DR. EUSÉBIO TAMAGNINI

A heterogeneidade da variação
Análise da variância

Questões de método

II

COIMBRA
TIPOGRAFIA DA ATLANTIDA
1938

A heterogeneidade da variação
Análise da variância

DR. EUSÉBIO TAMAGNINI

A heterogeneidade da variação
Análise da variância

Questões de método

II

COIMBRA
TIPOGRAFIA DA ATLANTIDA
1938

SEPARATA

DA

Revista da Faculdade de Ciências da Universidade de Coimbra

VOL. VI — N.º 4

A heterogeneidade da variação
A análise da variância

Quando se estuda qualquer carácter duma população dispersa numa área geográfica extensa reconhece-se imediatamente que as constantes (médias, desvios padrões, coeficientes de correlação, etc.) que caracterizam os vários grupos locais, ou as classes, em que a repartimos, manifestam, em regra, valores diferentes.

O conhecimento da significação estatística destas diferenças é indispensável para a formulação de qualquer juízo sôbre a homogeneidade ou heterogeneidade da população considerada, pois se muitas delas podem ser atribuíveis a causas de ordem genética ou paracinética, outras há que resultam de circunstâncias fortuitas, dependentes da forma como se constituiram as séries.

Já tivemos ocasião de expor alguns métodos adequados à resolução dêste problema (1). Tais métodos são porém um tanto laboriosos e estão sujeitos a certas restrições, muito particularmente os que se baseiam na noção de contingência.

Com efeito a equação fundamental da distribuïção dos valores de χ^2 não se pode utilizar quando a respectiva táboa de contingência, nalguns dos *loci*, contiver poucas observações, pois, nestes casos, não é possível determinar com segurança as probabilidades de ocorrência fortuita de valores de χ^2 iguais ou superiores aos obtidos, isto é, não é possível apreciar a significância das diferenças observadas.

---

(1) *A pigmentação dos portugueses. Revista da Faculdade de Ciências de Coimbra*, vol. VI, pág. 119.

1. Graus de liberdade — A *análise da variância* fornece um método elegante e cómodo para o estudo da variabilidade que, além disso, não é trabalhoso nem enferma das restrições apontadas.

Por variância de qualquer carácter duma população entende-se o valor de σ^2, isto é, o quadrado do respectivo desvio padrão. É evidente, por ser pràticamente impossível atingir a população na sua totalidade, que as nossas determinações da variância são meras *estimativas estatísticas,* como lhes chama o Prof. R. A. Fisher, que se afastam, mais ou menos, do valor exacto ou *parâmetro* da população respectiva. Para diferentes amostras de qualquer população obtemos geralmente variâncias diferentes, números que oscilam em tôrno do valor paramétrico.

Devemos ter sempre presente esta distinção entre os *parâmetros* duma população, que são absolutamente fixos, e as estimativas correspondentes, que são meras *estatísticas* cujo valor está dependente da constituïção das séries de que nos servimos para a sua determinação.

O grau de confiança que merecem as nossas *estatísticas* depende òbviamente do número dos casos incluidos nas séries que utilizamos para o seu cômputo. As diferentes estimativas do mesmo parâmetro duma população fazem parte duma distribuïção de freqüências cuja forma depende, até certo ponto, do número de indivíduos incluidos nas respectivas séries.

De facto, o que influi não é pròpriamente o número dos indivíduos, mas antes o número das *selecções fortuitas independentes* que, para formarmos as séries, se podem realizar na população.

Se uma caixa contiver duzentas esferas e com elas quisermos formar quatro lotes é evidente que, de cada vez, apenas três dêles podem ser arbitràriamente constituidos. Em cada caso, a constituïção do quarto lote está de antemão fixada pelo número total das esferas.

Semelhantemente uma amostra de n variantes, ou indivíduos, duma população nem sempre se pode considerar representativa de n selecções fortuitas e independentes, utilizáveis para a estimativa de qualquer parâmetro, pois uma delas, pelo menos, pode ser dependente da constituïção própria da amostra, isto é, dos pontos fixos, determinados em função dos valores das

variantes incluidas na amostra, em referência aos quais fazemos a estimativa dêsse parâmetro.

Êste facto tem importância pois reduz o número dos *graus de liberdade* disponíveis para a estimativa dos parâmetros.

Consideremos, por exemplo, o cálculo da variância dum carácter qualquer x numa amostra de grandeza n duma população, e seja m o valor médio exacto, ou paramétrico, dêsse carácter.

A melhor estimativa s do desvio padrão terá evidentemente a expressão

$$s = \sqrt{\sum (x-m)^2/n},$$

onde o somatório abrange todos os valores individuais do carácter incluidos na amostra.

Pràticamente, porém, como desconhecemos m (valor exacto da média da população geral), substituimo-lo por $\bar{x}$, isto é, pela média dos valores individuais do carácter na amostra escolhida.

