Bol. Zool. e Biol. Mar., N.S., n.º 30, pp. 809-816, São Paulo, 1973

TENDÊNCIAS MODERNAS NA APRESENTAÇÃO NUMÉRICA CONCISA DE DADOS DE OBSERVAÇÃO

(Reminiscências de um curso de estatística para biólogos ministrado na Cadeira do Prof. Paulo Sawaya)

A. M. PENHA
Instituto Biológico — São Paulo

RESUMO

Para resumir os dados de uma série de observações costuma-se apresentá-los de duas maneiras equivalentes:

1. n ——— m ——— s.d.;
2. n ——— m ——— s.e.m

ou,

1'. n ——— m ± s.d.;
2'. n ——— m ± s.e.m.

onde:

m representa a média, s.d. o desvio padrão e s.e.m. o erro padrão da média.

Em ambos os casos, o número *n* de observações deverá ser indicado a fim de obter os valores *t* de Student nas tabelas de estatística, necessário para o cálculo do intervalo de confiança da média. Nos trabalhos mais recentes, há uma preferência em favor do erro padrão, que substitui o antigo erro provável da média.

MODERN TENDENCY IN THE PRECISE NUMERICAL PRESENTATION OF OBSERVATION DATA

SUMMARY

To summarize the data of a series of observations it is customary to present them in either of two equivalent manners:

1. n ——— m ——— s.d.;
2. n ——— m ——— s.e.m.

or,

1' n ——— m ± s.d.;
2'. n ——— m ± s.e.m.,

Número especial em homenagem ao Prof. Dr. Paulo Sawaya, no ano jubilar de seu magistério.

where,

m stands for the mean, s.d. for the standard deviation, and s.e.m. for the standard error of the mean.

In both ways the number n of observations should be indicated in order to obtain the values of Student's t in the statistical tables, necessary for the calculation of the confidence intervals of the mean. In the most recent papers there is preference in favor of the standard error, which replaces the old probable error of the mean.

Nenhum pesquisador conscienciosо baseia suas conclusões experimentais numa só observação ou experiência. Devido à variabilidade natural inerente à maioria dos fenômenos observados, procura-se repetir a observação o maior número de vezes possível, a fim de se poder obter uma média realmente representativa do fenômeno estudado (*Experimentum unum, experimentum nulum*).

Daí surge um conjunto de dados, geralmente anotados num borrador, que é guardado para futuras interpretações ou novas manipulações numéricas. No decorrer do trabalho aparecem novas séries de observações, que terão o mesmo destino: arquivo ou publicação.

Devido à escassez de espaço, as revistas especializadas, nas quais são feitas as publicações de trabalho científicos, não aceitam mais tabelas completas contendo todos os dados originais observados; estes devem ficar arquivados com os autores, constando no texto do trabalho apenas o resumo das tabelas, contendo os resultados de cada grupo de observações, representados pelas respectivas médias aritméticas, que substituem de maneira concisa o valor do conjunto de elementos de cada grupo; e um segundo parâmetro, que mede a variação observada nos valores dos dados.

Esta maneira de proceder, relacionada à repetição de observações, data de Gauss, que publicou em 1809, em apêndice a um trabalho seu referente ao cálculo das trajetórias dos astros, a famosa lei de distribuição dos erros de observação, que traz o seu nome. Gauss deduziu a expressão dessa lei baseando-se na hipótese de que o valor mais provável de uma grandeza é a média aritmética de suas observações, supostas independentes.

Trata-se de uma função exponencial, de expoente quadrático, contendo duas constantes paramétricas — a "média aritmética" *mu* (μ) da população hipotética, da qual o grupo de observações é interpretado como sendo uma amostra casualizada — e um segundo parâmetro, denominado "medida de precisão" (h), ligado à variabilidade

dos dados. Devido às propriedades algébricas que goza, esta distribuição, hoje denominada "normal", é a base da maioria dos métodos estatísticos modernos empregados nos trabalhos experimentais, métodos estes devidos a R. A. Fisher.

Por influência principalmente de autores ingleses, a medida da precisão "h" de Gauss foi substituida pelo desvio padrão *sigma* (σ) — "standard deviation" dos referidos autores —, que mede a distância dos dois pontos de inflexão da curva de Gauss ao pé da ordenada máxima, cuja abscissa é a média da distribuição. Existe entre "h" e *sigma* a seguinte relação:

