemplo 2.37 *No Exemplo 1.1*

$$\begin{aligned} M(W^2) &= 0 \times 0.30 + 1 \times 0.22 + 4 \times 0.25 + 9 \times 0.14 + 16 \times 0.06 + 25 \times 0.03 \\ &= 0 + 0.22 + 1 + 1.26 + 0.96 + 0.75 = 4.19. \end{aligned}$$

Logo

$$V(W) = 4.19 - 1.53^2 = 1.85. \tag{2.81}$$

Exemplo 2.38 *No Exemplo 1.3*

$$M(X^2) = \int_0^1 x^2\, 3x^2\, dx = 3 \int_0^1 x^4\, dx = 3\frac{1}{5} = 0.6. \tag{2.82}$$

Logo

$$V(X) = 0.6 - 0.75^2 = 0.0375. \tag{2.83}$$

Observação 2.39 *Da não negatividade da variância e de (2.80), temos que, se X não for constante, então*

$$M(X^2) > [M(X)]^2. \tag{2.84}$$

Veja a Observação 2.36 acima. Isto mostra que a comutação entre a média e funções quadráticas não é válida em geral, isto é,

$$M(h(X)) \neq h(M(X)), \tag{2.85}$$

para h da forma $h(x) = a + bx + cx^2$, $c \neq 0$, ilustrando a discussão feita na Observação 2.36.

Comparação entre média e mediana revisitada

Com os conceitos de desvio absoluto e desvio quadrático em mãos, podemos fazer mais uma consideração sobre as diferenças entre média e mediana. A questão sobre qual delas é "mais central"poderia ser respondida usando aqueles conceitos, da seguinte maneira.

Comecemos estendendo a idéia dos desvios absoluto e quadrático. Para a um número real qualquer, consideremos os desvios médios absoluto e quadrático *em relação a a*.

$$\begin{aligned} DA(X;a) &= M(|X-a|); && (2.86) \\ DQ(X;a) &= M[(X-a)^2]. && (2.87) \end{aligned}$$

(Note que, comparando esta notação com (2.71) e (2.68), temos $DA(X) = DA(X; M(X))$ e $V(X) = DQ(X; M(X))$.)

Uma boa forma de definir o valor central da distribuição de X seria então aquele valor a que minimiza o desvio de X dentre todos os possíveis valores.

Temos a seguinte caracterização de média e mediana.

Proposição 2.40

$$\min_{a\in\mathbb{R}} DA(X;a) = M(|X - m(X)|); \tag{2.88}$$

$$\min_{a\in\mathbb{R}} DQ(X;a) = M[(X - M(X))^2] = V(X). \tag{2.89}$$

Podemos então dizer que, se usarmos o desvio absoluto médio como desvio, então a mediana é o valor central; se usarmos o desvio quadrático médio, então a média é o valor central.

Demonstração de (2.89)

Começamos somando e subtraindo $M(X)$ no interior da expressão quadrática em (2.87), obtendo desta maneira

$$\begin{aligned} & DA(X;a) = M[\{(X - M(X)) + (M(X) - a)\}^2] \\ = \; & M[(X - M(X))^2 + 2(M(X) - a)(X - M(X)) + (M(X) - a)\}^2] \\ = \; & M[(X - M(X))^2] + 2(M(X) - a)M[X - M(X)] + (M(X) - a)^2, \end{aligned} \tag{2.90}$$

onde expandimos o quadrado na segunda igualdade, e na terceira a linearidade (note que $(M(X) - a)$ é uma constante). Da linearidade segue também

$$M[X - M(X)] = M(X) - M(X) = 0. \tag{2.91}$$

Agora substituindo (2.91) em (2.90), obtemos

$$DA(X;a) = V(X) + (M(X) - a)^2 \geq V(X), \tag{2.92}$$

pois a expressão quadrática é não negativa. Concluimos a demonstração de (2.89) da observação que temos igualdade em (2.92) se e só se $a = M(X)$.

Propriedades da variância e desvio padrão

Proposição 2.41 *Seja X uma variável numérica e $a, b \in \mathbb{R}$ constantes. Então*

$$V(a + bX) = b^2\, V(X). \tag{2.93}$$

Em particular, fazendo $b = 0$ em (2.93), temos que a variância de uma constante se anula.