Mas êste valor $\bar{x}$ é determinado pela constituïção própria da amostra, e a sua escolha para ponto fixo, a partir do qual se calculam os desvios das variantes, reduz duma unidade o número dos graus de liberdade disponíveis para o cômputo de s, pois se demonstra fàcilmente que, nesta hipótese, o melhor valor de s será dado pela expressão

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

Com efeito:

Sejam $x_1, x_2 \ldots\ldots x_n$ a série dos valores do carácter x; $f_1, f_2 \ldots f_n$ as freqüências respectivas numa amostra qualquer de grandeza n; e representemos por $\bar{x}_0$ a média do carácter na amostra considerada.

Será evidentemente

$$\bar{x}_0 = \frac{\sum f_i x_i}{n},$$

para todos os valores de i de 1 até n.

Seja, além disso, m a média exacta do carácter x na população infinita a que pertence a amostra considerada; represen-

temos por $\varepsilon = m - \overline{x}_o$ o êrro da nossa estimativa da média, por δ_i o desvio duma classe qualquer x_i do carácter relativamente à média exacta m; e por d_i o desvio da mesma classe relativamente à média estimada $\overline{x}_o$.

Podemos escrever

$$\delta_i = m - x_i,$$

e

$$d_i = \overline{x}_o - x_i,$$

e, por conseguinte,

$$\delta_i = \varepsilon + d_i. \qquad (1)$$

O desvio de qualquer classe x_i *relativamente à média exacta é, pois, igual ao seu desvio relativamente à média estimada somado com o êrro dessa média.*

Elevando ao quadrado a expressão (1), temos

$$\delta_i^2 = \varepsilon^2 + 2\,\varepsilon\, d_i + d_i^2,$$

e, somando para tôdas as classes, será

$$\begin{aligned} \sum f_i \delta_i^2 &= \varepsilon^2 \sum f_i + 2\,\varepsilon \sum f_i d_i + \sum f_i d_i^2 \\ &= n\,\varepsilon^2 + \sum f_i d_i^2 \end{aligned} \qquad (2)$$

por ser, evidentemente, $\sum f_i = n$ e $\sum f_i d_i = o$.

O primeiro membro da igualdade (2) é a soma dos quadrados dos desvios em relação à média exacta; o primeiro termo do segundo membro é igual a n vezes o quadrado do êrro da média estimada, e o segundo termo é a soma dos quadrados dos desvios relativamente a essa média.

Como não conhecemos o valor da média exacta, m, não sabemos qual seja o valor do êrro (ε) da média estimada, mas podemos substituí-lo pela sua melhor estimativa, isto é, pelo *êrro médio da média* cuja expressão, como se sabe, é

$$\varepsilon_m = s \Big/ \sqrt{n},$$

onde s representa a estimativa do desvio padrão.

A igualdade (2) fornece assim a seguinte aproximação:

$$n s^2 \longrightarrow s^2 + \sum f_i d_i^2,$$

ou

$$(n-1) s^2 \longrightarrow \sum f_i d_i^2; \qquad (3)$$

e, portanto

$$s^2 \longrightarrow \frac{\sum f_i d_i^2}{(n-1)},$$

q. e. d.

Quando n é grande torna-se pràticamente indiferente dividir a soma dos quadrados dos desvios por n ou por $(n-1)$, mas quando n é pequeno as diferenças podem ser consideráveis e ter influência na apreciação dos resultados.

2. Propriedade aditiva da variância — A utilização da variância para o estudo da variabilidade das populações funda-se na sua propriedade aditiva.

Demonstra-se, com efeito, que se um carácter qualquer está sugeito à operação de várias causas independentes, cada uma das quais contribui com uma certa variância, a variância total é igual à soma das variâncias parciais (Cf. TIPPETT — *The methods of Statistics*, pág. 89).

Por conseguinte, representando por σ_x^2 a variância total do carácter x, sujeito à influência dos factores A e B, independentes, cada um dos quais determina, respectivamente, as variâncias σ_a^2 e σ_b^2, será

$$\sigma_x^2 = \sigma_a^2 + \sigma_b^2 + \sigma_r^2,$$

onde σ_r^2 representa a variância devida aos erros fortuitos.

3. Análise da variância — Consideremos uma população geral de grandeza N (número total dos indivíduos observados) que, a respeito dum carácter qualquer x, se acha repartida por m grupos de grandeza n; será evidentemente $N = n\,m$.

Representemos por $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_m$ e $\bar{x}$ as médias do carácter nos grupos correspondentes e na população geral.

É evidente que, em cada grupo, *o desvio de qualquer valor de* x, *relativamente à média geral da população, é igual ao seu*

desvio a repeito da média do grupo a que pertence somado com o desvio da média dêsse grupo em referência à média geral.

