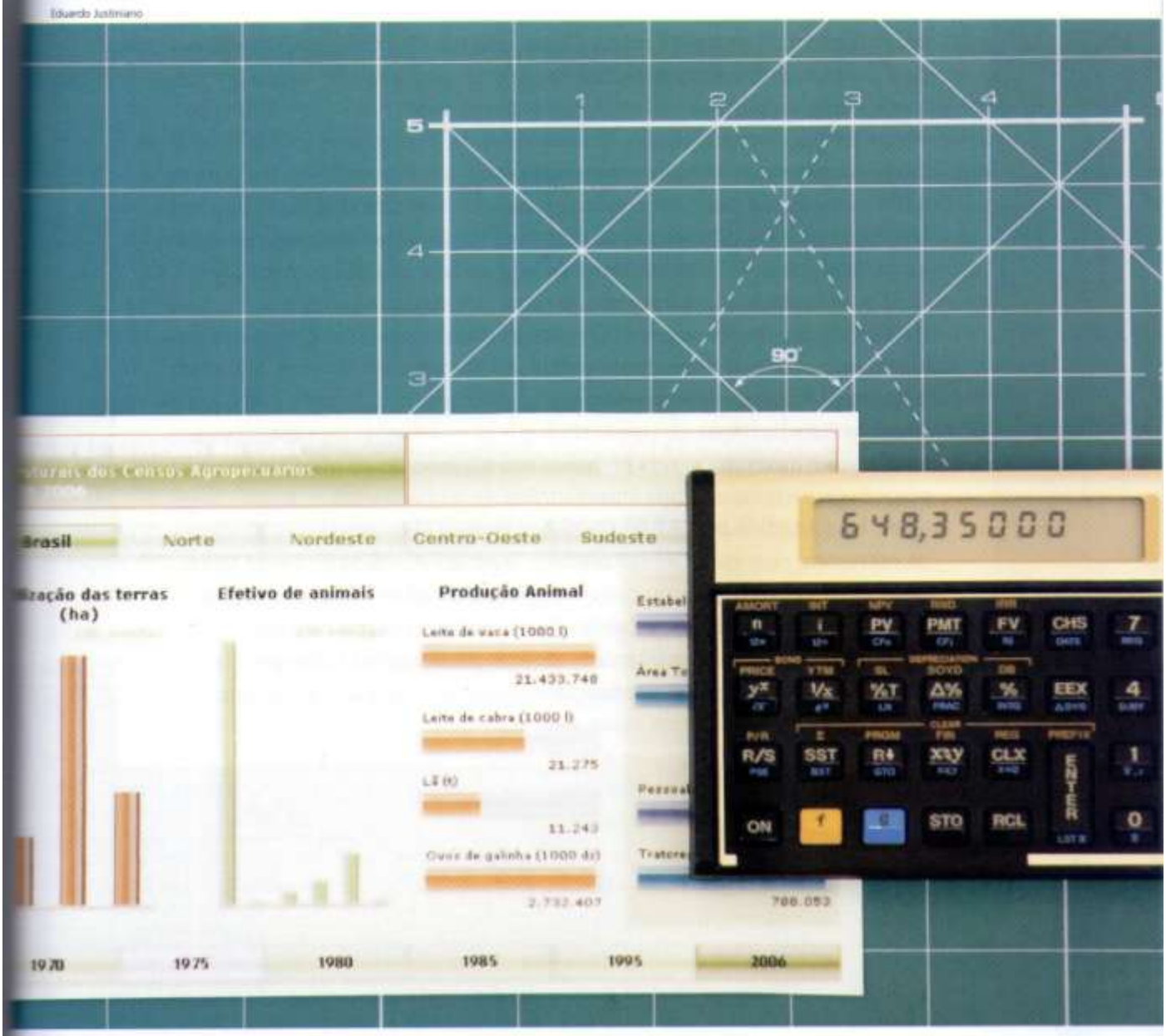
Estatística Descritiva em Sala de Aula

22

EMERSON GALVANI



INTRODUÇÃO

A probabilidade de acertar as seis dezenas (sena) na Mega-Sena é de 1 chance em 50.063.860, segundo informações do *site* da Caixa Econômica Federal. Isso mostra que a estatística faz parte do nosso dia a dia, embora muitas vezes sequer percebemos. Ela é utilizada também na Geografia. A graduação em Geografia exige, pela própria natureza do curso, um número significativo de trabalhos de campo. Essas "saídas" realizadas pelas diferentes áreas/disciplinas, cada qual com seu instrumental apropriado, produzem em cada trabalho de campo um volume de informações específicas e, quando retornamos para a sala de aula, a grande questão que se apresenta é: *"o que fazer com os dados quali-quantitativos coletados no trabalho de campo?"*. Tradicionalmente, os alunos de graduação em Geografia não são muito afeitos à área de exatas, o que pode limitar a análise e a interpretação dos dados observados em campo. Eventualmente, os resultados finais da pesquisa podem ser prejudicados por falta de uma análise mais numérica (estatística) dos dados observados, prevalecendo uma análise visual dos dados.

O que se pretende com este capítulo não é formar geógrafos especialistas em estatística, mas sim, fornecer os princípios básicos da estatística descritiva para promover melhor análise dos dados gerados nos trabalhos de campo e, por outro lado, desmistificar a análise numérica para o estudante de Geografia e de outras áreas das Ciências Humanas. Vale lembrar que as informações aqui apresentadas aplicam-se a qualquer tipo de pesquisa, seja ela produto de questionários ou medições específicas em cada área/disciplina do conhecimento.

Ao longo deste capítulo, o que se pretende não é elucidar todos os conceitos da estatística, nem desencorajar o estudo mais aprofundado do tema. Nossa intenção é a de fornecer os conceitos mínimos para melhor formação dos profissionais em áreas nas quais a "estatística" sempre é vista com preconceito – talvez pelo desconhecimento de suas potencialidades ou pela forma como geralmente é apresentada nos cursos de graduação que envolvem as áreas de humanas.

MEDIDAS DE TENDÊNCIA CENTRAL

A análise de um conjunto de dados com uso de tendência central permite avaliar para onde caminham os dados pesquisados. É uma espécie de "radiografia" inicial, que pode ser determinada com a utilização dos indicadores descritos a seguir.