Corolário 2.42 *Seja X seja uma variável numérica e $a, b \in \mathbb{R}$ constantes. Então*

$$DP(a + bX) = |b|\, DP(X). \tag{2.94}$$

Em particular, o desvio padrão de uma constante se anula.

Demonstração da Proposição 2.41 Da linearidade da média temos

$$\begin{aligned} V(a + bX) &= M\{[(a + bX) - M(a + bX)]^2\} = M\{[bX - b\, M(X)]^2\} \\ &= M\{[b(X - M(X))]^2\} = M[b^2(X - M(X))^2] \\ &= b^2 M[(X - M(X))^2] = b^2\, V(X). \end{aligned} \tag{2.95}$$

Demonstração do Corolário 2.42 De (2.93) e (2.74) temos

$$DP(a + bX) = \sqrt{V(a + bX)} = \sqrt{b^2\, V(X)} = |b|\sqrt{V(X)} = |b|\, DP(X). \tag{2.96}$$

Padronização de variáveis

A média $M(X)$ e desvio padrão $DP(X)$ de uma variável X qualquer são parâmetros que podem ser vistos respectivamente como a origem e a escala naturais para o conjunto $\mathcal{V}$ de valores possíveis de X.

A transformação

$$Z = \frac{X - M(X)}{DP(X)}, \tag{2.97}$$

denominada *padronização* de X, reduz a origem a 0 e a escala a 1.

De fato, usando a linearidade da média e (2.94), notando que $Z = a + bX$ com $a = -M(X)/DP(X)$ e $b = 1/DP(X)$, temos

$$M(Z) = 0, \quad DP(Z) = 1. \tag{2.98}$$

Observação 2.43 *Note que no caso de X ter distribuição normal, (2.94) é equivalente a (1.24), pois, como observado acima, $\mu = M(X)$ e $\sigma = DP(X)$.*

2.3 Uma medida de dependência entre variáveis

Nesta seção discutimos brevemente uma medida de dependência entre variáveis. Queremos saber quão dependentes são dadas variáveis. Vamos restringir a atenção ao caso de duas variáveis, digamos (X, Y).

Suponha que $P(X = \cdot, Y = \cdot)$ seja a função de distribuição conjunta de (X, Y). Da condição de independência (1.77) é natural considerarmos a seguinte *quantidade de dependência*.

$$QD(X,Y) = \frac{1}{2}\sum_{x,y} |P(X = x, Y = y) - P(X = x)P(Y = y)|, \qquad (2.99)$$

onde a soma é feita sobre todos os valores de X e Y.

QD pode ser vista como uma distância entre a situação de dependência de (X, Y) e a independência. Um aspecto crucial neste ponto de vista é que, se $QD(X, Y) = 0$, então (X, Y) satisfazem (1.77) e são portanto independentes; caso contrário, se $QD(X, Y) > 0$, então (X, Y) não satisfazem (1.77) e logo são dependentes.

Um outro aspecto é o seguinte. De (2.99) temos

$$\begin{aligned} QD(X,Y) &= \frac{1}{2}\sum_{x,y} |P(X = x, Y = y) - P(X = x)P(Y = y)| \\ &\leq \frac{1}{2}\sum_{x,y} [P(X = x, Y = y) + P(X = x)P(Y = y)] \\ &= \frac{1}{2}\left[\sum_{x,y} P(X = x, Y = y) + \sum_{x,y} P(X = x)P(Y = y)\right] \\ &= \frac{1}{2}[1 + 1] = 1, \end{aligned} \qquad (2.100)$$

onde a desigualdade segue do fato que para todo $a, b \geq 0$, temos que

$$|a - b| \leq a + b$$

e

$$\begin{aligned} &\sum_{x,y} P(X = x)P(Y = y) = \sum_x \sum_y P(X = x)P(Y = y) \\ = &\sum_x P(X = x)\left[\sum_y P(Y = y)\right] = \sum_x p_X(x)\,[1] = \sum_x p_X(x) = 1. \end{aligned}$$

QD então é limitada superiormente por 1. Será que pode acontecer de QD ser 1? Em que situação?