Com efeito, para qualquer indivíduo do grupo i, a expressão

$$(x - \overline{x}) = (x - \overline{x}_i) + (\overline{x}_i - \overline{x})$$

é uma identidade.

Elevando ao quadrado, teremos

$$(x - \overline{x})^2 = (x - \overline{x}_i)^2 + (\overline{x}_i - \overline{x})^2 + 2\,(x - \overline{x}_i)\,(\overline{x}_i - \overline{x}).$$

E a soma dêstes quadrados, extensiva a todos os indivíduos do grupo, será

$$\sum (x - \overline{x})^2 = \sum (x - \overline{x}_i)^2 + \sum (\overline{x}_i - \overline{x})^2 + 2\,(\overline{x}_i - \overline{x}) \cdot \sum (x - \overline{x}_i)$$
$$= \sum (x - \overline{x}_i)^2 + n\,(\overline{x}_i - \overline{x})^2,$$

por, em qualquer grupo, $(\overline{x}_i - \overline{x})^2$ ser constante para todos os indivíduos e o somatório incluir n parcelas, e o terceiro termo do segundo membro ser nulo visto a soma dos desvios individuais relativamente à média do grupo respectivo, $\sum (x - \overline{x}_i)$, ser igual a zero.

Somando membro a membro as expressões semelhantes correspondentes aos vários grupos, temos

$$\sum_i \sum (x - \overline{x})^2 = \sum_i \sum (x - \overline{x}_i)^2 + n \sum_i (\overline{x}_i - \overline{x})^2, \qquad (4)$$

para todos os valores de i de 1 até m.

E assim conseguimos dividir a soma total dos quadrados dos desvios em relação à média geral (primeiro membro da expressão (4)) em duas partes: *a*) soma dos quadrados dos desvios das observações individuais relativamente as médias dos grupos respectivos — *1.º termo do segundo membro; b*) n vezes a soma dos quadrados dos desvios das médias dos grupos relativamente à média geral da população — *2.º termo do segundo membro.*

Para podermos calcular as variâncias resta apenas determinar o número dos respectivos graus de liberdade.

Para a variância média total, como determinamos apenas a média $\overline{x}$, êsse número será $N - 1 = n\,m - 1$; para a *variância média dos grupos*, ou variância residual, como para cada grupo se determinou a média $\overline{x}_i$, e os grupos são m de grandeza n, será $m\,(n-1) = N - m$ [1]; para a *variância média entre os grupos*, será $m - 1$ por se terem calculado m desvios relativamente à média geral x.

4. Cômputo da variância — Quando se procede ao cálculo da variância de dados reünidos em séries ordenadas é conveniente, pára achar os valores em referência à média, quadrar primeiramente os desvios relativamente a uma origem arbitrária e efectuar depois a correcção dos resultados.

Para qualquer valor x duma série, o quadrado do desvio em relação à média $\overline{x}$ é dada pela expressão

$$(x - \overline{x})^2 = x^2 - 2\,\overline{x}\,.\,x + \overline{x}^2.$$

---

[1] Sejam $x_1, x_2 \ldots x_n$ as variantes do carácter x numa população de grandeza N que se dividiu em m grupos de grandeza $n_1, n_2, \ldots n_m$. Seja $f_i^{(j)}$ a freqüência da variante x_i, no grupo j e s_j^2 a variância correspondente.

Pelas considerações feitas a pág. 5 as melhores estimativas das variâncias s_j^2 dos grupos teem as seguintes expressões.

$$s_1^2 = \frac{\Sigma\, f_i'\, d_i'^2}{n_1 - 1}, \qquad s_2^2 = \frac{\Sigma\, f_i''\, d_i^{(2)2}}{n_2 - 1} \cdots \qquad s_m^2 = \frac{\Sigma\, f_i^{(m)}\, d_i^{(m)2}}{n_m - 1},$$

e o seu valor médio pesado será evidentemente

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \ldots + (n_m - 1)\, s_m^2}{n_1 - 1 + n_2 - 1 + \ldots + n_m - 1} =$$

$$= \frac{\Sigma\, f_i'\, d_i'^2 + \Sigma\, f_i''\, d_i''^2 + \ldots + \Sigma\, f_i^{(m)}\, d_i^{(m)2}}{N - m},$$

por ser $n_1 + n_2 + \ldots + n_m = N$.