Média aritmética ($\bar{x}$)

A média aritmética é o procedimento mais simples e comum passível de ser aplicado a um conjunto de dados. Essa medida de tendência central expressa o somatório de todos os elementos da série dividido pelo número total de elementos. Numericamente, a média aritmética é expressa por:

$$\bar{x} = \frac{\sum_i x_i}{n}$$

Nessa expressão, x_i é cada elemento da série e i varia de 1 a n; n é o número de elementos e o símbolo $\sum$ significa a somatória de todos os elementos da série. Resumindo, somam-se todos os elementos e divide-se pelo número total de elementos da série.

Moda (MO)

A moda ou modo (MO) é o valor presente que ocorre com maior frequência na série. Existem séries em que nenhum dado se repete, nesses casos, não existe a moda da série. Isso geralmente ocorre em séries reduzidas (menos de 50 elementos amostrados). De forma análoga, podem ocorrer séries com duas (bimodal) ou mais modas. Nesses casos, prevalece o valor de maior frequência de ocorrência ou, em caso de empate, a série pode apresentar mais de uma moda.

Mediana (ME)

A mediana é aplicável em séries extensas de dados (mais de mil informações) nas quais existam extremos que possam "contaminar" a média, ou seja, alguns dados que "fogem" da *tendência central*, podendo sub ou superestimar as análises. A mediana é determinada ordenando-se os dados de forma crescente ou decrescente e identificando-se a posição central da série. Em caso de séries com número ímpar de elementos, a mediana estará na posição central da série. Para séries com número par de elementos, a mediana será a média dos elementos que ocupam a posição central da série. O conceito de mediana gera algumas confusões: a mediana é simplesmente o valor que se situa na posição central do conjunto de dados ordenados. Assim, deve haver uma relação de ordem nos valores.

Valor Máximo (Vmax) e Mínimo (Vmin)

O valor máximo da série é aquele de maior magnitude, ou seja, o maior valor encontrado na série. O valor mínimo, por sua vez, é o menor valor encontrado na série. Em princípio, parece ser uma informação sem importância, contudo permite-nos visualizar em qual intervalo de medidas se encontra distribuído o conjunto de dados. Serve para evidenciar o *tamanho* dos dados que serão trabalhados. Em séries climatológicas de temperatura do ar, por exemplo, o Vmax equivale à temperatura máxima do ar e o Vmin à temperatura mínima do ar.