Pode ser mostrado que $QD < 1$ para todo (X, Y) (da forma como estamos considerando nestas notas), mas veremos no exemplo seguinte que QD pode ser arbitrariamente próximo de 1 (quando consideramos todos os pares (X, Y) em todas as populações), e que isto acontece numa situação de forte dependência entre X e Y.

Suponha que (X, Y) sejam tais que os valores possíveis de cada uma das variáveis sejam $\{1, 2, \ldots, k\}$, onde k é um número inteiro positivo arbitrário, e que a função de distribuição conjunta seja a seguinte.

$$P(X = i, Y = i) = \frac{1}{k} \tag{2.101}$$

para $i = 1, \ldots, k$.

Note que (2.101) implica que $P(X = Y) = 1$, pois

$$\begin{aligned} P(X = Y) &= P((X, Y) \in \{(1,1), (2,2), \ldots, (k,k)\}) \\ &= \sum_{i=1}^{k} P((X = i, Y = i) \\ &= \sum_{i=1}^{k} \frac{1}{k} = \frac{1}{k} \sum_{i=1}^{k} 1 = \frac{1}{k} k = 1. \end{aligned} \tag{2.102}$$

Podemos então dizer que X e Y são fortemente dependentes.

É fácil de ver (usando a Proposição 1.20, por exemplo) que as distribuições marginais são dadas por

$$P(X = i) = P(Y = i) = \frac{1}{k} \tag{2.103}$$

para $i = 1, \ldots, k$.

Temos então de (2.101-2.103) que

$$\begin{aligned}
2\,QD(X,Y) &= \sum_{x,y} |P(X=x,Y=y) - P(X=x)P(Y=y)| \\
&= \sum_{i=1}^{k} P(X=i,Y=i) - P(X=i)P(Y=i)|| \\
&+ \sum_{\substack{i=1\\ j\neq i}}^{k} |P(X=i,Y=j) - P(X=i)P(Y=j)| \\
&= \sum_{i=1}^{k} (1/k - (1/k)^2) + \sum_{\substack{i=1\\ j\neq i}}^{k} (1/k)^2 \\
&= k\,(1/k - (1/k)^2) + k(k-1)(1/k)^2 \\
&= 2(1-1/k). \qquad (2.104)
\end{aligned}$$

E concluímos que $QD(X,Y) = 1 - 1/k$, o que se aproxima arbitrariamente de 1 quando k cresce.

Em conclusão, diremos que quanto mais próximo de 0 for QD, mais independentes são (X,Y), e quanto mais próximo de 1 for QD, mais dependentes são (X,Y).

Exemplo 2.44 *Vamos considerar os pares de variáveis (X,W) e (Y,W) no Exemplo 1.1 usando a quantidade de dependência acima. Da Tabela 1.6 e (2.99), temos que*

$$\begin{aligned}
& 2QD(X,W) \\
=\ & |0.11 - 0.30\times 0.44| + |0.14 - 0.22\times 0.44| + \ldots + |0 - 0.03\times 0.44| \\
+\ & |0.19 - 0.30\times 0.56| + |0.08 - 0.22\times 0.56| + \ldots + |0.03 - 0.03\times 0.56| \\
=\ & 0.1672 \qquad (2.105)
\end{aligned}$$

No caso do par (Y,W), a distribuição conjunta é dada na Tabela 2.2, e temos portanto

$$\begin{aligned}
& 2QD(Y,W) \\
=\ & |0.14 - 0.30\times 0.33| + |0.08 - 0.22\times 0.33| + \ldots + |0 - 0.03\times 0.33| \\
+\ & |0.13 - 0.30\times 0.50| + |0.11 - 0.22\times 0.50| + \ldots + |0.03 - 0.03\times 0.50| \\
+\ & |0.03 - 0.30\times 0.17| + |0.03 - 0.22\times 0.17| + \ldots + |0 - 0.03\times 0.17| \\
=\ & 0.3738 \qquad (2.106)
\end{aligned}$$

W / Y	0	1	2	3	4	5	total
1	0.14	0.08	0.08	0.03	0	0	0.33
2	0.13	0.11	0.17	0.03	0.03	0.03	0.50
s	0.03	0.03	0	0.08	0.03	0	0.17
total	0.30	0.22	0.25	0.14	0.06	0.03	

Tabela 2.2 Distribuição conjunta de (Y, W)

De (2.105 e (2.106), temos que

$$QD(X, W) = 0.08, \quad QD(Y, W) = 0.19$$

(com aproximação na segunda casa decimal), e concluimos que ambos os pares são dependentes, o segundo par é mais dependente do que o primeiro.