A soma dêstes quadrados, extensiva a todos os valores individuais da série, será portanto

$$\sum (x-\overline{x})^2 = \sum (x^2) - 2\,\overline{x} \,.\, \sum (x) + n\,\overline{x}^2$$
$$= \sum (x^2) - n\,\overline{x}^2,$$

por ser, por definição da média,

$$\sum (x) = n \,.\, \overline{x} \,.$$

A soma dos quadrados dos desvios duma série de valores em relação à média é pois igual à diferença entre a soma dos quadrados dos valores individuais e n *vezes o quadrado da média.*

Mas, como os valores individuais se podem considerar desvios da variável em relação à origem das coordenadas, a média dos valores individuais será evidentemente igual ao desvio médio em relação a essa origem, e, por isso:

A soma dos quadrados dos desvios em relação à média é igual à soma dos quadrados dos desvios em relação a qualquer origem arbitrária menos n *vezes o quadrado do desvio médio referido à mesma origem.*

Representando por x' os desvios em relação a uma origem arbitrária e por $\overline{x}'_i$ o desvio médio de qualquer grupo i em referência à mesma origem, temos para o cálculo dos termos da expressão (4) as seguintes igualdades

$$\sum_i \sum (x-\overline{x}_i)^2 = \sum_i \sum (x'^2) - n \sum_i (\overline{x}_i'^2),$$
$$n \sum_i (\overline{x}_i - \overline{x})^2 = n \left[\sum_i (\overline{x}_i'^2) - m\,\overline{x}'^2 \right] \qquad (5)$$
$$= n \sum_i (\overline{x}_i'^2) - \mathrm{N}\,\overline{x}'^2 ;$$

e, portanto

$$\sum_i \sum (x-\overline{x})^2 = \sum_i (x'^2) - \mathrm{N}\,\overline{x}'^2 .$$

O trabalho de cálculo ainda se simplifica muito se, em vez de determinarmos directamente o desvio médio $\overline{x}'$, quadrarmos primeiramente o total $\sum (x')$ e dividirmos o resultado pelo número correspondente.

TABELA I

Largura nasal em m/m. Centros das classes	Pantera (Ngo)	Gafanhoto (Nsenene)	Bufalo (Mbogo)	Vaca sem cauda (Ente ya kikugu)	Dipnoico (Mamba Bakerekere)	Carneiro (Endiga)	Macaco cinzento (Nkima)	Oribi — Antiope (Mpewo)	Sementes — contas (Katinvum)	Passaro (Nyonys)	Rato (Musu)	Inhame (Kobe)	Cabrito montês (Ngabi)	Gineto (Kasimba)	Corações (Mutima)	Corvo (Namungona)	Dipnoico (Mamba — Munguya)	Totais
36								1										1
37								1		1		1						3
38					1			—		1		—		1				3
39			1	1	1			—	1	—		--	2	2				8
40	1		—	1	—		1	—	—	—		2	—	—	1			6
41	2	1	1	1	1	1	3	1	—	1	2	—	2	1	2	3		22
42	—	1	—	3	—	1	3	—	1	2	2	—	1	—	2	4	1	21
43	—	1	4	2	1	6	1	4	4	2	1	—	3	—	—	2	2	33
44	1	2	2	3	3	2	3	1	2	2	1	4	2	2	3	1	3	37
45	1	1	2	2	2	1	—	2	1	1	1	4	—	1	2	—	2	23
46	2	3			2		2	1	2	3	1		1	4	1	4	4	31
47	3	1	1	1	—	3	—	2	2	1	2	1	1	2	2	1	—	23
48	1	2	3	—	1	—	—	1	1	1	3	2	3	1	1		1	21
49	—	1	—	—	—	1	1	—	1		—	—		—	—		—	4
50	2	—	—	—	2			—			2	—		1	1		—	8
51	—	2	—	—	—			1				—					2	5
52	1		—	—	1							—						2
53	—		—	—								1						1
54	—		—	—														—
55	1		1	—														2
56				—														—
57				1														1
Totais	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	255
$\Sigma (x')$	+9	+1	−14	−32	−15	−26	−39	−31	−21	−38	−8	−21	−34	−24	−24	−38	−6	−361
$\Sigma (x'^2)$	247	127	214	312	233	116	189	279	125	234	142	231	212	218	154	162	102	3297
$n\bar{x}'^2$	5.4000	0.0667	13.0667	68.2667	15.0000	45.0667	101.4000	64.0667	29.4000	96.2667	4.2667	29.4000	77.0667	38.4000	38.4000	96.2667	2.4000	511.0621
$\Sigma (x'^2) - n\bar{x}'^2$	241.6000	126.9333	200.9333	243.7333	218.0000	70.9333	87.6000	214.9333	95.6000	137.7333	137.7333	201.6000	134.9333	179.6000	115.6000	65.7333	99.6000	2785.9379

Com efeito, por ser

$$\overline{x}' = \frac{\sum (x')}{N},$$

será evidentemente

$$\left[\sum (x')\right]^2 / N = N \cdot \overline{x}'^2 .$$

5. Exemplo I. — Sôbre os diferentes clans dos *Baganda* estão publicadas algumas séries de dados antropométricos (1).

O número de indivíduos observados em cada clan é muito pequeno, 15 em regra, de modo que os valores médios e as variabilidades que se podem determinar para os diferentes caracteres considerados não oferecem grande confiança e não se podem aproveitar para fins comparativos.