$$h = \frac{1}{\sigma \sqrt{2}}$$

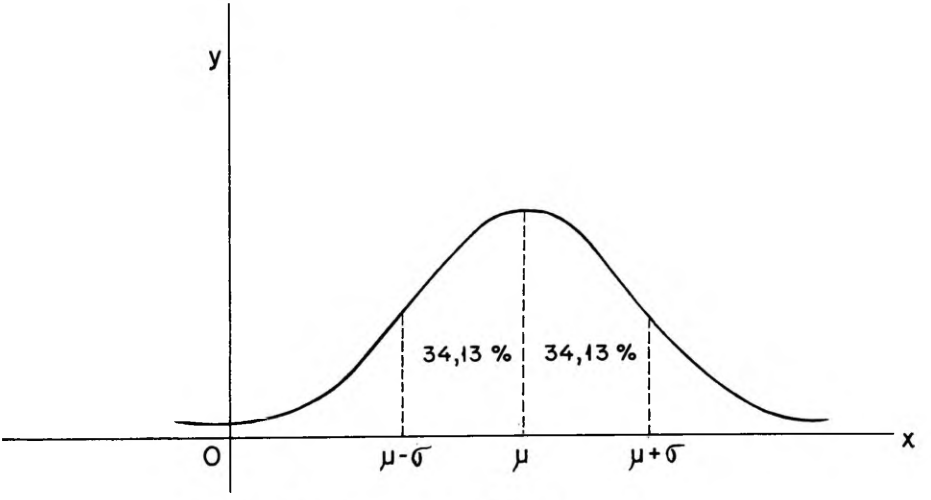

Fig. 1 Distribuição de Gauss:

Durante muito tempo, era usual, principalmente em astronomia e em física, expressar o resultado de um grupo de observações pela sua média aritmética "m" acrescida ou diminuida de seu desvio provável "d.pr.":

$$m \pm d.pr$$

O desvio provável delimita na distribuição normal uma faixa central, compreendendo a média aritmética, de probabilidade igual a 1/2, a área total delimitada pela distribuição normal inteira sendo igual a 1.

O cálculo do desvio provável depende do cálculo do desvio padrão:

$$d.pr = (0{,}6745)(d.p.)$$

O conceito de desvio padrão, introduzido pelos estatísticos ingleses, especialmente Karl Pearson, por ser de interpretação geométrica mais intuitiva, acabou prevalecendo: é definido como sendo a média quadrática dos quadrados dos desvios das observações em relação à média da população hipotética. Esta sendo, em geral desconhecida, deve-se compensar sua substituição pela estimativa "m" da média aritmética do grupo (ou amostra) mediante o chamado fator de correção de Bessel, igual a n/ (n-1), onde "n" é o número de dados observados. Sem esta correção, a estimativa do desvio padrão da população hipotética seria defeituosamente menor que a real. O fator (n-1), que aparece em denominador, representa os "graus de liberdade" introduzidos por Fisher, em estatística, por analogia a um outro conceito de restrição, existente em mecânica. A restrição referente ao problema estatístico, que nos interessa aqui, prende-se à determinação da média aritmética dos dados da amostra, obtida por meio da equação:

$$n\bar{x} = \text{Soma de todos os valores observados de "x"},$$

ou, simbolicamente,

$$n\bar{x} = Sx.$$

Equação semelhante nos dá a estimativa "s^2" do quadrado do desvio padrão, denominada "variância", expressa em função dos desvios do valor de cada dado de observação "x" em relação à média aritmética "$\bar{x}$" de todos os dados da amostra:

$$(n-1)\ s^2 = S\ (x-\bar{x}^2).$$

O cálculo da média aritmética "m", também representada habitualmente por "$\bar{x}$" ("x" com barra), é muito simples: é o quociente da soma de todos os dados de observação pelo número delas. Quanto à estimativa do desvio padrão, o cálculo é mais trabalhoso; dispondo-se de máquina de calcular, é preferivel obter-se a soma dos quadrados de desvios por meio da identidade algébrica:

$$S\ (x-\bar{x})^2 = Sx^2 - \frac{T^2}{n}$$

na qual,

$$T = Sx$$

é a soma de todos os valores observados de "x"; $S\,(x-\bar{x})^2$, a soma de quadrados de desvios em relação à média aritmética das observações; e Sx^2, a soma dos quadrados de cada observação. A parcela T^2/n é o chamado termo de correção para a média.

A média aritmética varia de amostra para amostra e admite, assim, um desvio padrão próprio que, para evitar confusão com o desvio padrão dos dados originais, se denomina "erro padrão da média" (e.p.m.). Relaciona-se ao desvio padrão pela expressão:

$$\text{e.p.m.} = \frac{\text{d.p.}}{\sqrt{n}};$$

isto é,

$$\text{e.p.m.} = \sqrt{\frac{S\,(x-\bar{x})^2}{n\,(n-1)}},$$

fórmula esta na qual os símbolos presentes têm os significados mencionados acima.

Embora ainda se encontre em trabalhos e tabelas físicas a apresentação de resultados de observação em termos de média e erro provável, a tendência moderna, particularmente em biologia, é apresentá-los em termos de média e desvio padrão, ou média e erro padrão da média:

1) n ——— m ——— d.p.