Amplitude (Δ)

A amplitude em um conjunto de dados expressa a diferença entre o Vmax e o Vmin.

Essa medida de tendência central expressa a variação máxima dos valores constituintes do conjunto de dados. Dois ou mais conjuntos de dados poderão ter a mesma média, porém diferentes Vmax , Vmin e Δ, evidenciando-se tratar de séries distintas.

A seguir, será apresentado um exemplo de cálculo das medidas de tendência central (média, moda, mediana, valor máximo, valor mínimo e amplitude) para um conjunto simples de dados.

Esses procedimentos podem ser efetuados facilmente por meio de programas como o *Excel*, da Microsoft[1], pelos seguintes passos: com o conjunto de dados dispostos em duas colunas, entrar na barra de ferramentas no atalho ***fx***; em seguida em *estatística* e selecionar a análise de tendência desejada; selecionar o intervalo de dados. O resultado é mostrado. Caso a barra de ferramentas não disponibilize o atalho ***fx***, clique em *inserir* e em seguida em ***fx***, seguindo os mesmos procedimentos descritos anteriormente.

Uma questão que geralmente surge entre os estudantes é sobre a diferença de interpretação entre a mediana e a média. Embora a média seja um valor mais fácil de ser entendido, ela tem restrições, pois pode nos induzir a um erro de tendência, se a amostra analisada apresentar valores de amplitude elevados. Por exemplo, na distribuição dos dados da Tabela 22.1, a média da variável A é 161 e a mediana é 163 (ver Tabela 22.2). Caso uma amostra tivesse apresentado valor 300 e não 121, isso faria com que a média saltasse para 187, ou seja, seria superior a todos os valores individuais, mas a mediana continuaria a ser 163. Se observarmos para todos os 7 valores individuais da amostra, verifica-se que o número 163 é o melhor representante da distribuição desse conjunto de dados. Assim, no caso das variáveis quantitativas, quando o valor da mediana é muito diferente da média, é aconselhável considerar sempre a mediana como o valor de referência mais importante.

Tabela 22.1 – Valores arbitrários para duas variáveis A e B. Dados brutos (à esquerda) e dados ordenados (à direita) em forma crescente

Dados Brutos		Dados Ordenados	
A	B	A	B
121	171	121	152
171	152	157	168
158	170	158	169
173	168	163	170
184	169	171	171
163	171	173	171
157	190	184	190

Tabela 22.2 – Resultados da análise de tendência central para o conjunto de dados da Tabela 22.1

Medida de tendência	Variável A	Variável B
Média ($\bar{x}$)	161	170
Moda (MO)	-	171
Mediana (ME)	163	170
Valor máximo (Vmax)	184	190
Valor mínimo (Vmin)	121	152
Amplitude (Δ)	63	38

1 A citação da marca comercial não implica a recomendação do referido programa por parte do autor. Existem mais programas estatísticos, como *Origin*, *MatLab*, *Estatística* e *SAS* – entre outros. Esses programas também efetuam os procedimentos discutidos nessa passagem. Contudo, com a massificação de uso do *Office*, da Microsoft, o *Excel* é facilmente encontrado em qualquer computador.

MEDIDAS DE DISPERSÃO

As medidas de dispersão são úteis quando diferentes conjuntos de dados apresentam mesma média e mediana, porém variabilidades distintas. Esse tipo de análise pode ser utilizado para comparar quantos conjuntos de dados forem necessários, pois os cálculos são efetuados individualmente para cada conjunto.

Desvio em relação à média (DM)

Essa medida de dispersão fornece-nos uma ideia da variabilidade dos dados em torno da média, sendo, portanto, a diferença entre o valor observado (x_i) e a média do conjunto ($\bar{x}$), representado numericamente por:

$$DM = x_i - \bar{x}$$

Determinados conjuntos de dados podem apresentar médias iguais, contudo com acentuados desvios em relação à média. Veja o exemplo na Tabela 22.3.

Neste exemplo, representado pela Figura 22.1, observa-se que, embora os conjuntos de dados A e B apresentem mesma média e mediana, a variabilidade do conjunto A é menos acentuada que em B. Desta forma, a análise somente da média e mediana pode levar a conclusões não satisfatórias. Vale lembrar que o somatório (Σ) dos desvios em relação à média deve ser igual a zero. O desvio em relação à média tem a desvantagem de não fornecer um único indicador da variabilidade dos dados, ficando restrito a uma análise visual dos dados; para tanto existem outros índices, como será visto a seguir.