Diz Roscoe que o tipo físico dos Bangada varia consideràvelmente; contudo a prática continuada da exogamia deve tender a distribuir uniformemente os factores genéticos por tôda a população. Será portanto interessante averiguar se, estatìsticamente, os números registados por Roscoe estabelecem qualquer distinção efectiva entre os clans, ou se as diferenças observadas se podem atribuir às flutuações do acaso. Nesta última hipótese poderiam juntar-se tôdas as observações numa única série, cujos parâmetros se poderiam considerar representativos da população geral.

Na exemplificação do emprêgo do método da análise da variância limitar-nos-hemos ao estudo da *largura nasal*.

Os valores individuais da largura nasal observados em 17 clans constam da Tabela I.

Na primeira coluna estão registados as classes do carácter com intervalos de 1 $^{m}/_{m}$, e na última a série total resultante da reunião das observações feitas nos diferentes clans.

Tomou-se para origem arbitrária a variante 46 e as quatro últimas linhas dão os valores de $\sum (x')$, $\sum (x'^2)$, $n \cdot \overline{x}'^2$ e $\sum (x'^2) - n\,\overline{x}'^2$, para os diferentes clans e para a série total.

---

(1) John Roscoe. — The Baganda. — *Their customs and beliefs — Mac Millan & C.°*, 1911.

As somas dos quadrados dos desvios em relação à média, os graus de liberdade e as variâncias médias vão indicados na Tabela II.

A soma dos quadrados dos desvios e a dos graus de liberdade das partes devem dar os valores correspondentes à série total; e as somas dos quadrados dos desvios divididas pelos graus de liberdade dão as estimativas das variâncias médias.

A variância média total é a variância ordinária da série constituida pelo conjunto de tôdas as observações individuais.

A variância média «*nos grupos*» é, neste exemplo, uma estimativa da verdadeira variância, σ_r^2, baseada em 238 graus de liberdade.

TABELA II

Origem da variação	Soma dos quadrados dos desvios	Graus de liberdade	Variâncias médias
Entre os grupos	213.1382	16	13.3211
Nos grupos	2572.7997	238	10.8101
Total	2785.9379	254	10.9683

A variância média *entre os grupos* é uma estimativa, baseada 16 graus de liberdade, de n vezes a variância σ_i^2 das médias dos grupos, como fàcilmente se verifica pelo exame da expressão (4).

Êste valor σ_i^2 é o parâmetro correspondente a uma população infinita, isto é, representa a variância das médias dos grupos na hipótese dum número infinitamente grande de grupos cada um constituido por um número indefinidamente grande de indivíduos.

Como, porém, na prática o número (m) dos grupos é finito, bem como o (n) dos indivíduos que os constituem, as médias dos grupos estão sujeitas a êrros fortuitos, resultantes da sua formação, que farão acrescer a sua variância.

Como representamos por σ_r^2 a variância real «nos grupos», o êrro médio das médias — que mais não é do que o desvio padrão da sua distribuïção fortuita, ou a raíz quadrada da variância correspondente — terá o valor $\sigma_r/\sqrt{n}$, e, por isso, a variância total

das médias dos grupos, pela propriedade aditiva da variância, será igual a

$$\sigma_i^2 + \sigma_r^2/n.$$

Representando por v_i^2 e v_r^2 as variâncias médias «*entre os grupos*» e «*nos grupos*» determinadas a partir dos dados, obtem-se as relações

$$\begin{array}{ll} v_i^2 & \longrightarrow \quad n\,\sigma_i^2 + \sigma_r^2; \\ v_r^2 & \longrightarrow \quad \sigma_r^2 \end{array} \qquad (6)$$

que nos permitem apreciar a homogeneidade da variação na população considerada.

Com efeito, se a variação entre os grupos é relativamente importante, σ_i^2 é grande, comparada com σ_r^2, e as duas estimatimas v_i^2 e v_r^2 serão muito diferentes. Se, pelo contrário a variação entre os grupos é nula, será $\sigma_i^2 = 0$ e por conseguinte v_i^2 tenderá para σ_r^2.

Neste caso, v_i^2 e v_r^2 representam duas estimativas independente da mesma variância σ_r^2, que estão sujeitas a êrros fortuitos como tôdas as estimativas da variância, e a significação das suas diferenças poderá apreciar-se por meio de testes adequados.

Introduzindo nas relações (6) os valores registados na Tabela II, temos

$$\begin{array}{ll} 13{,}3211 & \longrightarrow \quad 15\,\sigma_i^2 + \sigma_r^2 \\ 10{,}8101 & \longrightarrow \quad \sigma_r^2. \end{array}$$

Por conseguinte

$$2{,}5110 \longrightarrow 15\,\sigma_i^2$$

e

$$0{,}1674 \longrightarrow \sigma_i^2.$$

Vê-se pois que a variância «*entre os grupos*» é muito menor que a variância «*nos grupos*» e possìvelmente não existem, a respeito do carácter em questão, quaisquer diferenças significativas entre os clans.