2.4 Associação linear entre duas variáveis

Dado um conjunto de variáveis definidas em dada população, uma questão natural é se algum subconjunto delas pode ser explicado pelas demais, se é uma função delas, em que medida, e que tipo de função.

Nesta seção vamos considerar o caso mais simples em que o conjunto é formado por uma dupla de variáveis numéricas (X, Y), e o tipo de função investigado é a linear.

Em outras palavras, queremos estabelecer em que medida vale a relação

$$Y = a + bX \tag{2.107}$$

para a e b constantes.

Vamos considerar a dupla (V, Z) do Exemplo 1.1. Uma forma de avaliar a validade de (2.107) é analizando o diagrama de dispersão $V \times Z$. Veja a Figura 1.23 na página 27. Poderíamos sobrepor uma reta aos pontos, como na Figura 2.10.

Qual é a reta que melhor se ajusta a esta distribuição conjunta?

Esta pergunta é natural, mas a priori não é precisa: ajuste segundo qual critério, ou qual distância? (Ajuste pressupõe minimizar alguma distância.)

De novo, como na discussão ao redor da Proposição 2.40 na seção anterior, podemos considerar diversos tipos de distância. Aquela mais usada neste caso é uma distância quadrática que discutimos a seguir.

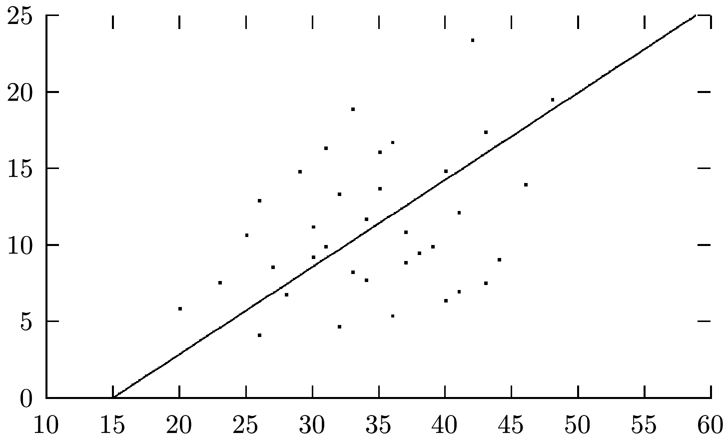

Figura 2.10 Diagrama $V \times Z$ com ajuste de reta

Seja

$$R(x; a, b) = a + bx \tag{2.108}$$

a reta com coeficientes a e b, ditos *intercepto* e *coeficiente angular*, respectivamente.

Queremos então medir a distância entre Y e $R(X; a, b)$. A distância que consideraremos será então o desvio quadrático médio entre as duas variáveis, isto é,

$$M\{[Y - R(X; a, b)]^2\} = M[(Y - a - bX)^2]. \tag{2.109}$$

Para responder a questão acima, precisamos então achar a e b que minimizem a expressão do lado direito de (2.109).

Temos o seguinte resultado.

Proposição 2.45

$$\min_{a,b \in \mathbb{R}} M[(Y - a - bX)^2] = M[(Y - \hat{a} - \hat{b}X)^2], \tag{2.110}$$

onde

$$\hat{b} = \frac{C(X,Y)}{V(X)}; \tag{2.111}$$

$$\hat{a} = M(Y) - \hat{b}M(X), \tag{2.112}$$

e

$$C(X,Y) = M(XY) - M(X)M(Y) \tag{2.113}$$

é, por definição, a covariância *entre X e Y.*

Definição 2.46 *Neste contexto, a reta $R(\cdot;\hat{a},\hat{b})$ será dita a* reta de regressão *de (X,Y), e será denotada por $\hat{Y}$, i.e.*

$$\hat{Y} = \hat{a} + \hat{b}X \tag{2.114}$$

é a reta de regressão de (X,Y).