Tabela 22.3 – Valores de A e B e desvio em relação à média

A	B	DM "A"	DM "B"
4	9	-1	4
6	1	1	-4
4	5	-1	0
6	5	1	0
5	1	0	-4
5	9	0	4
$\bar{x}=5$	$\bar{x}=5$	$\Sigma=0$	$\Sigma=0$

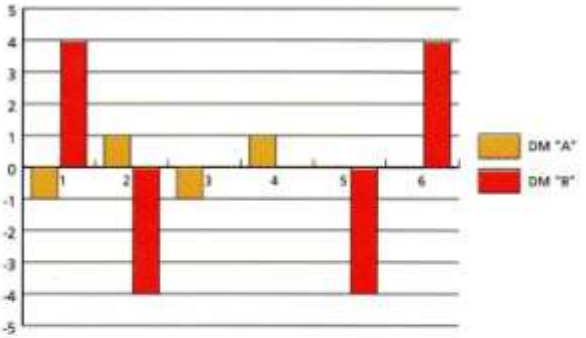


Figura 22.1. Desvio em relação à média para o conjunto de dados da Tabela 22.3.

Variância da amostra (S^2)

Para se avaliar a variabilidade da amostra faz-se uso da noção de variância. Numericamente, a variância é determinada pela somatória do quadrado do desvio em relação à média, dividida pela quantidade de elementos da série menos 1.

$$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

Um dos problemas que prejudicam a análise por meio da variância da amostra é justamente o fato de o resultado ser expresso na unidade de medida dos dados elevado ao quadrado. Por exemplo, se a unidade dos dados da Tabela 22.4 for metro, a variância será expressa em metros ao quadrado (m^2); se for quilograma, a variância será expressa em kg^2, o que causa dificuldade na interpretação da variância da variável.

Tabela 22.4 – Exemplo de cálculo de variância

x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-1	1
6	1	1
4	-1	1
6	1	1
5	0	0
5	0	0
$\bar{x} = 5$		$\sum(x - \bar{x})^2 = 4$

Então, a variância será calculada assim:

$$S^2 = \frac{4}{6-1}$$

$$S^2 = 0{,}8$$

Desvio-padrão (S)

Uma forma de eliminar o problema da interpretação da variância da amostra é extrair sua raiz quadrada. Tem-se assim o desvio-padrão. Essa é uma medida do grau de dispersão dos valores em relação ao valor médio (a média). É um erro dizer que o desvio-padrão é a média de todas as diferenças, mas pode-se "senti-lo" como algo aproximado. Ele é determinado numericamente pela raiz quadrada da variância:

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Coeficiente de variação (CV)

Por vezes, precisa-se comparar duas variáveis quantitativas quanto ao seu grau de dispersão, por exemplo, o peso (em kg) e a idade (em anos). Essa comparação não pode ser feita comparando-se simplesmente os respectivos desvios-padrão, porque eles estão expressos em unidades de medida diferentes, isto é, não se pode comparar a dispersão de massa (kg) com a de idade (anos). No entanto, é possível fazer esta comparação em termos relativos, se calcularmos o coeficiente de variação de cada conjunto de dados, na expressão abaixo, onde **CV** é o coeficiente de variação expresso em porcentagem e **S** é o desvio-padrão já definido anteriormente.

$$CV = \frac{100 \cdot S}{\bar{x}}$$

Veja na tabela a seguir um exemplo de desvio-padrão (**S**) e de coeficiente de variação (**CV**).

Tabela 22.5 – Variáveis A, B e C para cálculo de desvio-padrão e coeficiente de variação

A	B	C
4	9	9
6	1	1
4	5	1
6	5	2
5	1	8
5	9	9

Desvio-padrão de A = 0,9 $\quad CV_A = 18{,}0\%$
Desvio-padrão de B = 3,6 $\quad CV_B = 72{,}0\%$
Desvio-padrão de C = 4,0 $\quad CV_C = 80{,}0\%$

O coeficiente de variação expressa, portanto, a variabilidade de cada conjunto de dados normalizada em relação à média, em porcentagem. Assim, a variável ***A*** oscila, em média, 18,0%.