6. Testes de significância — Quando se consideram amostras diferentes da mesma população e se fazem estimativas de qualquer dos seus parâmetros obtem-se, em regra, para cada um dêles, uma série de valores diferentes. Para interpretar estas diferenças torna-se portanto indispensável o emprêgo de critérios que nos permitam ajuizar da probabilidade da sua ocorrência fortuita.

Consideremos o caso simples duma população infinita cujos caracteres variam normalmente, isto é, oscilam em tôrno das respectivas médias de harmonia com a *lei dos êrros*.

Se desta população infinita extrairmos uma série de amostras de igual grandeza (n) e para cada uma estimarmos a média de qualquer dos seus caracteres, obteremos uma série de valores que, à medida que aumenta o número das amostras e se reduz o valor do intervalo das classes, se distribuem segundo um diagrama que tende para uma curva de tipo gaussiano, cujo desvio padrão se demonstra fàcilmente ser igual a $\sigma / \sqrt{n}$, se σ fôr o desvio padrão do carácter na população infinita.

Êste desvio padrão $(\sigma / \sqrt{n})$ da distribuïção das médias da série das amostras da população infinita é o que se denomina *êrro médio da média*.

No caso de $n = 1$, o êrro médio da média coincide com o desvio padrão da distribuïção dos valores individuais do carácter, como é óbvio, e por conseguinte a designação «êrro» não tem nestes casos a significação usual, pois o êrro médio, como se vê, mais não é do que uma constante descritiva da distribuïção dos valores da média, que nos permite determinar as probabilidades da ocorrência fortuita de desvios iguais ou maiores que qualquer valor dado.

Sabendo que a curva de distribuïção das médias é uma curva normal, fácil se torna determinar as probabilidades de, numa amostra fortuita da mesma população, obtermos uma estimativa que difira da média verdadeira tanto ou mais que uma certa quantidade.

Basta, com efeito, marcar sôbre o eixo dos x da respectiva distribuïção, para um e para o outro lado da ordenada central, distâncias iguais ao desvio considerado, levantar pelos pontos respectivos as ordenadas correspondentes e calcular a fracção da área limitada pela curva, para além dessas ordenadas.

Se essas superfícies representarem uma fracção muito pequena

da área total limitada pela curva, também serão muito pequenas as probabilidades da ocorrência fortuita de desvios iguais ou maiores que o observado e, por isso, a diferença entre a média verdadeira e a estimada deve considerar-se *estatìsticamente significativa*.

Como, se costuma adoptar, embora arbitràriamente, o valor de $P = 0.05$ para limite ou *nível da significância*, considera-se real qualquer desvio da média cujas probabilidades de ocorrência fortuita sejam menores que 5 %.

Esta probabilidade, $P = 0.05$, corresponde aproximadamente a um desvio igual a duas vezes o desvio padrão, ou êrro médio, pois numa distribuïção normal, como a da média, as porções da área limitada pela curva que ficam para além das ordenadas levantadas pelos pontos $\pm x/\sigma = 2$ são iguais a 0.0227, e a sua soma é, por isso, aproximadamente igual a 0.05 da área total.

Não se deve confundir o nível da significância $P = 0.05$, com o desvio correspondente que é definido pelo ponto 0.025, em referência à área total. Se porém nos referimos apenas à área correspondente aos desvios positivos, ou negativos, o nível da significância de 5 % corresponderá, como é óbvio, ao ponto 0.05.

Quando as distribuïções são assimétricas, como sucede com as da maior parte das constantes estatísticas, a ordenada que passa pela média não divide a área limitada pelo curva em duas partes iguais. Contudo o critério para o estabelecimento do nível da significância de 5 % pode ainda estender-se a tais casos, considerando situados nesse nível os desvios que marcam ordenadas interceptando de cada lado da média áreas caudais iguais a 2,5 % da área total.

Neste caso, como é óbvio, os dois desvios — positivo e negativo — não são iguais. Pode, porém, por uma conveniente mudança de variável reduzir se sempre a curva à forma normal. A transformação altera as posições relativas da média e da moda mas não influe nos desvios definidos pelas relações em que dividem a área limitada pela curva, de modo que os valores da variável x' correspondentes aos pontos 2,5 % permanecem valores da variável x, correspondentes aos mesmos pontos da distribuïção assimétrica.

Portanto, transformando a curva, podemos fàcilmente deter-

minar o desvio correspondente ao nível de significância de 5 % em obediência à normalidade da distribuïção dos desvios.