Exemplo 2.47 *No caso de (V,Z) do Exemplo 1.1, temos (calculando todas as médias explicitamente como médias aritméticas sobre a população; veja (2.37)),*

$$\begin{aligned}
M(V) &\overset{(2.7)}{=} 34.58 \\
V(V) &= M(V^2) - [M(V)]^2 = \left\{\frac{1}{36}(26^2 + \ldots + 42^2)\right\} - [34.58]^2 \\
&= \frac{44645}{36} - 1195.78 = 44.36 \\
M(Z) &= \frac{1}{36}(4 + \ldots + 23.3) = \frac{400.8}{36} = 11.13 \\
M(VZ) &= \frac{1}{36}(26 \times 4 + \ldots + 42 \times 23.3) = \frac{14249}{36} = 395.81 \\
C(V,Z) &= 395.81 - 34.58 \times 11.13 = 10.93 \\
\hat{b} &= \frac{10.93}{44.36} = 0.246 \\
\hat{a} &= 11.13 - 0.246 \times 34.58 = 2.62.
\end{aligned}$$

Então, a reta de regressão de (V,Z) é dada por

$$\hat{Z} = 2.62 + 0.246\,V. \tag{2.115}$$

Na Figura 2.11, exibimos o diagrama $V \times Z$ com a reta de regressão.

Coeficiente de correlação linear

Suponha que queiramos estabelecer uma medida de intensidade da relação linear entre X e Y, ou em outras palavras, de quão bem a reta de regressão de (X,Y) se ajusta a sua distribuição conjunta. Uma tal medida surge do seguinte raciocínio. Uma primeira consideração é que a variabilidade de Y (em torno de sua média) tem dois componentes:

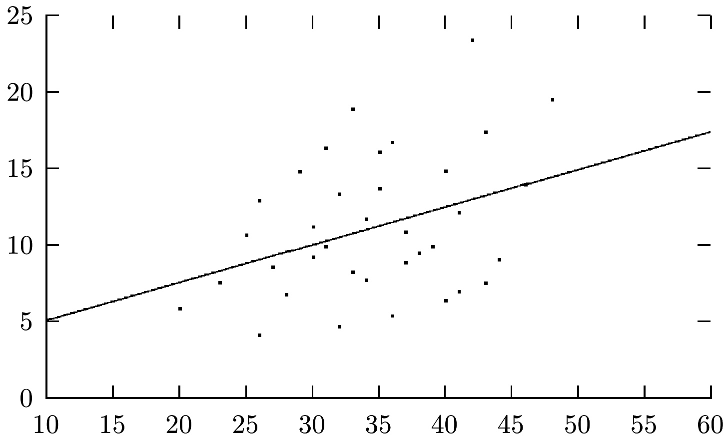

Figura 2.11 Reta de regressão superposta ao diagrama $V \times Z$

1. a variabilidade de Y em torno da reta de regressão $\hat{Y}$, e
2. a variabilidade de $\hat{Y}$ em torno de $M(Y)$.

Veja a Figura 2.12.

Mais precisamente, temos o seguinte resultado.

Proposição 2.48

$$V(Y) = M[(Y - \hat{Y})^2] + M[(\hat{Y} - M(Y))^2] \tag{2.116}$$

O primeiro termo à direita de (2.116) representa a variabilidade de Y *em torno* da reta de regressão; e o segundo termo representa a variabilidade de $\hat{Y}$ em torno de $M(Y)$. Esta última pode ser entendida como a variabilidade *dentro* da reta de regressão.

De 2.116 temos então que a expressão

$$\frac{M[(\hat{Y} - M(Y))^2]}{V(Y)} \tag{2.117}$$

pode ser vista como a proporção da variabilidade de Y devida à reta de regressão, o que portanto se constitui numa medida da intensidade da relação linear entre X e Y.