2) n ——— m ——— e.p.m.

ou então:

1') n ——— m ± d.p.

2') n ——— m ± e.p.m.

Em qualquer dos casos, convém indicar explicitamente o número "n" de observação de cada grupo, dado fundamental para se poder proceder certos testes de significância de interesse estatístico. A este respeito, a segunda maneira de apresentar os resultados de observação, mencionando-se diretamente o erro padrão da média, tem a vantagem de permitir calcular rapidamente o chamado "intervalo de confiança" com o grau de probabilidade que for desejado. Para isto, soma-se ou subtrai-se o erro padrão da média (e.p.m.) multiplicado

pelo valor de "t" lido nas tabelas de distribuição de "Student", correspondente à probabilidade desejada, e ao número de graus de liberdade (g.l. = n-1) utilizado no cálculo do desvio padrão. Assim, por exemplo, se,

$$n = 12$$

$$m = 5{,}60$$

$$e.p.m. = 0{,}26,$$

o intervalo de confiança de probabilidade igual a 95% estará compreendido entre os limites:

$$L_1 = 5{,}60 - (2{,}201)\,(0{,}26) = 5{,}03$$

$$L_2 = 5{,}60 + (2{,}201)\,(0{,}26) = 6{,}17,$$

onde t = 2,201 é o valor de "t", com 11 (= n-1) graus de liberdade, correspondente à probabilidade de 95% (ou à sua probabilidade complementar igual a 5%) das tabelas. O intervalo de confiança estende-se, portanto, de

5,03 a 6,17.

Isto significa que há 95% de probabilidade de que a verdadeira média da população hipotética da qual a média 5,60 da amostra é uma estimativa, esteja dentro do intervalo de confiança calculado, e apenas 5% de probabilidade que esteja fora dele.

Consultando 72 números recentes (1971-1972) da revista *Science,* publicação oficial da *American Association for the Advancement of Science,* verifica-se que das tabelas e gráficos lá constantes, as preferências foram:

59 vezes a favor de S.D. (Standard Deviation)

154 vezes a favor de S.E.M. (Standard Error of the Mean).

Acentua-se, portanto, a preferência a favor do *erro padrão da média,* em substituição ao *erro provável* adotado antigamente pelos físicos e astrônomos, na apresentação de dados de observação. O erro provável delimita o intervalo de probabilidade 50%; o erro padrão, o intervalo de probabilidade 68,26%; há entre eles a mesma relação mencionada a propósito do desvio provável e desvio padrão, isto é: e.pr. = (0,6745) (e.p.m.).

Das outras estatísticas de interesse computacional, constam explicitamente em tabelas e gráficos:

Mediana	5 vezes
Intervalo interquartil	1 vez
Variância	1 vez
"Range"	19 vezes

Apesar de seu interesse em diversas operações estatísticas, o número "n" de observações consta explicitamente apenas 49 vezes, contra um total de 213 para desvio padrão e erro padrão da média. A variância consta uma vez, e os quadros de análise da variância, 5 vezes.

Em resumo: Desejando-se apresentar de maneira concisa os dados de uma série de observações, pode-se fazê-lo de duas maneiras equivalentes:

1) n ——— m ——— d.p.

2) n ——— m ——— e.p.m.

ou então:

1') n ——— m ± d.p.

2') n ——— m ± e.p.m.

Em ambas, deve constar o número "n" de observações feitas, dado fundamental para se obter o valor de "t" nas tabelas de "Student", necessário para o cálculo dos intervalos de confiança das respectivas médias. Nos trabalhos recentes nota-se preferência a favor do erro padrão da média, como medida da variabilidade dos dados. Entre a estimativa do desvio padrão e do correspondente erro padrão da média existe a relação:

$$\text{d.p.} = (\text{e.p.m.}) \sqrt{\text{n.}}$$

BIBLIOGRAFIA

1. BAILLAUD, B. (1893) — *Cours D'Astronomie.* 1re. Partie. Gauthier-Villars, Paris.

2. CZUBER, E. (1932) — *Wahrscheinlichkeitsrechnung.* 1. Band. B. G. Teubner, Leipzig und Berlin, 4. Auflage.

3. DELTHEIL, R. (1930) — *Erreurs et Moindres Carrés.* Gauthier-Villars, Paris.

4. FISCHER, R. A. (1954) — *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, 12th ed.

5. GAUSS, K. F. — *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections* (Theoria Motus). English trans. (1857). Dover Publications, New York.

6. RIETZ, H. L. (1930) — *Handbuch der mathematishen Statistik.* Deutsche Ausgabe. B. G. Teubner, Leipzig und Berlin.

7. ———— — *Science.* n.os 3983-4057 (1971-72). American Association for the Advancement of Science.