7. **Significação das diferenças entre variâncias** — Consideremos duas amostras da mesma população de grandezas N_1 e N_2 e sejam s_1 e s_2 as estimativas dos respectivos desvios padrões de qualquer dos seus caracteres.

A apreciação da diferença $(s_1 - s_2)$ entre os desvios padrões das duas amostras costuma fazer se *supondo que a forma da distribuïção de tais diferenças é normal* e que o seu êrro médio é

$$\sqrt{\frac{\sigma^2}{2N_1} + \frac{\sigma^2}{2N_2}},$$

onde σ representa o desvio padrão da população infinita que forneceu as amostras que, por nos ser desconhecido, substituímos por s_1 e s_2.

Quando as amostras são pequenas, isto é, pouco numerosos os indivíduos constituintes, os êrros resultante desta suposição podem ser grandes, e, por isso, R. A. Fisher introduziu, para a interpretação das diferenças, outro critério definido pela equação

$$z = \frac{1}{2}\left[\log_e s_1^2 - \log_e s_2^2\right] = \log_e \frac{s_1}{s_2}.$$

onde s_1 e s_2 são as variâncias calculadas na base dos correspondentes graus de liberdade.

Fisher demonstrou também que a forma da distribuïção dos valores de z é dada pela equação

$$df = k \frac{e^{n_1 z}}{\left(n_1 e^{2z} + n_2\right)^{\frac{1}{2}(n_1 + n_2)}} dz,$$

onde k é uma constante e $n_1 = N_1 - 1$ e $n_2 = N_2 - 1$ são os graus de liberdade.

Esta equação que encerra apenas as variáveis z, n_1 e n_2 é independente do desvio padrão σ da população e, por conseguinte, utilizável no caso de pequenas amostras.

A grandeza z, que oscila entre $\mp \infty$, é negativa quando

$s_1/s_2 < 1$, positiva se $s_1/s_2 > 1$, e a distribuïção dos seus valores é assimétrica a não ser que $n_1 = n_2$.

Como porém, para qualquer combinação dos graus de liberdade, a parte positiva da curva, correspondente aos valores de s_1/s_2, é igual à parte negativa correspondente aos valores s_2/s_1, basta considerar os valores do integral das probabilidades relativos aos desvios positivos, pois os restantes se podem obter trocando n_1 por n_2.

Mas, por ser mais prático trabalhar com valores positivos de z, tomando n_1 como número dos graus de liberdade correspondente à variância maior, a diferença dos logarítmos será sempre positiva.

Fisher elaborou para diferentes combinações de n_1 e n_2 uma táboa dos valores de z correspondentes às ordenadas limitantes de áreas caudais com 5% e 1% da área total da curva.

A distribuïção dos valores de z, em regra, é assimétrica e por isso, empregando o critério definido a pág. 13, consideram-se situados no nível da significância os valores de z cujas ordenadas separam áreas caudais que representam cada uma 0.025 da área total. Por conseguinte o nível de significância de 5% ficará situado entre os pontos 5% e 1% da Táboa de Fisher.

No nosso exemplo, a variância da média dos grupos é muito pequena, como vimos (cfr. pág. 11); trata-se agora de saber se contudo é significativamente diferente de zero, ou se podemos considerar os grupos, quanto à variabilidade da largura nasal, amostras fortuitas da mesma população.

Os dados são:

$$z_1 = 13.3211, \qquad n_1 = 16,$$
$$z_2 = 10{,}8101, \qquad n_2 = 238;$$

e, por conseguinte

$$z = \frac{1}{2}\left[\log_e 13.3211 - \log_e 10.8101\right] = 0.1044.$$

A Táboa de Fisher dá, para esta combinação de graus de liberdade, como correspondente ao nível 5% da significância,

o valor de $z = 0\,2573$, que excede mais duas vezes e meia o valor encontrado (1).

Não se pode por isso considerar significativa a diferença das variâncias e os diversos clans, quanto aos valores registados para a largura nasal, podem tomar-se como amostras fortuitas da mesma população.

8. Exemplo II. — No exemplo I os diferentes grupos são constituídos por igual número de indivíduos. A análise da variância pode todavia efectuar-se semelhantemente ainda que os grupos sejam numèricamente desiguais.

Com efeito, se os diferentes grupos foram constituídos por n_1, $n_2 \ldots n_m$ indivíduos, respectivamente, a expressão (4) toma a forma

$$\sum_i \sum (x - \bar{x})^2 = \sum_i \sum (x - \bar{x}_i)^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2,$$

e nas equações (5) em vez de

$$n \sum_i (\bar{x}_i'^2),$$

temos de escrever

$$\sum_i n_i \bar{x}'^2.$$

Vê-se assim que, no caso geral, o cômputo das somas dos quadrados dos desvios se pode efectuar da seguinte maneira:

1) A «total» é a soma dos quadrados dos valores individuais menos N vezes o quadrado da média geral;
2) A «entre os grupos» é igual à sôma, extensiva a todos os grupos, dos produtos do quadrado da média de cada grupo pela freqüência (n_i) correspondente, menos N vezes o quadrado da média geral;
3) A «nos grupos» é a diferença entre 1) e 2).