Desenvolvendo o numerador em (2.117), obtemos (omitindo detalhes)

$$\frac{M[(\hat{Y} - M(Y))^2]}{V(Y)} = \frac{[C(X,Y)]^2}{V(X)\,V(Y)} = R^2, \tag{2.118}$$

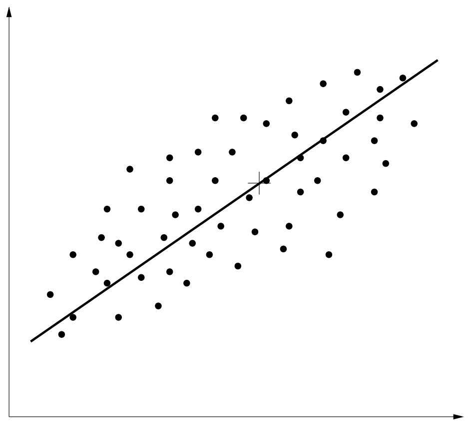

Figura 2.12 Diagrama $X \times Y$ (hipotético); cruz (+) indica $(M(X), M(Y))$

onde

$$R = R_{X,Y} = \frac{C(X,Y)}{DP(X)\, DP(Y)} \tag{2.119}$$

é, por definição, o *coeficiente de correlação linear* entre X e Y.

Devido a (2.117,2.118), R^2 é chamado *coeficiente de explicação.*

Exemplo 2.49 *Vamos calcular os coeficientes de correlação linear e de explicação de (V, Z) do Exemplo 1.1. Além dos números obtidos no Exemplo 2.47, precisaremos ainda dos seguintes, alguns dos quais baseados naqueles.*

$$\begin{aligned} DP(V) &= \sqrt{V(V)} = \sqrt{44.36} = \\ M(Z^2) &= \frac{1}{36}(4^2 + \ldots + 23.3^2) = \frac{5196}{36} = 144.33 \\ V(Z) &= M(Z^2) - [M(Z)]^2 = 20.46 \\ DP(Z) &= \sqrt{20.46} = 4.52 \end{aligned}$$

Agora podemos substituí-los em (2.119) para obter

$$R = \frac{10.93}{6.66 \times 4.52} = 0.363 \tag{2.120}$$

e logo

$$R^2 = 0.132. \tag{2.121}$$

Temos então que por volta de apenas 13% da variabilidade do salário é explicado pela reta de regressão, ou em outras palavras por uma relação linear com a idade. Seria razoável dizer neste caso que se trata de uma relação linear fraca.

O coeficiente de correlação (linear) tem uma informação a mais do que R^2, qual seja, o *sinal*. Discutimos este ponto agora.

Consideremos as padronizações de X e Y

$$X' = \frac{X - M(X)}{DP(X)}, \quad Y' = \frac{Y - M(Y)}{DP(Y)}, \tag{2.122}$$

respectivamente (veja (2.97)). Da definição de $C(X,Y)$ (veja (2.113)) e da linearidade da média, podemos rescrever R da seguinte forma.

$$R = \frac{M(XY) - M(X)M(Y)}{DP(X)\,DP(Y)} = M\left[\left(\frac{X - M(X)}{DP(X)}\right)\left(\frac{Y - M(Y)}{DP(Y)}\right)\right]$$

Logo,

$$R = M(X'Y'). \tag{2.123}$$

O diagrama de dispersão de (X', Y') tem aspecto como na Figura 2.13.

Note neste diagrama que temos mais pontos no primeiro e terceiro quadrantes, onde $X' \times Y'$ é positivo, do que no segundo e quarto quadrantes, onde $X' \times Y'$ é negativo. Isto implica em $R > 0$, e logo uma relação crescente entre X e Y. Neste caso também falamos em *associação positiva* entre X e Y.

Suponha agora que o diagrama de $X \times Y$ fosse como na Figura 2.14. Neste caso, o diagrama de $X' \times Y'$ teria o aspecto da Figura 2.15. Podemos ver que a distribuição de (X', Y') se concentra segundo e quarto quadrantes, o que leva a $R < 0$.

Observação 2.50 *Uma outra forma de entender o papel de R é a seguinte. De novo, vamos considerar as variáveis padronizadas X', Y' e tomar a reta de regressão $\hat{Y}'$ associada a elas. De (2.114), (2.112) e (2.111), substituindo (X,Y) por (X',Y'), podemos verificar que ela é dada por*

$$\hat{Y}' = RX', \tag{2.124}$$

isto é, o intercepto da reta de regressão de (X', Y') se anula e seu coeficiente angular é R.

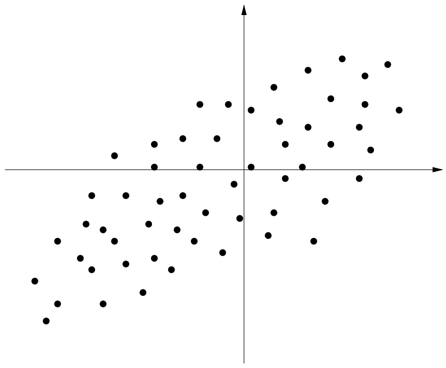

Figura 2.13 Diagrama $X' \times Y'$.

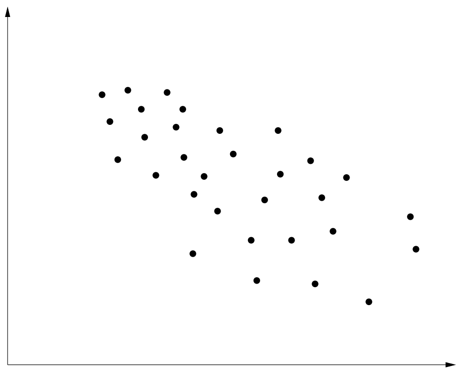

Figura 2.14 Outro caso.

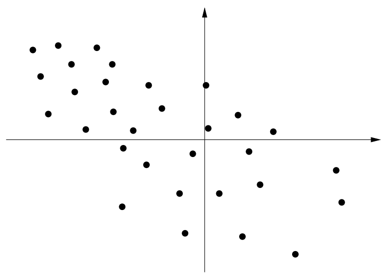

Figura 2.15 Variáveis padronizadas

Em conclusão, $R > 0$ indica associação positiva ou crescente entre X' e Y', e logo entre X e Y, enquanto $R < 0$ indica associação negativa ou decrescente.

Análise da relação linear entre duas variáveis numéricas
Resumindo o que vimos acima, dadas duas variáveis numéricas X e Y, a *intensidade* da relação lienar entre X e Y é medida pelo coeficiente de explicação R^2:

- se R^2 estiver próximo de 1, então podemos dizer que a relação linear é forte;
- se R^2 estiver próximo de 0, então podemos dizer que a relação linear é fraca;
- se R^2 assumir valores intermediários, então podemos dizer que a relação linear é moderada.

E o sinal da relação linear entre X e Y é medida pelo sinal de R:

- se $R > 0$, então a relação linear é positiva;
- se $R < 0$, então a relação linear é negativa.

Exemplo 2.51 *De volta a (V, Z) do Exemplo 1.1 (veja os Exemplos 2.47, 2.49), com $R = 0.363$ e $R^2 = 0.132$, podemos dizer que há uma relação linear positiva mas fraca entre a idade e o salário dos funcionários da companhia em questão.*

Observação 2.52 *Mesmo com R^2 relativamente próximo de 1 (mas não muito próximo), pode ocorrer de a relação entre as variáveis ser preponderantemente de outra forma que não a linear. Veja a Figura 2.16.*

Reciprocamente, podemos ter R^2 próximo de 0 (ou igual a 0), e, apesar de a relação linear entre as varíáveis ser fraca ou inexistente, outro tipo de relação pode existir, e pode ser uma relação forte. Veja a Figura 2.17.

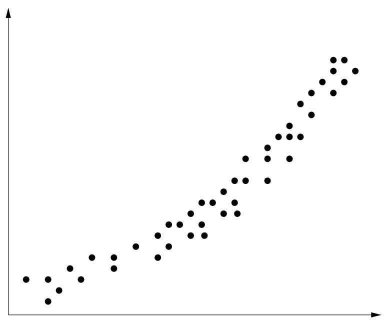

Figura 2.16 R^2 é bastante próximo de 1, mas a relação entre as variáveis está mais para quadrática.

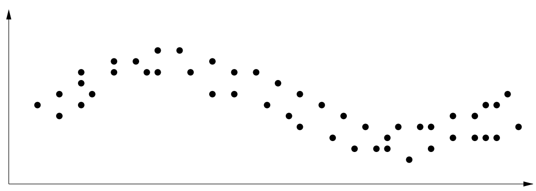

Figura 2.17 R^2 está próximo de 0, mas há uma relação bastante forte do tipo senóide entre as variáveis.