---

(1) A Táboa de Fisher dá os valores de z correspondentes aos pontos 5% e 1% para $n_1 = 1, 2, 3, 4, 5, 6, 8, 12, 24, \infty$, e $n_2 = 1, 2, \ldots 30, 60, \infty$. Para outros valores de n_1 e n_2 é necessário fazer interpolações, podendo empregar-se o método indicado por C. H. Goulden — *Methods of statistical analysis*, págs. 77.

Os desvios são calculados em relação a uma origem arbitrária qualquer e N é o número total dos indivíduos.

Neste caso, não são válidas as relações (6) porque, quando se calculam os quadrados dos desvios das médias dos grupos em relação à média geral, se dá a cada grupo um pêso n_i diferente.

Em todo o caso, se a variância verdadeira entre os grupos (σ_i^2) fôr nula, v_i^2 ainda tenderá para v_r^2, e o teste de homogeneidade consistirá também em verificar se v_i^2 e v_r^2 são significativamente diferentes.

Na Tabela III estão registados os valores da largura nasal numa série de tribus da África oriental. Os números foram extraídos do estudo de NORMAN M. LEYS e T. A. JOYCE — *Note on a series of physical measurements from East Africa — The Journal of the Royal Anthropological Institute of Great Britain and Ireland.* — Vol. XLIII, pág. 195 e seg.

Os autores, na apreciação do grau de semelhança física entre as tríbus consideradas duas a duas, empregaram um critério a que chamaram *índice diferencial*, cuja expressão analítica é

$$\sum \Delta = \sum \left[(M_1 - M_2) \Big/ \sqrt{\sigma_1^2 + \sigma_2^2} \right],$$

onde M_1 e M_2 são as médias correspondentes a cada carácter, σ_1^2 e σ_2^2 as estimativas dos desvios padrões respectivos, e o somatório $\sum$ abrange os vários caracteres estudados.

O grau de divergência entre as tribus, apreciado por êste critério, é considerado tanto maior quanto mais elevado fôr o valor do índice diferencial.

Na Tabela 9 do citado estudo (cf. pág. 211), estão ordenadas as tribus duas a duas consoante os valores de $\sum \Delta$; por ela se vê que os Akamba, Akikuyu, Embu, Suk e Kachamega estão interrelacionados por índices diferenciais cujo valor é inferior a 2. Os autores afirmam que, por isso, os devemos considerar constituindo «um grupo, no qual as tribus de Kenia estão em relações mais estreitas entre si do que as restantes» (*Op. cit.*, pág. 200).

TABELA III

Largura nasal em m/m. Centros das classes	Designação das tribus					Totais
	Akikuyu	Embu	Kachamega	Akamba	Suk	
28	1					1
29						—
30						—
31						—
32	1			1		2
33	1			1		2
34	4			—		4
35	8	2	1	1	2	14
36	13	3	1	5	—	22
37	16	7	7	4	1	35
38	41	6	15	3	2	67
39	37	14	18	13	1	83
40	36	16	16	6	2	76
41	45	17	15	7	1	85
42	26	12	7	5	1	51
43	18	9	11	3	3	44
44	12	4	3	3	1	23
45	7	2	2	3	—	14
46	5	3	3		1	12
47	—	3	—			3
48	1	—	—			1
49	1	1	1			3
50	1	1				2
Totais	274	100	100	55	15	544
$\Sigma(x')$	+ 255	+ 183	+ 133	+ 37	+ 19	+ 627
$\Sigma(x')^2$	2493	1161	777	475	175	5081
$n_s \bar{x}'^2$	273.3175	334.8900	176.8900	24.8909	24.0667	722.6636
$\Sigma(x'^2) - n_s \bar{x}'^2$	2219.6825	826.1100	600.1100	450.1091	150.9333	4358.3364

Na Tabela IV estão registadas as somas dos quadrados, os graus de liberdade e as variâncias médias respeitantes à largura nasal, calculadas a partir das observações do Leys e Joyce.

O valor de z, calculado pela fórmula a pág. 15 é 0.6312.

TABELA IV

	Somas dos quadrados	Graus de liberdade	Variâncias médias
Entre os grupos	111 3915	4	27.8479
Nos grupos	4246.9449	539	7.8792
Total	4358.3364	543	—

Pela Táboa de Fisher, para $n_1 = 4$ e $n_2 = 539$, os valores de z correspondentes aos pontos de 5% e 1% são respectivamente .4667 e .6052. O valor achado é maior e, por isso, quanto à largura nasal, a heterogeneidade da população geral constituída pelas referidas tribus, deve considerar-se real.