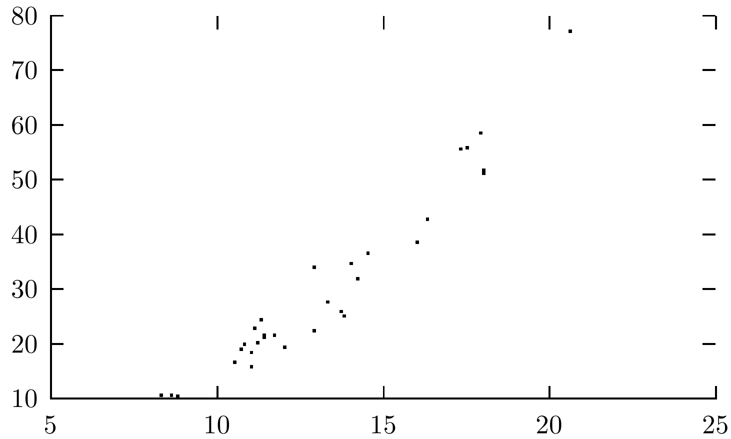

Figura 2.18 $D \times V$

Exemplo 2.53 *Os dados da Tabela 2.3 na página 74 se referem a árvores. Qual variável está mais bem associada linearmente com o volume: o diâmetro ou a altura?*

Para responder a esta questão, vamos analisar os diagramas $D \times V$ *e* $A \times V$, *e calcular* $R_{D,V}$ *e* $R_{A,V}$.

Veja os diagramas nas Figuras 2.18 e 2.19

Temos

$$R_{D,V} = 0.967; \quad R_{A,V} = 0.598, \tag{2.125}$$

e logo

$$R^2_{D,V} = 0.935; \quad R^2_{A,V} = 0.358. \tag{2.126}$$

Espécime	Diâmetro	Altura	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11.0	66	15.6
8	11.0	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21.0
13	11.4	76	21.4
14	11.7	69	21.3
15	12.0	75	19.1
16	12.9	74	22.2
17	12.9	85	33.8
18	13.3	86	27.4
19	13.7	71	25.7
20	13.8	64	24.9
21	14.0	78	34.5
22	14.2	80	31.7
23	14.5	74	36.3
24	16.0	72	38.3
25	16.3	77	42.6
26	17.3	81	55.4
27	17.5	82	55.7
28	17.9	80	58.3
29	18.0	80	51.5
30	18.0	80	51.0
31	20.6	87	77.0

Tabela 2.3 Árvores

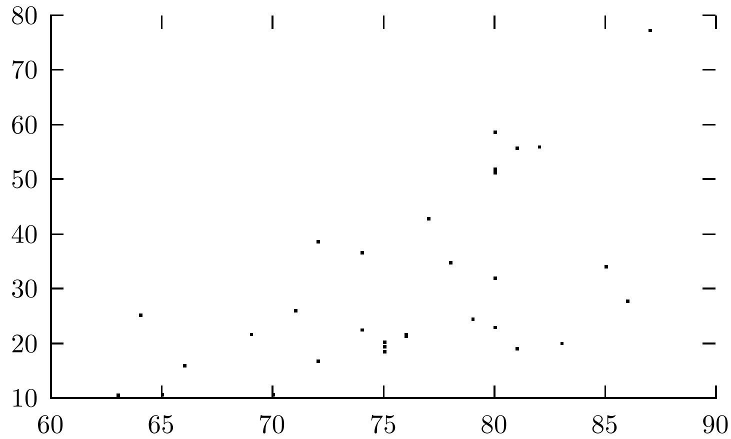

Figura 2.19 $A \times V$

De forma que no caso de (D, V)*, 93.5% da variabilidade do volume das árvores são explicados pela relação linear com o diâmetro das árvores, e no caso de* (A, V)*, apenas 35.8% da variabilidade do volume das árvores são explicados pela relação linear com a altura das árvores.*

Não há dúvida que temos uma relação linear forte, positiva, entre diâmetro e volume. A relação entre altura e volume, também positiva, é no máximo moderada, por isto podemos concluir com segurança que a resposta à pergunta inicial é o diâmetro.

Podemos então calcular a reta de regressão de (D, V)*, obtendo*

$$\hat{V} = -36.9 + 5.07\, D. \tag{2.127}$$